

**MODELOS DE INTELIGÊNCIA ARTIFICIAL NA ANÁLISE DA
MONITORIZAÇÃO DE EVENTOS CLÍNICOS ADVERSOS,
DISFUNÇÃO/FALÊNCIA DE ÓRGÃOS E PROGNÓSTICO DO
DOENTE CRÍTICO**

ÁLVARO JOSÉ BARBOSA MOREIRA DA SILVA

Dissertação de doutoramento em Ciências Médicas

2007

ÁLVARO JOSÉ BARBOSA MOREIRA DA SILVA

**MODELOS DE INTELIGÊNCIA ARTIFICIAL NA ANÁLISE DA
MONITORIZAÇÃO DE EVENTOS CLÍNICOS ADVERSOS,
DISFUNÇÃO/FALÊNCIA DE ÓRGÃOS E PROGNÓSTICO DO
DOENTE CRÍTICO**

Dissertação de candidatura ao grau de Doutor em Ciências
Médicas apresentada ao Instituto de Ciências Biomédicas de Abel
Salazar da Universidade do Porto

Orientador – Professor Doutor João José Lopes Gomes
Professor Catedrático Convidado
Instituto de Ciências Biomédicas Abel Salazar
Universidade do Porto

Co-orientadores
Professor Doutor José Carlos Ferreira Maia Neves
Professor Catedrático
Departamento de Informática, Escola de Engenharia, Universidade
do Minho
Professor Doutor Manuel Filipe Vieira Torres dos Santos
Professor Auxiliar
Departamento de Sistemas de Informação, Escola de Engenharia,
Universidade do Minho



J1849

Trabalho realizado nas seguintes instituições:

Health Services Research Unit do Hospital Universitário de Groningen, Holanda

Serviço de Cuidados Intensivos do Hospital Geral de Santo António, Porto

Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto

Escola de Engenharia da Universidade do Minho

Índice

CAPÍTULO I – INTRODUÇÃO.....	9
Agradecimentos	9
1.Âmbito e motivação	11
2. Objectivos.....	17
3.Organização da dissertação	19
CAPÍTULO II – A AVALIAÇÃO DO PROGNÓSTICO E DA FALÊNCIA DE ÓRGÃOS EM MEDICINA INTENSIVA	29
1.Introdução.....	29
2. O prognóstico na Medicina Intensiva.....	31
3. Índices gerais de gravidade da doença e modelos de prognóstico	33
3.1 Descrição e classificação dos índices gerais de gravidade da doença e modelos prognósticos.....	35
3.2 Cálculo do índice de gravidade e da probabilidade de mortalidade	50
3.3 Validação dos índices gerais de gravidade e modelos prognósticos	52
3.4 Comparação dos modelos	55
4. Índices de quantificação da falência múltipla de órgãos	57
4.1 Descrição e classificação dos índices de quantificação da falência de órgãos	60
4.2 Comparação dos índices de avaliação da falência múltipla de órgãos e sistemas	66
5. Limitações dos modelos prognósticos e dos índices de quantificação da falência múltipla de órgãos	68
6. Actualizações dos modelos prognósticos	73
7. Conclusões.....	75
8. Bibliografia	77
CAPÍTULO III - A DESCOBERTA DE CONHECIMENTO EM MEDICINA INTENSIVA.....	93
1. Introdução.....	93
2. O Processo da Descoberta de Conhecimento em Base de Dados	95
2.1 Definição.....	95
2.2 Fases do Processo de Descoberta de Conhecimento em Bases de Dados	97
2.3 <i>Data Mining</i>	99
2.4 Avaliação dos modelos	117
3. Aplicação das redes neuronais artificiais em Medicina Intensiva	122
4. Vantagens e limitações das redes neuronais artificiais	126
5.Potencialidades da Descoberta de Conhecimento em Bases de Dados na Medicina Intensiva	128

6. Bibliografia	130
-----------------------	-----

CAPÍTULO IV- INDUÇÃO DE MODELOS DE PREVISÃO DE MORTALIDADE E DA FALÊNCIA DE ÓRGÃOS E SISTEMAS NO DOENTE GRAVE: CONTRIBUIÇÃO PESSOAL 141

1. Introdução.....	141
--------------------	-----

2. População e métodos.....	142
-----------------------------	-----

3. Indução de modelos de previsão da mortalidade intra-hospitalar.....	146
--	-----

3.1. Preâmbulo	146
----------------------	-----

3.2 <i>Mortality assessment in intensive care units via adverse events using artificial neural networks...</i>	152
--	-----

4. Indução de modelos de previsão de falência de órgãos e sistemas.....	173
---	-----

4.1. Preâmbulo	173
----------------------	-----

4.2 <i>Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks</i>	177
--	-----

4.3 Preâmbulo	186
---------------------	-----

4.4 <i>Rating organ failure via adverse events using data mining in intensive care medicine</i>	187
---	-----

4.5 Preâmbulo	205
---------------------	-----

4.6 <i>Organ Failure Prediction Based On Clinical Adverse Events: A Cluster Model Approach</i>	209
--	-----

5. Viabilização de sistemas de suporte à decisão inteligente em Medicina Intensiva	222
--	-----

5.1. Preâmbulo	222
----------------------	-----

5.2 <i>INTCARE: A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine</i>	224
---	-----

6. Bibliografia	247
-----------------------	-----

CAPÍTULO V- DISCUSSÃO GLOBAL..... 251

1. Bibliografia.....	263
----------------------	-----

CAPÍTULO VI - CONCLUSÕES..... 269

Resumo	271
--------------	-----

Resumé	274
--------------	-----

Abstract	278
----------------	-----

CAPÍTULO I

Introdução

CAPÍTULO I – Introdução

Agradecimentos

Ao Prof. Doutor Lopes Gomes, meu orientador de doutoramento, a quem devo o estímulo para a realização deste trabalho de investigação.

Ao Prof. Doutor Maia Neves e Prof. Doutor Filipe Santos por terem co-orientado este trabalho, num esforço de complementaridade entre as áreas científicas da Medicina Intensiva e da Informática.

Aos Professores Doutores Joaquim Maia e Henrique Barros por me terem iniciado na investigação científica e ensinado a trilhar este caminho enriquecedor da prática clínica.

À Direcção do Serviço de Cuidados Intensivos do Hospital Geral de Santo António por me ter proposto a elaboração desta tese de doutoramento.

Ao Dr. Reis Miranda pelo convite que me dirigiu, conjuntamente com o então director de serviço, Dr. Manuel Brandão, para participar activamente no projecto EURICUS-II e aceder à base de dados resultante desse projecto.

Ao Prof. Dr. Vítor Ribeiro, pelo seu incondicional apoio e disponibilidade.

Ao Prof. Doutor Nuno Grande, pela sua isenção, sentido de justiça e apoio incomensurável e imprescindível.

Ao Prof. Dr. Caetano Pereira por ter investido, com amizade, no ultrapassar de contrariedades.

Ao meu colega e amigo Dr. Pedro Amorim que acreditou sempre na possibilidade da concretização deste projecto e me apoiou desde o primeiro momento.

Ao amigo Dr. Pedro Sousa e Silva, por ter tornado possível a concretização deste projecto. Sem a sua amizade, disponibilidade e investimento profissional, nada teria sido possível.

Ao amigo Eng.º Pedro Gago, “companheiro de estrada”, pela sua enorme disponibilidade em apoiar este trabalho através dos seus comentários e incentivos. A sua cumplicidade e amizade foram sempre estímulos para percorrer este caminho.

Ao amigo Eng.º Fillipe Pinto, pela disponibilidade e interesse demonstrados.

Ao Prof. Doutor Paulo Cortez, pela forma empenhada como disponibilizou os seus conhecimentos para que este trabalho tivesse sucesso.

Aos meus colegas do Serviço de Cuidados Intensivos do Hospital Geral de Santo António.

Aos meus pais, à minha mulher e aos meus filhos pelo partilhar de emoções e pelas capacidades inesgotáveis de incentivo, apoio e carinho.

1.Âmbito e motivação

Apresenta-se a introdução à dissertação, motivos que levaram à escolha do tema, objectivos propostos e organização do documento.

Entre os desenvolvimentos mais marcantes da Medicina do Século XX surgiu a Medicina Intensiva, definida como uma área multidisciplinar das Ciências Médicas que aborda especificamente a prevenção, diagnóstico e tratamento de situações de doença aguda potencialmente reversíveis, em doentes que apresentam falência de uma ou mais funções vitais, eminente (s) ou estabelecida (s).

A Medicina Intensiva aumentou a capacidade de salvar vidas em risco e a sobrevida de doentes portadores de doença grave.

Resultado dos desenvolvimentos da tecnologia, temos assistido na Medicina Intensiva ao crescimento sustentado da quantidade de dados¹ disponíveis à cabeceira do doente grave.

A dimensão dos recursos, a complexidade crescente das doenças, a introdução de novas tecnologias de monitorização do doente e a progressiva informatização das Unidades de Cuidados Intensivos (UCIs), via implementação de sistemas de informação² clínica, têm contribuído para que a UCI seja um ambiente gerador de um volume exponencial de informação³ e à formação de bases de dados extensas (Hanson & Marshall, 2001).

Existem bases de dados com mais de duas mil variáveis sobre um doente (Imhoff, 1998) e estima-se que um intensivista, durante a visita clínica, possa ser confrontado com o desafio de analisar e integrar mais de duzentas variáveis respeitantes a um doente grave (Morris & Gardner, 1992). Mesmo um intensivista experiente muitas

¹ Representação da informação sob uma forma convencional adequada à comunicação, à interpretação ou ao processamento. APDSI-Associação para a Promoção e Desenvolvimento da Sociedade de Informação. (2007). "Glossário da Sociedade da Informação – Versão 2007."

² Sistema constituído por recursos humanos, recursos materiais (o equipamento) e procedimentos que possibilitam a aquisição, o armazenamento, o processamento e a difusão da informação pertinente ao funcionamento de uma organização, quer o sistema esteja informatizado ou não. Ibid.

³ Dados e factos que foram organizados e comunicados de forma coerente e com significado e a partir dos quais se podem tirar conclusões. Ibid.

vezes tem grandes dificuldades em elaborar uma resposta sistematizada perante um problema que envolva mais do que sete variáveis (Miller, 1956).

Nas UCIs, o doente está rodeado por múltiplos equipamentos para monitorização, diagnóstico e terapêutica, que obviamente contribuem para a geração de mais dados. Entre estes equipamentos incluem-se os monitores dos sinais vitais. A monitorização dos sinais vitais é uma das actividades fundamentais na UCI e tem por objectivo detectar as alterações clínicas, o mais precocemente possível, com o intuito de prevenir qualquer deterioração adicional do estado clínico do doente. A importância desta prática é inquestionável e é enfatizada em muitos manuais e regulamentação publicada (Makivirta et al., 1991, Medicine, 1988a, Medicine, 1988b, Medicine., 1997). Estes monitores de sinais vitais têm associada a função alarme, gerando sinais acústicos e visuais quando os valores dos sinais vitais monitorizados ultrapassam os limites de normalidade estabelecidos e, consequentemente, alertando os prestadores de cuidados de saúde para uma alteração do estado clínico do doente. Estes desvios, relativamente aos limites de normalidade dos valores de quatro parâmetros monitorizados por rotina nas UCIs (ex.: frequência cardíaca, saturação arterial de O₂, tensão arterial e débito urinário horário), que nesta dissertação são denominados por eventos clínicos adversos, poderão proporcionar a detecção precoce de falência de órgãos.

A eficácia destes sistemas depende da sensibilidade e especificidade dos alarmes, assim como da resposta dos prestadores de cuidados de saúde aos alarmes.

Apesar da relevância atribuída à monitorização destes eventos clínicos adversos (Jones et al., 1991, Lawless, 1994, Tsien & Fackler, 1997), a sua interpretação e valorização não tem resultado em práticas normalizadas e também não se tem traduzido em conhecimento⁴ “novo” (Chambrin et al., 1999). Chambrin et al. revelaram que apenas 6% dos eventos clínicos adversos eram valorizados pelos clínicos, e que em apenas 2/3 destes casos conduziram a alterações terapêuticas.

Acresce ainda, que à medida que o número de aparelhos de monitorização aumenta nas UCIs, o número de falsos alarmes ou clinicamente interpretados como não relevantes aumenta (O'Carroll, 1986).

Estudos prévios, da década de noventa, revelaram que em algumas situações mais de 90% dos eventos clínicos adversos eram falsos ou clinicamente irrelevantes, não determinando, por isso, qualquer intervenção clínica (Biot et al., 2000, Chambrin et al., 1999, Kacmarek & Tobin, 1998, Koski et al., 1990, Lawless, 1994, Tsien &

⁴ Colecção de factos, acontecimentos e regras, organizadas para uso sistemático. Ibid.

Fackler, 1997). Este problema é ainda mais notório na avaliação do efeito mensurável de uma intervenção terapêutica. O viés pessoal, a experiência e determinadas expectativas relativas a uma determinada estratégia terapêutica podem distorcer a capacidade de avaliação objectiva.

Podemos ainda pressupor que o desempenho clínico de uma UCI poderá ser influenciado pelas diferentes formas de integração da informação provida da monitorização nas actividades de cuidados diárias.

A falta de uma integração efectiva dos eventos clínicos adversos e da sua representação sob a forma de conhecimento útil limita a sua utilidade no suporte à decisão clínica.

Uma vez que os intensivistas quando confrontados com informação clínica idêntica (ex.: mesmo tempo de ocorrência de um evento adverso clínico), agem de formas diferentes, existe a necessidade de introduzir na prática clínica métodos e ferramentas para apoiar a homogeneidade científica e a mensurabilidade das acções e decisões clínicas.

Os benefícios potenciais desta acção incluem a melhoria da qualidade do processo de cuidados de saúde do doente grave e uma redução dos custos associados.

Diariamente, a previsão do risco e da gravidade do doente admitido na UCI é feita de forma empírica. Na Medicina Intensiva, a previsão do risco desde sempre assumiu uma relevância específica, sendo uma das áreas da Medicina em que o desenvolvimento de índices de gravidade de doença e modelos de previsão do *outcome*⁵ final mais atenção mereceu por parte dos investigadores.

Embora os modelos de previsão do risco, como o SAPS II (Lemeshow & Le Gall, 1994), tenham sido introduzidos nas UCIs na década de oitenta e a sua utilização tenha vindo a ser cada vez mais adoptada, a sua aceitação na prática clínica é ainda controversa e não aceite de forma universal (Hyzg, 1995, Le Gall, 1995, Meyer et al., 1992, Teres & Lemeshow, 1994).

Por exemplo, a aplicação de diferentes modelos de previsão do risco de mortalidade hospitalar, num determinado doente, pode resultar em diferentes valores de probabilidades (Teres & Lemeshow, 1994), resultado da natureza probabilística destes métodos, assentes na modelação por regressão logística de um conjunto de variáveis independentes predefinidas. Devido à sua natureza probabilística,

⁵ Para efeitos deste trabalho, é entendido como um resultado mensurável do estado de saúde de um indivíduo ou de um grupo de indivíduos. Trata-se de uma consequência externa atribuída ao decorrer da evolução do estado do doente. Pode, por outro lado, ser interpretado como o conjunto mudanças no estado de saúde do doente. Podem considerar-se dois tipos de outcome: Intermédio - resultados que se seguem aos resultados iniciais; e final - resultado final (estado vital à data da alta hospitalar) Everitt, B. (2006). Medical Statistics from A to Z, Cambridge University Press.

oferecem a probabilidade de ocorrência de um determinado *outcome* e não a sua previsão absoluta, sendo a sua utilidade como sistema de suporte ao processo de decisão no doente individualizado limitada e não recomendada (Fahkry et al., 1994, Suter et al., 1994).

Para além destes modelos de previsão do risco, surgiram os índices de quantificação do grau de disfunção múltipla de órgãos, como o *SOFA score* (Vincent, de Mendonça et al. 1998), visando complementar a informação dos modelos de previsão de risco e monitorizar a evolução da gravidade do doente durante a sua estadia na UCI. Contudo, estes índices quantificam o grau de disfunção de órgão, de acordo com diferentes definições, incorporam a informação de forma descontínua por oposição a uma aquisição de dados em tempo útil e, por outro lado, não possuem capacidade preditiva.

Na prática clínica, a informação disponibilizada pela monitorização dos eventos clínicos adversos e pelos sistemas e modelos acima referidos é interpretada de forma desarticulada, não tirando proveito do conhecimento inerente às sinergias potencialmente existentes.

Outra questão, merecedora de destaque, é o fluxo de informação que habitualmente é adoptado na UCI. Apesar da riqueza de dados disponíveis à cabeceira do doente, na maioria dos casos a síntese e interpretação desta informação é feita de forma compartimentalizada e, conseqüentemente, subaproveitada no processo de decisão clínica (Hanson & Marshall, 2001).

Esta realidade complexa (volume e diversidade de dados) coloca desafios aos intensivistas para a integração e extracção de conhecimento ou informação potencialmente útil relativa ao prognóstico e à evolução da gravidade da (s) doença (s) com o objectivo do desenvolvimento e aplicação de instrumentos de suporte ao processo de decisão clínica.

Neste contexto, a evolução corrente no processo de decisão clínica tem mostrado uma abertura crescente relativamente à introdução de técnicas de análise inteligente de dados para a extracção de conhecimento, identificação de padrões, modelos de previsão e classificação, a partir de bases de dados clínicas (Kononenko, 2001).

A análise inteligente de dados (Berthold & Hand 2003) disponibiliza um conjunto de métodos e técnicas no âmbito de um processo semi-automático designado por Descoberta de Conhecimento em Bases de Dados, viabilizando a construção de sistemas de apoio à decisão inteligente (Fayyad et al., 1996b, Santos & Azevedo, 2005).

Um sistema de suporte à decisão clínica define-se como um sistema computacional desenhado para apoiar uma ou mais fases do processo de tomada de decisão, seja

ele individual, em equipa, organizacional ou até inter-organizacional (Forgionne et al., 2000).

A inclusão de técnicas cuja ênfase reside na aquisição de conhecimento a partir dos dados, de forma automática, através da aplicação de técnicas de *Data Mining*⁶, introduziu um dos componentes inteligentes dos sistemas de suporte à decisão (Hanson & Marshall, 2001).

Quase desde o aparecimento dos primeiros computadores que têm sido propostas ferramentas das tecnologias da informação para o suporte ao processo de decisão (Ledley & Lusted, 1959).

Os primeiros sistemas foram desenhados com o objectivo de sistematizar a investigação sob um espectro de cenários possíveis, com o intuito de gerar hipóteses diagnósticas, a partir dos dados introduzidos pelos clínicos no computador. Um exemplo clássico é o sistema de diagnóstico de alterações electrolíticas de Bleich (Bleich, 1969).

Outra estratégia, cujos pioneiros foram Warner et al (Warner et al., 1961), baseava-se na realização de raciocínios probabilísticos. Dombal desenvolveu um sistema de suporte à decisão diagnóstica de apendicite aguda, que foi validado e utilizado na prática clínica (De Dombal et al., 1969).

Na década de oitenta foram desenvolvidos programas que pretendiam simular as capacidades diagnosticas dos clínicos, e que foram denominados de sistemas periciais, com a pretensão de que iriam ser capazes de superar a capacidade dedutiva dos clínicos. Como exemplos destes sistemas periciais surgiram o MEDAS, orientado para o diagnóstico de múltiplas patologias (Ben-Bassat et al., 1980) e o HEME para o diagnóstico de doenças hematológicas desenvolvido por Lipkin et al (Lipkin & Hardy, 1957). Surgiram ainda outros sistemas baseados na aproximação heurística como o INTERNIST – 1 de Pople (Pople, 1982) que abarca cerca de seiscentos diagnósticos e uns quatro mil e quinhentos sintomas e sinais, ou o MYCIN de Shortliffe (Shortliffe, 1976) que sugere terapêuticas antimicrobianas.

Outros tipos de sistemas baseados em regras heurísticas são os que geram avisos sugerindo ao clínico a necessidade de realizar uma tarefa concreta, quando um doente cumpre uma série de critérios. Um exemplo destes sistemas é o HELP de Gardner et al (Gardner et al., 1999).

O fracasso da introdução na prática clínica destes sistemas periciais, mesmo quando demonstravam ser capazes de gerar hipóteses correctas, revelou muitos

⁶ Extração não trivial de informação implícita, previamente não conhecida e potencialmente útil a partir dos dados Ibid.

aspectos tanto práticos como formais que condicionam a tomada de decisão clínica (Bonis et al., 2004).

Na década de noventa, o entendimento comum era que estes sistemas periciais, baseados na aquisição de informação obtida de peritos não eram suficientes para os problemas complexos da prática clínica (Lavrac et al., 1997). Foram-lhe reconhecidas deficiências graves, não eram capazes de lidar com variações da apresentação clínica das patologias, em termos do espectro de achados bem como da sua gravidade. Não conseguiam lidar com a evolução da doença ao longo do tempo, por exemplo: face à resposta à terapêutica instituída; reconhecer como uma patologia poderia influenciar a apresentação de outra; ou disponibilizar informação com base na fisiopatologia (Bonis et al., 2004).

O trabalho para “alimentar” o sistema com os dados requeridos era demorado e de difícil execução, o que na prática fazia com que o clínico demorasse mais tempo a introduzir toda a informação do que a estabelecer o diagnóstico pelos seus próprios meios (Bonis et al., 2004).

Mais ainda, o conhecimento novo não era rapidamente incorporado sem interrupções potenciais do modelo, violando o princípio da adaptabilidade, característica dos sistemas inteligentes (Schwartz et al., 1987).

Perante este cenário, foi induzida uma viragem no desenho destes sistemas, onde a principal ênfase foi colocada na aquisição de conhecimento directamente dos dados, usando uma análise inteligente dos dados (Kalogeropoulos et al., 2003).

No final da década de noventa, esta abordagem foi ganhando um interesse crescente, particularmente na área da Medicina Intensiva, devido ao volume e complexidade dos dados clínicos.

Actualmente os sistemas de suporte à decisão clínica são menos ambiciosos mas procuram aproveitar sinergicamente as potencialidades dos sistemas computacionais e dos peritos médicos (Hanson & Marshall, 2001).

A possibilidade de construir sistemas de suporte à decisão clínica inteligentes, baseados em modelos de previsão induzidos a partir dos dados gerados à cabeceira do doente através da Descoberta de Conhecimento em Bases de Dados, é bastante atractiva e constitui a principal motivação deste trabalho.

Esta dissertação é um exemplo de um trabalho transversal dado o seu carácter interdisciplinar, incorporando conhecimentos da:

Medicina Intensiva – mais concretamente as áreas de investigação do *outcome* clínico e da previsão do risco de disfunção de órgãos e mortalidade, entendido como um processo continuo onde a actuação pró-activa é resultado da identificação

atempada do compromisso fisiopatológico do doente grave, reflectindo potencialmente a qualidade dos processos de cuidados de saúde;

Descoberta de Conhecimento em Bases de Dados e *Data Mining* – a partir de bases de dados desenvolvidas para esta dissertação, seguindo a metodologia inerente ao processo de *Data Mining* e recorrendo a técnicas de Aprendizagem Automática⁷. Por exemplo, redes neuronais artificiais (Haykin, 1999), árvores de decisão (Quinlan, 1986) e indução de regras de decisão (induzir modelos de classificação, regressão e segmentação).

2. Objectivos

Esta dissertação ambiciona a investigação de abordagens holísticas, originais e multidisciplinares na Medicina Intensiva para:

1. Induzir modelos de previsão do *outcome* final (mortalidade intra-hospitalar) do doente grave admitido na UCI;
2. Induzir modelos de evolução da gravidade da doença crítica, através da previsão da falência de órgãos e sistemas, durante a estadia do doente na UCI;
3. Avaliar o impacto do grau de disfunção fisiológica de quatro parâmetros monitorizados por rotina na UCI: frequência cardíaca, tensão arterial sistólica, saturação arterial de O₂ e débito urinário horário, aqui definidos como eventos clínicos adversos sob evolução na falência de órgãos e *outcome* final do doente grave internado na UCI;
4. Aferir a contribuição ou influência que cada um dos eventos clínicos adversos tem nos modelos de previsão do *outcome* e falência de órgãos;

Este trabalho pretende assim viabilizar a construção de sistemas inteligentes de apoio à decisão clínica, baseados nestes modelos de previsão, potenciando uma

⁷ Processo pelo qual uma unidade funcional melhora o seu desempenho, adquirindo novos conhecimentos e aptidões ou reorganizando os já existentes. APDSI-Associação para a Promoção e Desenvolvimento da Sociedade de Informação. (2007). "Glossário da Sociedade da Informação – Versão 2007."

actuação atempada e pró-activa dos prestadores de cuidados de saúde e contribuindo para o incremento da qualidade do processo de cuidados de saúde.

Estes modelos podem ser considerados holísticos na medida em que, contrariamente às visões mais clássicas, exploram o problema como um todo. Trata-se de um novo paradigma que surge como alternativa às aproximações baseadas na fragmentação física ou lógica dos sistemas. O resolver e analisar o complexo como um todo torna-se hoje possível graças a abordagens holísticas disponibilizadas pela inteligência artificial⁸.

Estas são entendidas como inovadoras no sentido em que faz uso duma abordagem baseada na descoberta de conhecimento que recorre a técnicas oriundas da área da Aprendizagem Automática e do *Data Mining*. Trata-se de uma forma de criar os modelos que contrasta com as aproximações mais empregues na investigação médica.

Por último, a interdisciplinaridade deste trabalho é um facto que deve ser realçado na medida em que foi desenvolvido um esforço muito grande na busca e integração de conhecimentos oriundos de áreas tão diversas como a Medicina Intensiva e a inteligência artificial.

Para a prossecução destes objectivos, foi dada particular atenção aos seguintes objectivos específicos:

Dinamismo – abordagem dinâmica, por oposição aos métodos habituais caracterizados pela análise única (à data da admissão à UCI ou sequencial de variáveis predefinidas), permitindo a incorporação de dados biométricos monitorizados por rotina;

Adaptabilidade – a abordagem do problema em estudo possui a característica da adaptabilidade, isto é, baseia-se nos dados adquiridos ao longo do tempo na UCI, de forma a preservar automaticamente a acuidade dos modelos, evitando assim a perda de calibração.

⁸ Disciplina que procura compreender o comportamento inteligente (raciocínio, pensamento e aprendizagem) no sentido lato, utilizando computadores para reproduzi-lo e para construir máquinas que se comportem de forma inteligente, independentemente do mecanismo subjacente Everitt, B. (2006). Medical Statistics from A to Z, Cambridge University Press.

Previsão – capacidade de identificação e previsão precoces, atempadas e ajustadas da disfunção de órgãos de forma a proporcionar a estratégia terapêutica preventiva.

Para o desenvolvimento dos modelos foram utilizadas as técnicas de amostragem baseadas na validação cruzada e na partição da amostra. Os resultados alcançados foram avaliados através dos seguintes processos e técnicas:

Modelos de previsão do *outcome* final – a avaliação do desempenho e poder discriminativo destes modelos baseou-se em: matriz de confusão, precisão, sensibilidade, especificidade e capacidade discriminativa⁹ [área sob *receiver operating characteristic curve* (ROC)] (Hanley & Mc Neil, 1982).

A validação dos modelos foi feita através da comparação com os resultados obtidos pela regressão logística múltipla e pela pontuação obtida pelo modelo Simplified Acute Physiologic Score (SAPS II)(Le Gall et al., 1993).

Modelos de previsão da falência de órgãos e sistemas – a avaliação do desempenho e poder discriminativo destes modelos baseou-se em: matriz de confusão, precisão, sensibilidade, especificidade e a capacidade discriminativa [área sob *receiver operating characteristic curve* (ROC)] (Hanley & Mc Neil, 1982). A validação dos modelos foi feita através da comparação com os resultados obtidos por regressão logística múltipla e pela pontuação do *Sequential Organ Failure Assessment* (SOFA Score) (Vincent et al., 1998).

3.Organização da dissertação

Para além deste capítulo introdutório, esta dissertação inclui quatro capítulos cujo conteúdo é apresentado da seguinte forma:

Capítulo II – Através de um trabalho sistematizado de revisão, são apresentados os principais índices gerais de gravidade de doença, modelos de previsão do risco de mortalidade, bem como os sistemas de quantificação de falência múltipla de órgão e

⁹ Capacidade do modelo distinguir entre as observações com um *outcome* positivo ou negativo (sobreviventes e falecidos). É habitualmente avaliada pela área sob a *receiver operating characteristic* (ROC) *curve*, que desenha um gráfico da sensibilidade (proporção dos correctamente classificados como falecidos) versus 1 – especificidade (proporção dos correctamente classificados como sobreviventes Ibid).

respectivas limitações. São também abordados os conceitos, objectivos e as técnicas habitualmente utilizadas no desenvolvimento dos modelos prognósticos;

Capítulo III – O objectivo deste capítulo é o de abordar o processo de Descoberta de Conhecimento em Bases de Dados e técnicas do *Data Mining*, focalizando a atenção nas redes neuronais artificiais, as árvores de decisão e indução de regras de decisão. Apresentam-se ainda as técnicas de validação dos modelos;

Capítulo IV – A contribuição pessoal e os resultados alcançados neste trabalho materializam-se em três trabalhos já publicados aos quais se acrescenta um já submetido. A metodologia utilizada e as condições para concretizar a investigação são expostas em cada artigo;

Artigos originais publicados em revistas, capítulo de livro e actas de conferências internacionais:

Mortality Assessment in Intensive Care Units via Adverse Events Using Artificial Neural Networks

Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, José Neves.
Artificial Intelligence in Medicine (2006) 36, pp 223 – 234. Elsevier, ISSN: 0933-3657.

Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks

Á. Silva, P. Cortez, M.F. Santos, L. Gomes and J. Neves.
In I. Seruca, J. Cordeiro, S. Hammoudi and J. Filipe (Eds.), book chapter of Enterprise Information Systems VI (2006), pp. 127-134, Springer, The Netherlands,. ISBN: 1-4020-3674-4.

Rating Organ Failure via Adverse Events Using Data Mining In Intensive Care Medicine

Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, José Neves.
Submetido à revista Artificial Intelligence in Medicine em 2 de Maio de 2007.

Organ Failure Prediction Based On Clinical Adverse Events: A Cluster Model Approach

Álvaro Silva, João Pereira, Manuel Santos, Lopes Gomes, José Neves.
In Artificial Intelligence and Applications (AIA2003), 695-700, ACTA Press, 2003.

INTCare: A Knowledge discovery based intelligent decision support system for Intensive Care Medicine

Pedro Gago, Manuel Filipe Santos, Álvaro Silva, Paulo Cortez, José Neves, Lopes Gomes.

In Journal of Decision Systems, VOL 14 Nº3/2005, GUPTA Jatinder N.D., FORGIONNE Guisseppi, GARRIDO Leonardo, MORA Manuel (Eds.), Hermes Science publications, Lavoisier, 2005, pp.241-259, ISBN: 2-7462-1405-9.

Capítulo V – Discussão global mediante um processo de análise dos resultados obtidos dos trabalhos de investigação elaborados.

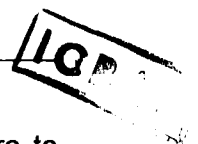
Capítulo VI – Conclusão e trabalho futuro. São apresentadas as conclusões do trabalho desenvolvido bem como as possíveis direcções de trabalho futuro que permitam complementar o trabalho iniciado nesta dissertação.

4. Bibliografia

- Ben-Bassat, M., Carlson, R.W., Puri, V.K., Davenport, M.D. & Schriver, J.A. (1980) Pattern-based interactive diagnosis of multiple disorders: the MEDAS system, *IEEE Trans Pattern Anal Mach Intel* PAM1-2, pp. 148-160.
- Berthold, M. & Hand, D.J. (2003) *Intelligent Data Analysis an Introduction*
- Biot, L., Carry, P.Y. & Perdrix, J.P. (2000) Évaluation clinique de la pertinence des alarmes en réanimation, *Ann Fr Réanim* 19, pp. 459-466.
- Bleich, H.L. (1969) Computer evaluation of acid-base disorders, *J Clin Invest* 1, pp. 145-148.
- Bonis, J., Sancho, J.J. & Sanz, F. (2004) Sistemas informáticos de soporte a la decisión clínica, *Med Clin (Barc)* 122(Supl 1), pp. 39-44.
- Chambrin, M.C., Ravaux, P., Calvelo-Aros, D., Jaborska, A., Chopin, A. & Boniface, B. (1999) Multicentric study of monitoring alarms in the adult intensive care unit (ICU): a descriptive analysis, *Intensive Care Med*, 25, pp. 1360-1366.
- De Dombal, F.T., Hartley, J.R. & Sleeman, D.H. (1969) A computer-assisted system for learning clinical diagnosis, *Lancet* 1, pp. 145-148.
- Fahkry, S.M., Rutledge, R., Rutherford, E., Albrecht, R., O'Meeghan, R. & Udekwu, O. (1994) Unexpected ICU survival in patients with high APACHE II scores: A multi-institutional study of 8,408 surgical patients from 5 hospitals, *Crit Care Med* 22(1), pp. 44.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996) The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, 39(11), pp. 27-34.
- Forgionne, G.A., M., M., F., C. & Kohli, R. (2000) Development of integrated decision making support systems: practical approach *Proceedings of the AIS Conference* (Long Beach, CA,
- Gardner, R.M., Pryor, T.A. & Warner, H.R. (1999) The HELP hospital information system, *Int J Med Inf* 54, pp. 169-182.
- Hanley, J.A. & Mc Neil, B.J. (1982) The meaning and use of the area under a receiver-operating characteristic (ROC) curve, *Radiology* 143, pp. 29-36.
- Hanson, C.W.r. & Marshall, B.E. (2001) Artificial intelligence applications in the intensive care unit, *Crit Care Med* 29, pp. 427-435.
- Haykin, S.S. (1999) *Neural networks : A comprehensive foundation*. (Upper Saddle River, NJ, Prentice Hall International).

- Hyzg, R.C. (1995) ICU scoring and clinical decision making [Editorial], *Chest*, 107(2), pp. 1482–1483.
- Imhoff (1998) Clinical data acquisition: what and how?, *Journal fur Anesthesie und Intensivmedizin* 5, pp. 85 -86.
- Jones, D., Lawson, A. & Holland, R. (1991) Integrated monitor alarms and alarm 'overload', *Anaesth Intensive Care* 19(1), pp. 101-102.
- Kacmarek, R. & Tobin, M.J. (1998) Alarms, in: I. McGraw-Hill (Ed) *Principle and Practice of Intensive Care Monitoring* (New York,
- Kalogeropoulos, D., A. , Carson, E.R. & Collinson, P.O. (2003) Towards knowledge-based systems in clinical practice: Development of an integrated clinical information and knowledge management support system, *Computer Methods and Programs in Biomedicine* 72, pp. 65-80.
- Kononenko, I. (2001) Machine learning for medical diagnosis: history, state of the art and perspective, *Artif Intell Med* 23, pp. 89-109.
- Koski, E.M., Makivirta, A., Sukuvaara, T. & Kari, A. (1990) Frequency and reliability of alarms in the monitoring of cardiac postoperative patients, *Int J Clin Monit Comput* 7, pp. 129–133.
- Lavrac, N., Keravnou, E. & Zupan, B. (1997) *Intelligent Data Analysis in Medicine and Pharmacology* (Boston,
- Lawless, S.T. (1994) Crying wolf: false alarms in a paediatric intensive care unit, *Crit Care Med* (22(6)), pp. 981-985.
- Le Gall, J. (1995) Customized probability models for early severe sepsis in adult intensive care patients, *JAMA* 273(8), pp. 644-650.
- Le Gall, J.R., Lemeshow, S. & Saulnier, F. (1993) A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study,, *JAMA*, 270, pp. 2957–2963.
- Ledley, R.S. & Lusted, L.B. (1959) Reasoning foundations of medical diagnosis, *Science*, 130, pp. 9-21.
- Lemeshow, S. & Le Gall, J.R. (1994) Modelling the severity of illness of ICU patients, A system update, *JAMA*, 272 (13), pp. 1049-1055.
- Lipkin, M. & Hardy, J.D. (1957) Differential diagnosis of hematological diseases aided by mechanical correlation of data, *Science* 125, pp. 551-552.
- Makivirta, A., Koski, E., A, K. & Sukuvaara, T. (1991) The median filter as a pre-processor for a patient monitor limit alarm system in intensive care. *Comput Methods Programs Comput Methods Programs Biomed* 34, pp. 139-144.
- Medicine, T.F.o.t.E.S.o.I.C. (1988a) Reports, Guidelines and Recommendations of the European Society of Intensive Care Medicine.Recommendations for

- critical care unit design, *Crit Care Med* 16(Reports, Guidelines and Recommendations of the European Society of Intensive Care Medicine), pp. 796-806.
- Medicine, T.F.o.t.E.S.o.I.C. (1988b) Reports, Guidelines and Recommendations of the European Society of Intensive Care Medicine.Recommendations for services and personnel for delivery of care in a critical care setting, *Crit Care Med* 16, pp. 809-811.
- Medicine., T.F.o.t.E.S.o.I.C. (1997) Reports, Guidelines and Recommendations of European Society of Intensive Care Medicine.Minimal requirements for Intensive Care Departments, *Intensive Care Med* 23, pp. 226-232.
- Meyer, A.A., Messick, W.J., Young, R., Backer, C.C., Fakhry, S., Muakkassa, F., Rutherford, E.J., Napolitano, L.M. & Rutledge, R. (1992) Prospective comparison of clinical judgement and APACHE II score in predicting the outcome in critically ill surgical patients, *J Trauma* 32, pp. 747-753.
- Miller, G.A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychological Review*, 63, pp. 81 - 97.
- Morris, A. & Gardner, R. (1992) Computer applications, in: M.-. Hill (Ed) *Principles of Critical Care* (New York, Hall, J,Schmidt, G Wood, L).
- O'Carroll, T.M. (1986) Survey of alarms in an intensive therapy unit., *Anaesthesia*, 41, pp. 742-744.
- Pople, H.E. (1982) Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnostics, in: e. Szolovits P (Ed) *Artificial intelligence in medicine* Boulder: Westview Press).
- Quinlan, J. (1986) Induction of Decision Trees. , *Machine Learning*, 1(1), pp. 81 -106.
- Santos, M. & Azevedo, C. (2005) *Data Mining – Descoberta de Conhecimento em Bases de Dados* FCA editores).
- Schwartz, W.B., Patil, R.S. & Szolovits, P. (1987) Board: artificial intelligence - where do we stand ? , *N. Engl. J. Med*, 316, pp. 685-688.
- Shortliffe, E.H. (1976) Computer-based medical consultations, in: E.C.S. Library (Ed) *MYCIN Artificial Intelligence Series* (New York,
- Suter, P., Argmandis, A. & Beaufils, F. (1994) Predicting outcome in ICU patients.Consensus Conference organized by the ESICM and SRLF., *Intensive Care Med*, 20, pp. 390-397.
- Teres, D. & Lemeshow, S. (1994) Why severity models should be used with caution, *Crit Care Clin* 10, pp. 93-110.
- Tsien, C.L. & Fackler, J.C. (1997) Poor prognosis for existing monitors in the intensive care unit, *Crit Care Med* 25(4), pp. 614-619.



- Vincent, J.L., de Mendonça, A. & Cantraine, F. (1998) Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study, *Crit Care Med*, 26, pp. 1793–1800.
- Warner, H.R., Toronto, A.F., Veasey, L.G. & Stephenson, R.A. (1961) Mathematical approach to medical diagnosis, *JAMA*, 177, pp. 75-81.

CAPÍTULO II

A avaliação do prognóstico e da falência de órgãos em Medicina Intensiva

CAPÍTULO II – A avaliação do prognóstico e da falência de órgãos em Medicina Intensiva

São apresentados os principais índices gerais de gravidade da doença, modelos prognósticos, sistemas de quantificação de falência de órgãos e as suas limitações.

1.Introdução

Os avanços tecnológicos, uma melhor compreensão dos processos fisiopatológicos do doente grave e a consequente expansão das UCIs têm contribuído para uma complexidade crescente da Medicina Intensiva e um incremento dos seus custos associados. Estes factos, entre outros, têm motivado o desenvolvimento e implementação de índices da gravidade da doença e modelos de prognóstico utilizáveis como sistemas de suporte à tomada de decisão clínica, para a avaliação do desempenho das UCIs (através da comparação das taxas de mortalidade normalizadas) e para a investigação(Ridley, 2002).

Os índices gerais de gravidade da doença e os modelos de prognóstico têm sido utilizados nas UCIs desde 1981, com a publicação do índice *Acute Physiology and Chronic Health Evaluation Score* (APACHE)(Knaus et al., 1981), a que se seguiram outros.

Inicialmente designados por índices gerais de gravidade da doença, estes instrumentos estratificam os doentes graves de acordo com a gravidade da sua doença, atribuindo a cada doente uma pontuação crescente assim que a gravidade da doença aumenta. O desenvolvimento posterior destes índices possibilitou que para além desta função de estratificação da gravidade da doença, fosse associada a capacidade de previsão do risco de mortalidade intra-hospitalar, passando a designar-se por modelos de prognóstico. Para este propósito, os modelos prognósticos utilizam um conjunto de variáveis prognósticas (predefinidas) e uma equação modeladora, geralmente obtida por técnicas de regressão logística múltipla.

Estes modelos estimam o risco da mortalidade à data da alta hospitalar de uma população de doentes, com um determinado grau de disfunção fisiológica, como se tivessem sido tratados numa UCI virtual usada para o desenvolvimento do modelo, independentemente da terapêutica recebida (Moreno & Afonso, 2006).

No momento actual, vários modelos prognósticos são utilizados tais como: o *Acute Physiology and Chronic Health Evaluation Score* (APACHE)(Knaus et al., 1981), o *Simplified Acute Physiology Score* (SAPS) (Le Gall et al., 1984) (Le Gall et al., 1993) e o *Mortality Probability Model* (MPM)(Teres et al., 1987).

Inicialmente desenvolvidos para a aplicação no doente individualizado, rapidamente foi constatado que estes modelos prognósticos de facto apenas deveriam ser utilizados em grupos heterogéneos de doentes (Teres et al., 1987).

A maioria dos modelos prognósticos é estática, dado que se baseiam no grau de disfunção fisiológica aguda avaliado à admissão ou nas primeiras 24 horas de internamento na UCI. Toda a informação após a admissão é excluída. Por outro lado face à sua natureza probabilística, a sua adopção como instrumento de apoio à decisão clínica à cabeceira do doente, é limitada. Na realidade, a sua utilização na prática clínica como suporte à decisão no doente individual é controversa, e não é recomendada (Suter et al., 1994).

Uma vez que os processos fisiopatológicos do doente grave são dinâmicos, porque dependentes da evolução da sua doença de base e das consequências destes nos vários órgãos e sistemas, surgiu a necessidade de desenvolver instrumentos que permitissem monitorizar a evolução clínica do doente ao longo do seu internamento na UCI.

Neste contexto, foram desenvolvidos os índices de quantificação de falência múltipla de órgãos, tais como: o *Sequential Organ Failure Assessment* (SOFA score) (Vincent, 1998)(Vincent, 1998), o *Multiple Organ Dysfunction Score* (MODS) (Marshall, Cook et al. 1995) e o *Logistic Organ Dysfunction Score* (LOD) (Le Gall, Klar et al. 1997) (Le Gall, Klar et al. 1996) entre outros.

A motivação subjacente ao desenvolvimento destes índices foi a necessidade de encontrar uma forma objectiva para descrever a falência de órgãos no doente individualizado. A maioria destes índices utiliza uma associação de variáveis fisiológicas com parâmetros analíticos (creatinina, bilirrubina).

Para além da quantificação da falência de órgãos e sistemas e da sua consequente correlação com o *outcome* final (mortalidade intra-hospitalar), estes sistemas têm sido utilizados pelos clínicos e investigadores para definir critérios de inclusão em ensaios clínicos.

Contudo, estes índices têm algumas limitações como instrumentos de apoio à decisão clínica, resultantes das diferentes definições utilizadas de falência de órgão, do critério de selecção dos órgãos e sistemas a avaliar, da dependência de algumas variáveis laboratoriais (não tão específicas quanto deveriam), da não disponibilidade da

avaliação em tempo real e da não previsão da evolução do grau de disfunção de órgão ou sistema.

2. O prognóstico na Medicina Intensiva

Os objectivos da Medicina Intensiva são os de diagnosticar, monitorizar e tratar doentes com doença grave e recuperá-los para o seu estado de saúde e de qualidade de vida prévios (Suter et al., 1994)..

Uma melhor compreensão da história natural e dos mecanismos fisiopatológicos do doente grave e a evolução contínua das técnicas terapêuticas e diagnósticas têm tornado o processo de cuidados de saúde cada vez mais complexo. Além disso, o uso crescente das tecnologias de monitorização e de suporte de função de órgãos vitais tem fomentado a procura e a consequente expansão das UCIs tendo como resultado o aumento dos custos associados (David, 1991).

Dados europeus e norte-americanos revelam que as UCIs representam pelo menos 7% das camas hospitalares e os seus custos consomem 15 a 20% dos orçamentos hospitalares (Hospital Statistics, 1990)

Acresce a constatação de que uma proporção crescente do orçamento da saúde é gasta numa pequena percentagem de doentes internados nas UCIs, durante os seus últimos dias de vida (Suter, 1994).

Na realidade, os avanços tecnológicos permitem hoje o suporte da vida de doentes considerados em estado terminal, por semanas ou meses, confrontando os intensivistas com novas questões de carácter ético e social para além da questão económica.

A avaliação precisa da gravidade da doença e a previsão do *outcome* poderão otimizar a utilização das UCIs, promovendo o uso apropriado dos recursos disponíveis.

Pelos motivos acima expostos, impõe-se cada vez mais uma alocação racional dos recursos, que são limitados, e um suporte eficaz e inteligente ao processo de decisão clínica no doente grave.

É cada vez mais reconhecido que o diagnóstico e as opções terapêuticas não se devem basear apenas em eventos isolados, mas também ser ponderados tendo em consideração os *outcomes* esperados, quer para o doente quer para a comunidade como um todo (Lucas & Abu-Hanna, 1999).

O curso e o *outcome* de uma doença são habitualmente expressos em termos de doente ou grupos de doentes, e não em termos de doença ou terapêutica (Abu-Hanna & Lucas, 2001)

A qualidade e a esperança de vida, a recorrência da doença, os custos económicos associados aos procedimentos e as limitações orçamentais deverão no seu conjunto ser avaliados no processo de decisão, atribuindo a cada um uma ponderação ou importância relativa.

Na avaliação da relação custo/benefício de um determinado procedimento clínico é de extrema importância o conhecimento sobre o que irá acontecer ao doente, isto é, o prognóstico (Lucas & Abu-Hanna, 1999).

Prognóstico (do grego *prognostikós*, indício do que deve acontecer) significa literalmente, conjectura sobre o que há-de suceder (Abu-Hanna & Lucas, 2001).

A perspectiva de recuperação de uma doença ou de um estado patológico é definida pela palavra prognóstico (Ohno-Machado et al., 2006).

A previsão do curso e do *outcome* de uma doença desempenha um papel importante nas tarefas associadas ao processo de cuidados de saúde do doente, tais como o diagnóstico e o planeamento terapêutico. Consequentemente, os modelos prognósticos constituem uma parte integrante de vários sistemas de suporte a estas tarefas (Lucas & Abu-Hanna, 1999).

O conceito chave subjacente ao prognóstico é a previsão de um evento antes da sua possível ocorrência. A elaboração de um prognóstico ocorre, por exemplo, quando um clínico estima um *outcome*, sem o conhecimento do *outcome* observado.

De qualquer forma, a essência do prognóstico é a de que o evento previsto ocorra num futuro relativo à informação disponível aquando do momento da previsão. O tempo está assim inerente ao conceito de prognóstico e distingue-o, por exemplo, do diagnóstico onde o futuro desempenha um papel menos relevante.

O prognóstico médico é aqui definido como: *a previsão do curso futuro de vários estados patológicos associados ou isolados, resultante das intervenções terapêuticas ou das suas histórias naturais* (Abu-Hanna & Lucas, 2001).

A complexidade do corpo humano e do ambiente que o rodeia (incluindo os efeitos das intervenções médicas) estão para além do que conseguimos compreender e prever. Contudo, a classificação e previsão dos casos em determinados padrões de progressão da doença poderá ajudar os prestadores de cuidados de saúde, assim como os doentes e as suas famílias, a seleccionar estratégias terapêuticas que considerem os custos e os benefícios potenciais (Ohno-Machado et al., 2006).

A informação diagnóstica e dos actos médicos, principalmente os terapêuticos, influenciam fortemente o prognóstico (Hilden & Habbema, 1987). Nesta perspectiva, o

valor da informação diagnóstica que não contribui para o prognóstico pode ser questionado e o mesmo se aplica ao valor da informação prognóstica que não contribui para a escolha da terapêutica (Abu-Hanna and Lucas 2001).

Devido ao interesse crescente da medicina baseada na evidência e ao papel central que o prognóstico desempenha nos cuidados do doente grave, será de esperar que o interesse dos modelos prognósticos continuará a crescer no futuro (Abu-Hanna & Lucas, 2001) e que o clínico veja o prognóstico como uma ferramenta de suporte à decisão (Windeler, 2000).

3. Índices gerais de gravidade da doença e modelos de prognóstico

O processo de classificação e previsão dos doentes em padrões particulares de progressão da doença é sistematizado e avaliado pelos denominados índices de gravidade e modelos prognósticos.

Há mais de duas décadas que os intensivistas e os investigadores têm desenvolvido vários índices de gravidade de doença e modelos prognósticos, num esforço para uma melhor utilização da experiência clínica colectiva na UCI e para responder às questões de eficácia, eficiência, equidade e qualidade da Medicina Intensiva.

Diferentes motivações têm conduzido os investigadores a elaborarem modelos prognósticos para vários *outcomes*, e muito trabalho tem sido desenvolvido na área da Medicina Intensiva para chegar a um consenso sobre quais os *outcomes* mais relevantes (Rubenfeld et al., 1999).

Entre os *outcomes* relevantes incluem-se: a mortalidade, o tempo de internamento na UCI ou no hospital, diferentes definições de disfunção de órgão e a qualidade de vida pós-UCI. A escolha destes *outcomes* resulta da importância atribuída pelos clínicos e pela facilidade com que são mensuráveis (Rosenberg, 2002).

Os índices gerais de gravidade de doença e os modelos prognósticos visam estratificar os doentes de acordo com a sua gravidade e estimar um determinado *outcome* (probabilidade de mortalidade intra-hospitalar), com base num conjunto de variáveis preditivas e numa equação modeladora. Permitem assim o cálculo do risco ajustado para o *case mix*¹⁰ e a descrição objectiva das populações de doentes admitidas nas UCIs, com a finalidade da normalização do *outcome* final, isto é, a razão entre a mortalidade observada e a estimada pelo modelo prognóstico (Rosenberg,

¹⁰ Conjunto de características dos doentes que descreve o resultado das unidades prestadoras de cuidados de saúde. Fetter(1980).

2002). A premissa fundamental destes sistemas consiste em: *as variáveis clínicas aquando ou próximo do momento de admissão à UCI estão associadas ao risco de mortalidade hospitalar* (Teres & Pekow, 2000).

Embora os nomes destes sistemas difiram (ex. previsão de risco, previsão de *outcome*, índices de gravidade da doença), têm objectivos semelhantes. Tentam quantificar objectivamente os distúrbios fisiológicos e comorbilidades do doente grave, a partir dos quais as estimativas da mortalidade são calculadas (Rosenberg, 2002)

Os modelos prognósticos tomam em consideração determinadas características dos doentes que potencialmente interferem com o *outcome* (mortalidade intra-hospitalar), independentemente da terapêutica recebida. Concebidos para serem aplicados a grupos heterogêneos de doentes, prevêm qual será a mortalidade agregada aquando da alta hospitalar, de um grupo de doentes com um determinado grau de disfunção fisiológica, como se fossem tratados numa UCI virtual, utilizada para o desenvolvimento dos diferentes modelos (Moreno & Afonso, 2006).. Como tal, estratificam os doentes em categorias de risco para mortalidade intra-hospitalar.

Apesar de denominados de gerais, nenhum dos modelos existentes é aplicável a todos os doentes graves. Estes sistemas excluem da sua análise os doentes queimados, os jovens (com menos de 16 ou 18 anos de idade), os doentes com doença coronária aguda (enfarte agudo do miocárdio), os doentes no pós-operatório de cirurgia de *by-pass* coronário e os com tempo de internamento na UCI muito curto (menos de 4 horas) (Moreno, 2000).

Um dos primeiros objectivos destes sistemas foi o de melhorar a qualidade dos cuidados de saúde (Hunt & Meyer, 1997) (Randolph et al., 1998a) e proporcionar uma estratificação do risco para estudar a ampla variação das taxas de mortalidade não-ajustadas entre os doentes graves (Hunt & Meyer, 1997).

Estes instrumentos contribuem para a compreensão da eficácia dos cuidados intensivos porque proporcionam estimativas do risco de mortalidade hospitalar, podendo ser um indicador para auditorias comparativas (isto é, comparação das taxas de mortalidade normalizadas), desde que interpretadas no contexto do sistema de cuidados de saúde, nas quais estão a ser utilizadas (Mackenzie et al., 2000).

Para além dos objectivos acima citados, os custos elevados associados ao processo de cuidados de saúde do doente grave (Oye & Bellamy, 1991) e o constatar de que muitos doentes viriam a morrer independentemente da terapêutica levou outros investigadores a usar estes modelos para definir futilidade terapêutica (Esserman et al., 1995); (Mendez-Tellez & Dorman, 2005).

Alguns investigadores usaram ainda os modelos de prognóstico para aperfeiçoar os critérios de selecção de doentes para ensaios clínicos (Knaus et al., 1984, Sibbald &

Vincent, 1995)(Vincent, 1998) e outros usaram-nos para suporte à decisão clínica, nos doentes em fim de vida (Jayes et al., 1993); (Teno et al., 1997); (Rubenfeld & Randall, 2001). Os índices de gravidade gerais e os modelos prognósticos para o doente grave estão entre os mais amplamente utilizados e testados no âmbito da Medicina (Ohno-Machado et al., 2006).

3.1 Descrição e classificação dos índices gerais de gravidade da doença e modelos prognósticos

Sendo o objectivo desta secção a descrição dos índices de gravidade e modelos prognósticos gerais, não serão discutidos os modelos aplicáveis apenas a populações específicas de doentes graves ou patologias.

O desenvolvimento dos índices gerais de gravidade de doença e dos modelos prognósticos teve início em 1981 com a publicação por Knaus et al do *Acute Physiology and Chronic Health Evaluation* (APACHE),(Knaus et al., 1981).

Desde então, foram surgindo vários outros, tais como o *Simplified Acute Physiology Score* (SAPS) (Le Gall et al., 1993) e o *Mortality Prediction Model* (MPM) (Teres et al., 1987). A avaliação destes sistemas confirmou a relação entre o grau de disfunção fisiológica aguda aquando ou próximo da admissão do doente na UCI e a previsão da mortalidade à data da alta hospitalar (Ohno-Machado et al., 2006). No início da década de 90, surgiram novas iterações dos modelos existentes utilizando bases de dados extensas oriundas da Europa e dos Estados Unidos da América, tendo sido desenvolvida a terceira versão do APACHE (APACHE III), (Knaus et al., 1981) a segunda versão do SAPS (SAPS II) (Le Gall et al., 1993) e o MPM II (Lemeshow et al., 1993). Foi demonstrado que estas versões mais recentes têm um melhor desempenho do que as anteriores (Castella et al., 1995).

Os desenvolvimentos mais recentes destes modelos ocorreram no início do século XXI, com a publicação de novos modelos prognósticos: o APACHE IV (Zimmerman et al., 2006), o SAPS III (Metnitz et al., 2005) (Moreno et al., 2005) e o MPM III (Higgins et al., 2005). Estes modelos mais recentes estão ainda em fase de implementação.

De uma forma geral, podemos classificar os índices de gravidade e modelos prognósticos em subjectivos e objectivos, de acordo com a forma de selecção e combinação das variáveis preditivas do *outcome* de interesse (probabilidade mortalidade hospitalar). Nos sistemas subjectivos, estas escolhas são da responsabilidade de painéis de peritos. Nos modelos objectivos, a técnica de regressão logística é utilizada para a selecção de variáveis preditivas e/ou para a

construção da equação preditiva de mortalidade (descrita na secção 3.2) (Le Gall, 2005). A maioria dos índices de gravidade gerais e modelos prognóstico são calculados a partir dos dados registados durante o 1º dia do doente na UCI. Neste grupo estão incluídos: o *Acute Physiology and Chronic Health Evaluation* (APACHE); o *Simplified Acute Physiology Score* (SAPS) e o *Mortality Prediction Model* (MPM).

Estes índices e modelos prognósticos estratificam a gravidade da doença e estimam o risco de mortalidade intra-hospitalar com base nas variáveis preditivas registadas durante o 1º dia de internamento na UCI.

3.1.1 Índices de gravidade gerais e modelos prognósticos do 1º dia – subjectivos

Os índices subjectivos foram desenvolvidos por grupos de peritos que seleccionaram as variáveis preditivas e os pesos respectivos, utilizando o julgamento clínico e as relações fisiológicas documentadas na literatura.

Para cada variável é definida uma escala de normalidade. Se dentro dos limites estabelecidos é pontuada com 0. Quanto mais anormal o resultado, maior a pontuação atribuída. O somatório de pontos atribuídos às diferentes variáveis constitui o índice de gravidade para cada doente (Le Gall, 2005)..

Estes sistemas têm por objectivo a classificação do doente grave em categorias de gravidade de doença de acordo com o valor do índice.

Por ordem cronológica de publicação os principais índices de gravidade deste grupo são: o APACHE, O SAPS I e o APACHE II. (Quadro 1)

3.1.1.1 Acute Physiology and Chronic Health Evaluation (APACHE)

O APACHE foi desenvolvido por Knaus *et al* no *George Washington University Medical Center* em 1981, utilizando uma base de dados com 805 doentes. (Quadro 1)

Foi o primeiro índice de gravidade, expressamente desenvolvido para as UCIs, a utilizar um conjunto de variáveis fisiológicas registadas por rotina, constituindo uma tentativa de criação de um instrumento objectivo que permitisse a comparação de *outcomes*, a avaliação de novas terapêuticas e do desempenho das UCIs (Ohno-Machado *et al.*, 2006).

O APACHE contempla sete sistemas: cardiovascular, renal, gastrointestinal, respiratório, hematológico, metabólico e neurológico. As variáveis foram seleccionadas

por um painel de peritos que desenvolveram uma escala de gravidade de 0 a 4, de acordo com o grau de anormalidade (incorporando trinta e quatro variáveis). Para o cálculo do índice são considerados os piores valores de cada variável registados nas primeiras 32 horas após a admissão na UCI.

O sistema APACHE demonstrou ser capaz de estratificar de forma fiável e generalizável a gravidade dos doentes graves (Knaus et al., 1982) (Wagner et al., 1984). Evidenciou ainda uma boa correlação com a mortalidade e um bom desempenho na modelação de *outcomes* para um conjunto de diagnósticos específicos: cardiovasculares, neurológicos, respiratórios e gastrointestinais (Wagner et al., 1983).

3.1.1.2 Simplified Acute Physiologic Score (SAPS)

O primeiro modelo *Simplified Acute Physiology Score* (SAPS) foi desenvolvido por Le Gall et al em 1984, utilizando um conjunto de parâmetros fisiológicos similares aos do APACHE (Le Gall et al., 1984). (Quadro 1)

O SAPS engloba treze variáveis fisiológicas e a idade. São registados os piores valores das primeiras 24 horas após a admissão na UCI. Para a atribuição da pontuação é usada uma escala de 0 a 4, similar à do APACHE.

O índice foi validado numa amostra de validação. A sua capacidade discriminativa avaliada pela área sob a curva *ROC – receiver operating characteristic curve* (Hanley & Mc Neil, 1982) foi de 0,85. Não há referência aos testes de calibração (Ohno-Machado et al., 2006). As definições de capacidade discriminativa e calibração dos modelos estão descritas na secção 3.3.

3.1.1.3 Acute Physiology and Chronic Health Evaluation - APACHE II

Em 1985, os investigadores do APACHE publicaram a segunda iteração do modelo, o APACHE II, desenvolvido a partir de uma base de dados com 5815 admissões consecutivas colhidas em treze hospitais norte-americanos (Knaus et al., 1985a) (Quadro 1). Representou um passo na tentativa de simplificar o modelo original.

O APACHE II utiliza o pior valor nas primeiras 24 horas de doze variáveis fisiológicas (pontuáveis entre 0 e 4 - o *Acute Physiology Score*), a idade do doente, a situação de insuficiência orgânica ou de imunodeficiência prévias à admissão hospitalar, e a principal categoria de admissão na UCI.

O APACHE II, para além de permitir o cálculo de um índice de gravidade, introduziu a possibilidade de prever a probabilidade de mortalidade do doente grave, impondo a

selecção da causa principal de admissão na UCI a partir de uma lista de cinquenta diagnósticos e recorrendo a uma equação de regressão logística que transforma a pontuação obtida numa probabilidade de mortalidade intra-hospitalar. (Knaus, A. et al. 1985)

O APACHE II evidenciou uma boa capacidade discriminativa na amostra de validação (área sob a curva *ROC* = 0.863) (Ohno-Machado et al., 2006).

Um número elevado de estudos avaliou a validade externa do APACHE II (Jacobs, 1987); (Giangiuliani et al., 1989); (Chisakuta & Alexander, 1990) (Teskey et al., 1991); (Turner et al., 1991); (Wong et al., 1995)..

Globalmente, o poder discriminativo do modelo era bastante bom no contexto de um conjunto extenso e heterogéneo de comorbilidades, dados demográficos e geográficos, mas vários estudos revelaram uma variabilidade acentuada da sensibilidade e da especificidade do modelo face à presença de diferentes comorbilidades (Lloyd-Thomas et al., 1988) (Abbott et al., 1991); (Edwards et al., 1991); (Headley et al., 1992); (Ludwigs & Hulting, 1995); (Van, 1995) ;(Chiang et al., 2001);(Chiavone & Sens, 2003).

Quadro 1: Índices gerais de gravidade de doença e modelos prognósticos do 1º dia na UCI – subjectivos. Adaptado de (Moreno, 2000).

Características	APACHE¹	SAPS²	APACHE II³
Ano	1981	1984	1985
Países participantes	1	1	1
UCIs participantes	2	8	13
Número de doentes	805	679	5815
Seleção das variáveis e pesos atribuídos	Painel de peritos	Painel de peritos	Painel de peritos
Variáveis:			
Idade	Não	Sim	Sim
Origem do doente	Não	Não	Não
Estado cirúrgico	Não	Não	Sim
Estado de saúde crónico	Sim	Não	Sim
Variáveis fisiológicas agudas	Sim	Sim	Sim
Diagnóstico agudo	Não	Não	Sim ^a
Tempo de registo	Nas primeiras 32 horas após a admissão à UCI	Nas primeiras 24 horas após a admissão à UCI	Nas primeiras 24 horas após a admissão à UCI
Número de variáveis	34	14	17
Pontuação	Sim	Sim	Sim
Equação preditiva de mortalidade	Não	Não	Sim

¹: APACHE I, Acute Physiology and Chronic Health Evaluation.

²: SAPS I, Simplified Acute Physiologic Score.

³: APACHE II, Acute Physiology and Chronic Health Evaluation, segunda iteração.

^a: escolhido de uma lista de 50 diagnósticos de admissão.

3.1.2 Índices de gravidade gerais e modelos prognósticos do 1º dia – objectivos

No desenvolvimento dos índices objectivos, são utilizadas técnicas de regressão logística múltipla e o parecer de peritos para a selecção das variáveis, dos limites de normalidade e dos respectivos pesos atribuídos. Todos resultam de um modelo de regressão logística que estima a probabilidade de mortalidade intra-hospitalar, através de uma equação preditiva. No seu desenvolvimento foram utilizadas bases de dados multicêntricas.

De uma forma geral, as variáveis registadas podem ser classificadas em quatro grupos: idade, comorbilidades, grau de disfunção fisiológica aguda e o diagnóstico de admissão. Alguns sistemas introduziram variáveis destinadas a minimizar o impacto do

*lead-time bias*¹¹ (Le Gall, 2005). Os principais modelos objectivos são: o APACHE III (Knaus et al., 1991), o SAPS II (Le Gall et al., 1993), o MPM II (Lemeshow et al., 1994), o SAPS III (Metnitz et al., 2005); (Moreno et al., 2005), o APACHE IV (Zimmerman et al., 2006) e o MPM III (Higgins et al., 2005).

Quadro 2 – Principais características dos modelos de prognóstico objectivos (do 1º dia na UCI) Adaptado de (Suter, Argmandis et al. 1994).

Modelo	Tempo de registo	Variáveis	Base de dados utilizada	Avaliação
APACHE III	Nas primeiras 24 horas após a admissão à UCI	Diagnóstico principal (79 categorias) 16 variáveis fisiológicas Idade Estado de saúde crónico (7 categorias) Tipo de admissão (3 categorias) Origem do doente (9 categorias)	17440 doentes de 40 hospitais norte-americanos Amostra de desenvolvimento: 8720 casos Amostra de validação: 8720 casos	Calibração: não publicada Discriminação: área sob a curva ROC = 0,90 para a população total Especificidade: 96,3%
MPM II ₀	Aquando da admissão à UCI	3 variáveis fisiológicas Doença crónica (3 categorias) Diagnóstico agudo (5 categorias) Idade PCR pré-admissão à UCI Ventilação mecânica Tipo de admissão (3 categorias)	19124 doentes dos EUA, Canadá e Europa Amostra de desenvolvimento: 12610 casos Amostra de validação: 6514 casos	Calibração, <i>goodness-of-fit tests</i> ¹² . Amostra de desenvolvimento: p= 0,327 Amostra de validação: p= 0,623 Discriminação: área sob a curva ROC: Amostra de desenvolvimento: p= 0,837 Amostra de validação: p= 0,824 Especificidade e sensibilidade: não publicada

¹¹ O efeito da terapêutica prévia no grau de disfunção fisiológica presente aquando da admissão na UCI. Habitualmente avaliado indirectamente pela localização do doente no hospital antes da admissão à UCI. Everitt, B. (2006). *Medical Statistics from A to Z*, Cambridge University Press.

¹² Testes de qui-quadrado propostos por Hosmer e Lemeshow para a avaliação formal da calibração dos modelos de previsão. Traduz o grau de concordância entre um conjunto de observações e os valores correspondentes previstos pelo modelo. Ibid.

Quadro 2 – Principais características dos modelos de prognóstico objectivos (do 1º dia na UCI) Adaptado de (Suter et al., 1994). (cont.)

Modelo	Tempo de registo	Variáveis	Base de dados utilizada	Avaliação
SAPS II	Nas primeiras 24 horas após a admissão à UCI	12 variáveis fisiológicas Diagnóstico crónico (3 categorias) Idade Tipo de admissão (3 categorias)	13152 doentes dos EUA e Europa Amostra de desenvolvimento: 8369 casos Amostra de validação: 4783 casos	Calibração, <i>goodness of fit tests</i> : Amostra de desenvolvimento: $p = 0,883$ Amostra de validação: $p = 0,104$ Discriminação: área sob a curva ROC ^b : Amostra de desenvolvimento: $p = 0,88$ Amostra de validação: $p = 0,86$ Especificidade e sensibilidade: não publicada
SAPS III	Entre 1 hora antes e 1 hora após admissão	10 variáveis fisiológicas Idade Estado de saúde crónico Fármacos vasoactivos pré-admissão UCI Origem do doente Tempo de internamento pré-admissão à UCI Motivo de admissão à UCI (10 categorias) Local anatómico da cirurgia (5 categorias) Infecção à admissão (2 categorias)	16784 doentes de vários países	Calibração: teste C ¹³ de HL ^a = 14,29, $p = 0,16$; Teste H ¹⁴ = 10,56, $p = 0,39$ Discriminação: área sob a curva ROC = 0,848
APACHE IV	Nas primeiras 24 horas após a admissão à UCI	Idade 16 Variáveis fisiológicas Doença crónica (7 categorias) Diagnóstico de admissão (116 categorias) Origem do doente (8 categorias) Tempo de internamento pré-admissão à UCI Tipo de cirurgia (2 categorias) Terapêutica trombolítica (2 categorias) Ventilação mecânica Escala de Coma de Glasgow modificada	110558 doentes dos EUA	Calibração: teste C de HL = 16,9, $p = 0,08$ Discriminação: área sob a curva ROC = 0,88 Especificidade e sensibilidade: não publicada
MPM IIIO	Aquando da admissão à UCI	3 variáveis fisiológicas Doença crónica (3 categorias) Diagnóstico agudo (5 categorias) Idade PCR pré-admissão à UCI Ventilação mecânica Tipo de admissão (3 categorias) Tempo antes da admissão à UCI	124855 doentes dos EUA, Canadá e Brasil	Calibração: teste HL = 11,62, $p = 0,31$; Discriminação: área sob a curva ROC = 0,823

^a Teste de Hosmer-Lemeshow (Hosmer & Lemeshow, 1989).^b Receiver operating characteristic curve (Hanley & Mc Neil, 1982).¹³ O teste C é semelhante ao H, mas os doentes são distribuídos por 10 grupos baseados nos decis das probabilidades estimadas.Ibid.¹⁴ No teste H, os doentes são divididos por 10 grupos, de acordo com as suas probabilidades estimadas. Então, é utilizado o qui-quadrado para a comparação do número de mortes e sobreviventes, observados e esperados em cada um dos 10 grupos.Ibid.

3.1.2.1 Acute Physiology and Chronic Health Evaluation (APACHE III)

Em 1991, Knaus et al publicaram uma terceira iteração do sistema APACHE, o APACHE III (Knaus et al., 1991) (quadro 2), desenvolvido a partir de uma base de dados com 17440 doentes provindos de quarenta UCIs dos Estados Unidos da América (EUA).

Foram excluídos do modelo os doentes com uma permanência na UCI inferior a quatro horas, os doentes com menos de 16 anos de idade, os admitidos por queimaduras, enfarte agudo de miocárdio ou após cirurgia de *by-pass* coronário.

O APACHE III compreende: a pontuação APACHE baseada nas variáveis fisiológicas agudas, idade, estado de saúde crónico e a equação preditiva APACHE III. Utiliza as mesmas doze variáveis que o APACHE II, mas regista de uma forma diferente os dados neurológicos, usando uma versão modificada do *Glasgow Coma Score* (GCS) (Teasdale and Jennett 1974). Inclui cinco variáveis fisiológicas adicionais: ureia, débito urinário, albumina sérica, bilirrubina e a glicose. Os pesos das variáveis foram modificados de acordo com os pareceres de peritos e dos resultados da análise univariada. De um modo geral, estas alterações foram feitas através do encurtamento dos intervalos de normalidade e aumentando a pontuação para os valores extremos de cada variável.

A variável idade foi reavaliada e o estado de saúde crónico reduzido a sete condições: SIDA (síndrome de imunodeficiência adquirida), linfoma, insuficiência hepática, neoplasia metastizada, leucemia ou mieloma múltiplo, cirrose hepática e imunossupressão. Foi ainda incorporada no modelo a localização do doente antes da admissão à UCI.

Para todas as variáveis fisiológicas são registados os piores valores das primeiras 24 horas de internamento na UCI.

O diagnóstico agudo de admissão é tido em conta sendo seleccionado através da regressão logística univariada entre setenta e oito categorias nosológicas.

A equação preditiva utiliza a pontuação APACHE III, os dados de referência sobre as categorias nosológicas, o tipo de admissão (médica ou cirúrgica programada ou urgente) e a localização do doente antes da admissão à UCI para a conversão numa probabilidade de mortalidade, para cada uma das setenta e oito categorias diagnosticas e das nove possibilidades de origem do doente.

O APACHE III tem tido uma penetração muito mais lenta nas UCIs devido ao facto dos coeficientes da regressão logística não serem do domínio público. (Teres, 1994)

O poder discriminativo do APACHE III avaliado na amostra de desenvolvimento foi de 0,90 (área sob a curva ROC).

Todos os estudos para a validação externa do APACHE III mostraram boa discriminação, tais como os realizados nos EUA (área sob a curva ROC = 0,89) (Zimmerman et al., 1998), no Reino Unido (área sob a curva ROC = 0,89) (Pappachan et al., 1999), em Portugal (área sob a curva ROC = 0,830) (Carneiro et al., 1997), na Alemanha (área sob a curva ROC = 0,873) (von Bierbrauer et al., 1998), no Brasil (área sob a curva ROC = 0,82) (Bastos et al., 1996), na Coreia do Sul (área sob a curva ROC = 0,905) (Ihnsook et al., 2003), em Espanha (área sob a curva ROC = 0,820) (Rivera-Fernandez et al., 1998) e na Austrália (área sob a curva ROC = 0,92) (Cook, 2000).

Os resultados dos testes de calibração foram díspares. Os estudos coreanos e alemães mostraram uma boa calibração a partir do modelo de referência. Os estudos australianos (Cook et al., 2002) e os espanhóis mostraram uma boa calibração com ajustes para a população local. O estudo inglês mostrou uma fraca calibração, mesmo após ajuste, e os estudos norte-americanos e brasileiros mostraram uma calibração muito fraca.

Além disso, a selecção de um único diagnóstico agudo de admissão é muitas vezes difícil no doente grave (Holt et al., 1992) (Teres & Lemeshow, 1994).

3.1.2.2 Simplified Acute Physiologic Score (SAPS II) e SAPS II expandido

Em 1993, Le Gall et al apresentaram uma segunda iteração do modelo SAPS, o SAPS II (Le Gall et al., 1993) (Quadro 2) desenvolvida a partir de uma base de dados com 13152 doentes de 137 UCIs de doze países europeus e norte-americanos, entre 1991 a 1992.

Este modelo inclui dezassete variáveis preditivas: doze variáveis fisiológicas, idade, tipo de admissão (cirúrgica programada, cirúrgica urgente ou médica) e três diagnósticos prévios (SIDA, neoplasia metastizada e neoplasia hematológica). Os pesos para cada variável foram estimados utilizando a regressão logística múltipla.

O SAPS II utiliza os piores valores registados durante as primeiras 24 horas na UCI e não inclui a escolha de um diagnóstico agudo de admissão para o cálculo da probabilidade de mortalidade hospitalar (Quadro 3).

A pontuação SAPS II varia entre 0 e 163 pontos (até 116 para as variáveis fisiológicas, até 17 para a idade e até 30 para os diagnósticos prévios).

No desenvolvimento da equação de regressão logística do SAPS II, foi utilizada uma amostra de desenvolvimento de 8369 casos. A área sob a curva ROC foi de 0,88 na amostra de desenvolvimento e de 0,86 na amostra de validação. Foi avaliada a calibração através do *goodness-of-fit test*, resultando num $p = 0,883$ na amostra de desenvolvimento e num valor de $p = 0,104$ na amostra de validação.

Tal como nos outros modelos, a capacidade discriminativa e particularmente a calibração do modelo SAPS não se ajustam quando é aplicado a populações diferentes daquelas em que foi desenvolvido (Le Gall, 2005).

O modelo pode ser adaptado a uma população específica através do recurso a um processo de ajuste ou através da expansão do modelo (inclusão de novas variáveis).

Aergerter et al propuseram uma revisão do SAPS II (Aergerter et al., 2005). Fizeram uma análise retrospectiva de 33471 dados colhidos prospectivamente de 32 UCIs da área de Paris e desenvolveram dois modelos de regressão logística baseados no SAPS II. O primeiro modelo derivou da utilização de uma estratégia de ajuste (usando apenas a pontuação do SAPS II como uma variável independente, e consequentemente mantendo todas as variáveis originais e os seus pesos), e um segundo modelo que reavaliou os itens originais do SAPS II e a integração de localização pré-admissão e comorbilidade crónica. Estes autores demonstraram uma melhor calibração para a mortalidade intra-hospitalar do que o SAPS II original. Le Gall et al elaboraram outra proposta (Le Gall et al., 2005), a partir de uma base de dados com 77490 admissões em 106 UCIs francesas e adicionaram seis variáveis de admissão ao SAPS II: idade, sexo, tempo de internamento hospitalar pré-admissão à UCI, localização do doente pré-admissão à UCI, categoria clínica e presença ou não de intoxicação medicamentosa. As qualidades estatísticas do SAPS II expandido são melhores do que as do original, e mesmo do que do SAPS II ajustado (Le Gall, 2005). A inclusão de dados simples colhidos por rotina, no SAPS II original, determinou uma melhor capacidade discriminativa e de calibração do modelo.

Quadro 3- New simplified Acute Physiology Score (SAPS II) (Moreno, 2000)

Idade, anos						
< 40	40-59	60-69	70-74	75-79	≥ 80	
0	7	12	15	16	18	
Frequência cardíaca, batimento/minuto						
< 40	40-69	70-119	120-159	≥ 160		
11	2	0	4	7		
Pressão arterial sistólica, mmHg						
<70	70-99	100-199	≥ 200			
13	5	0	2			
Temperatura, °C						
<39	≥ 39					
0	3					
Pao ₂ /FiO ₂ (apenas se ventilação mecânica ou CPAP), mm Hg						
<100	100-199	≥ 200				
11	9	6				
Débito urinário, litro/dia ¹						
<0,500	0,500-0,0999	≥1,000				
0		0				
Ureia sérica, mg/dL						
<28	28-83	≥ 84				
0	6	10				
Leucócitos						
<1,0	1,0-19,9	≥ 20,0				
12	0	3				
Potássio sérico, mEq/L						
<3,0	3,0-4,9	≥ 5,0				
3	0	3				
Sódio sérico, mEq/L						
<125	125-144	≥ 145				
5	0	1				
Bicarbonato sérico (mEq/L)						
<15	15-19	≥20				
6	3	0				
Bilirrubina (mg/dL)						
<4,0	4,0-5,9	≥ 6,0				
0	4	9				
Escala de comas de Glasgow						
< 6	6-8	9-10	11-13	14-15		
26	13	7	5	0		
Neoplasia sólida metastizada						
		9				
Neoplasia hematológica						
		10				
SIDA						
		17				
Tipo de doente						
Cirurgia programada	Médico	Cirurgia urgente				
0	6	8				

* Se o doente estiver na UCI menos de 24 horas calcule o débito previsto para esse período, por exemplo, 1 litro em 8 horas = 3 litros em 24 horas.

1. Para o cálculo da probabilidade de mortalidade hospitalar (PR) utilize a seguinte equação:

$$PR = \frac{e^{-7,7631+(0,0737 \times \text{pontuaçãoSAPSII})+(0,9971[\ln(\text{PontuaçãoSAPSII}+1)])}}{1 + e^{-7,7631+(0,0737 \times \text{pontuaçãoSAPSII})+(0,9971[\ln(\text{PontuaçãoSAPSII}+1)])}}$$

2. No doente com sépsis grave precoce, pode ser utilizada a adaptação do SAPS II para este tipo de doentes:

$$PR = \frac{e^{-3,5524+(0,0694 \times \text{pontuaçãoSAPSII})}}{1 + e^{-3,5524+(0,0694 \times \text{pontuaçãoSAPSII})}}$$

O cálculo do SAPS II faz-se através da utilização do quadro, atribuindo a cada variável a pontuação respectiva. Os valores utilizados devem ser avaliados e registados como os valores mais alterados das primeiras 24 horas após a admissão na UCI. Não devem ser utilizados valores prévios à admissão na UC1, considerando-se todos os valores em falta como normais.

3.1.2.3 MPM - Mortality Prediction Model (MPM e MPM II)

Em 1987, Teres et al desenvolveram o modelo Mortality Prediction Model (MPM) (Teres et al., 1987) incluindo variáveis preditivas em número reduzido e de fácil colheita.

O MPM foi o primeiro modelo a ser inteiramente desenvolvido de forma objectiva (Ohno-Machado et al., 2006).

Com o recurso a análise univariada para a selecção de variáveis a incorporar no modelo MPM, estas foram reduzidas a vinte e seis para a admissão e quarenta e quatro para as 24 horas. Cada conjunto de variáveis preditivas entrou num modelo de regressão logística. As variáveis com significado estatístico foram incluídas nos modelos finais de regressão logística, para cada período de tempo (admissão e 24 horas na UCI).

O modelo final de admissão (MPM₀) utilizou apenas sete variáveis preditivas (coma/esturpor, admissão urgente, cancro, infecção, número de insuficiência de órgãos/sistemas, idade e pressão arterial sistólica). O modelo às 24 horas (MPM₂₄) incluía sete variáveis (coma/esturpor, infecção, ventilação mecânica, choque, admissão urgente, idade, e o número de insuficiência de órgãos/sistemas).

No modelo MPM, não é calculado um índice de gravidade mas apenas uma probabilidade de mortalidade hospitalar.

Para a definição da resposta do modelo, é necessário definir um ponto de corte, de tal forma que todos os doentes com uma probabilidade de mortalidade superior ao ponto de corte, são classificados como "1" (ex.: morto) e todos com probabilidades iguais ou inferiores são classificados como "0" (ex.: sobrevive). Com um ponto de corte de 50% do cálculo da probabilidade de mortalidade, a taxa de total de classificações correctas foi de 87% à admissão e 85% às 24 horas. A área sob a curva ROC foi de 0.85, e a calibração avaliada pelo teste H de Hosmer-Lemeshow obteve um valor de $p = 0.38$ e o teste C de Hosmer-Lemeshow obteve um valor de $p = 0.56$ (Ohno-Machado et al., 2006).

O modelo MPM II (Lemeshow et al., 1993) foi publicado em 1993 por Stanley Lemeshow et al, baseado nos mesmos dados que o SAPS II, aos quais foram adicionados dados colhidos em seis UCIs de quatro hospitais universitários dos EUA (Quadro 2).

Os procedimentos para o desenvolvimento do modelo MPM II foram semelhantes aos utilizados para o MPM original. Os critérios de exclusão de doentes utilizados para o desenvolvimento do MPM II foram idênticos aos do SAPS II.

No MPM II, não obtemos uma pontuação ou índice de gravidade mas um modelo (por regressão logística) cujo resultado final é expresso apenas como uma probabilidade estimada de mortalidade hospitalar.

O MPM II inclui modelos para prever a mortalidade à admissão na UCI (MPM II₀) e às 24 horas de internamento (MPM II₂₄). Posteriormente, foram ainda descritos modelos para as 48 horas (MPM II₄₈) e para as 72 horas (MPM II₇₂) na UCI, desenvolvidos com base numa amostra de doentes menor. (Lemeshow et al., 1994). O modelo MPM II₀ inclui quinze variáveis: idade, três variáveis fisiológicas (coma ou esturpor, frequência cardíaca e tensão arterial sistólica) três doenças crónicas (insuficiência renal crónica, cirrose hepática e neoplasia metastática), cinco diagnósticos agudos (insuficiência renal aguda, disritmia cardíaca, acidente vascular cerebral, hemorragia gastrointestinal, efeito de massa intracraniano), tipo de admissão (médica ou cirúrgica urgente), ventilação mecânica e paragem cardio-respiratória prévia à admissão hospitalar. Todas as variáveis são avaliadas com base nos dados registados na primeira hora antes e depois da admissão na UCI.

O modelo MPM II₂₄ compreende treze variáveis: idade, seis variáveis fisiológicas (coma ou esturpor, creatinina, PaO₂, infecção confirmada e débito urinário), três variáveis avaliadas à admissão (cirrose hepática, efeito de massa intracraniano e neoplasia metastática), tipo de admissão (médica ou cirúrgica urgente), ventilação mecânica e utilização de fármacos vasoactivos. As variáveis fisiológicas são avaliadas com base nos piores valores registados durante as primeiras 24 horas na UCI.

O modelo MPM II₄₈ e o MPM II₇₂ utilizam as mesmas variáveis que o MPM II₂₄, com pesos diferentes no cálculo do risco de mortalidade. Ambos são baseados nos piores valores registados nas 24 horas precedentes.

O poder discriminativo (área sob a curva ROC) do MPMII₀ nas amostras de desenvolvimento e de validação foi de 0,837 e de 0,824 respectivamente. Os valores correspondentes para o MPM II₂₄ foram respectivamente de 0,844 e 0,836. A calibração foi adequada para os modelos MPM II₀ e MPM II₂₄ (Quadro 2).

3.1.2.4 Simplified Acute Physiologic Score (SAPS III)

O modelo SAPS III foi publicado em 2005 por Metnitz P. G et al e encontra-se em fase de investigação e implementação. (Quadro 2 e 3) (Metnitz et al., 2005) ((Moreno et al., 2005). Para o desenvolvimento do modelo, foi utilizada uma base de dados extensa que incluía 19577 doentes admitidos consecutivamente em 307 UCIs espalhadas pelo mundo, entre Outubro e Dezembro de 2002.

O SAPS III regista os dados aquando da admissão na UCI ($\pm$ 1 hora) e compreende três componentes: estado de saúde crónico, circunstâncias relacionadas com a admissão e variáveis fisiológicas agudas.

Os autores desenvolveram coeficientes de regressão através da regressão logística múltipla para estimar a probabilidade de morte hospitalar. O modelo final, que engloba vinte variáveis, evidenciou um bom poder discriminativo (área sob a curva ROC = 0,848), sem diferenças significativas entre subgrupos de doentes, a calibração foi também satisfatória (teste H de Hosmer-Lemeshow = 10,56, com valor de $p = 0,39$ e teste C = 14,29, com $p = 0,16$) (Moreno and Afonso 2006).

Curiosamente, os determinantes da mortalidade hospitalar modificaram-se em relação às do início da década de 90 (Knaus, Wagner et al. 1991), sendo o estado de saúde crónico e as circunstâncias da admissão do doente à UCI responsáveis por $\frac{3}{4}$ do poder prognóstico do modelo (Moreno and Afonso 2006).

3.1.2.5 APACHE IV - Acute Physiology and Chronic Health Evaluation (2006)

Jack E. Zimmerman et al publicaram em 2006 a última iteração do modelo APACHE, o APACHE IV (Zimmerman , 2006) (Quadro 2).

O modelo foi desenvolvido durante dois anos a partir de uma base de dados com 110558 admissões consecutivas a 104 UCIs, de quarenta e cinco hospitais norte-americanos.

O modelo APACHE IV usa os piores valores das variáveis fisiológicas agudas ocorridos durante as primeiras 24 horas na UCI. Para estimar a probabilidade de morte hospitalar recorre a equações preditivas obtidas por regressão logística múltipla.

Para além das variáveis preditivas presentes no APACHE III, foram incluídas novas variáveis.

A capacidade preditiva do APACHE IV foi analisada na globalidade da base de dados e em subgrupos de doentes.

O APACHE IV evidenciou uma boa calibração e um bom poder discriminativo. Em 90% dos 116 diagnósticos de admissão, a razão da mortalidade observada/prevista (razão da mortalidade normalizada) calculada não foi significativamente diferente de 1,0.

Comparativamente às predições obtidas pelo modelo APACHE III, constatou-se que a variação no poder discriminativo era mínima, mas a mortalidade agregada era sistematicamente sobrestimada com o decorrer do tempo (Moreno & Afonso 2006).

Relativamente ao impacto dos diagnósticos agudos na acuidade do modelo APACHE IV, o comportamento da capacidade preditiva era heterogéneo, reflectindo potencialmente modificações na terapêutica ou na prática clínica (Moreno & Afonso, 2006).

3.1.2.6 Mortality Probability Admission Model (MPM III₀)

O MPM III₀ (quadro 2) foi descrito em 2005 e publicado em 2007 por Tom Higgins et al. (Higgins, Teres et al. 2005) Foi desenvolvido usando os dados provenientes de UCIs dos EUA participantes no projecto IMPACT (2004).

Os autores actualizaram o modelo MPM II₀, recorrendo a uma análise retrospectiva de uma base de dados com 124855 doentes admitidos em 135 UCIs de 98 hospitais norte-americanos participantes no projecto IMPACT, entre 2001 a 2004.

Para o desenvolvimento do modelo foram utilizadas as quinze variáveis preditivas do MPM II₀, o tempo antes da admissão à UCI e a ausência de todos os factores de risco do MPM II₀, excepto a idade. O MPM III₀ demonstra uma boa calibração e discriminação (Quadro 2).

3.2 Cálculo do índice de gravidade e da probabilidade de mortalidade

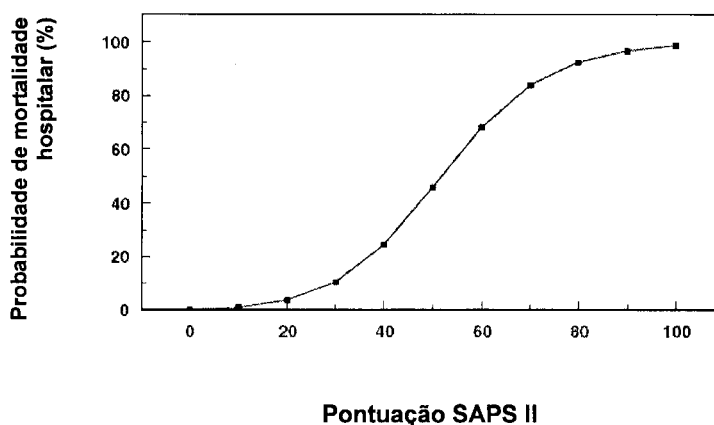
A aplicação destes sistemas a uma determinada população de doentes graves deverá cumprir várias etapas: a selecção de doentes, a avaliação e registo das variáveis preditivas, o cálculo do índice de gravidade e a transformação do índice numa probabilidade de mortalidade.

O cálculo do índice de gravidade para cada doente é obtido a partir do somatório da pontuação atribuída a cada variável, dependente do seu grau de afastamento dos valores normais. No caso do modelo MPM, as variáveis utilizadas são directamente transformadas numa probabilidade de mortalidade através de uma equação de regressão logística múltipla.

A partir da pontuação resultante do índice de gravidade, é calculada a probabilidade de mortalidade intra-hospitalar que varia entre 0 e 1.

Como resultado, estes modelos disponibilizam uma probabilidade de ocorrência de um determinado *outcome* e não uma previsão absoluta. Para o prognóstico de um doente individualizado, o modelo tem de gerar um *outcome* binário (isto é, vivo ou morto), o que implica a escolha de uma pontuação limiar. Este valor limiar determina o nível aceitável de incerteza. Por exemplo, um valor limiar de 95% implica que as probabilidades estimadas iguais ou acima desse valor são assumidas como associadas a determinada previsão. Devido à natureza dicotómica da mortalidade, o gráfico da probabilidade de mortalidade versus um determinado preditor tende a ser sigmoide, e não linear (Figura 1).

Figura 1 -Curva de configuração sigmoide do risco de mortalidade hospitalar, variando com o valor crescente da pontuação SAPS II. Risco de mortalidade hospitalar gerada pela equação: $\text{logit} = 7,7631 + 0,0737 * (\text{pontuação SAPS II}) + 0,9971 * \ln [(\text{pontuação SAPS II}) + 1]$. Probabilidade de mortalidade = $e^{(\text{logit})} / (1 + e^{(\text{logit})})$ (Ridley, 2002).

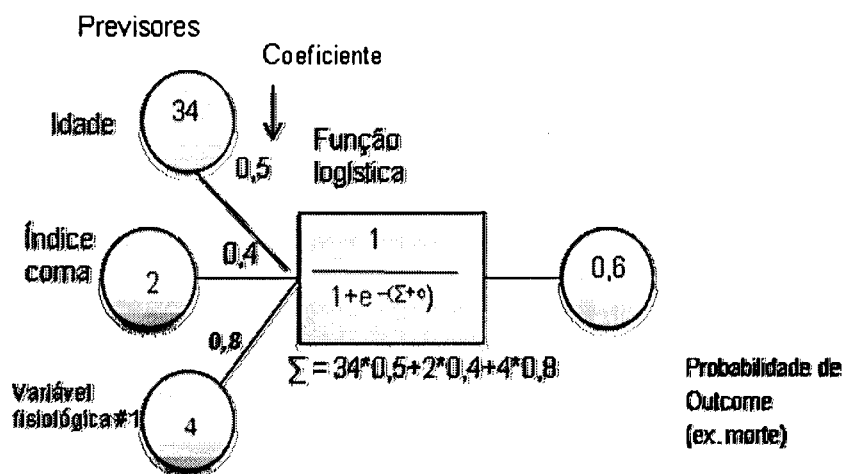


Matematicamente, a transformação do valor do índice de gravidade numa probabilidade de mortalidade hospitalar é feita utilizando uma equação de regressão logística, na maioria dos modelos utilizados nas UCIs (Lemeshow and Le Gall 1994). No modelo de regressão logística a função logística relaciona o conjunto das variáveis preditivas (i) ou variáveis independentes, designadas por x_i que formam a matrix x , contendo informação sobre j indivíduos, relativas ao *outcome* de interesse (ex.: mortalidade intra-hospitalar), ou variável dependente designada por y na seguinte equação:

$$y = \frac{1}{1 + e^{(bx+c)}}$$

A Figura 2 ilustra um modelo de regressão logística. Na maioria dos modelos revistos neste capítulo e nos modelos da Aprendizagem Automática revistos adiante, y é uma variável binária representando o estado do doente como morto ($y = 1$) ou vivo ($y = 0$) aquando da alta hospitalar, e a matrix x representa o conjunto das variáveis preditivas (ex.: x_{11} representa a idade do doente número um, x_{21} representa a saturação de oxigénio, e assim sucessivamente). O vector b representa os coeficientes estimados para cada preditor e c é a constante. Os parâmetros da função logística são geralmente determinados pela estimativa da máxima verosimilhança através de algoritmos iterativos (Hosmer and Lemeshow 1989).

Figura 2 – Modelo de regressão logística para a classificação de casos em *outcomes* categóricos binários. (Ohno-Machado et al., 2006).



Os parâmetros correspondem aos coeficientes de cada variável independente na função logística, como designada pelo vector **b**, o qual tem uma interpretação directa em termos de *odds ratio* (OR)¹⁵ correspondente a cada variável. O parâmetro **c** calibra o modelo para a taxa basal do *outcome* de interesse (Hosmer & Lemeshow, 1989).

Estas características fazem com que o modelo seja de fácil interpretação, uma vez que o sinal e o valor dos coeficientes (quando normalizados) podem fornecer uma indicação directa da força da associação de cada variável preditiva com o risco de um determinado *outcome*. Em certos dados, os preditores podem ter de ser combinados em termos de interacção ou transformados de forma a conseguirem-se melhores resultados. Contudo, embora a adição de interacções e de transformações de variáveis aumente a flexibilidade do modelo também diminui a sua interpretabilidade (Ohno-Machado et al., 2006).

3.3 Validação dos índices gerais de gravidade e modelos prognósticos

Todos os modelos prognósticos necessitam de uma validação formal para a avaliação do seu desempenho (Lemeshow & Le Gall, 1994). Inicialmente, o modelo deve ser validado na base de dados onde foi desenvolvido. A amostra de validação pode ser obtida da divisão aleatória da base de dados disponível em duas partes: uma usada para a construção do modelo e a outra para a validação do modelo.

Posteriormente, o desempenho de um modelo deve ser demonstrado numa amostra de doentes independente da utilizada para a construção do modelo (Lemeshow & Le Gall, 1994).

A discriminação e a calibração são dois conceitos fundamentais na avaliação do modelo.

A discriminação refere-se à capacidade do modelo distinguir se um doente morre ou sobrevive, com base nas estimativas de morte intra-hospitalar. Pode ser expressa como a sensibilidade e a especificidade para um determinado valor de decisão limite, em que o limite escolhido afectará os valores da sensibilidade e especificidade do modelo (Ridley, 2002).

Habitualmente, o poder discriminativo é medido pela área sob a curva ROC (*receiver-operating characteristic*), (Hanley and Mc Neil 1982) que representa graficamente a sensibilidade (ou a taxa de verdadeiros positivos) *versus* a 1 – especificidade (taxa dos

¹⁵ Quociente das probabilidades do resultado para uma variável binária. Mede a probabilidade da ocorrência de um evento num grupo relativamente a outro grupo. Ibid.

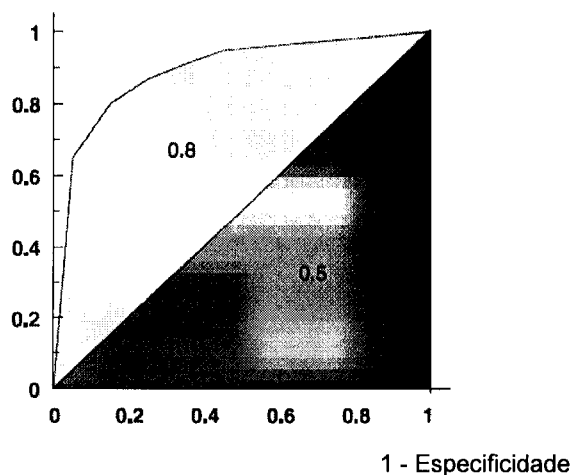
falsos positivos) do modelo, à medida que desloca o ponto de corte para a mortalidade (Figura 3).

Para construir a curva ROC, é especificada uma sequência de pontos de corte, e construída uma tabela de classificação de dois por dois do *outcome* previsto e observado, para cada um dos pontos de corte definidos (Le Gall, 2005).

Para cada tabela, são calculadas a taxa de verdadeiros positivos (ou sensibilidade) e a taxa de falsos positivos (ou 1 menos especificidade). O gráfico final de todos os pares de taxas de verdadeiros positivos versus falsos positivos é a representação visual da curva ROC. Por exemplo, se o ponto de corte é estabelecido a 0,35, prevê-se que qualquer doente cuja probabilidade de morte é igual ou maior do que 0,35 irá morrer, enquanto que o modelo prevê que qualquer doente cuja probabilidade de morte seja inferior a 0,35 sobreviva.

Figura 3 – Exemplo de curva *receiver operating characteristic* (ROC). A linha a 45° tem um área sob a curva ROC de 0,5, indicando que a discriminação não é melhor do que o acaso. Adaptado de (Ridley, 2002); (Rosenberg, 2002).

Sensibilidade



Quanto mais elevada é a taxa de verdadeiros-positivos relativamente à taxa de falsos-positivos, maior é a área sob a curva ROC.

Assim que a área sob a curva ROC aumenta de 0,5 (escolha aleatória) até 1 (discriminação perfeita), a acuidade e o poder preditivo do modelo aumentam.

A área sob a curva ROC dá a probabilidade de um grupo de não-sobreviventes aleatoriamente seleccionados terem uma pontuação preditiva mais elevada do que um grupo de doentes sobreviventes aleatoriamente seleccionados.

Geralmente, os valores da área sob a curva ROC de 0,8 ou mais elevados são esperados para a previsão da mortalidade nos modelos correntes (Knaus et al., 1991); (Le Gall et al., 1993) e os valores de 0,7 ou inferiores são considerados inaceitáveis (Hanley & Mc Neil, 1982) (Zweig & Campbell, 1993); (Glance et al., 2000).

A calibração de um modelo mede o grau de concordância entre as probabilidades do *outcome* (mortalidade) estimadas pelo modelo e o *outcome* (mortalidade) observada ao longo de todo o intervalo de probabilidades de morte (Lemeshow & Hosmer, 1982). São geralmente empregues dois métodos nesta análise: as razões da mortalidade observada/prevista (razão da mortalidade normalizada) e os testes de Hosmer-Lemeshow (C ou H) (Lemeshow & Hosmer, 1982).

Um modelo bem calibrado elabora estimativas que se aproximam dos valores observados ou da probabilidade do *outcome* em causa, para um determinado doente (ex.: probabilidade de morte à data da alta hospitalar) (Ohno-Machado et al., 2006). As razões de *outcome* observado/previsto são calculadas dividindo a taxa do *outcome* observado (ou seja o número de mortes) pela taxa de *outcome* previsto (ou seja a soma das probabilidades de mortalidade de todos os doentes analisados). Num modelo perfeitamente calibrado este valor deve ser igual a 1 (Moreno, 2000).

A razão normalizada da mortalidade embora seja uma medida grosseira de calibração, é um método fiável para determinar a falta de calibração se o intervalo de confiança para 95% não incluir o valor 1 (Ohno-Machado et al., 2006).

Como acima referido, a calibração pode ser estatisticamente avaliada através de testes formais de *Hosmer-Lemeshow* (Hosmer & Lemeshow, 1989).

Para avaliar a calibração do modelo, os doentes são ordenados de acordo com a sua probabilidade de morte e agrupados em dez categorias de risco, cada uma com aproximadamente o mesmo número de doentes. Para obter o número de mortes previstas numa categoria, são somadas as probabilidades de morte de todos os doentes incluídos nessa categoria. Os testes de Hosmer-Leshow comparam o número de mortes observadas com as previstas, e o número de sobreviventes observados com o número de sobreviventes previsto, para cada categoria (Le Gall, 2005).

Esta análise pode ser apresentada sob variadas formas, tais como curvas de calibração e/ou através de uma estatística tipo qui-quadrado desenvolvida por Hosmer – Lemeshow ((Hosmer & Lemeshow, 1989) ; (Rosenberg et al., 2000).

É desejável que o modelo se aproxime o mais possível do *outcome* observado e o teste estatístico de Hosmer-Lemeshow deve ter um valor de $p > 0,05$ (isto é, a diferença entre o modelo e o *outcome* deve ser não significativa). De uma forma geral, os valores absolutos da estatística de Hosmer-Lemeshow deverão estar próximos do número de categorias de risco de mortalidade menos dois (os graus de liberdade)

(Angus et al., 2000). Uma estatística de Hosmer-Lemeshow inferior a 15 e valores de p entre 0,2 e 0,8 são considerados bons (Hunt & Meyer, 1997).

Embora estejam sujeitos a críticas (Champion et al., 1996), estes testes são hoje considerados parte integrante da avaliação da calibração do modelo. A importância relativa da calibração e da discriminação varia com a utilização que for dada ao modelo.

De uma forma geral, um modelo deverá ter uma boa capacidade discriminativa se o objectivo é o de identificar não-sobreviventes, e uma boa calibração se o objectivo é o de calcular taxas de mortalidade normalizada e outras medidas de desempenho com base na previsão da mortalidade (Schuster, 1992).

Para o processo de decisão envolvendo o doente individualizado, ambos os parâmetros são muito importantes.

3.4 Comparação dos modelos

Existem múltiplos estudos comparando o desempenho dos índices gerais de gravidade e modelos prognósticos (Castella et al., 1991); (Rowan et al., 1994) (Castella et al., 1995); (Del Bufalo et al., 1995); (Wilairatana et al., 1995); (Moreno et al., 1998a); (Nouira et al., 1998); (Tan, 1998); (Patel & Grant, 1999); (Vassar et al., 1999); (Capuzzo et al., 2000); (Katsaragakis et al., 2000) (Livingston et al., 2000); (Markgraf et al., 2000); (Beck et al., 2003).

Estes estudos evidenciaram de uma forma geral a variabilidade da capacidade discriminativa (Quadro 3), associada a uma fraca calibração dos modelos prognósticos entre diferentes populações de doentes (Ohno-Machado et al., 2006). Foi demonstrado que os modelos APACHE III, MPM II e SAPS II apresentavam um desempenho equivalente ou superior aos modelos antecessores.

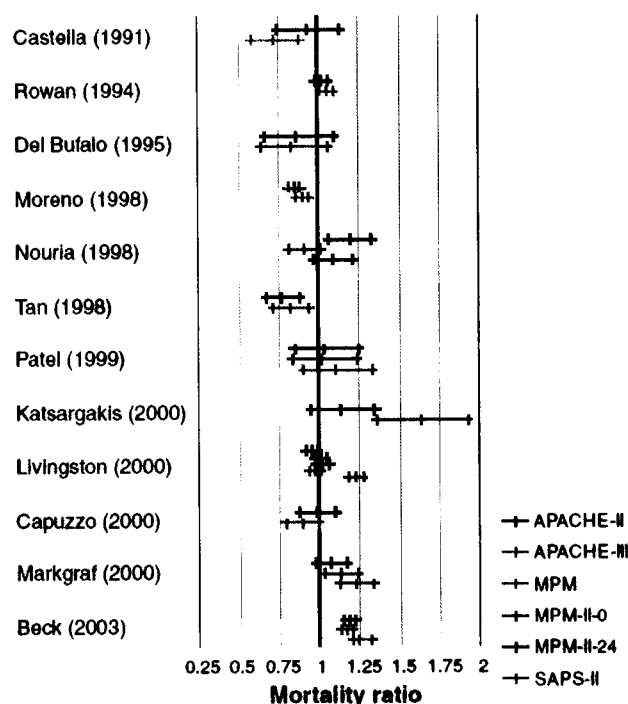
Desde então, a comparação de modelos tem geralmente sido feita entre os modelos de 3ª geração com o modelo APACHE II, e confirmaram a melhoria do poder discriminativo dos modelos de 3ª geração, embora nenhum modelo se tenha mostrado claramente superior (Rowan et al., 1994); (Castella et al., 1995). Globalmente, não foi encontrada uma melhoria da calibração dos modelos de 2ª geração em relação aos de 3ª geração (Ohno-Machado et al., 2006) (quadro 3).

Quadro 3 - Sumário dos estudos comparativos em termos de capacidade discriminativa (área sob a curva ROC) dos modelos prognósticos.

Autor	APACHE II	APACHE III	SAPS II	MPM II₀	MPM II₂₄
(Castella, Gilabert et al. 1991)	0,867				
(Rowan, Kerr et al. 1994)	0,830				
(Del Bufalo, Morelli et al. 1995)	0,808		0,735		
(Castella, A. et al. 1995)	0,852	0,866			
(Castella, A. et al. 1995)	0,857		0,855	0,815	0,833
(Moreno, Apolone et al. 1998)			0,822	0,785	
(Nouira, Belghith et al. 1998)	0,820		0,840	0,850	0,882
(Tan 1998)	0,880		0,870		
(Patel and Grant 1999)	0,702		0,672		0,695
(Vassar, Lewis et al. 1999)	0,870	0,890			
(Katsaragakis, Papadimitropoulos et al. 2000)	0,839		0,870		
(Livingston, MacKirdy et al. 2000)	0,763	0,795	0,784	0,741	0,791
(Capuzzo, Valpondi et al. 2000)	0,805		0,816		
(Markgraf, Deutschinoff et al. 2000)	0,832	0,846	0,846		
(Beck, Smith et al. 2003)	0,835	0,867	0,852		

Os valores da taxa de mortalidade normalizada e os intervalos de confiança de vários estudos são revelados na Figura 4. Os diferentes estudos evidenciam a variabilidade das taxas de mortalidade normalizada quando os modelos são aplicados a diferentes populações de doentes graves, revelando a variabilidade do desempenho dos modelos prognósticos em populações distintas.

Figura 4 – Razões da mortalidade normalizada e intervalos de confiança a 95% em estudos comparativos. Adaptado de(Ohno-Machado et al., 2006)



4. Índices de quantificação da falência múltipla de órgãos

A síndrome de disfunção múltipla de órgãos (SDMO) permanece como a principal causa de mortalidade nas UCIs (Beal & Cerra, 1994). As taxas de mortalidade referidas têm variado entre 20% e 100% (Knaus et al., 1985b); (Hebert et al., 1993); (Tran et al., 1993); (Zimmerman et al., 1996), dependendo do número, da gravidade, da duração, do tipo, combinação e definição de disfunção de órgão.

A maioria dos índices gerais de gravidade de doença e modelos prognósticos tais como o APACHE, o SAPS e o MPM usam a informação do grau de disfunção fisiológica e o diagnóstico agudo obtido nas primeiras 24 horas de admissão à UCI para a previsão do *outcome* (mortalidade intra-hospitalar) do doente grave.

O facto de estes se basearem no grau de disfunção fisiológica avaliado à admissão ou nas primeiras 24 horas na UCI limita a sua capacidade de acompanhar a evolução da gravidade do processo fisiopatológico, porque ignoram os múltiplos factores que podem influenciar o *outcome* do doente durante o internamento na UCI (Ferreira et al., 2001) (Le Gall, 2005). Estes factos motivaram o desenvolvimento de sistemas que permitam acompanhar o doente ao longo do seu internamento na UCI, assumindo que

os processos fisiopatológicos do doente grave são dinâmicos, dependentes da evolução da doença de base, das suas consequências nos vários órgãos e sistemas, da resposta à terapêutica e do aparecimento de complicações.

O princípio subjacente aos índices de quantificação de falência múltipla de órgãos é que a presença e o grau de disfunção de órgãos influencia o *outcome* (Ridley, 2002).

O grau de disfunção de órgãos descreve a gravidade da doença de uma forma mais facilmente compreensível que um simples índice de gravidade de doença.

Neste contexto, foram propostas várias estratégias para a quantificação da falência múltipla de órgãos.

Uma das estratégias advogada por vários investigadores foi a utilização diária dos índices de gravidade gerais (corrigida eventualmente para a presença e grau de falências de órgãos) para monitorizar da evolução da gravidade da doença (Chang, 1988 #312); (Sicignano et al., 1996, Wagner et al., 1994); (Guiguet et al., 1998).

Chang et al, utilizando o modelo APACHE II corrigido para o grau de falência múltipla de órgão aplicado diariamente, obtiveram um incremento da taxa de classificação correcta (para a previsão de mortalidade) de 76% para 83%, em comparação com o APACHE II da admissão. A validação desta metodologia foi considerada insuficiente em estudos independentes.

Wagner et al descreveram o APACHE III adaptado para a avaliação prognóstica durante os primeiros sete dias de internamento na UCI (Wagner et al., 1994). Esta metodologia revelou-se complexa e pouco eficaz (Knaus et al., 1995).

Lemeshow et al (Lemeshow et al., 1994)descreveram a adaptação do MPM II para a sua utilização às 48 horas e 72 horas na UCI. Problemas de calibração limitaram a aplicação desta metodologia (Moreno et al., 1998a).

Alguns investigadores têm proposto a utilização diária do SAPS II. Utilizando esta metodologia, Sicignano et al (Sicignano et al., 1996) observaram que o poder discriminativo do SAPS II diminui ao longo do tempo, nos doentes que permanecem na UCI por mais de cinco dias. Estes investigadores argumentam que factores clínicos não incluídos no modelo e outras variáveis não clínicas são responsáveis pela redução da capacidade discriminativa do modelo ao longo dos dias.

Um achado consistente nestes estudos é que a gravidade aquando da admissão permanece como variável significativa durante vários dias. Acresce, que o factor mais importante na previsão da mortalidade hospitalar, nas avaliações sequenciais é o valor do índice no dia corrente (Lemeshow et al., 1994); (Wagner et al., 1994).

Num estudo englobando 1441 doentes de 15 UCIs espanholas, Rué et al reexaminaram este tópico (Rue et al., 2001). Recorrendo a quatro modelos de

regressão logística diferentes, os investigadores avaliaram o impacto da previsão da mortalidade hospitalar do dia corrente. O modelo mais preciso era obtido quando incluía o valor do índice do próprio dia. Esta abordagem, face aos resultados não conclusivos (Lemeshow et al., 1994), necessita todavia de mais investigação e tem como maior limitação o facto de não disponibilizar uma informação dinâmica sobre a evolução e quantificação da disfunção individual de órgão.

Em 2001, Timsit et al publicaram um novo modelo prognóstico composto, usando o SAPS II, o índice de disfunção de órgão LOD e a evolução dos dois sistemas durante os três primeiros dias de internamento na UCI, a partir de uma base de dados oriunda de 28 UCIs francesas, denominado de Three days Recalibrated ICU Outcome Score (TRIOS) (Timsit et al., 2001).

Utilizaram um modelo de regressão logística múltipla, na amostra de desenvolvimento, para a selecção das variáveis independentes associadas à mortalidade (registadas durante os três primeiros dias de internamento na UCI). As variáveis seleccionadas foram: SAPS II do primeiro dia [OR: 1,04 (intervalo de confiança (IC) a 95%: 1,027 – 1,053) por ponto]; LOD do primeiro dia [OR: 1,16 (IC de 95%: 1,085 – 1,253) por ponto]; transferência da enfermaria geral [OR: 1,74 (IC de 95%: 1,25 – 2,42)]; diferença entre SAPS do terceiro dia e do segundo dia [OR 1,516 (IC de 95%: 1,04 – 2,22)]; e a diferença entre o LOD do terceiro dia e o do segundo dia [OR: 2,00 (IC de 95%: 1,29 -3,11)].

O modelo final na amostra de validação revela boa capacidade discriminativa e calibração (área sob a curva ROC: 0,826 e teste C de Hosmer-Lemeshow: 7,14, p = 0,5).

Outra abordagem constou do desenvolvimento de escalas práticas para a quantificação da SDMO. O primeiro índice foi publicado em 1985 por Knaus W et al (Knaus et al., 1985b).

Desde então, foram desenvolvidos vários índices, nomeadamente (Fagon et al., 1993) o *Organ System Failure* (OSF) (Fry et al., 1980) (Pine et al., 1983); (Goris et al., 1985); (Knaus et al., 1985b)), o *Organ Dysfunction and Infection System* (ODIN) (Fagon et al., 1993), o *Sequential Organ Failure Assessment* (SOFA Score) {Vincent, 1998 #240}, o *Multiple Organs Dysfunction Score* (MODS) (Marshall et al., 1995) e o *Logistic Organ Dysfunction Score* (LOD) (Le Gall et al., 1996).

Os índices de falência múltipla de órgão são escalas de quantificação sequencial da presença, gravidade e número de falência de órgãos e foram desenvolvidos com os seguintes pressupostos: (Vincent, 1996)

- serem capazes de descrever a disfunção crescente de órgão, como um processo contínuo, em vez de um *outcome* binário (ausente/presente);
- necessidade de avaliação repetida, dado a SDMO não ser um fenómeno estático e o grau de disfunção variar com o tempo, no decurso da evolução do processo patológico;
- incorporarem variáveis objectivas e disponíveis por rotina em todas as UCIs, específicas para cada órgão ou sistema em questão e independentes das características basais dos doentes.

As variáveis escolhidas devem ser idealmente independentes da terapêutica, embora este facto seja virtualmente impossível de atingir na prática (Marshall et al., 1995).

Os índices de avaliação da disfunção de órgão têm várias utilizações potenciais:

- a avaliação do doente através da monitorização do curso da doença ao longo do tempo, possibilitando um melhor conhecimento dos padrões de disfunção de órgão;
- a avaliação de novas terapêuticas nos ensaios clínicos;
- a caracterização de doentes para inclusão em ensaios clínicos ou estudos epidemiológicos (Vincent 2000).

O princípio subjacente aos índices de quantificação de falência múltipla de órgãos é que a presença e o grau de disfunção do órgão influenciam o *outcome* (Ridley, 2002).

Os índices de avaliação da falência múltipla de órgão, embora tenham como primeiro objectivo a descrição e quantificação da morbilidade, evidenciaram um relacionamento estreito entre a disfunção de órgão e a mortalidade estudos (Turner et al., 1991); (Deitch, 1992); (Moreno et al., 1999); (Ferreira et al., 2001). Estes índices podem também ser usados para a previsão do risco de mortalidade, aquando da admissão na UCI (Le Gall et al., 1996).

Permitem avaliar a gravidade da doença ao logo do tempo, à medida que os órgãos e sistemas vão demonstrando sinais de disfunção ou falência (Vincent, 1998).

À semelhança dos índices de gravidade e modelos prognósticos, estes instrumentos podem ser úteis para a estratificação de doentes a incluir em ensaios clínicos.

4.1 Descrição e classificação dos índices de quantificação da falência de órgãos

Os índices de avaliação da disfunção de órgão destinam-se a complementar os modelos prognósticos gerais (APACHE, SAPS e MPM) e não a sua substituição. Foram desenhados para descrever a evolução do doente grave recorrendo a um número reduzido de variáveis e aplicáveis à globalidade dos doentes (Vincent, 1998).

Os índices de disfunção de órgãos quantificam a disfunção fisiológica através da avaliação de seis órgãos e sistemas: respiratório, renal, cardiovascular, hematológico, sistema nervoso central e hepático.

Os vários sistemas de avaliação do SDMO diferem em alguns aspectos, particularmente no tipo de variáveis utilizadas (quadro 4).

Estes sistemas diferem também nas pontuações que utilizam e no modo como são derivadas.

Quadro 4 – Variáveis utilizadas nos sistemas de quantificação da falência múltipla de órgãos. Adaptado de (Knaus et al., 1985b)

Órgão ou Sistema	Variáveis	Variáveis terapêuticas
Respiratório	PaO_2^a / FiO_2^b	Ventilação mecânica Valor da PEEP ^c
Cardiovascular	Pressão arterial	Uso de fármacos vasoactivos
Hematológico	Plaquetas Leucócitos Hematócrito	Uso de sangue ou derivados
Renal	Débito urinário Ureia sérica Creatinina sérica	Uso de técnicas dialíticas
Neurológico	Escala de comas de Glasgow	Uso de sedação
Hepático	Bilirrubina Tempo de protrombina	

^a: pressão arterial de oxigénio

^b: fracção inspiratória de oxigénio

^c: pressão positiva no fim da expiração

Os índices de avaliação da falência múltipla de órgãos podem ser classificados em subjectivos e objectivos, de acordo com as técnicas de selecção e atribuição de pesos das variáveis que os compõem (selecção por peritos versus regressão logística) (Le Gall, 2005).

4.1.1 Índices de quantificação de falência múltipla de órgãos – subjectivos

4.1.1.1 Organ System Failure (OSF)

O índice OSF foi descrito em 1985 por Knaus et al. (Quadro 6)(Knaus et al., 1985b). Constitui a primeira escala de avaliação diária da falência múltipla de órgão (deve ser registado por períodos de 24 horas, durante a estadia do doente na UCI), avaliando esta síndrome como uma variável binária (ausente ou presente), não permitindo a quantificação de graus de intermédios de disfunção de órgão.

O índice OSF inclui dados sobre cinco órgãos e sistemas: respiratório, cardiovascular, renal, hematológico e neurológico.

Foram publicadas algumas modificações, como a de Garden et al, que adicionaram a falência hepática (Garden et al., 1985).

Os principais factores prognósticos compreendem o número e a duração das falências de órgãos.

A mortalidade aproxima-se dos 100% quando a insuficiência de três órgãos persiste por cinco ou mais dias (Le Gall, 2005).

4.1.1.2 Organ Dysfunction and/or Infection (ODIN)

Fagon et al propuseram em 1993 o sistema ODIN (Fagon et al., 1993). Este sistema foi desenvolvido numa população de 1070 doentes. Inclui dados referentes à presença de infecção e falência de seis órgãos e sistemas (respiratório, cardiovascular, renal, hematológico, hepático e neurológico).

Através da análise univariada, este sistema revelou uma correlação entre a taxa de mortalidade e o número, tipo de falência de órgãos e presença de infecção: 2,6; 9,7; 16,7; 32,3; 64,9; 75,9; 94,4 e 100% para pontuação de ODIN de 0, 1, 2, 3, 4, 5, 6 e 7, respectivamente; ($p < 0,001$).

As taxas de mortalidade mais elevadas associaram-se à falência hepática (60,8%), hematológica (58,1%) e renal (54,8%).

4.1.1.3 Sequential Organ Failure Assessment (SOFA score)

Inicialmente designado por *sepsis-related organ failure assessment – SOFA score* (Vincent, 1996) foi desenvolvido por um painel de peritos da *European Society of Intensive Care Medicine*.

Após a publicação por Vincent et al em 1998 (Vincent, 1998), este índice passou a ser denominado de *Sequential Organ Failure Assessment – SOFA score* (Quadro 5).

O índice SOFA foi inicialmente avaliado em 1449 doentes. Quantifica a disfunção ou falência de seis órgãos e sistemas: respiratório, hematológico, hepático, cardiovascular, neurológico e renal (quadro 5), pontuados de 0 (função normal) 1 e 2 (disfunção) e 3 a 4 (falência). Apresenta portanto, uma pontuação máxima de 24 pontos.

O sistema SOFA compreende um índice baseado no grau de disfunção ou falência de seis órgãos e sistemas, sendo a pontuação para cada um deles variável entre 0 e 4 pontos (falência). As piores pontuações registadas para cada órgão ou sistema num determinado período de 24 horas são depois somadas para o cálculo do SOFA total diário, e assim consecutivamente durante o tempo de internamento na UCI.

Quadro 5 – O Sequential Organ Failure Assessment (SOFA score)

	0	1	2	3	4
Respiratório	> 400	≤ 400	≤ 300	≤ 200 ^a	< 100 ^a
PaO ₂ /FiO ₂ , mmHg					
Hematológico	> 150	≤ 150	≤ 100	≤ 50	≤ 20
Contagem de plaquetas (x 10 ³ /mm ³)					
Hepático	< 1,2	1,2 – 1,9	2,0 – 5,9	6,0 – 11,9	> 12,0
Bilirrubina sérica, mg/dL (μmol/l)	(< 20)	(20 – 32)	(33 – 101)	(102 – 204)	> 204
Cardiovascular	PAM ^b ≥ 70	PAM ^b < 70	Dopamina ≤ 5 ou dobutamina (em qualquer dose) ^c	Dopamina > 5 ou adrenalina ≤ 0,1 ou noradrenalina ≤ 0,1 ^c	Dopamina > 15 ou adrenalina > 0,1 ou noradrenalina > 0,1 ^c
Hipotensão					

Quadro 5 – O Sequential Organ Failure Assessment (SOFA score). (cont.)

	0	1	2	3	4
Neurológico	15	13 -14	10 – 12	6 – 9	<6
Escala de Coma de Glasgow					
Renal	< 1,2	1,2 – 1,9	2,0 – 3,4	3,5 – 4,9	> 5,0
Creatinina sérica, mg/dL	(< 110)	(110 -170)	(171 – 299)	(300 – 400)	(> 440)
(μ mol/L) ou débito urinário L/dia				ou < 0,5	ou < 0,2

^a Valores com suporte ventilatório mecânico

^b PAM: pressão arterial média, em mmHg

^c Os agentes adrenérgicos devem ser administrados em perfusão contínua endovenosa, de duração superior a uma hora. As doses indicadas estão em μ g/Kg/minuto.

Embora o índice SOFA tenha sido desenvolvido primariamente com o objectivo de descrever a disfunção de órgão, vários estudos demonstram a sua relação com a mortalidade (Moreno et al., 1999) (Ferreira et al., 2001); (Amaral et al., 2005). Nestes estudos, os valores do SOFA score da admissão foram comparados com o SOFA score máximo, o SOFA score médio e a diferença dos SOFA scores (Δ SOFA). Os três últimos apresentam um valor preditivo mais elevado para a mortalidade (Le Gall, 2005).

Ferreira et al (Ferreira et al., 2001) demonstraram que os valores do SOFA médio apresentavam uma correlação significativa com a mortalidade (OR = 3,06; intervalo de confiança a 95%, 2,36 – 3,97) e mais elevada do que o valor do SOFA à admissão. O valor do SOFA máximo evidenciou a maior área sob a curva ROC (0,90) comparado com o valor do SOFA da admissão (área sob a curva ROC = 0,79). Estes investigadores revelaram que a diminuição dos valores de SOFA durante as primeiras 48 horas está associada ao *outcome* com uma redução da taxa de mortalidade de 50% para 27% (Moreno et al., 1999).

Duas falências de órgão (respiratório e outro) estão correlacionadas com um risco de mortalidade menor, enquanto que todas as outras combinações geram um risco de mortalidade entre 65 a 74% (Le Gall, 2005).

Vários investigadores têm demonstrado que a avaliação sequencial do SOFA tem melhorado a capacidade de previsão do *outcome* final.

Os valores do SOFA score superiores a 15 estão associados a uma taxa de mortalidade superior a 90% (Vincent, 1998); (Ferreira et al., 2001).

4.1.2 Índices de quantificação de falência múltipla de órgãos – objectivos

4.1.2.1 Multiple Organ Dysfunction Score (MODS)

O índice MODS (Quadro 6) foi descrito por Marshall et al em 1995 (Marshall et al., 1995) após uma extensa revisão crítica da literatura sobre critérios definidos para falência de órgãos. Foi testado e validado numa população de 692 doentes (Le Gall, 2005).

O índice MODS compreende um índice baseado no grau de disfunção ou falência de seis órgãos e sistemas (respiratório, renal, hepático, cardiovascular, hematológico e neurológico), sendo a pontuação para cada um deles variável entre 0 (normalidade) e 4 pontos (falência). As piores pontuações registadas para cada órgão ou sistema num determinado período de 24 horas são depois somadas para o cálculo do MODS diário, e assim consecutivamente durante o tempo de internamento na UCI.

O índice MODS difere dos outros sistemas pelo método escolhido para a avaliação do sistema cardiovascular. Enquanto que nos outros índices é avaliado um conjunto de parâmetros fisiológicos e/ou intervenções terapêuticas, no MODS a avaliação da falência cardiovascular é-nos dada por uma variável composta: a frequência cardíaca ajustada para a pressão arterial, o que torna o seu cálculo mais complexo, embora apresente uma excelente capacidade discriminativa quando comparada com a de outros índices (Moreno et al., 1997).

Os autores do sistema MODS demonstraram que a mortalidade dependia não só da pontuação registada à admissão na UCI, mas também da sua evolução (Le Gall, 2005).

Foi demonstrado que este sistema apresenta uma melhor capacidade discriminativa do que o APACHE II ou o OSF diários (Jacobs, 1999).

4.1.2.2 Logistic Organ Dysfunction System (LOD)

Proposto por Jean-Roger Le Gall et al (Le Gall et al., 1996); (Le Gall et al., 1997), o índice LOD foi obtido por regressão logística, a partir da base de dados europeia e norte-americana utilizada para o desenvolvimento do SAPS II e do MPM II.

Foram testadas doze variáveis e definidas seis falências de órgãos ou sistemas: respiratório, renal, hepático, cardiovascular, hematológico e neurológico.

Para cada variável, o valor escolhido é o mais anormal no período de 24 horas, sendo os valores em falta considerados normais para o cálculo da pontuação.

A originalidade do modelo LOD é a de dar uma pontuação de 0 a 5 pontos a cada disfunção e permitir o cálculo da probabilidade de mortalidade hospitalar com base no valor obtido às 24 horas de admissão na UCI, com o recurso à equação preditiva LOD. O intervalo para a probabilidade estimada varia desde 3,2% nos doentes sem qualquer disfunção/falência de órgão até 99,7% nos doentes com pontuação máxima (falência dos seis órgãos/sistemas analisados).

A diferença entre a pontuação LOD do 3º dia e do 1º dia de internamento é altamente preditiva do outcome hospitalar (Le Gall, 2005).

4.2 Comparação dos índices de avaliação da falência múltipla de órgãos e sistemas

As principais semelhanças e diferenças entre os índices encontram-se descritas no Quadro 6.

Os índices MODS, SOFA e LOD calculam um valor sumário. Estes índices diferem nas pontuações que atribuem e na forma como são derivadas. O LOD usa apenas a pontuação do primeiro dia na UCI. O índice SOFA é calculado diariamente. Essencialmente, os índices de disfunção de órgão expressam a disfunção de órgãos como um contínuo em oposição à variável binária (presente ou ausente) ou uma escala (ligeira, moderada, severa).

A diferença principal entre os índices mais utilizados reside no modo como definem a função cardiovascular: o MODS utiliza uma variável composta, isto é, uma frequência cardíaca ajustada para a pressão arterial.

O LOD emprega um princípio diferente, mais complexo, combinando o nível de disfunção de todos os órgãos numa pontuação global, calculada apenas no dia de admissão. Apesar das diferenças na estratégia do desenvolvimento (empírico versus técnicas estatísticas), dos objectivos para o desenvolvimento (previsão versus descrição), da forma de registo dos dados (pior pontuação para cada dia versus pior pontuação ocorrida durante a estadia na UCI) e dos pesos atribuídos a cada disfunção, os índices MODS, SOFA e LOD têm muitas similaridades, tais como a

escolha das variáveis, a escala de pontuação e o poder preditivo de mortalidade hospitalar dificultando a sua comparação.

Num estudo comparativo dos índices MODS, SOFA e LOD, para a previsão da mortalidade hospitalar por SDMO, foi demonstrado que os três índices apresentam capacidades discriminativas (valores das áreas sob a curva ROC entre 0,805 e 0,695) e calibração semelhantes e comparáveis com o APACHE III (Pettilä et al., 2002).

Quadro 6 – Variáveis utilizadas pelos índices de quantificação da falência múltipla de órgão mais utilizados. Adaptado de (Moreno, 2000).

Órgão/sistema	SOFA	MODS	LOD	OSF
Respiratório	PaO ₂ /FiO ₂ Ventilação mecânica	PaO ₂ /FiO ₂	PaO ₂ /FiO ₂ Ventilação mecânica	Frequência respiratória PaCO ₂ DAaO ₂ ^a Ventilação mecânica > 72 horas
Cardiovascular	Pressão arterial média Uso de vasopressores	PAR ^b	Pressão arterial sistólica Frequência cardíaca	Frequência cardíaca Pressão arterial média Taquicardia ventricular Fibrilação ventricular pH
Renal	Creatinina Débito urinário	Creatinina	Creatinina Ureia Débito urinário	Creatinina Ureia Débito urinário
Hematológico	Plaquetas	Plaquetas	Leucócitos Plaquetas	Hematócrito Leucócitos Plaquetas
Neurológico	Escala de Coma de Glasgow	Escala de Coma de Glasgow	Escala de Coma de Glasgow	Escala de Coma de Glasgow
Hepático	Bilirrubina	Bilirrubina	Bilirrubina Tempo de protrombina	-----

^a: diferença alvéolo-arterial de oxigênio

^b: frequência cardíaca ajustada para a pressão arterial

5. Limitações dos modelos prognósticos e dos índices de quantificação da falência múltipla de órgãos

Nesta secção, abordamos primeiro os modelos prognósticos e seguidamente os índices de quantificação de falência múltipla de órgãos.

Um sistema para comparar o desempenho das UCIs entre instituições deve estratificar de forma precisa os doentes aquando da admissão, e ser independente da terapêutica, enquanto que um sistema para suporte à decisão deve reflectir a resposta à terapêutica (Mackenzie, 2000)

Os modelos prognósticos gerais são utilizados nas UCIs na avaliação do desempenho da UCI (comparação das taxas normalizadas de mortalidade), no suporte à decisão clínica e na estratificação de doentes para ensaios clínicos (Le Gall, 2005).

As aplicações dos modelos prognósticos gerais a sua utilidade podem ser analisadas separadamente a nível agregado e individual.

A avaliação do desempenho da UCI tem estado focalizada na taxa de mortalidade normalizada, isto é, a razão entre a mortalidade observada e a mortalidade estimada pelo modelo prognóstico, com o pressuposto de que os modelos existentes podem tomar em consideração os principais determinantes da mortalidade (idade, estado de saúde prévio, diagnóstico de admissão e grau de disfunção fisiológico agudo) (Knaus et al., 1986). A interpretação da razão da mortalidade normalizada (RMN) é fácil, pois uma razão menor que 1 implica um desempenho melhor do que no grupo de referência e uma razão superior a 1 implica um desempenho pior do que no grupo de referência.

Não devemos concluir que o desempenho clínico de diferentes UCIs é necessariamente pior quando a mortalidade observada é maior do que a prevista (Knaus et al., 1986).

Para usar estas razões de forma eficaz, devemos saber qual a extensão do impacto na RMN de factores não clínicos. A RMN deve ser interpretada como um indicador que devemos analisar mais profundamente para identificar os factores associados ao diferencial da mortalidade observada. Estas probabilidades em si mesmo não controlam todos os factores que podem ter impacto no *outcome* final. Não controlam por exemplo as diferenças no *case mix* ou as disparidades nos recursos técnicos e terapêuticos disponíveis e as diferenças de gestão e organizacionais. Só considerando todos estes factores para além do valor da RMN, podemos estabelecer comparações com razoável fiabilidade (Hosmer & Lemeshow, 1989 , Le Gall, 2005, Teres & Lemeshow, 1994).

Esta metodologia tem sido utilizada para diversos propósitos como: comparações internacionais de UCIs (Knaus et al., 1982); (Rowan et al., 1993), comparações de hospitais (Le Gall et al., 1983); (Zimmerman et al., 1994), avaliações da UCI (Oh, Hutchinson et al. 1993), e avaliação da gestão (Zimmerman et al., 1994).

Contudo, são múltiplos os factores responsáveis pela variação significativa das taxas de mortalidade normalizada. Entre estes incluem-se o padrão de cuidados na UCI e no hospital, a validade dos modelos (discriminação e calibração) e o rigor na sua aplicação, as características do sistema de saúde, as características das populações e a dimensão da população em estudo.

O carácter geral destes sistemas é simultaneamente uma vantagem e uma desvantagem. Os grupos heterogéneos de doentes graves admitidos nas UCIs requerem modelos gerais que por definição podem ser menos sensíveis em relação a subgrupos de doentes (Ridley, 2002).

Os modelos gerais desenvolvidos nas UCIs europeias e norte-americanas não têm um pior desempenho do que em outros países, devido a diferenças no *case mix* (Parikh & Karnad, 1999).

As diferenças na qualidade dos cuidados prestados nas UCIs não são apercebidas pelos modelos obtidos por regressão logística múltipla (Ridley, 2002).

Cada um destes pressupostos tem sido alvo de investigação. Assim a taxa de mortalidade normalizada disponibiliza um indicador do desempenho, mas deve ser interpretada no contexto do sistema de cuidados de saúde onde é utilizada.

Uma área em que os índices de gravidade gerais e os modelos prognósticos têm ampla aceitação é a dos ensaios clínicos, quer no sentido de assegurar a comparabilidade de grupos quer na definição de critérios de inclusão e para aleatorização (Gatinoni et al., 1995).

A utilização dos modelos prognósticos gerais na previsão do *outcome* e no suporte à decisão do doente individualizado é controversa e as aplicações adequadas não são universalmente aceites (Teres & Lemeshow, 1994).

Para a previsão de *outcomes* dicotómicos (ex.: mortalidade), a regressão logística emergiu como a técnica estatística de eleição. Todos os modelos prognósticos gerais actuais usam equações de regressão logística para estimar as probabilidades de um determinado *outcome* num doente a partir de um conjunto de variáveis preditivas estabelecidas. A regressão logística múltipla fundamenta-se na capacidade do investigador que desenvolve o modelo, ter conhecimento a priori para seleccionar as variáveis relevantes com maior probabilidade de serem preditivas para o *outcome* de interesse (Buchman et al., 1994).

Contudo, esta técnica tem dois pressupostos que podem não ser completamente cumpridos quando se considera a previsão do *outcome* do doente grave. Idealmente, a variável dependente (*outcome*) na análise de regressão deveria ser contínua. Isto claramente não ocorre num *outcome* dicotómico como “morto” ou “vivo”. Segundo, é assumida uma relação linear entre as variáveis independentes e a variável dependente.

Matematicamente, é necessária a transformação do valor do índice de gravidade num *outcome* binário, e a transformação mais utilizada é a transformação *logit*: isto comprime o valor da pontuação num intervalo de 0 a 1. Esta compressão torna a relação entre o preditor independente e o *outcome* dependente mais sigmoide (Ridley, 2002).

Para além disto, a regressão logística pode não modelar adequadamente as relações muitas vezes complexas, não lineares e interdependentes encontradas nos sistemas biológicos (Rosenberg, 2002).

Por último, a presença de dados omissos e amostras de tamanho reduzido podem gerar modelos com fraco desempenho (Tu & Guerriere, 1993).

A grande vantagem desta técnica estatística é a sua fácil interpretação. A regressão logística disponibiliza os *odd ratios* e respectivos intervalos de confiança, assim como indica de forma transparente quais as variáveis mais importantes para calcular o *outcome* de interesse (Hosmer & Lemeshow, 1989).

Alguns erros no desenvolvimento do modelo podem advir do processo de selecção de populações e variáveis que são incluídas no modelo. A selecção de doentes pode conduzir a um fraco desempenho do modelo em populações distintas (Ridley, 2002)..

O *case mix* pode não ser adequadamente diverso (se limitado a uma única UCI). Por outro lado, os modelos prognósticos gerais não incluem todos os doentes admitidos nas UCIs, como por exemplo os com estadia de horas, os queimados, etc.

O *outcome* avaliado é habitualmente o estado vital à data da alta hospitalar. Contudo, é conceptualmente difícil de entender como o *outcome* à data da alta hospitalar está exclusivamente dependente da gravidade da doença aquando da admissão na UCI. Não sendo empregue uma avaliação sequencial, o impacto de *outcomes* intermédios, após o primeiro dia, é sub-valorizado (Ridley, 2002).

As variáveis preditivas resultam habitualmente da combinação das características demográficas do doente, da doença aguda e crónica, da disfunção fisiológica aguda, da terapêutica prévia e da localização pré-admissão à UCI.

Os modelos prognósticos são desenhados para disponibilizarem o melhor valor explicativo, utilizando o menor número de preditores, através da eliminação de

variáveis que estatisticamente não influenciam o *outcome*. Embora as técnicas estatísticas sejam robustas na execução deste processo de selecção, podem não conseguir capturar influências indirectas que afectem o *outcome* (Ridley, 2002).

Um exemplo possível é a estrutura organizativa e a qualidade de cuidados prestados numa determinada UCI. Tais factores são raramente capturados nos modelos de regressão. Mais ainda, pela sua natureza, estas técnicas matemáticas excluem variáveis que têm um menor contributo para o *outcome*. É possível que a influência das variáveis excluídas possam exceder o erro mínimo aceitável. Estas exclusões podem ser responsáveis pelos limites de sensibilidade corrente destes modelos e daí a sua inadequação para uso exclusivo destas ferramentas no processo de tomada de decisão no doente individualizado.

Na sua forma actual, a interpretação das estimativas prognósticas destes sistemas não deve ser utilizada para o prognóstico individualizado do doente, mas apenas para a avaliação de grupos de doentes, devido à sua natureza probabilística (Suter et al., 1994);(Gunning & Rowan, 1999)

Um bom modelo prognóstico disponibiliza uma estimativa fiável do número de doentes que irão morrer entre um grupo de doentes similares (Le Gall, 2005).

Por exemplo, num modelo bem calibrado, se em 100 doentes cada um tem uma probabilidade de mortalidade de 0,2, isto significa que é expectável que morram cerca de 20 destes doentes. Contudo, não nos indica quais os doentes individualizados que irão morrer ou os que irão sobreviver.

Os modelos prognósticos recentes têm uma especificidade elevada (90%), ou seja são capazes de prever a sobrevida, mas uma menor sensibilidade (50 a 70%), portanto menos fiáveis na previsão de morte (Herridge, 2003).

Para além disto, a aplicação de diferentes modelos no mesmo doente resulta frequentemente em predições diferentes (Lemeshow et al., 1995).

Estes factos limitam o seu uso nos processos de limitação ou suspensão terapêutica e na definição de futilidade médica(Atkinson et al., 1994); (Suter et al., 1994); (Herridge, 2003). Potencialmente, técnicas oriundas da Aprendizagem Automática, tais como as redes neuronais artificiais, poderão ultrapassar alguns dos problemas matemáticos inerentes à análise da regressão (Clermont et al., 2001); (Hanson & Marshall, 2001).

Apesar das definições estritas das variáveis preditivas, pode ser introduzido um viés¹⁶ aquando do cálculo da pontuação do índice de gravidade (Fery-Lemonnier et al., 1995); (Rowan, 1996); (Polderman et al., 2001b); (Polderman et al., 2001a).

¹⁶ Desvio dos resultados ou inferências do observado ou processos que levam a tal desvio.Ibid.

O viés de selecção resulta de diferentes métodos de registo das variáveis (manual ou automático), da variação na interpretação de um sintoma ou sinal clínico e do método para categorizar e pesar as variáveis contínuas.

Bosman *et al* (Bosman et al., 1998) registaram valores de pontuações mais altos dos índices de gravidade e consequentes predições de mortalidade mais elevadas quando são utilizados sistemas de informação automáticos. A mortalidade prevista aumentava cerca de 15% para o APACHE II, 25% para o SAPS II, e 24% para o MPM.

A classificação do diagnóstico agudo de admissão é particularmente difícil em muitos doentes graves, devido à presença concomitante de múltiplos problemas médicos. A selecção de um diagnóstico errado pode conduzir ao cálculo errado do risco de mortalidade.

A selecção de dados constitui não só um problema no desenho dos modelos mas também na sua aplicação. Dados acessíveis e de fácil registo são um pré-requisito para um modelo eficaz. A qualidade do registo dos dados é quantificada pela consistência inter e intra-observador. A consistência inter-observadora é definida pela capacidade de vários observadores registarem a mesma observação ao longo do tempo. A consistência observadora é definida pela capacidade de um único observador quantificar o mesmo parâmetro adequada e repetidamente ao longo do tempo. A consistência inter e intra-observadora podem ser quantificadas pelo coeficiente k para as observações binárias e pelo coeficiente de correlação intra-classe para as variáveis contínuas. O coeficiente K e o coeficiente de correlação intra-classe são designados por índices de concordância (Ridley, 2002).

Para o modelo SAPS II, os coeficientes de correlação intra-classe variam entre 0,81 e 0,95 e o teste k entre 0,67 e 1 (valor máximo para os dois testes é de 1, indicativo de concordância perfeita) (Le Gall et al., 1993).

Embora esta concordância seja muito boa em termos estatísticos, vários investigadores sugerem que usando os mesmo valores das variáveis, o risco de mortalidade calculado por diferentes indivíduos pode não ser o mesmo, particularmente para os doentes com um risco de mortalidade entre os 20 e os 80% (Chen et al., 1999).

Goldhill *et al* (Goldhill & Withington, 1996) demonstraram o quanto pequenas variações no registo de dados fisiológicos modificavam a pontuação do APACHE II bem como a razão da mortalidade normalizada.

Dado que os modelos prognósticos gerais destinam-se a prever o *outcome* à data da alta hospitalar, a sua utilização para prever outros *outcomes*, nomeadamente o *outcome* na UCI pode resultar numa sub-estimação grosseira da mortalidade hospitalar (Moreno, 2000).

Os índices de quantificação de órgão passíveis de serem obtidos diariamente possibilitam uma melhor monitorização da evolução da gravidade da doença durante a estadia do doente na UCI, em oposição ao período de 24 horas inicial. Permitem ainda uma melhor compreensão do que é verdadeiramente a disfunção de órgão adquirida na UCI, contudo não têm a capacidade de previsão.

O desenvolvimento de índices de disfunção/falência de órgãos disponibilizou instrumentos de descrição de morbilidade diária individualizada e têm uma boa correlação com a mortalidade (Ridley, 2002).

Os métodos de desenvolvimento dos índices de quantificação da disfunção de órgão são semelhantes aos usados no desenvolvimento dos modelos prognósticos. Daí que estes índices tenham as mesmas limitações que os modelos prognósticos (Ridley 2002).

São exemplo dessas limitações: a definição de disfunção/falência de órgão não é universal, algumas variáveis incluídas não são específicas de órgão, os tempos de registo (intervalos de 24 horas) e as variáveis laboratoriais não disponibilizam informação em tempo real, e têm custos associados. Para além disso, o agravamento de uma pontuação de disfunção de órgão pode reflectir o impacto da terapêutica (ex.: diminuição do *Glasgow Coma Score* devido a sedação do doente) (Cook et al., 2001).

O facto do tempo de registo destes sistemas ser de intervalos de 24 horas, não permite de alguma forma a valorização de sinais precoces de deterioração do estado clínico do doente (Rosenberg, 2002).

A previsão do desenvolvimento de disfunção de órgãos pode ser melhorada através de sistemas que prevejam a disfunção eminente, substituindo a medição do estado funcional diário dos órgãos e sistemas.

A inclusão de *outcomes* intermédios, baseados exclusivamente no grau de disfunção fisiológica aguda (ex.: duração e intensidade de anormalidade dos parâmetros vitais rotineiramente registados hora a hora nas UCIs), em tempo real permitiria uma intervenção precoce de forma a evitar a instalação de falência de órgão. Embora estes *outcomes* intermédios pareçam clinicamente intuitivos e racionais, esta abordagem não tem ainda validação clínica.

6. Actualizações dos modelos prognósticos

A utilização do modelo prognóstico fora da população onde foi desenvolvido pode implicar uma modificação (ajuste) do modelo utilizado (Moreno, 2000).

Uma das limitações de qualquer modelo é que o seu desempenho em populações diferentes daquelas em que foi desenvolvido é muitas vezes fraco (reflectindo uma fraca calibração) (Le Gall, 2005).

Na maioria dos casos, esta falta de calibração torna-se evidente em subgrupos de doentes e é geralmente acompanhada por uma sub-estimação da mortalidade em doentes de baixo risco e por uma sobre-estimação da mortalidade dos doentes de elevado risco (Metnitz et al., 2005).

Este padrão foi observado em todos os modelos de prognóstico publicados em vários países (Apolone et al., 1996);(Bastos et al., 1996); (Moreno & Morais, 1997); (Moreno et al., 1998b); (Rivera-Fernandez et al., 1998); (Zimmerman et al., 1996); (Metnitz et al., 1999); (Glance et al., 2000) e parece piorar ou agravar-se com o decorrer do tempo.(Popovich, 2002).

Embora os modelos mantenham um bom poder discriminativo, a sua calibração tem vindo a deteriorar-se ao longo do tempo, introduzindo diferenças crescentes entre a mortalidade observada e a prevista pelos modelos(Moreno & Matos, 2000).

A perda de calibração dos modelos existentes (mesmo os revistos na década de 90) pode também resultar de modificações das características das populações de doentes graves ao longo do tempo (Zhu et al., 1996), distribuição etária, distribuição de doenças e o desenvolvimento de novas terapêuticas, que podem afectar o prognóstico. Estes modelos podem tornar-se obsoletos ao longo do tempo, necessitando de actualizações (Herridge, 2003).

As actualizações e as modificações permitem que os modelos possam ser aplicados a grupos, condições e países específicos(Aergerter et al., 2005); (Le Gall et al., 2005) .

A primeira abordagem para melhorar a calibração de um modelo é o ajuste (Moreno & Apolone, 1997). Vários métodos têm sido propostos (Zhu et al., 1996), tais como o ajuste do logit, modificando os coeficientes da equação logística (sem modificar os pesos das variáveis)(Apolone et al., 1996, Le Gall, 1995, Rivera-Fernandez et al., 1998) e o ajuste de todos os coeficientes de todas as variáveis do modelo (Moreno & Apolone, 1997). Ambos os métodos têm sido usados com sucesso parcial no incremento da capacidade prognóstica dos modelos (Moreno and Apolone 1997; Metnitz, Valentin et al. 1999).

Os melhores resultados são obtidos quando é feita o ajuste para o ambiente local e o *case mix*. Contudo, estes processos impedem a comparação entre diferentes instituições e limita a sua validação numa nova população de doentes (Ridley, 2002).

Mesmo quando um modelo prognóstico é ajustado para utilização local, a sua sensibilidade e especificidade não se aproximam dos valores necessários para a previsão fiável do *outcome* individualizado (Moreno & Apolone, 1997).

7. Conclusões

Os modelos de prognóstico habitualmente em uso foram desenvolvidos recorrendo a várias técnicas, que incluem métodos estatísticos (ex.: regressão logística) e o julgamento clínico elaborado por peritos clínicos (Breiman et al., 1984); (Watson et al., 1985).

Durante os últimos quase trinta anos, os índices gerais quantitativos de gravidade de doença, os modelos de prognóstico e posteriormente as escalas de avaliação da falência múltipla de órgão utilizados na área da Medicina Intensiva têm tido um desenvolvimento e aperfeiçoamento significativos.

Na década de 90, os modelos prognósticos foram submetidos às últimas revisões e actualizações, de que resultaram novos modelos como o SAPS II, o MPM II e o APACHE III. Estes modelos foram desenvolvidos com o recurso a dados provenientes de bases de dados multicêntricas e à técnica da regressão logística múltipla.

Para a monitorização da evolução da gravidade da doença crítica e complementar o ajuste do risco, surgiram vários sistemas de avaliação da falência múltipla de órgãos, tais como o MODS, o SOFA score e o LODS.

Todos estes sistemas foram desenvolvidos para quantificar objectivamente o grau de disfunção fisiológica e comorbilidades a partir das quais são estimadas a probabilidade de mortalidade intra-hospitalar (Teres & Pekow, 2000)..

Estes sistemas têm como principais objectivos para além da estratificação do risco de mortalidade ajustado para o *case mix*, a normalização dos *outcomes* (taxa de mortalidade normalizada) e a possibilidade de auditar a prática clínica, no contexto do controlo de qualidade institucional. Acresce ainda, a sua ampla utilização para fins investigacionais (ex.: definição de critérios de inclusão em ensaios clínicos).

O desempenho destes instrumentos deverá ser avaliado por testes formais (avaliação da capacidade discriminativa e de calibração do modelo).

Todos estes modelos têm problemas potenciais de erros de classificação e de recolha de dados.

A natureza probabilística destes modelos limita a sua utilização como suporte à decisão no doente individualizado, comprometendo a sua utilização universal, à cabeceira do doente. Na sua forma actual, estes sistemas não devem ser usados para o prognóstico individualizado nem para definir futilidade médica.

Apesar das actualizações sucessivas dos modelos prognósticos, tem-se verificado que os seus desempenhos têm vindo a deteriorar-se com o tempo (perda de capacidade discriminativa e particularmente da calibração). Várias tentativas foram já feitas para a recalibração e expansão dos "velhos" modelos, com parcial sucesso.

A avaliação do desempenho das diferentes UCIs tem sido outro tópico de interesse da utilização dos modelos. Porém, continua controverso se é possível através destes instrumentos comparar a qualidade de cuidados de saúde.

Todos os modelos devem ser vistos como em permanente evolução. Muitas das características dos doentes e das UCIs ainda não foram incorporadas nos modelos actuais (ex.: a estrutura organizacional das UCIs, o dados da monitorização contínua dos parâmetros do doentes, etc.) (Herridge, 2003).

A obtenção de potenciais avanços na previsão do risco poderá passar pelo desenvolvimento e aplicação de novas técnicas de modelação, assim como pela incorporação de novas variáveis preditivas.

8. Bibliografia

- Abbott, R.R., Setter, M., Chan, S. & Choi, K. (1991) APACHEII, prediction of outcome of 451 ICU oncology admissions in a community hospital, *Ann. Oncol.*, 2(8), pp. 571– 574.
- Abu-Hanna, A. & Lucas, P.J.F. (2001) Prognostic models in medicine AI and statistical approaches, *Special issue of Methods Inf Med*, 40, pp. 1-5.
- Aegerter, P., Boumendil, A., Retbi, A., Minvielle, E., Dervaux, B. & Guidet, B. (2005) SAPS II revisited, *Intensive Care Med.*, 31, pp. 416–423.
- Amaral, A.C., Kajdacsy-Balla, Andrade, F.M. & Moreno, R. (2005) Use of the Sequential Organ Failure Assessment score as a severity score, *Intensive Care Med*, 31, pp. 243–249.
- Angus, D.C., Clermont, G. & Kramer, D.J. (2000) Short-term and long-term outcome prediction with the Acute Physiology and Chronic Health Evaluation II system after orthotopic liver transplantation, *Crit. Care Med*, 28, pp. 150–156.
- Apolone, G., D'Amico, R. & Bertolini, G. (1996) The performance of SAPS II in a cohort of patients admitted in 99 Italian ICUs: results from the GiViTI, *Intensive Care Med*, 22, pp. 1368–1378.
- Atkinson, S., Bihari, D. & Smithies, M. (1994) Identification of futility in intensive care, *Lancet*, 344, pp. 1203–1206.
- Bastos, P.G., Sun, X., Wagner, D.P., Knaus, W.A. & Zimmerman, J.E. (1996) Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study, *Intensive Care Med.*, 22 (6), pp. 564–570.
- Beal, A.L. & Cerra, F.B. (1994) Multiple organ failure syndrome in the 1990s: Systemic inflammatory response and organ dysfunction, *JAMA*, 271, pp. 226–233.
- Beck, D.H., Smith, G.B., Pappachan, J.V. & Millar, B. (2003) External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study, *Intensive Care Med*, 29(2), pp. 249–256.
- Bosman, R.J., Oudemans van Straaten, H.M. & Zandstra, D.F. (1998) The use of intensive care information systems alters outcome prediction, *Intensive Care Medicine*, 24, pp. 953–958.
- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and Regression Trees* (Belmont, CA, Wadsworth International).

- Buchman, T.G., Kubos, K.L., Seidler, A.J. & Siegforth, M.J. (1994) A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit, *Crit Care Med* 22, pp. 750-762.
- Capuzzo, M., Valpondi, V., Sgarbi, A., Bortolazzi, S. & Pavoni, V. (2000) Validation of severity scoring systems SAPS II and APACHE II in a single-center population, *Intensive Care Med*, 26(12), pp. 1779–1785.
- Carneiro, A.V., Leitão, M.P., Lopes, M.G. & De Pádua, F. (1997) Risk stratification and prognosis in critical surgical patients using the Acute Physiology, Age and Chronic Health III System, APACHE III, *Acta Med. Port.* , 10(11), pp. 751–760.
- Castella, X., A., A. & Bion, J. (1995) A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter multinational study. The European/North American Severity Study Group. , *Crit Care Med*, 23, pp. 1327 - 1335.
- Castella, X., Gilabert, J., Torner, F. & Torres, C. (1991) Mortality prediction models in intensive care, acute physiology and chronic health evaluation II and mortality prediction model compared, *Crit. Care Med*, 19(2), pp. 191–197.
- Champion, H.R., Copes, W.S. & Sacco, W.J. (1996) Improved predictions from a severity characterization of trauma (ASCOT) over trauma and injury severity score (TRISS), results of an independent evaluation, *J. Trauma*, 40, pp. 42-49.
- Chang, R.W., Jacobs, S. & Lee, B. (1988) Predicting outcome among intensive care unit patients using computerized trend analysis of daily Apache II scores corrected for organ system failure, *Intensive Care Med*, 14, pp. 558–566.
- Chen, L.M., Martin, C.M., Morrison, T.L. & Sibbald, W.J. (1999) Interobserver variability in data collection of the APACHE II, score in teaching and community hospitals, *Critical Care Medicine*, 27, pp. 1999–2004.
- Cher, D.J. & A., L.L. (1997) Method of Medicare reimbursement and the rate of potentially ineffective care of critically ill patients, *JAMA*, 278(12), pp. 1001–1007.
- Chiang, H.T., Lin, S.L., Hsu, H.C., Wann, S.R., Kung, M.H. & Liu, C.P. (2001) Prediction of in-hospital mortality in patients with myocardial infarction using APACHE II system, *Chung Hua I Hsueh Tsa Chih Chin, Med. J.* , 64(9), pp. 501–506.
- Chiavone, P.A. & Sens, Y.A. (2003) Evaluation of APACHEII system among intensive care patients at a teaching hospital, *São Paulo Med. J.*, 121(2), pp. 53–57.
- Chisakuta, A.M. & Alexander, J.P. (1990) Audit in intensive care, The APACHE II classification of severity of disease, *Ulster Med. J.*, 59(2), pp. 161–167.

- Clermont, G., Angus, D.C., DiRusso, S.M., Griffin, M. & Linde-Zwirble, W.T. (2001) Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models, *Crit Care Med* 29, pp. 291-296.
- Cook, D.A. (2000) Performance of APACHE III models in an Australian ICU, *Chest* 118(6), pp. 1732–1738.
- Cook, D.A., Joyce, C.J., Barnett, R.J., Birgan, S.P. & Playford, H. (2002) Prospective independent validation of APACHE III models in an Australian tertiary adult intensive care unit, *Anaesth. Intensive Care*, 30(3), pp. 308–315.
- Cook, R., Cook, D. & Tilley, J. (2001) Multiple organ dysfunction: baseline and serial component scores, *Crit. Care Med*, 29, pp. 2046–2050.
- David, T. & Knaus, W.A. (1991) Predicting outcome in critical care: the current status of the APACHE prognostic scoring system, *Can J Anaesth*, 38(3), pp. 374-383.
- Deitch, E.A. (1992) Multiple organ failure: pathophysiology and potential future therapy, *Ann. Surg*, 216, pp. 117–134.
- Del Bufalo, C., Morelli, A., Bassein, L., Fasano, L. & Quarta, C.C. (1995) Severity scores in respiratory intensive care, APACHE II predicted mortality better than SAPS II, *Respir. Care*, 40(10), pp. 1042–1047.
- Edwards, A.T., Ng, K., Shandall, A.A. & Price-Thomas, J.M. (1991) Experience with the APACHE II severity of disease scoring system in predicting outcome in a surgical intensive therapy unit, *J.R. Coll. Surg. Edinb.*, 36 (1), pp. 37–40.
- Esserman, L., J., B. & L., L. (1995) Potentially ineffective care, A new outcome to assess the limits of critical care, *JAMA*, 274(19), pp. 1544–1551.
- Fagon, J.Y., Chastre, J. & Novara, A. (1993) Characterization of intensive care unit patients using a model based on the presence or absence of organ dysfunctions and/or infection: The ODIN model. , *Intensive Care Med*, 19, pp. 137–144.
- Ferreira, F.L., Bota, D.P. & Bross, A. (2001) Serial Evaluation of the SOFA score to predict outcome in critically ill patients, *JAMA*, 286(1754-1758).
- Fery-Lemonnier, E., Landais, P., Loirat, P., Kleinknecht, D. & Brivet, F. (1995) Evaluation of severity scoring systems in ICUs—translation, conversion and definition ambiguities as a source of inter-observer variability in Apache II, SAPS and OSF, *Intensive Care Med.*, 21(4), pp. 356–360.
- Fry, D.E., Pearlstein, R. & Fulton, R.L. (1980) Multiple system organ failure: The role of uncontrolled infection, *Arch. Surg*, 115, pp. 136–140.
- Garden, O.J., Motyl, H. & Gilmour, W.H. (1985) Prediction of outcome following acute variceal haemorrhage, *Br. J. Surg*, 72, pp. 91-95.

- Gatinoni, L., Brazzi, L. & Pelosi, P. (1995) A trial of goal oriented hemodynamic therapy in critically ill patients, *NEJM*, 333, pp. 1025–1032.
- Giangiuliani, G., Mancini, A. & Gui, D. (1989) Validation of a severity of illness score, APACHE II, in a surgical intensive care unit, *Intensive Care Med*, 15(8), pp. 519–522.
- Glance, L.G., Osler, T.M. & Papadakos, P. (2000) Effect of mortality rate on the performance of the Acute Physiology and Chronic Health Evaluation II, a simulation study, *Crit Care Med.*, 28, pp. 3424–3428.
- Goldhill, D.R. & Withington, P.S. (1996) Mortality predicted by APACHE II, The effect of changes in physiological values and post-ICU hospital mortality, *Anaesthesia*, 51, pp. 719–723.
- Goris, R.J.A., Boekhorst, T.P.A. & Nuytinck, J.K.S. (1985) Multiple-organ failure, *Arch. Surg*, 120, pp. 1109–1115.
- Guiguet, M., Blot, F. & Escudier, B. (1998) Severity-of-illness scores for neutropenic cancer patients in an intensive care unit: which is the best predictor? Do multiple assessment times improve the predictive value?, *Crit Care Med*, 26, pp. 488–493.
- Gunning, K. & Rowan, K. (1999) Outcome data and scoring systems, *Br. Med. J. Clin. Res*, 319, pp. 241–244.
- Hadorn, D.C., Keeler, E.B. & Rogers, W.H. (1993) Assessing the performance of mortality prediction models RAND/ UCLA/Harvard Center for Health Care Financing Policy Research).
- Hanley, J.A. & Mc Neil, B.J. (1982) The meaning and use of the area under a receiver-operating characteristic (ROC) curve, *Radiology* 143, pp. 29–36.
- Hanson, C.W.r. & Marshall, B.E. (2001) Artificial intelligence applications in the intensive care unit, *Crit Care Med* 29, pp. 427–435.
- Headley, J., Theriault, R. & Smith, T.L. (1992) Independent validation of APACHE II severity of illness score for predicting mortality in patients with breast cancer admitted to the intensive care unit, *Cancer* 70(2), pp. 497–503.
- Hebert, P.C., Drummond, A.J. & Singer, J. (1993) A simple multiple system organ failure scoring system predicts mortality of patients who have sepsis syndrome., *Chest*, 104, pp. 230–235.
- Herridge, M.S. (2003) Prognostication and intensive care unit outcome: the evolving role of scoring systems, *Clin. Chest Med*, 24 (4), pp. 751–762.
- Higgins, T., Teres, D. & Copes, W. (2005) Preliminary update of the Mortality Prediction Model (MPM0), *Crit Care Clin*, 9, pp. 97.

- Hilden, J. & Habbema, J.D.F. (1987) Prognosis in medicine: an analysis of its meaning and role, *Theor Med*, 8, pp. 249-365.
- Holt, A.W., Bury, L.K. & Bersten, A.D. (1992) Prospective evaluation of residents and nurses as severity score data collectors, *Crit Care Med.*, 20, pp. 1688-1691.
- Hosmer, D.W. & Lemeshow, S. (1989) *Applied logistic regression* (New York, John Wiley & Sons).
- Hunt, J.P. & Meyer, A.A. (1997) Predicting survival in the intensive care unit,, *Curr. Prob. Surg.*, 34(7), pp. 527-599.
- Ihnsook, J., Myunghee, K. & Jungsoon, K. (2003) Predictive accuracy of severity scoring system: a prospective cohort study using APACHE III in a Korean intensive care unit, *Int J. Nurs. Stud.*, 40 (3), pp. 219-226.
- Jacobs, S., Arnold, A., Clyburn, P.A. & Willis, B.A. (1992) The Riyadh intensive care program applied to a mortality analysis of a teaching hospital intensive care unit, *Anaesthesia*, 47, pp. 775-780.
- Jacobs, S., Chang, R.W. & Lee, B. (1987) One year's experience with the APACHE II severity of disease classification system in a general intensive care unit, *Anaesthesia* Jul, 42 (7), pp. 738-744.
- Jacobs, S., Zuleika, M. & Mphansa, T. (1999) The multiple organ dysfunction scores as descriptor of patient outcome in septic shock compared with two other scoring systems, *Crit. Care Med*, 27, pp. 741.
- Jayes, R.L., Zimmerman, J.E., Wagner, D.P., Draper, E.A. & Knaus, W.A. (1993) Do-not resuscitate orders in intensive care units, Current practices and recent changes, *JAMA*, 270(18), pp. 2213-2217.
- Katsaragakis, S., Papadimitropoulos, K., Antonakis, P., Strergopoulos, S., Konstadoulakis, M.M. & Androulakis, G. (2000) Comparison of Acute Physiology and Chronic Health Evaluation II, APACHE II, and Simplified Acute Physiology Score II, SAPS II, scoring systems in a single Greek intensive care unit, *Crit. Care Med*, 28(2), pp. 426-432.
- Knaus, W.A., A., D.E., Wagner, D.P. & Zimmerman, J.E. (1985a) APACHE II, a severity of disease classification system, *Crit. Care Med.*, 13(10), pp. 818-829.
- Knaus, W.A., Draper, E.A. & Wagner, D.P. (1982) Evaluating outcome from intensive care, A preliminary multihospital comparison, *Crit Care Med.*, 10, pp. 491-496.
- Knaus, W.A., Draper, E.A. & Wagner, D.P. (1985b) Prognosis in acute organ-system failure, *Ann Surg*, 202, pp. 685-693.
- Knaus, W.A., Draper, E.A. & Wagner, D.P. (1986) An evaluation of outcome from intensive care in major medical centers, *Ann. Intern. Med.*, 104, pp. 410-418.

- Knaus, W.A., Harrell, F.E. & Lynn (1995) The SUPPORT prognostic model, objective estimates for seriously ill hospitalized adults, study to understand prognoses and preferences for outcomes and risks of treatments, *Ann. Intern. Med.*, 122, pp. 191-203.
- Knaus, W.A., J., Z., D.P., W., Draper, E.A. & Laurence, D.E. (1981) APACHE – Acute Physiology And Chronic Health Evaluation: a physiologically based classification system, *Crit Care Med*, 8, pp. 591-597.
- Knaus, W.A., Wagner, D.P. & Draper, E.A. (1984) The value of measuring severity of disease in clinical research on acutely ill patients, *J. Chronic Dis.*, 37(6), pp. 455–463.
- Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., M., B., G., B.P., Sirio, C.A., J., M.D., Lotring, T. & Damiano, A. (1991) The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults, *Chest* 100, pp. 1619–1636.
- Kollef, M.H. & Schuster, D.P. (1994) Predicting intensive care unit outcome with scoring systems, Underlying concepts and principles, *Crit Care Clin*, 10, pp. 1-18.
- Le Gall, J.-R. (2005) The use of severity scores in the intensive care unit,, *Intensive Care Med*, 31, pp. 1618-1623.
- Le Gall, J. (1995) Customized probability models for early severe sepsis in adult intensive care patients, *JAMA* 273(8), pp. 644-650.
- Le Gall, J.R., Klar, J. & Lemeshow, S. (1997) How to assess organ dysfunction in the intensive care unit? The logistic organ dysfunction (LOD) system, *Sepsis*, 1, pp. 45-47.
- Le Gall, J.R., Klar, J., Lemeshow, S., Saulnier, F. & Alberti, C. (1996) The logistic organ dysfunction system. A new way to assess organ dysfunction in the intensive care unit, *JAMA*, 276, pp. 802–810.
- Le Gall, J.R., Lemeshow, S. & Saulnier, F. (1993) A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study,, *JAMA*, 270, pp. 2957–2963.
- Le Gall, J.R., Loirat, P. & Alperovitch, A. (1984) A simplified acute physiology score for ICU patients, *Crit Care Med*, 12, pp. 975 – 977.
- Le Gall, J.R., Loirat, P. & Nicolas, F. (1983) Utilisation d'un indice de gravité dans huit service de réanimation multidisciplinaire, *Presse Médicale*, 12, pp. 1757–1761.
- Le Gall, J.R., Neumann, A., Hemery, F., Bleriot, J.P., Fulgencio, J.P., Garrigues, B., Gouzes, C., Lepage, E., Moine, P. & Villers, D. (2005) Mortality prediction using the SAPS II, an update for French ICUs, *Critical Care Med*, 9, pp. 645–652.

- Lemeshow, S. & Hosmer, D.W.J. (1982) A review of goodness of fit statistics for use in the development of logistic regression models, *Am J Epidemiol*, 115, pp. 92–106.
- Lemeshow, S., J., K. & Teres, D. (1995) Outcome prediction for individual intensive care patients: useful, misused, or abused?, *Intensive Care Med*, 21, pp. 770–776.
- Lemeshow, S., Klar, J., Teres, D., Avrunin, J.S., Gehlbach, S.H., Rapoport, J. & Rue (1994) Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective multicenter study, *Crit Care Med*, 22, pp. 1351–1358.
- Lemeshow, S. & Le Gall, J.R. (1994) Modelling the severity of illness of ICU patients, A system update, *JAMA*, 272 (13), pp. 1049-1055.
- Lemeshow, S., Teres, D. & Klar, J. (1993) Mortality Probability Models, MPM II, based on an international cohort of intensive care unit patients,, *JAMA*, 270, pp. 2478–2486.
- Livingston, B.M., MacKirdy, F.N., Howie, J.C., Jones, R. & Norrie, J.D. (2000) Assessment of the performance of five intensive care scoring models within a large Scottish database, *Crit. Care Med*, 28(6), pp. 1820–1827.
- Lloyd-Thomas, A.R., Wright, I., Lister, T.A. & Hinds, C.J. (1988) Prognosis of patients receiving intensive care for lifethreatening medical complications of haematological malignancy, *Br. Med. J. Clin. Res*, 296 (6628), pp. 1025–1029.
- Lucas, P.J.F. & Abu-Hanna, A. (1999) Prognostic methods in medicine, *Artif Intell Med* 15(2), pp. 105-119.
- Ludwigs, U. & Hulting, J. (1995) Acute physiology and chronic health evaluation II scoring system in acute myocardial infarction: a prospective validation study, *Crit. Care Med.*, 23(5), pp. 854–859.
- Mackenzie, S.J., Kendrick, S.W. & Howie, J.C. (2000) From severity scores to health gain-a difficult road but one worth travelling, *Current Opinion in Critical Care*, 6, pp. 181–186.
- Markgraf, R., Deutschinoff, G., Pientka, L. & Scholten, T. (2000) Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit, *Crit. Care Med*, 28 (1), pp. 26–33.
- Marshall, J., Cook, D.J., Christou, N.U., Bernard, G.R., Sprung, C.L. & Sibbald, W.J. (1995) Multiple organ dysfunction score: a reliable description of a complex clinical outcome, *Crit Care Med*, 23, pp. 1638–1652.

- Mendez-Tellez, P.A. & Dorman, T. (2005) Predicting patient outcomes, futility, and resource utilization in the intensive care unit: the role of severity scoring systems and general outcome prediction models, *Mayo Clin. Proc.*, 80(2), pp. 161–163.
- Metnitz, P.G., Moreno, R.P. & Almeida, E. (2005) SAPS 3, From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description,, *Intensive Care Med*, 31, pp. 1336–1344.
- Metnitz, P.G., Valentin, A. & Vesely, H. (1999) Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study, *Intensive Care Med*, 25, pp. 192–197.
- Moreno, R. (2000) A avaliação do prognóstico e da gravidade da doença, in: Permyer (Ed) *Gestão e Organização Medicina Intensiva*
- Moreno, R. & Afonso, S. (2006) Ehtical, legal and organizational issues in the ICU: prediction of outcome, *Curr Opin Crit Care*, 12, pp. 619-623.
- Moreno, R. & Apolone, G. (1997) The impact of different customization strategies in the performance of a general severity score, *Crit. Care Med*, 25, pp. 2001–2008.
- Moreno, R., Apolone, G. & Miranda, D.R. (1998a) Evaluation of the uniformity of fit of general outcome prediction models, *Intensive Care Med*, 24(1), pp. 40–47.
- Moreno, R. & Matos, R. (2000) The new scores: what problems have been fixed, and what remain, *Curr. Opin. Crit. Care*, 6, pp. 158 –165.
- Moreno, R. & Morais, P. (1997) Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study, *Intensive Care Med*, 23, pp. 177–186.
- Moreno, R., Pereira & Matos, E.R. (1997) The evaluation of cardiovascular dysfunction/failure in multiple organ failure (abstract), *Intensive Care Med*, 23, pp. 153.
- Moreno, R., Reis Miranda, D. & Fidler, D.V. (1998b) Evaluation of two outcome predictors on an independent database, *Crit Care Med*, 26, pp. 50-61.
- Moreno, R., Vincent, J.L., Matos, A., de Mendonça, A., Cantraine, F., Thijs, J., Takala, J., Sprung, C., Antonelli, M., Bruining, H. & Willatts, S. (1999) The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study, *Intensive Care Med*, 25, pp. 686–696.
- Moreno, R.P., Metnitz, P.G. & Almeida, E. (2005) SAPS 3, From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission, *Intensive Care Med*, 31, pp. 1345–1355.

- Nouira, S., Belghith, M., Elatrous, S., Jaafoura, M. & Ellouzi, M. (1998) Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units, *Crit. Care Med*, 26(5), pp. 852–859.
- Ohno-Machado, L., Resnic, F.S. & Mathne, M.E. (2006) Prognosis in critical care, *Annu Ver. Biomed Eng.*, 8, pp. 567-599.
- Oye, R.K. & Bellamy, P.E. (1991) Patterns of resource consumption in medical intensive care, *Chest*, 99(3), pp. 685–689.
- Pappachan, J.V., Millar, B., Bennett, E.D. & Smith, G.B. (1999) Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system, *Chest* 115(3), pp. 802–810.
- Parikh, C.R. & Karnad, D.R. (1999) Quality, cost, and outcome of intensive care in a public hospital in Bombay, India, *Critical Care Medicine*, 27, pp. 1754–1759.
- Patel, P.A. & Grant, B.J. (1999) Application of mortality prediction systems to individual intensive care units, *Intensive Care Med*, 25(9), pp. 977–982.
- Pettilä, V., Pettilä, M. & Sarna, S. (2002) Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill, *Crit Care Med*, 30, pp. 1705–1711.
- Pine, R.W., J., W.M. & Lennard, E.S. (1983) Determinants of organ malfunction or death in patients with intra-abdominal sepsis, *Arch. Surg*, 118, pp. 242–249.
- Polderman, K.H., Christiaans, H.M., Wester, J.P., Spijkstra, J.J. & Girbes, A.R. (2001a) Intra-observer variability in APACHE II scoring, *Intensive Care Med*, 27(9), pp. 1550–1552.
- Polderman, K.H., Girbes, A.R., Thijs, L.G. & Strack van Schijndel, R.J. (2001b) Accuracy and reliability of APACHE II scoring in two intensive care units. Problems and pitfalls in the use of APACHE II and suggestions for improvement, *Anaesthesia*, 56 (1), pp. 47–50.
- Popovich, M.J. (2002) If most intensive care units are graduating with honors, is it genuine quality or grade inflation?, *Crit. Care Med*, 30, pp. 2145–2146.
- Randolph, A.G., Guyatt, G.H. & Carlet, J. (1998) Understanding articles comparing outcomes among intensive care units to rate quality of care, *Evidence Based Medicine in Critical Care Group*, *Crit. Care Med*, 26(4), pp. 773–781.
- Ridley, S.A. (2002) Uncertainty and scoring systems, *Anaesthesia*, 57, pp. 761-767.
- Rivera-Fernandez, R., Vazquez-Mata, G., Bravo, M., Aguayo-Hoyos, E. & Zimmerman, J. (1998) The Apache III prognostic system: customized mortality predictions for Spanish ICU patients, *Intensive Care Med.*, 24(6), pp. 574–581.
- Rosenberg, A.L. (2002) Recent innovations in the intensive care unit risk-prediction models, *Current Opinion in Critical Care*, 8, pp. 321-330.

- Rosenberg, A.L., Zimmerman, J.E. & Alzola, C. (2000) Intensive care unit length of stay, recent changes and future challenges, *Crit Care Med*, 28, pp. 3465–3473.
- Rowan, K. (1996) The reliability of case mix measurements in intensive care, *Curr. Opin. Crit. Care*, 2, pp. 209–213.
- Rowan, K.M., Kerr, J.H. & Major, E. (1993) Intensive care society's APACHE II study in Britain and Ireland II: outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method, *Br. Med. J. Clin. Res*, 307, pp. 977-981.
- Rowan, K.M., Kerr, J.H., Major, E., Mc Pherson, K., Short, A. & Vessey, M.P. (1994) Intensive Care Society's Acute Physiology and Chronic Health Evaluation, APACHE II, study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients, *Crit. Care Med*, 22(9), pp. 1392–1401.
- Rubinfeld, G.D., Angus, D.C., Pinsky, M.R., Curtis, J.R., Connors, A.F.J. & Bernard, G.R. (1999) Outcomes research in critical care: results of the American Thoracic Society Critical Care Assembly Workshop on Outcomes Research, *Am. J. Respir. Crit. Care Med*, 160 (1), pp. 358–367.
- Rubinfeld, G.D. & Randall, C.J. (2001) End-of-Life Care in the ICU Working Group., End-of-life care in the intensive care unit: a research agenda., *Crit. Care Med.*, 29(10), pp. 2001–2006.
- Rue, M., Quintana, S. & Alvarez, M. (2001) Daily assessment of severity of illness and mortality prediction for individual patients, *Crit Care Med*, 29, pp. 45–50.
- Schuster, D.P. (1992) Predicting outcome after ICU admission, The art and science of assessing, *Chest*, 102, pp. 1861-1870.
- Sibbald, W.J. & Vincent, J.L. (1995) Round table conference on clinical trials for the treatment of sepsis, *Crit. Care Med.*, 23(2), pp. 394–399.
- Sicignano, A., Carozzi, C. & Giudici, D. (1996) The influence of length of stay in the ICU on power of discrimination of a multipurpose severity score, SAPS, ARCHIDIA, *Intensive Care Med*, 22, pp. 1048–1051.
- Suter, P., Argmandis, A. & Beaufils, F. (1994) Predicting outcome in ICU patients. Consensus Conference organized by the ESICM and SRLF., *Intensive Care Med*, 20, pp. 390-397.
- Tan, I.K. (1998) APACHE II and SAPS II are poorly calibrated in a Hong Kong intensive care unit, *Ann. Acad. Med. Singapore*, 27(3), pp. 318–322.
- Teno, J., Lynn, J., Connors, A.F.J., Wenger, N. & Phillips, R.S. (1997) The illusion of end-of-life resource savings with advance directives, SUPPORT Investigators,

- Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment, *J. Am. Geriatr. Soc.*, 45(4), pp. 513–518.
- Teres, D. (1994) Comment on "The Case for Using Objective Scoring Systems to Predict Intensive Care Unit Outcome, *Crit. Care Clin.*, 10, pp. 91–92.
- Teres, D. & Lemeshow, S. (1994) Why severity models should be used with caution, *Crit Care Clin* 10, pp. 93–110.
- Teres, D., Lemeshow, S., Avrunin, J.S. & Pastides, H. (1987) Validation of the mortality prediction model for ICU patients, *Crit. Care Med*, 15(3), pp. 208–213.
- Teres, D. & Pekow, P. (2000) Assessing data elements in a severity scoring system, *Intensive Care Med*, 26, pp. 263–264.
- Teskey, R.J., Calvin, J.E. & McPhail, I. (1991) Disease severity in the coronary care unit, *Chest*, 100 (6), pp. 1637–1642.
- Timsit, J.F., Fosse, J.P., Troch, G., De Lassence, A., Alberti, C., Garrouste-Orgeas, M., Azoulay, E., Chevret, S., Moine, P. & Cohen, Y. (2001) Accuracy of a composite score using daily SAPS II and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 hours, *Intensive Care Med*, 27, pp. 1012–1021.
- Tran, D.D., Miguel, M.A. & van Leeuwen, P.A.M. (1993) Risk factors for multiple organ system failure and death in critically injured patients., *Surgery*, 114, pp. 21–30.
- Tu, J.V. & Guerriere, M.R.J. (1993) Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery, *Comput Biomed Res* 26, pp. 220–229.
- Turner, J.S., Mudaliar, Y.M., Chang, R.W. & Morgan, C.J. (1991) Acute physiology and chronic health evaluation, APACHE II, scoring in a cardiothoracic intensive care unit, *Crit. Care Med*, 19 (10), pp. 1266–1269.
- Van, L.L., Fakhry, S., Walton, L.A., Moore, D.H., Fowler, W.C. & Rutledge, R. (1995) Use of the APACHE II scoring system to determine mortality of gynaecologic oncology patients in the intensive care unit, *Obstet. Gynecol.*, 85(1), pp. 53–56.
- Vassar, M.J., Lewis, F.R.J., Chambers, J.A., Mullins, R.J. & O'Brien, P.E. (1999) Prediction of outcome in intensive care unit trauma patients: a multicenter study of Acute Physiology and Chronic Health Evaluation APACHE, Trauma and Injury Severity Score (TRISS), and a 24-hour intensive care unit (ICU) point system, *J. Trauma-Injury Infection Crit. Care*, 47(2), pp. 324–329.
- Vincent, J.L. & Berre, J. (1998) An evidence-based approach to round tables and consensus conferences, *Crit. Care Clin.*, 14(3), pp. 539–547.

- Vincent, J.L., de Mendonça, A. & Cantraine, F. (1998) Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study, *Crit Care Med*, 26, pp. 1793–1800.
- Vincent, J.L., Ferreira, F. & Moreno, R. (2000) Scoring systems for assessing organ dysfunction and survival, *Crit. Care Clin.*, 16, pp. 353–366.
- Vincent, J.L., Moreno, R. & Takala, J. (1996) The SOFA, Sepsis-related organ failure assessment, score to describe organ dysfunction/failure, *Intensive Care Med*, 22, pp. 707–710.
- von Bierbrauer, A., Riedel, S., Cassel, W. & von Wichert, P. (1998) Validation of the acute physiology and chronic health evaluation, APACHE III scoring system and comparison with APACHE II in German intensive care units, *Anaesthesist*, 47(1), pp. 30–38.
- Wagner, D.P., Draper, E.A. & Campos, A.R. (1984) Initial international use of APACHE: an acute severity of disease measure, *Med Decis Making*, 4, pp. 297.
- Wagner, D.P., Knaus, W.A. & Draper, E.A. (1983) Statistical validation of a severity of illness measure, *Am. J. Public Health*, 73 (8), pp. 878–884.
- Wagner, D.P., Knaus, W.A. & Harrell, F.E. (1994) Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis, *Crit. Care Med*, 22, pp. 1359–1372.
- Watson, J.H., Sax, H.C., Neff, R.K. & Goldman, L. (1985) Chemical prediction rules: Applications and methodological standards, *N Engl J Med* 313, pp. 793-799.
- Wilairatana, P., Noan, N.S., Chinprasatsak, S., Prodeengam, K., Kityaporn, D. & Looareesuwan, S. (1995) Scoring systems for predicting outcomes of critically ill patients in north-eastern Thailand, *Southeast Asian, J. Trop. Med. Public Health*, 26(1), pp. 66–72.
- Windeler, J. (2000) Prognosis, What does the clinician associate with this notion? , *Stat Med*, 19, pp. 425-430.
- Wong, D.T., Crofts, S.L., Gomez, M., McGuire, G.P. & Byrick, R.J. (1995) Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients, *Crit. Care Med*, 23(7), pp. 1177–1183.
- Zhu, B.-P., Lemeshow, S., Hosmer, D.W., Klarm, J., Avrunin, J. & Teres, D. (1996) Factors affecting the performance of the models in the mortality probability model and strategies of customization: a simulation study, *Crit. Care Med*, 24, pp. 57– 63.
- Zimmerman, J.E., Knaus, W.A. & Wagner, D.P. (1996) A comparison of risks and outcomes for patients with organ system failures 1982–1990, *Crit Care Med*, 24, pp. 1633–1641.

- Zimmerman, J.E., Kramer, A.A., McNair, D., S. & Malila, F.M. (2006) Acute Physiology and Chronic Health Evaluation, APACHE, IV, Hospital mortality assessment for today's critically ill patients, *Crit Care Med*, 34, pp. 1297–1310.
- Zimmerman, J.E., Rousseau, D.M. & Duffy, J. (1994) Intensive care at two teaching hospitals: an organizational case study, *Am. J. Crit. Care*, 3, pp. 129-138.
- Zimmerman, J.E., Wagner, D.P., Draper, E.A., Wright, L., Alzola, C. & Knaus, W.A. (1998) Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database, *Crit. Care Med.*, 26(8), pp. 1317–1326.
- Zweig, M.H. & Campbell, G. (1993) Receiver-operating characteristic (ROC) plots, a fundamental evaluation tool in clinical medicine, *Clin Chem*, 39, pp. 561–577.

CAPITULO III

A Descoberta de Conhecimento em Medicina Intensiva

CAPÍTULO III - A Descoberta de Conhecimento em Medicina Intensiva

Abordam-se o processo de Descoberta de Conhecimento em Bases de Dados e algumas técnicas de Data de Mining (redes neuronais artificiais, árvores de decisão e indução de regras de decisão) como um conjunto de soluções metodológicas e técnicas para a análise dos dados clínicos, desenvolvimento de modelos de previsão e ferramentas de suporte à tomada de decisão .

1. Introdução

Na área da Medicina Intensiva, o desenvolvimento de modelos gerais para a estratificação do risco de mortalidade e a quantificação da falência de órgãos e sistemas, utilizáveis como ferramentas de suporte à decisão clínica, continua a ser objecto de investigação. A maioria destes modelos foi desenvolvida utilizando o parecer de peritos e técnicas estatísticas. Estes modelos são caracterizados pela utilização de um número reduzido de variáveis, no centro dos quais existe uma variável de "pontuação", reflectindo o estado da gravidade da doença e pela confiança no modelo logístico probabilístico.

Na área da Aprendizagem Automática têm sido desenvolvidas várias técnicas para geração de modelos de previsão, muitas das quais caracterizadas pela utilização de um número extenso de variáveis, pela sua natureza não-paramétrica e muitas vezes simbólica (Abu-Hanna & Lucas, 2001, Lucas & Abu-Hanna, 1999).

O processo de tomada de decisão é um ponto-chave na prática clínica, quer no processo de cuidados de saúde quer na avaliação do prognóstico do doente grave admitido na UCI. Alguns estudos publicados revelam que a taxa do erro médico poderá ser a oitava causa de mortalidade nos países industrializados (Kohn et al., 2000).

Em qualquer situação, este processo de tomada de decisão deve estar suportado por critérios de evidência e de experiência. Para isso, é fundamental reconhecer e processar adequadamente toda a informação referente ao estado clínico do doente. A informação clínica disponível distribui-se por múltiplas variáveis que reflectem a

condição clínica do doente. A partir destas variáveis, é necessário extrair uma conclusão adequada sobre a possível evolução da (s) doença (s). Considerando a complexidade do problema, a avaliação e integração da informação devem ser realizadas por forma a que se considerem simultaneamente todas as variáveis implicadas.

Apesar do enorme volume de dados acessíveis, a síntese e a interpretação da informação continua a ser feita através de ferramentas de análise e estatística, na maioria das UCIs (Hanson & Marshall, 2001).

Esta circunstância pode conduzir a situações de sub-utilização da informação disponível ou à sua interpretação potencialmente errónea (Rosenfeld et al., 2000).

É provável que haja mais informação nos dados do que aquela que é correntemente extraída destes. Este facto constitui um “novo” desafio para o intensivista: como extrair e integrar toda a informação relativa à monitorização do estado clínico para o estabelecimento do prognóstico dos doentes no contexto do suporte à tomada de decisão.

O crescimento exponencial dos dados e as limitações das técnicas estatísticas habituais conduziram à necessidade do desenvolvimento de novas ferramentas e tecnologias para o processamento de dados em informação útil, de forma inteligente e automática (Halpern & Pastores, 1999) (Lee & Siau, 2001, Lucas et al., 2000, Rosenfeld et al., 2000).

Para a transformação dos dados em informação útil tem-se assistido a um influxo de novos métodos oriundos da área da Estatística, da Aprendizagem Automática, do Reconhecimento de Padrões¹⁷ e das técnicas de Bases de Dados (Lucas, 2004).

O conjunto de técnicas associadas ao *Data Mining*, fase de um processo mais amplo denominado como Descoberta de Conhecimento em Bases de Dados (Fayyad et al., 1996a), disponibiliza um conjunto de soluções para a análise semi-automática de dados presentes em bases de dados extensas e complexas, para a extracção de conhecimento¹⁸, previamente não conhecido e potencialmente útil (Bellazzi & Zupan, 2006).

Ao contrário dos métodos estatísticos habituais (ex.: regressão logística múltipla), em que a estrutura matemática dos modelos é fornecida e a informação probabilística é apreendida a partir dos dados, a característica de algumas abordagens da Aprendizagem Automática, é que a estrutura dos modelos é apreendida automaticamente (Lucas, 2004).

¹⁷ Requisitos obrigatórios utilizados e impostos para atingir uma abordagem disciplinada e uniforme.

¹⁸ Colecção de factos, acontecimentos, crenças e regras, organizadas para uso sistemático.

Nestes sistemas, os dados individuais dos doentes graves presentes nas bases de dados são apresentados aos algoritmos¹⁹ de *Data Mining* para a construção de modelos de previsão clínica, tais como as árvores de decisão e as redes neuronais artificiais.

2. O Processo da Descoberta de Conhecimento em Base de Dados

Os sistemas de informação clínica, implementados no contexto das unidades de prestação de cuidados de saúde (ex.: UCIs), possuem, de uma forma crescente, quantidades consideráveis de conhecimentos de nível factual relativamente aos procedimentos médicos, aos problemas dos doentes, aos diagnósticos e aos planos terapêuticos (Halpern & Pastores, 1999, Hanson & Marshall, 2001, Lucas et al., 2000). Começa a dar-se particular atenção ao conhecimento clínico onde, para além de se incluir a informação clínica do doente, também se regista informação sobre a investigação clínica, isto é, sobre os processos de decisão dos intensivistas.

As bases de dados assim construídas constituem uma excelente fonte de exemplos ou casos que podem ser, e começam já a ser, utilizados para extracção de conhecimento sobre os processos de raciocínio médico e sobre o comportamento dos doentes (ex.: previsão do seu estado de saúde no futuro, probabilidade de sobrevida, previsão do estado funcional dos seus sistemas orgânicos).

A “nova” geração de técnicas e ferramentas computacionais capazes de apoiar o intensivista tarefa de extracção de conhecimento útil das bases de dados. Esta metodologia e técnicas são o objecto do processo de Descoberta de Conhecimento em Bases de Dados (Fayyad et al., 1996a) e têm tido sido cada vez mais referidos na literatura médica (Bellazzi & Zupan, 2006).

2.1 Definição

A Descoberta de Conhecimento em Bases de Dados é um processo de selecção, exploração e modelação de grandes quantidades de dados com o objectivo de descobrir padrões úteis e relações previamente não conhecidos, utilizáveis para o suporte à tomada de decisão (Giudici, 2003).

¹⁹ Conjunto ordenado e finito de regras bem definidas destinadas à resolução de um problema. APDSI-Associação para a Promoção e Desenvolvimento da Sociedade de Informação. (2007). "Glossário da Sociedade da Informação – Versão 2007."

A Descoberta de Conhecimento em Bases de Dados evoluiu da intersecção da investigação provinda de várias áreas científicas tais como a Estatística, a Inteligência Artificial e a Aprendizagem Automática. A Aprendizagem Automática tem por objectivo disponibilizar métodos computacionais para acumular, alterar e actualizar a informação em sistemas inteligentes e em mecanismos específicos de aprendizagem (treino) de forma a induzir conhecimento a partir de dados.

O que diferencia o processo de Descoberta de Conhecimento em Bases de Dados das áreas referidas é a sua focalização no processo global de descoberta de conhecimento a partir dos dados, incluindo a forma como os algoritmos podem ser aplicados a bases de dados com eficácia, e como a interacção utilizador-máquina pode ser apoiada (Fayyad et al., 1996b).

A aquisição automática do conhecimento a partir de casos ou bases de dados tornou-se possível graças ao desenvolvimento dos algoritmos oriundos da Aprendizagem Automática.

O mecanismo de aquisição de conhecimento consiste na aplicação de um algoritmo ou conjunto de algoritmos na base de dados. O conhecimento inferido por esses algoritmos dá-se sob a forma de padrões ou modelos.

As primeiras tentativas de aplicação dos métodos de Aprendizagem Automática em bases de dados tornaram evidente a necessidade de uma abordagem sistematizada (Frawley et al., 1991). Esta abordagem sistematizada, integrando fases de pré-processamento dos dados e pós-processamento dos resultados é designada por Descoberta de Conhecimento em Bases de Dados.

Fayad et al (Fayyad et al., 1996a) definem a Descoberta de Conhecimento em Bases de Dados como: *um processo não trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados existentes.*

Pode-se assim definir a Descoberta de Conhecimento em Bases de Dados como um processo computacional interactivo não trivial (na medida em que se trata de um processo com alguma autonomia de pesquisa) para identificar novos padrões nos dados que sejam válidos, potencialmente úteis e interpretáveis (Lee & Siau, 2001).

Examine-se cada um destes termos:

- dados são um conjunto de factos F (casos de uma base de dados);
- padrão é uma expressão E numa linguagem L que descreve um subconjunto de factos FE do conjunto F ; os padrões descobertos devem ser válidos em novos dados com algum grau de confiança;

- o grau de confiança associado a um padrão E pode ser definido como a função $c = C(E, F)$; os padrões devem ser novos para o sistema;
- novidade pode ser medida em relação às alterações verificadas ao nível dos dados ou do conhecimento, representada por uma função $N(E, F)$;
- utilidade, representa até que ponto o padrão contribui para os objectivos inerentes ao processo, a utilidade pode ser definida pela função $u = U(E, F)$;
- interpretável deve ser entendido na perspectiva que um dos objectivos do processo é gerar padrões que sejam compreendidos, isto é, contribuir para uma melhor compreensão dos dados; assume-se que o grau de interpretação de um padrão é definido pela função $s = S(E, F)$.

O conhecimento descoberto pode ser quantificado, seja $i = I(E, F, C, N, U, S)$ o grau de interesse num dado padrão E , diz-se que o padrão E é conhecimento se para um valor dado i , $I(E, F, C, N, U, S) > i$ (Giudici, 2003).

O processo de Descoberta de Conhecimento em Bases de Dados não deve ser entendido como um simples cálculo de quantidades predefinidas, mas também como a procura de modelos ou padrões. Um padrão é uma expressão genérica formal que descreve um subconjunto de dados, induzindo o desenvolvimento de um modelo a partir de um conjunto de dados (Fayyad et al., 1996b).

Um exemplo de padrão potencialmente útil, expresso em forma de regra, é:

SE PRESSÃO ARTERIAL MÉDIA > 70 ENTÃO BOA PERFUSÃO.

Este padrão embora restritivo é facilmente entendido pelo utilizador e poderá ser utilizado para a definição de acções.

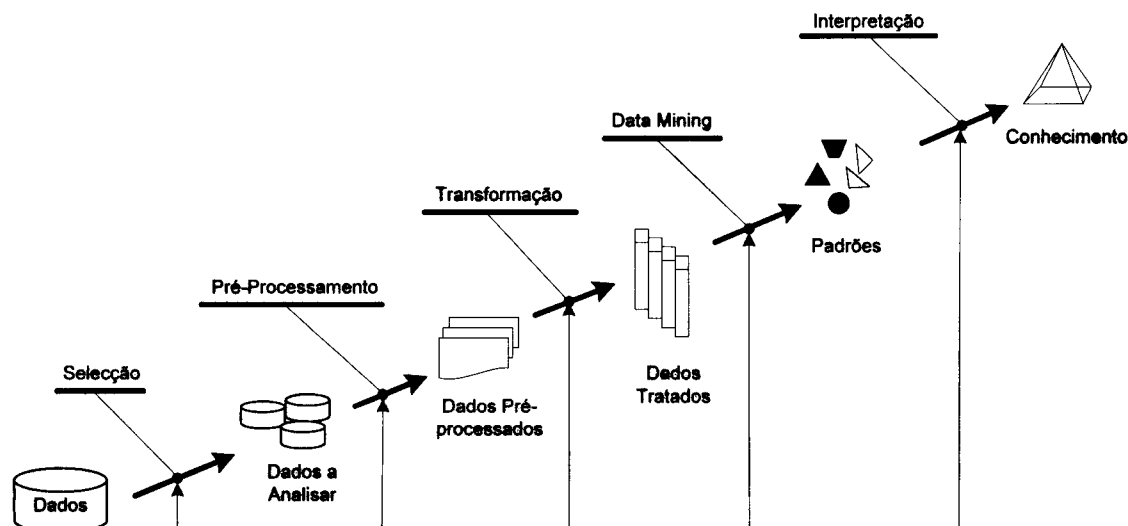
A Descoberta de Conhecimento em Bases de Dados é um processo não linear, isto é, composto por diversas fases, que podem ser repetidas em múltiplas iterações, e que envolve uma interacção com o utilizador (Fayyad et al., 1996b).

2.2 Fases do Processo de Descoberta de Conhecimento em Bases de Dados

Em termos genéricos, o processo de Descoberta de Conhecimento em Bases de Dados (Figura 1) consiste em: definição dos objectivos do processo; estudo e

preparação dos dados; escolha e aplicação dos algoritmos de extracção de padrões e avaliação dos resultados (Fayyad et al., 1996b).

Figura 1 – Fases do processo de Descoberta do Conhecimento em Bases de Dados. Adaptado de (Fayyad, Shapiro et al. 1996).



O processo de Descoberta de Conhecimento em Bases de Dados inicia-se com a definição dos objectivos e a selecção de dados, bem como a criação de um conjunto de dados alvo (subgrupo de variáveis) sobre os quais a “descoberta” será aplicada. O pré-processamento de dados inclui a análise e a transformação dos dados para que se apresentem sob a forma adequada ao algoritmo ou algoritmos de *Data Mining* que irão ser utilizados (Weiss & Indurkha, 1998).

Neste processo pode ser necessária a redução dos dados. O que se pretende é eliminar variáveis que não sejam importantes para o resultado final. Os métodos mais usados para a escolha das variáveis a serem eliminadas incluem métodos estatísticos de verificação de independência das variáveis, métodos heurísticos e árvores de decisão (Weiss & Indurkha, 1998).

É ainda na fase de pré-processamento dos dados que é feito o eventual enriquecimento da base de dados, com variáveis derivadas ou com medidas de sumarização.

Prosseguindo o processo, chega-se à fase de *Data Mining*, que começa com a escolha dos métodos e técnicas a serem aplicados (ex.: redes neuronais artificiais, árvores de decisão, técnicas estatísticas) tanto isoladamente como combinados. Essa escolha depende fundamentalmente do objectivo do *Data Mining* e do tipo de dados disponíveis. O *Data Mining* será abordado na secção seguinte deste capítulo.

Na fase de interpretação, são apresentados os padrões identificados e interpretados como conhecimento, o qual pode ser usado como suporte à tomada de decisão. Este conhecimento traduz-se na compreensão dos conteúdos de uma base de dados e na avaliação da sua utilidade/validade (Fayyad et al., 1996b). O conhecimento encontrado deve ser submetido a processos de validação ou verificação, entendidos como uma avaliação do desempenho do mecanismo de aprendizagem (ex.: ao analisar os padrões, o utilizador pode chegar à conclusão de que não são válidos e fazer com que o processo volte à fase de pré-processamento). Estas iterações estão representadas na Figura 1 pelas setas de menor espessura.

2.3 Data Mining

2.3.1 Definição

O *Data Mining* é a fase central do processo de Descoberta de Conhecimento em Bases de Dados e a que tem recebido mais atenção por parte dos investigadores (Benoit, 2002, Fayyad, 1991, Frawley et al., 1991, Weiss & Indurkha, 1998).

De uma forma geral, o *Data Mining* não tem estado associado a uma definição precisa, mas sim a um entendimento comum do seu significado: o uso de “novos” métodos e ferramentas para análise de grandes quantidades de dados (Bellazzi & Zupan, 2006).

O *Data Mining* é entendido como a etapa da extracção dos padrões (modelos) existentes nos dados (Goebel & Gruenwald, 1999), resultante da aplicação de algoritmo(s) específico(s) (Berry & Linoff, 1997, Bigus, 1996), de acordo com os objectivos previamente definidos, para o suporte de tomada de decisão (Benoit, 2002).

O *Data Mining* é uma etapa da Descoberta de Conhecimento em Bases de Dados, situada entre a transformação de dados e a interpretação dos resultados, que se caracteriza pela escolha e aplicação do (s) algoritmo (s), provenientes de diversas áreas do conhecimento, particularmente da área da Aprendizagem Automática, Visualização de Dados, Bases de Dados Multidimensionais e da Estatística (Bellazzi & Zupan, 2006).

Quando se fala do *Data Mining*, não se está a considerar apenas consultas complexas de bases de dados com o intuito restritivo de validar uma hipótese em função dos relacionamentos existentes entre os dados, mas também a procura de novas regularidades, restrições, padrões e relacionamentos (Fayyad et al., 1996a).

O *Data Mining* implica uma abordagem sistematizada para a identificação de associações previamente desconhecidas (padrões e relações), podendo envolver a geração automática de novas hipóteses (Pendharkar et al., 1999).

Assim, o *Data Mining* envolve um conjunto de técnicas seleccionadas com base na sua adequação para uma tarefa específica (Richards et al., 2001).

O *Data Mining* pressupõe que o investigador tenha um bom conhecimento dos dados e uma ideia precisa, obtida pelo conhecimento prévio, sobre a utilidade potencial da informação que poderá ser conseguida (Bath et al., 2000).

Por exemplo, o *Data Mining* procura identificar as variáveis independentes para a ocorrência de um *outcome* de interesse, complementando os métodos estatísticos que testam a existência de uma associação significativa entre factores de risco e o *outcome*. Esta abordagem terá potencialmente mais sucesso, quando no *Data Mining* for envolvido o conhecimento de um perito do domínio em estudo (Bath et al., 2000).

Nos últimos anos, o termo *Data Mining* tem sido usado de forma crescente na literatura médica. Contudo, a sua aplicação na análise de dados clínicos, apesar das expectativas elevadas, tem sido relativamente limitada (Bellazzi & Zupan, 2006). Isto é particularmente verdade em relação à sua aplicação prática em Medicina Intensiva, a qual poderá beneficiar de abordagens de *Data Mining* específicas, capazes de desenvolver modelos de previsão explorando o conhecimento clínico disponível e explicando as decisões propostas, uma vez utilizado o modelo construído para o suporte à tomada da decisão clínica (Richards et al., 2001).

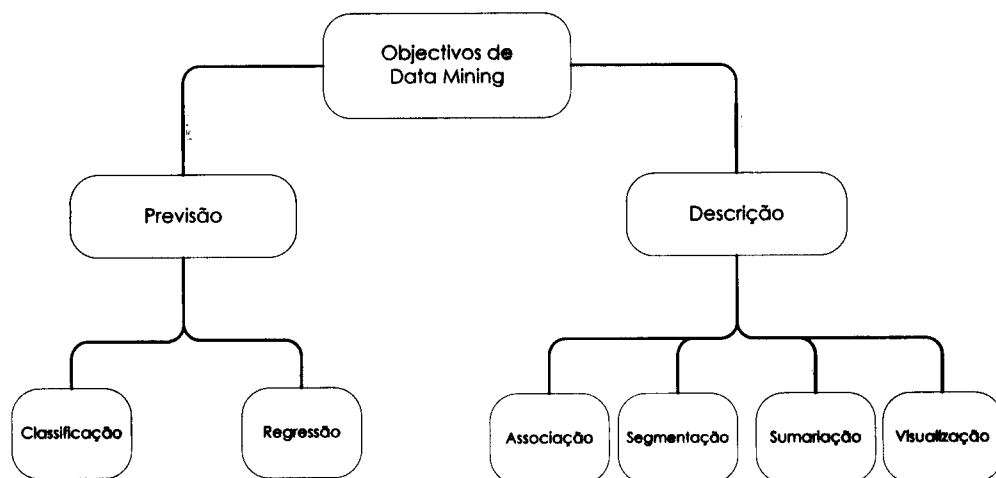
2.3.2 Objectivos do *Data Mining*

De uma forma geral as tarefas do *Data Mining* podem ser classificadas em tarefas de descrição e previsão (Figura 2) (Dietterich, 1997, Fayyad et al., 1996a, Kibler & Langley, 1988, Schaffer, 1994). A descrição procura descobrir padrões e associações interpretáveis, após considerar os dados como um todo. Embora os objectivos da descrição e da previsão se possam sobrepor, a principal diferença é que a previsão pressupõe que os dados incluam uma variável de resposta particular (Bellazzi and Zupan 2006).

A previsão tem como objectivos a classificação e a regressão e consiste em utilizar algumas das variáveis da base de dados para prever valores desconhecidos de outras variáveis de interesse. A classificação consiste em construir um modelo que possa ser

aplicado aos dados não classificados (variáveis discretas) visando categorizá-los em classes (ex.: *outcomes*) (Frawley et al., 1991, Freitas, 2000, Harrison, 1998).

Figura 2 – Objectivos do *Data Mining*. Adaptado de (Rezende, 2003)



A regressão tem como objectivo a construção de uma função que permita conhecer o valor de uma variável em função dos valores das variáveis dessa função (Fayyad et al., 1996a), isto é, definir um valor para uma variável contínua desconhecida (Harrison, 1998).

Pode-se apontar, como exemplo de tarefa de regressão, estimar a probabilidade de um doente falecer baseando-se nos resultados de um conjunto de variáveis clínicas.

A descrição tenta encontrar padrões interpretáveis que descrevam os dados e tem como objectivos a associação, a segmentação, a sumariação e a visualização.

A associação determina as relações entre as variáveis na base de dados.

Na segmentação ou *clustering*, o conjunto de registos é dividido em subconjuntos. Trata-se de um processo de partição de uma população heterogénea em vários subgrupos ou segmentos (*clusters*) homogéneos (Schwabacher et al., 1995). Na segmentação não há classes predefinidas e os registos são agrupados de acordo com a sua semelhança, o que a diferencia da tarefa de classificação (Fayyad et al., 1996a).

Num estudo de prognóstico, por exemplo, pode ser útil utilizar a segmentação para agrupar doentes e depois estudar cada um dos grupos obtidos em particular.

Segundo Fayyad et al (Fayyad et al., 1996a) a tarefa de sumariação envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Os métodos de sumariação envolvem a criação de regras de resumo e a descoberta de relações funcionais entre variáveis. As técnicas de sumariação são frequentemente aplicadas à

análise exploratória dos dados e à geração automática de relatórios (Fayyad et al., 1996a).

A visualização trata da apresentação dos dados, geralmente através de gráficos ou diagramas, de forma a facilitar a sua interpretação (Benoit, 2002).

O objectivo do *Data Mining* na previsão em Medicina Intensiva é o de desenvolver modelos que possam utilizar informação específica do doente para a previsão do *outcome* de interesse e consequentemente apoiar o processo de decisão (Bath et al., 2000, Michalski et al., 1997).

Os métodos do *Data Mining* podem ser aplicados para a construção de modelos de prognóstico, os quais, uma vez validados, podem ser incorporados nos sistemas de informação clínica (Bellazzi & Zupan, 2006).

2.3.3 Técnicas de *Data Mining*

As técnicas utilizadas na fase do *Data Mining* (redes neuronais artificiais, árvores de decisão e indução de regras de decisão) têm origem em várias áreas da ciência e usam abordagens diferentes de modelação (Bellazzi & Zupan, 2006) podendo ser comparadas com base nos seguintes pressupostos:

- o manuseamento de dados omissos e ruído;
- o tratamento de diferentes tipos de variáveis (categóricas, ordinais, contínuas); o tipo de apresentação dos modelos (que podem permitir ou não que o perito no domínio examine e perceba o processamento interno);
- a redução do número de testes (Lavrac, Keravnou et al. 1997), isto é, a redução do número de variáveis necessárias para derivar a conclusão;
- a capacidade de explicar as decisões obtidas quando os modelos são usados no processo de tomada de decisão;
- a generalização (isto é, capacidade de bom desempenho em dados ou casos novos).

A escolha dos algoritmos a utilizar deverá ser feita de acordo com os objectivos. No processo de escolha do algoritmo a aplicar deve começar-se por exprimir o problema em termos de uma das tarefas do *Data Mining* (Frawley et al., 1991).

As ferramentas de *Data Mining* usam quer a aprendizagem supervisionada quer a não supervisionada (Peña-Reyes & Sipper, 2000).

As técnicas de *Data Mining* que usam a aprendizagem supervisionada são treinadas para reconhecer classes diferentes de dados expondo-os a exemplos para os quais

tem respostas alvo (amostra de dados de treino), e depois testa-os numa nova amostra de dados que classifica (amostra de dados de teste) (Fayyad et al., 1996a).

A aprendizagem não supervisionada não requer informação inicial acerca da classificação correcta dos dados (Fayyad et al., 1996a).

Nas subsecções seguintes faz-se uma apresentação das técnicas de *Data Mining* empregues nesta dissertação, nomeadamente as redes neuronais artificiais, as árvores de decisão e a indução de regras de decisão.

2.3.3.1 Redes Neuronais Artificiais

As redes neuronais artificiais são técnicas de *Data Mining* eficazes na resolução das tarefas de reconhecimento de padrões, previsão de *outcomes*, classificação e segmentação de dados (Bath & Philp, 1998, Haykin, 1999).

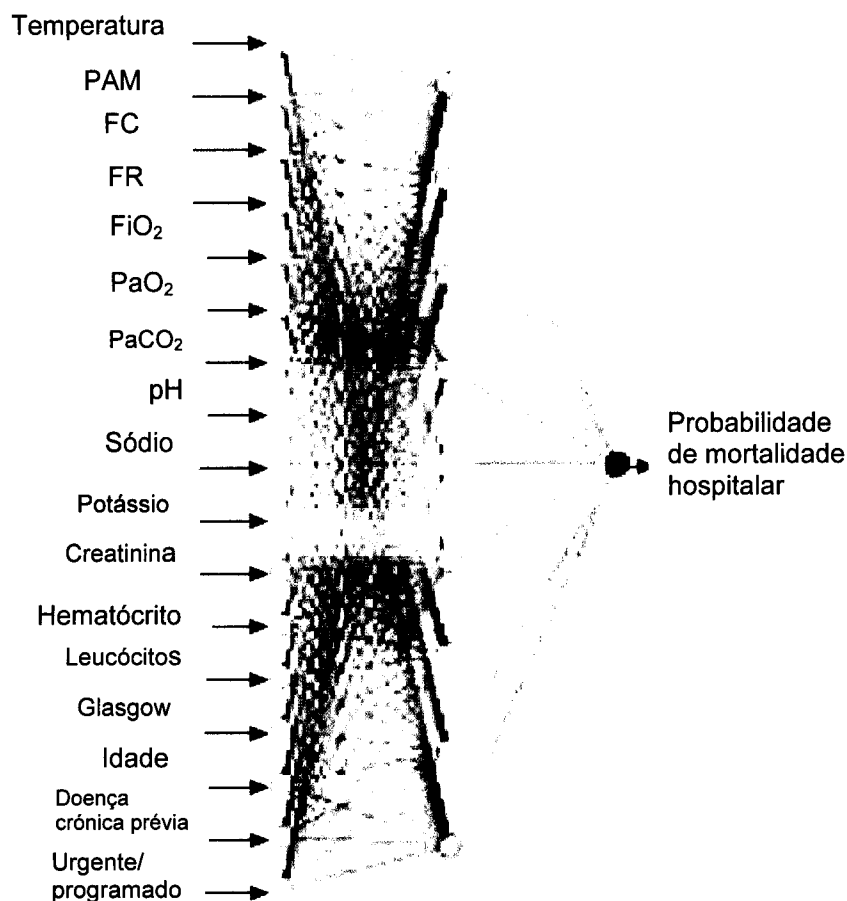
A primeira utilização prática das redes neuronais artificiais foi descrita por McCulloch e Pitts em 1946 (McCulloch & Pitts, 1946).

As redes neuronais artificiais são modelos computacionais não lineares para o reconhecimento de padrões, cujo conhecimento é armazenado nas interligações e adquirido a partir de um ambiente de dados, através de um processo de aprendizagem (treino) (Guerriere & Detsky 1991, Hanson & Marshall, 2001, Patterson, 1996, Penny & Frost, 1996, Wei et al., 1998).

Uma rede neuronal artificial é um algoritmo de cálculo que tem como metáfora o sistema nervoso (Hinton, 1992). A ideia geral consiste em simular a arquitectura e a capacidade de aprendizagem do sistema nervoso, de forma a que a rede neuronal artificial aprenda a identificar um padrão de associação entre os valores de um conjunto de variáveis preditivas (entradas) e as variáveis que se consideram dependentes dos ditos valores (saídas) (Trujillano et al., 2004).

A estrutura típica de uma rede neuronal artificial consiste em três camadas: nodos de entrada, uma ou mais camadas internas ou ocultas e os nodos de saída (Figura 3).

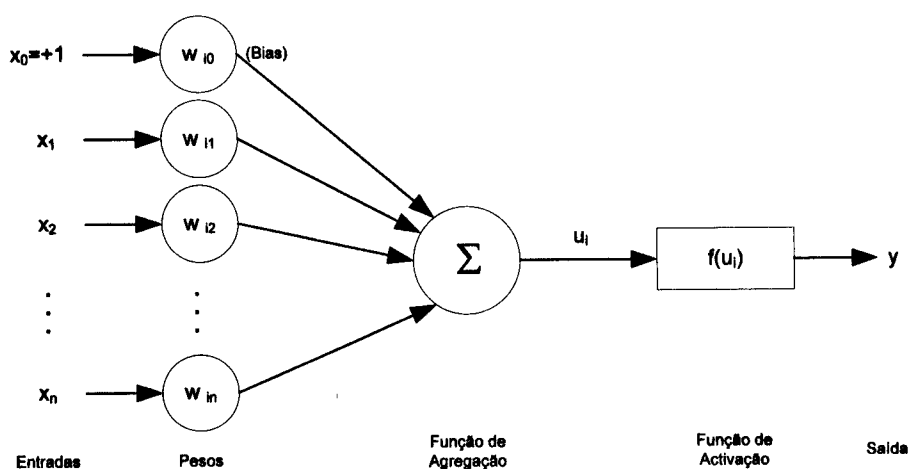
Figura 3 - Rede neuronal artificial. A camada de entrada recebe os valores das variáveis independentes, as camadas internas realizam os processos de ajuste de pesos, e a camada de saída disponibiliza os resultados da rede (saída) Adaptado de (Trujillano et al., 2003)



A rede neuronal artificial consiste num grupo de unidades de processamento computacionais (nodos) que se assemelham aos neurónios, por estarem interconectados através de uma rede de interligações (pesos) análogas ao conceito das sinapses do sistema nervoso(Trujillano et al., 2004).

A partir dos nodos de entrada, o sinal progride através da rede até proporcionar uma resposta sob a forma do nível de activação dos nodos de saída (Cross et al., 1995). Os valores de saída proporcionam uma previsão do resultado em função das variáveis de entrada (Patterson, 1996).

Esta situação modela-se matematicamente como o somatório dos pesos de todos os sinais de chegada ao nodo que se compara com um limiar característico. Se esse limiar é superado, então o nodo “dispara”, mandando um sinal aos outros nodos, que por sua vez processam essa informação juntamente com a que recebem dos nodos adjacentes (Figura 3).

Figura 4 – Rede neuronal artificial. Adaptado de (McCulloch & Pitts, 1946)

Como ilustrado na Figura 4, o estado de activação dos nodos é definido através da função de activação e representa a situação destes nodos dentro da rede neuronal artificial.

A função de activação é calculada a partir de um somatório envolvendo os estímulos de entrada de um nodo e a intensidade das interligações associadas a cada uma destas entradas, gerando um estado de activação. Os estímulos de entrada de um nodo são representados através de um vector de entrada $X = [x_1; x_2; \dots; x_n]$, e a intensidade através de um vector de pesos $W = [w_1; w_2; \dots; w_n]$. Os estímulos de entrada do neurónio são ponderados pelo vector de pesos através da equação:

$$u_i = \sum_{i=1}^n W_i X_i - \theta \quad \theta = \text{bias} \quad \text{ou} \quad u_i = \sum_{i=0}^n W_i X_i$$

A saída é representada por:

$$y = f(u_i)$$

Existe uma grande variedade de funções de activação (limiar, linear, logística) (McCulloch & Pitts, 1946). A função de saída de um neurónio $f(ui)$ é computada sobre o valor da função de activação e geralmente é uma função identidade. Cada conexão de um neurónio, ui , possui um peso associado e pode ser classificada em três tipos distintos: excitatória ($w_i > 0$); inibitória ($w_i < 0$) e inexistente ($w_i = 0$) (McCulloch & Pitts, 1946).

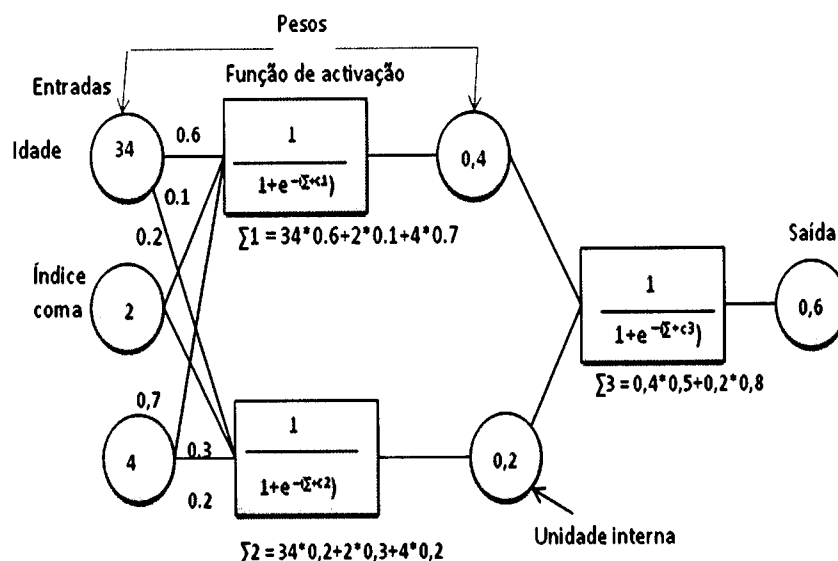
Os nodos estão interligados por parâmetros internos (pesos) cujos valores se modificam de acordo com os valores que os nodos recebem e segundo a função de rede que se lhes aplica. A complexidade de uma rede neuronal artificial depende da quantidade de parâmetros internos que contenha (Trujillano Cabello et al., 2005). Evidentemente, a resposta de cada nodo dependerá do valor das interações com os nodos precedentes dentro da estrutura da rede. Como no caso do sistema nervoso, o poder computacional de uma rede neuronal artificial não deriva da complexidade de cada unidade de processamento mas da densidade e da complexidade das suas interligações (Sarle, 1997).

Uma rede neuronal artificial toma como entrada um vector de números, por exemplo: as características dos doentes e, como saída, um número ou vector numérico. (ex.: probabilidade de morte intra-hospitalar) (Lucas & Abu-Hanna, 1999).

A modelação matemática de uma rede neuronal artificial depende dos seus nodos, da sua arquitectura e do seu método de aprendizagem (treino) (Trujillano Cabello et al., 2005). Os nodos estão interligados através de diferentes arquitecturas, mas normalmente nas arquitecturas unidireccionais, os nodos de entrada estão interligados aos nodos das camadas internas, e estes estão interligados aos nodos de saída. À medida que a rede neuronal aprende (ou é treinada), a partir do conjunto de dados, os pesos das interligações são ajustados. As interligações mais importantes vêem o seu peso aumentado, sucedendo o inverso para as interligações menos importantes. Isto é, os dados são fornecidos aos nodos de entrada, processados pelas camadas internas e os pesos das interligações aos nodos de saída são ajustados (ver Figura 5) (Hanson & Marshall, 2001). O estado de activação dos nodos é definido através da função de activação. A função de activação é calculada a partir de um somatório envolvendo os estímulos de entrada de um nodo e a intensidade das interligações associadas a cada uma destas entradas, gerando um estado de activação (McCulloch and Pitts 1946).

Cada nodo processa a informação que recebe de acordo com a sua função de activação. A função de activação mais comum é a logística (Haykin, 1999).

Figura 5 - Rede neuronal artificial para a previsão da probabilidade de ocorrência de um *outcome* categórico (ex.: probabilidade de mortalidade) Adaptado de (Ohno-Machado et al., 2006)



As unidades internas das redes neuronais artificiais operam entre as unidades de entrada e de saída para processamento da informação. Consideremos que a rede neuronal artificial tem a função logística como a função de activação. Seja B a matriz de pesos entre a camada de entrada e a interna; seja c um vector de constantes, uma por cada unidade interna e seja w o vector de pesos entre as unidades interna e a unidade de saída; se f for a função de activação da unidade interna, podemos representar o funcionamento da rede neuronal artificial para a entrada x como:

$$y = \frac{1}{1 + e^{-[wf(Bx+c)+k]}}$$

onde k é o viés da unidade de saída.

Para classificar as redes neuronais artificiais podemos considerar dois critérios básicos: o modo de aprendizagem e o fluxo de informação.

O fluxo de informação que manuseiam pode ser unidireccional, quando a informação segue uma direcção única desde os nodos de entrada aos de saída; ou realimentado, quando o fluxo de informação não é único e incorpora circuitos de realimentação entre as camadas da rede neuronal artificial (Trujillano et al., 2004).

Entre as redes neuronais artificiais mais utilizadas no *Data Mining*, estão as redes *Perceptron Multi-camada* (Haykin, 1999) e as redes de Kohonen (Kohonen, 1989), que são descritas nas subsecções seguintes.

2.3.3.1.1 *Perceptron* multi-camadas

As unidades do tipo *Perceptron* foram criadas por Frank Rosenblatt em 1959, (Rosenblatt, 1959) sendo a base de diversos tipos de redes neuronais artificiais com aprendizagem supervisionada e utilizando uma adaptação por correcção de erros.

Dentro das redes supervisionadas unidireccionais, a estrutura mais utilizada é o *Perceptron* de múltiplas camadas, sendo na maior parte das vezes associado ao algoritmo de aprendizagem de retropropagação (*BackPropagation*) (Jodouin, 1994, Widrow & Lehr, 1990) (ver Figura 6).

A arquitectura típica deste tipo de rede neuronal artificial é constituída por várias camadas de nodos com interligação completa entre elas. A topologia mais simples deste tipo de redes neuronais consiste em apenas duas camadas de nodos, a de entrada e a de saída, permitindo obter um modelo adequado para a solução de problemas lineares do tipo da regressão linear múltipla (Trujillano et al., 2004).

Se queremos analisar problemas não lineares, é necessário incorporar mais camadas internas. Neste tipo de rede neuronal artificial, um nodo recebe várias entradas e activa uma função de rede (ou regra de propagação) de acordo com os pesos de entrada associados. A entrada dos nodos da primeira camada (entrada) são os valores das variáveis preditivas e os níveis de activação dos nodos da última camada (saída) são os resultados da rede.

Figura 6- Rede neuronal artificial para a previsão da probabilidade de ocorrência de um *outcome* categórico. Adaptado de (Ohno-Machado et al., 2006).

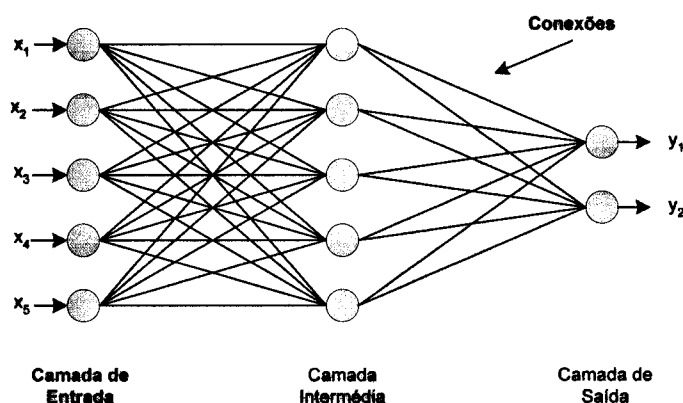


Figura 7 – unidireccional.

Adaptado de (Rumelhart et al., 1986)

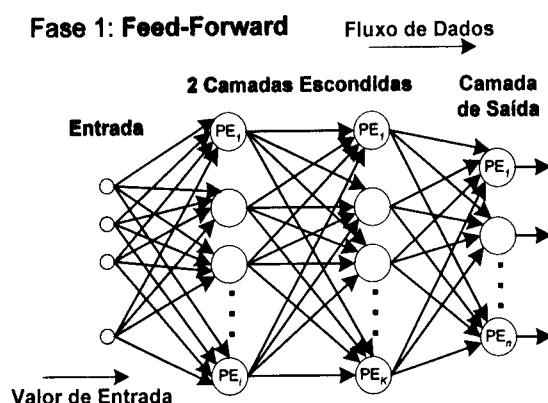
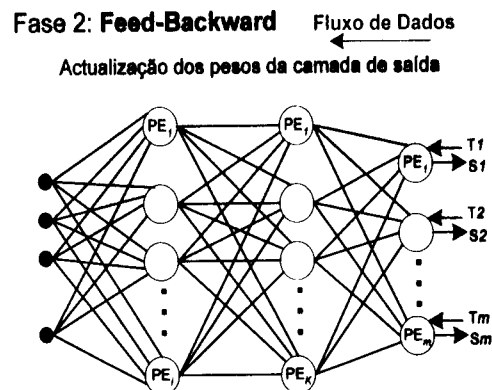


Figura 8 – retropropagação.



O procedimento de aprendizagem (Figuras 7 e 8) é repetido, até que para todos os processadores da camada de saída e para todos os padrões de treino, o erro seja menor do que o especificado.

A grande vantagem das redes *Perceptron* multi-camadas treinadas por retropropagação é a sua capacidade de identificar os diferentes tipos de padrões existentes nos dados (Rumelhart et al., 1986).

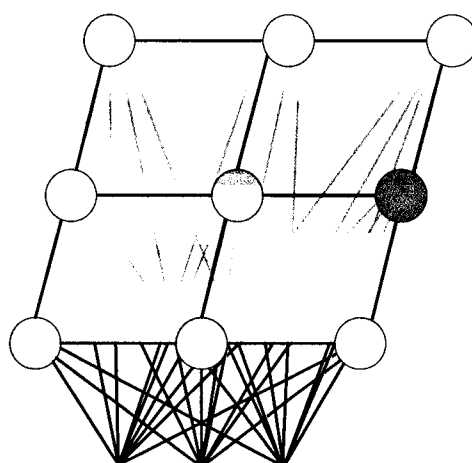
Nalguns casos, o processo de treino pode levar à perda da capacidade de generalização da rede, isto é, à sobre-especialização (a rede aprende com os exemplos mas é incapaz de dar uma resposta adequada em casos novos). Para evitar este problema, no processo de treino devem ser consideradas estratégias adicionais.

No contexto da Medicina Intensiva, o treino consistiria em apresentar à rede neuronal artificial, de forma iterativa, os valores de diferentes variáveis clínicas (na forma de valores da camada de entrada) de cada doente e conseguir que a rede fosse capaz de prever o *outcome* observado em cada doente ou população (indicado pelo estado das camadas de saída da rede) da forma o mais precisa possível.

2.3.3.1.2 Redes de Kohonen

As redes neurais artificiais de Kohonen (Kohonen, 1982, Kohonen, 1989)), também conhecidas como mapas de Kohonen, apresentam uma arquitectura relativamente simples composta por duas camadas de nodos (uma camada de entrada e uma de saída) interligados, como ilustrado na Figura 9.

Figura 9 - Arquitectura básica de uma rede de Kohonen. Adaptado de (Kohonen, 1982).



Entrada = (x1, x2,...,xn)

Na rede de Kohonen, a aprendizagem é não supervisionada (Kohonen, 1995). Os valores fornecidos à camada de entrada podem ser discretos ou contínuos, mas os valores de saída são usualmente binários (Kohonen, 1982).

A principal característica das redes de Kohonen é a de conseguirem preservar a semelhança entre os vectores de entrada nos dados de saída (Kohonen, 1995).

Como normalmente existem mais dimensões nos dados de entrada do que nos de saída, estas redes conseguem fazer a redução de dimensões salvaguardando a proximidade dos vectores. Os vectores de entrada semelhantes serão representados por nodos próximos na camada de saída (Kohonen, 1982).

Na rede de Kohonen, o sinal de entrada, representado pelo vector $x = \{x_1; x_2; \dots; x_n\}$, está composto por n valores escalares. Ao mesmo tempo, cada unidade da rede tem um vector de pesos com o mesmo número de componentes que o sinal de entrada. Isto é, $w = \{w_1; w_2; \dots; w_n\}$, onde w_i representa o peso associado à i -ésima componente do sinal de entrada x_i (Kohonen, 1982).

Este tipo de rede permite detectar similaridades entre vários sinais, agrupando-os em segmentos ou *clusters* (Lipmann, 1987).

Não há propriamente um reconhecimento do padrão como nas outras redes neurais artificiais, mas a classificação de um padrão juntamente com outros que têm características congruentes, formando segmentos (Hagan et al., 1996).

A saída de uma rede Kohonen é um conjunto de duas coordenadas (X, Y) que pode ser usada para visualizar grupos de registos e, conseqüentemente, combinar os mesmos para criar ou associar um número de segmentos. É esperado que os grupos

de segmentos sejam diferentes entre si, mas que contenham registos que sejam similares em algum aspecto.

As redes de Kohonen formam uma família de redes, com grande plausibilidade biológica que se auto-organizam através de mecanismos de competição (Kohonen, 1982).

No fim da aprendizagem, cada unidade responde selectivamente a um segmento, isto é, para uma determinada configuração de entrada, uma unidade encontra-se mais activa que as outras. Quando uma unidade vence a competição, são modificados os pesos das ligações que existem nesta unidade e nas unidades vizinhas (Chester, 1993). Um problema importante das redes de Kohonen é a necessidade de se definir a priori a arquitectura da rede, isto é, o número de segmentos que ela deverá ter (Hagan et al., 1996).

2.3.3.1.3 Tamanho e Arquitectura de uma Rede Neuronal Artificial

A arquitectura de uma rede neuronal artificial é definida com base no número de camadas e de nodos e na forma como estes estão organizados (Haykin 1999).

A complexidade da rede neuronal artificial é determinada pelo número de interligações que contém. Não é fácil estabelecer qual será a arquitectura ideal para cada aplicação (Hagan et al., 1996).

O aumento do número de camadas e nodos internos permite solucionar problemas mais complexos (Trujillano et al., 2004). Mas, a arquitectura óptima deve ser descoberta na prática, através de um processo iterativo, em que se vai validando a capacidade preditiva das diferentes arquitecturas consideradas, no conjunto de validação (Schwarzer et al., 2000). Por outro lado, quanto mais complexa é a rede neuronal artificial maior o número de pesos a estimar, sendo portanto necessário um maior volume de dados e tempo de treino, de forma a treinar adequadamente a rede neuronal artificial.

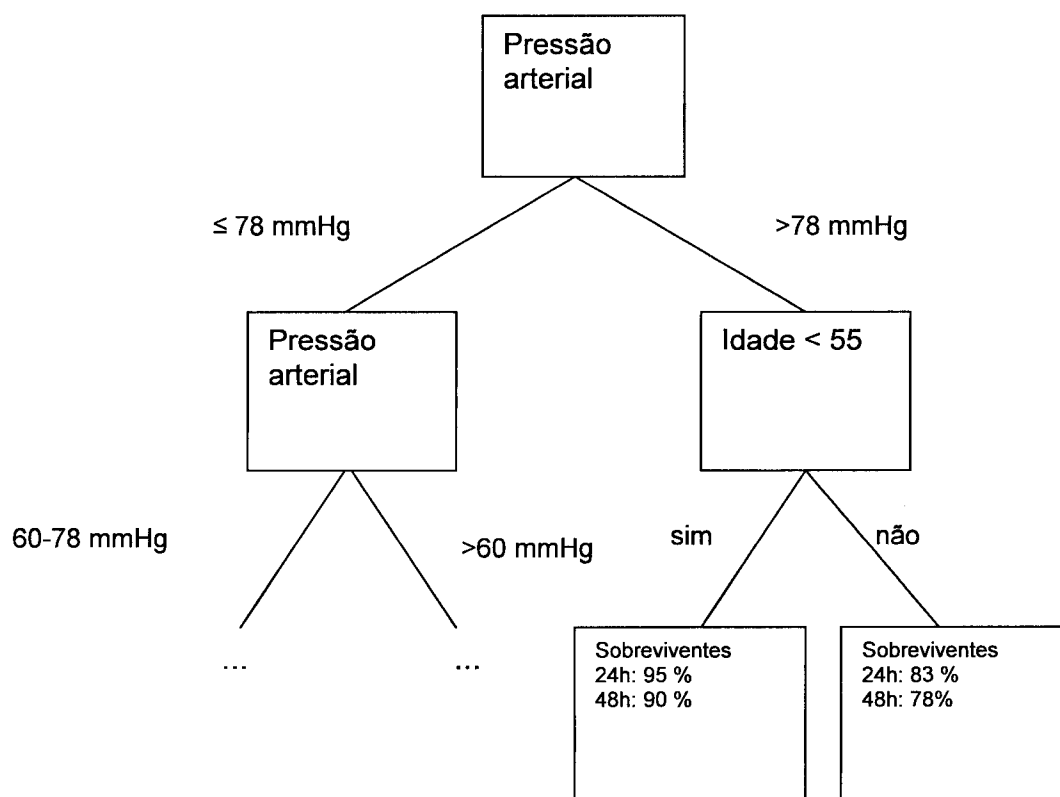
A utilização de redes neuronais artificiais demasiado complexas conduz normalmente à sobre-especialização (Hagan et al., 1996).

Como regra geral, para reduzir o número de parâmetros de uma rede neuronal artificial, é conveniente seleccionar apropriadamente as variáveis de entrada, excluindo variáveis pouco informativas e recorrendo a metodologias específicas (ex.: manipulação dos pesos, poda da rede)

2.3.3.2 Árvores de Decisão

Uma árvore de decisão é uma forma de representação de um conjunto de regras de que resulta numa hierarquia de classes ou valores, resultantes da partição recorrente dos dados e que graficamente são semelhantes a uma árvore (Figura 10)(Lucas, 2004).

Figura 10 - Árvore de Decisão para a previsão da sobrevida com base na pressão arterial e na idade.



Uma árvore de decisão é uma representação de um conjunto de regras de classificação, que classificam os registos, desde o nó da raiz até aos nós terminais (folhas). Cada nó da árvore especifica um teste para as variáveis, e cada ramo descendente desse nó corresponde a um dos valores possíveis para essa variável. Um registo é classificado começando por testar a variável especificada pelo nó da raiz, e depois seguindo o ramo correspondente ao valor da variável no registo (Patterson, 1990).

Este processo é então repetido para a sub-árvore com raiz no novo nó. O objectivo é gerar a árvore que melhor se adequa ao problema, ou seja que melhor classifique os registos do domínio considerado (Abu-Hanna & de Keizer, 2003).

A partição dos dados é feita através de testes contidos em cada um dos nós, sobre o valor de variáveis independentes (contínuas ou discretas) e escolhidos no sentido de minimizar uma determinada estimativa do erro (Quinlan, 1997).

As folhas são os nós terminais e estão associadas a uma classe que contém as previsões feitas pelo modelo. O conjunto de caminhos da raiz da árvore a cada uma das suas folhas constitui o conjunto de regras pelas quais a árvore faz as suas previsões (Abu-Hanna & de Keizer, 2003).

As árvores de decisão podem gerar regras para a classificação do conjunto de dados, isto é disponibilizam um conjunto de regras que podem ser aplicadas a um grupo de dados novo (não classificado) fazendo a previsão do *outcome* de cada registo (Lee & Siau, 2001) e podem ser usadas com objectivos meramente descritivos ou para a elaboração de previsões (Ennis et al., 1998).

Os modelos baseados em árvores de decisão constituem um conjunto de regras de fácil interpretação e compreensão para o utilizador, onde são destacadas as variáveis para a previsão mais relevantes no domínio, e a sua representação pode ser facilmente transformada em regras de decisão do tipo: "se...então" (Abu-Hanna & de Keizer, 2003). Na aprendizagem através de árvores de decisão, as variáveis são usadas para dividir os dados em grupos homogéneos, em relação a uma classe (ex.: sobreviventes, não-sobreviventes). As árvores de decisão tem uma estrutura em que cada nodo representa uma variável, cada ramo representa um valor ou um intervalo de valores para a variável correspondente e cada folha representa uma classe (dominante) (Abu-Hanna & de Keizer, 2003). Os algoritmos de indução da árvore de decisão constroem a árvore a partir da amostra de dados de treino, de uma forma recorrente, subdividindo este conjunto de dados até que seja formada apenas por "nós" puros, em que cada nó represente apenas uma única classe, ou a satisfação de um único critério (Quinlan, 1997).

Os objectivos na aprendizagem ou treino por árvores de decisão são o de encontrar atributos ou variáveis no conjunto de dados que contribuem significativamente para a identificação de grupos homogéneos no conjunto de dados (Lucas & Abu-Hanna, 1999).

Existem diversos algoritmos capazes de induzirem árvores de decisão nomeadamente o C5 (Quinlan, 1987), o CHAID (Chi-Squared Automatic Interaction Detection) (Kass, 1980) e o CART (Classification And Regression Tree) (Breiman et al., 1984).

Os algoritmos de indução de árvores de decisão trabalham com os dados sob a forma de folha de cálculo, sem existir a necessidade de normalização dos valores. As árvores de decisão são rápidas quando comparadas com outros métodos como as redes neuronais artificiais, que necessitam de muito tempo de aprendizagem. Uma das suas maiores vantagens é o facto de serem facilmente interpretadas. Outras das vantagens consiste na facilidade que têm no tratamento de dados não numéricos.

Os algoritmos de indução da árvore de decisão funcionam partindo do conjunto inicial de dados e usam os valores de uma das variáveis para efectuar a partição dos mesmos em vários subconjuntos. O processo de partição é iterativamente aplicado a cada um dos subgrupos até que cada um dos subgrupos obtidos seja constituído por registos de uma só classe (Lee and Siau 2001).

Os passos básicos para a construção de uma árvore de decisão são os seguintes: (Quinlan, 1993)

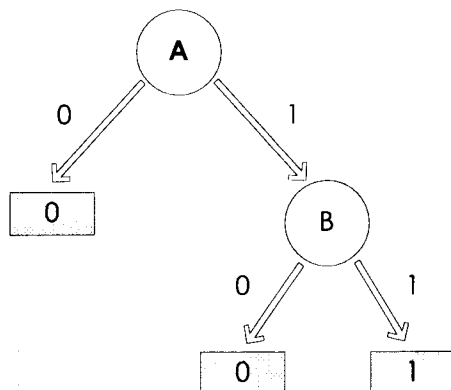
1. Escolher um atributo;
2. Estender a árvore adicionando um ramo para cada valor do atributo;
3. Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);
4. Para cada folha:
 - 4.1. Se todos os exemplos são da mesma classe, associar essa classe à folha;
 - 4.2. Senão, repetir os passos 1 a 4.

Como exemplo, apresenta-se uma árvore de decisão quando se utilizam variáveis binárias, neste caso na função E (quadro 1 e figura 11).

Quadro 1 - Função E

A	B	A E B
0	0	0
0	1	0
1	0	0
1	1	1

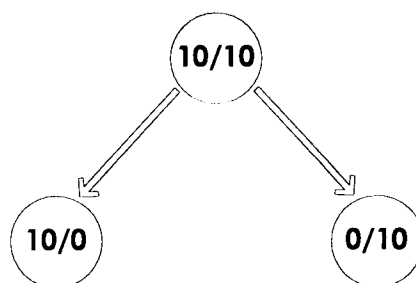
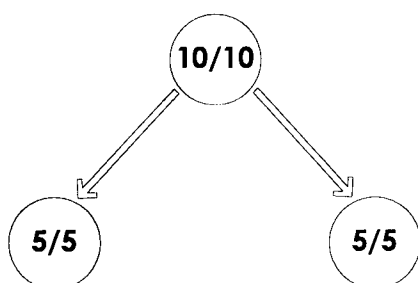
Figura 11 – Árvore de decisão para a função E. Sendo A a variável representada na raiz da AD, e se o seu valor for 0, a folha (resultado do modelo) será de 0, contudo se o valor de A for 1, o resultado do modelo depende do valor da variável B.



Uma das questões importantes que se coloca aquando da utilização de árvores de decisão é a de que forma se pode medir a habilidade de uma dada variável para discriminar as classes (critério de escolha da variável). Existem várias medidas para esse efeito, mas todas concordam em dois aspectos: uma divisão que preserve as proporções de classes em todas as partições é inútil e uma divisão onde em cada partição todos os exemplos são da mesma classe tem utilidade máxima (Lee & Siau, 2001).

Esta situação é demonstrada pelas árvores apresentadas nas Figuras 12 e 13:

Figura 12 – Árvore com Proporções iguais. **Figura 13** – Árvore com exemplos da mesma classe.



2.3.3.2.1 Sobre-especialização

Um modelo é útil se for possível a sua utilização num conjunto de dados novos (diferentes dos utilizados na construção do modelo). Pelo que deve ser evitada a

excessiva otimização do modelo, ajustando a condição de paragem do treino (Quinlan 1988).

Considera-se que uma árvore de decisão d está sobre-treinada se existir uma árvore de decisão d' em que: d tem menor erro que d' no conjunto de treino, mas d' tem menor erro na população.

O algoritmo de partição recursiva do conjunto de dados gera estruturas que podem obter um ajuste aos exemplos de treino perfeito (Quinlan, 1988)

As técnicas de poda (*pruning*) são utilizadas para evitar a sobre-especialização (*overfitting*) da árvore de decisão (Quinlan, 1988).

Um algoritmo básico de poda consiste em percorrer a árvore em profundidade e para cada nó de decisão calcular o erro no nó e a soma dos erros nos nós descendentes. Se o erro do nó é menor ou igual à soma dos erros dos nós descendentes, o nó é transformado em folha. Cada nó que compõe a árvore é candidato a poda. O nó é removido se a árvore resultante se comporta igual ou melhor que a árvore original (Quinlan, 1987).

O balanceamento²⁰ dos *outcomes* numa população de estudo constitui um factor determinante na qualidade dos resultados obtidos por árvores de decisão, que pode se ultrapassado com o recurso a técnicas específicas.

Apresentam-se de seguida algumas das vantagens apontadas às árvores de decisão (Lee & Siau, 2001).

Trata-se de um método não-paramétrico, uma vez que não assume nenhuma distribuição particular para os dados e pode construir modelos para qualquer função desde que o número de exemplos de treino seja suficiente. A estrutura da árvore de decisão é independente da escala das variáveis, de forma que as transformações monótonas das variáveis ($\log x$, $2*x$, ...) não alteram a estrutura da árvore. Possuem um elevado grau de interpretabilidade, em que uma decisão complexa (prever o valor da classe) é decomposta numa sucessão de decisões elementares. Acresce que as árvores de decisão são eficientes na construção de modelos e são robustas perante presença de pontos extremos e variáveis redundantes ou irrelevantes, uma vez que possuem um mecanismo de selecção de variáveis.

As árvores de decisão não impõem um viés na expressividade do modelo. Isto contrasta com alguns métodos estatísticos que restringem a forma do modelo (ex.: pré-assumindo um modelo linear), os pressupostos na distribuição de probabilidade ou a (in) dependência entre variáveis (Lucas & Abu-Hanna, 1999).

²⁰ Termo habitualmente aplicado a qualquer estudo experimental no qual o mesmo número de observações é tido para cada combinação dos factores experimentais. Everitt

As árvores de decisão têm sido utilizadas para modelação prognóstica no âmbito da Medicina (Andreassen et al., 1987) (Aitchison et al., 1995, Dusseldorp & J., 2001, Kuo et al., 2001).

2.3.3.3 Indução de Regras de Decisão

A indução de regras de decisão (Quinlan, 1988) é uma das técnicas de mais comuns em *Data Mining*. A indução pode ser feita a partir das árvores de decisão, ou directamente a partir dos dados de acordo com o algoritmo aplicado (Berson et al., 2000).

Esta técnica consiste na descoberta de regras de previsão do tipo “se – então”, a partir de uma amostra de treino, que são usadas para generalização, e então aplicadas para classificação na amostra de teste (Lavraç, 1999).

Uma regra é uma estrutura sob a forma “Se condições *Então* conclusão” em que condições são uma conjunção de variáveis e conclusão é uma classe. Cada termo da conjunção assume a forma “variável símbolo valor”, em que símbolo será: =, >, <, ≥ e ≤ (Quinlan, 1988).

A partir de uma árvore de decisão é possível, de forma implícita, extrair um conjunto de regras representativas do modelo. Geralmente, utilizam-se as regras de decisão para representar o conhecimento representado nas árvores de decisão. As regras geradas são simples e de fácil interpretação.

Contudo, esta técnica tem limitações no manuseamento de informação incompleta e de valores inesperados (Quinlan, 1988).

A abordagem da indução de regras de decisão tem sido utilizada no âmbito da medicina nomeadamente no suporte ao diagnóstico de doenças reumatológicas (Lavraç 1999), no prognóstico após os resultados de exames complementares no âmbito da cardiologia (Lavraç 1999), na análise de dados em doentes com meningite (Zhong & Dong, 2002) e na previsão da mortalidade em relação à primeira consulta hospitalar (Richards et al., 2001).

2.4 Avaliação dos modelos

Existem vários métodos para estimar a capacidade de generalização de um modelo, como são o caso da validação cruzada, da validação com divisão da amostra, da

matriz de confusão (Kohavi & Provost, 1998) e da capacidade discriminativa [*receiver operating characteristic curve* (ROC)] (Hanley & Mc Neil, 1982).

2.4.1 Validação cruzada

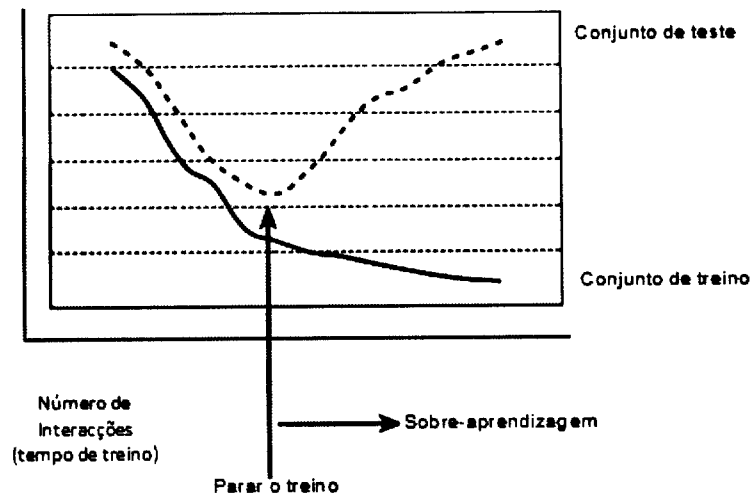
A flexibilidade das redes neurais artificiais tem custos, nomeadamente o risco de sobre-especialização em que a rede neuronal se adapta muito bem à amostra em estudo, mas perde capacidade de generalização para responder a casos novos (Ohno-Machado et al., 2006).

Para elaborar um modelo neuronal artificial eficaz é conveniente dividir os dados em três conjuntos, atendendo que cada um deles mantenha a representatividade da população de origem: conjunto de treino, conjunto de verificação (teste) e um conjunto de validação (Chester, 1993).

O conjunto ou amostra de treino é usado para o ajuste dos pesos durante a fase de treino (aprendizagem), enquanto que o conjunto de verificação é utilizado para a decisão da paragem do processo de treino. Como critério geral, o treino deve ser parado quando o erro do conjunto de verificação é mínimo (Duh et al., 1998).

Desta forma, asseguramos que a rede neuronal artificial é capaz de prever correctamente os *outcomes* de um conjunto de dados, que não fazem parte dos exemplos de treino. Esta técnica denomina-se de validação cruzada. Se continuarmos o treino para além deste ponto, a rede neuronal artificial começa memorizar os dados do conjunto de treino, mas por outro lado, a capacidade de generalização (Figura 14) (Duh et al., 1998).

Figura 14 - Critério de paragem do processo de treino. RMS: raiz quadrada do erro quadrado médio. Adaptado de (Trujillano et al., 2004).



A generalização óptima reflecte a capacidade da rede neuronal artificial em proporcionar uma resposta correcta perante padrões que não foram utilizados no seu treino. Uma vez finalizado o treino, a rede neuronal artificial treinada avalia o conjunto de validação e gera as previsões correspondentes dos dados, com dados não utilizados nem no treino nem na validação cruzada (Duh et al., 1998).

Esta prova fornece-nos uma estimativa independente da capacidade de generalização da rede neuronal artificial diminui. (Trujillano et al., 2004).

2.4.2 Validação com divisão de amostra

Na construção de um modelo definem-se as principais características do sistema, que devem representar o mais fielmente possível a realidade. Recolhem-se os dados necessários para a sua construção e para a consequente validação, sendo necessária uma divisão do conjunto de dados em dois subconjuntos, um para a geração do modelo, chamado de conjunto de treino, e outro para a validação do modelo, chamado de conjunto de teste.

A validação com divisão da amostra (Flexer, 1996) é o método mais comum para a estimação do erro de generalização, que se baseia numa divisão dos dados em dois conjuntos ou amostras: de treino e de teste (validação), normalmente em 2/3 e 1/3 respectivamente.

Enquanto o conjunto de treino se destina à aprendizagem do modelo, o conjunto de teste permite estimar o erro de validação. Este último é um conjunto de exemplos independentes que não tiveram nenhum papel na construção do classificador. De realçar que se deve ter em conta o facto de que nos conjuntos de treino e teste deverem ser consideradas amostras representativas do problema em questão.

Após a geração dos modelos, é necessário avaliar o desempenho dos mesmos no conjunto de teste ou validação. Uma medida de desempenho para problemas de classificação é a taxa de erro, cujo sucesso significa que a classe da instância é prevista correctamente. Erro significa que a classe da instância é prevista incorrectamente e a taxa de erro é a proporção dos erros em relação ao conjunto de exemplos.

2.4.3 Matriz de Confusão

Os diferentes tipos de erros e acertos realizados podem ser sintetizados numa matriz de confusão. Esta matriz é uma das técnicas de avaliação mais utilizadas em problemas de classificação. No Quadro 1 é apresentada uma matriz de confusão para um problema que possui duas classes rotuladas como classe positiva e classe negativa.

Quadro 1 - Matriz de confusão

	Previsão Positiva	Previsão Negativa
Classe Positiva	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (TN)

A partir desta matriz de confusão é possível extrair diversas medidas de desempenho para sistemas de aprendizagem, tal como a taxa de erro (Err) e a acuidade preditiva ou precisão (Pa).

$$Err = \frac{FP + FN}{TP + FN + FP + TN}$$

$$Pa = \frac{TP + TN}{TP + FN + FP + TN}$$

A taxa de erro e a precisão são duas medidas utilizadas para medir o desempenho de sistemas de aprendizagem.

Em domínios nos quais o custo de classificação incorrecta é relevante, uma matriz de custo pode ser utilizada. Uma matriz de custo define os custos de classificação incorrecta, isto é, uma penalidade para cada tipo de erro que o classificador pode cometer. Nesse caso, o objectivo do classificador deve ser minimizar o custo total de classificação incorrecta.

Para conjuntos de dados com classes não balanceadas, uma medida de desempenho mais apropriada deve associar os erros, ou acertos, ocorridos para cada classe. A partir do Quadro 1, é possível derivar quatro medidas de desempenho que medem o desempenho de classificação nas classes negativa e positiva independentemente, nomeadamente as seguintes quatro taxas:

Taxa de Falso Negativo

$$FNR = \frac{FN}{TP + FN}$$

É a percentagem de casos positivos classificados incorrectamente como pertencentes à classe negativa.

Taxa de Falso Positivo

$$FPR = \frac{FP}{FP + TN}$$

É a percentagem de casos negativos classificados incorrectamente como pertencentes à classe positiva.

Especificidade (Taxa de Verdadeiro Negativo)

$$TNR = \frac{TN}{FP + TN} \times 100\%$$

É a percentagem de casos negativos classificados correctamente como pertencentes à classe negativa.

Sensibilidade (Taxa de Verdadeiro Positivo)

$$TPR = \frac{TP}{TP + FN} \times 100\%$$

É a percentagem de casos positivos classificados correctamente como pertencentes à classe positiva.

Essas quatro taxas de desempenho possuem a vantagem de serem independentes do custo e das probabilidades das classes. O principal objectivo de qualquer classificador é minimizar as taxas de falso positivo e de falso negativo, ou, de forma similar, maximizar as taxas de verdadeiro positivo e verdadeiro negativo.

2.4.4 Capacidade discriminativa

Outra medida de avaliação de classificadores pode ser obtida pela área sob a *receiver operating characteristic curve* (ROC] que traduz a sua capacidade discriminativa (descrita na secção 2.3 do capítulo II).

Esta curva permite a avaliação do desempenho de um classificador, sendo apropriada quando existem duas classes. A curva ROC estabelece a relação entre a TNR (Especificidade) e a TPR (Sensibilidade). Na situação ideal, o algoritmo deveria possuir indicadores máximos de sensibilidade e especificidade, ambos iguais a 1.

3. Aplicação das redes neuronais artificiais em Medicina Intensiva

Nos últimos quinze anos a aplicação de redes neuronais artificiais na área da Medicina tem-se traduzido pelo número crescente de artigos publicados.

As redes neuronais artificiais consistem em modelos altamente flexíveis e têm sido usadas no âmbito da Medicina Intensiva para explorar relações entre várias variáveis fisiológicas e para a construção de modelos de previsão (Buchman et al., 1994, Dybowski et al., 1996, Fraser & Turney, 1990, Frize et al., 2001, Kayaalp et al., 2000, Tu & Guerriere, 1993).

As redes neuronais artificiais têm sido utilizadas no ambiente da UCI com vários propósitos, mas mais extensivamente para a previsão de *outcomes* (Hanson &

Marshall, 2001). A vantagem das redes neurais artificiais pode estar na possibilidade de encontrar associações, ou melhorar a capacidade de descoberta de variáveis independentes para outros *outcomes* clínicos importantes (ex.: disfunção de órgãos) e na monitorização da evolução da gravidade da doença na UCI (Penny & Frost, 1996).

Para além disso, alguns estudos revelaram que as redes neurais artificiais podem ter um papel importante na estimativa da sobrevida do doente individualizado (Ohno-Machado, 1996, Ohno-Machado & Musen, 1997). No Quadro 2, apresentam-se algumas das aplicações dos modelos neurais artificiais. Segundo os resultados publicados, esses modelos obtidos por redes neurais são 5 a 10% mais precisos do que os obtidos por técnicas estatísticas, uma vez que não trabalham com as restrições dos modelos lineares e são capazes de incorporar automaticamente as relações entre as variáveis preditivas (Levine, 2001).

Quadro 2 -Aplicações de modelos neuronais artificiais em Medicina Intensiva

Adaptado de (Trujillano Cabello et al., 2005).

	Referência	Ano	Valida.	Estat.	Result.	N.º
Scores						
APACHE II	Wong e Young	1999	Sim	CS	=	8796
APACHE III	Clermont, Angus et al	2001	Sim	CS	=	1647
TRISS-IS	DiRusso, Sullivan et al	2000	Sim	CS	+	10609
ICISS-TRISS-ISS	Rutledge, Osler et al	1998	Sim	CS	+	7276
Prognóstico						
Sepsis	Warner, Bencosme et al	1996	Sim	Não	+	68
SIRS-Choque	Dybowski, Weller et al	1996	Sim	Sim	++	258
TCE	Lang, Pitts et al	1997	Sim	Sim	=	1066
Cirurgia cardíaca	Orr	1997	Sim	Sim	=	1112
Bypass Ao-Cor	Lippmann e Shahian	1997	Sim	Sim	=	80606
Hemorragia cerebral	Edwards, Hollingsworth et al	1999	Sim	Sim	++	81
Transplante hepático	Dvorchik et al	1994	Não	Não	++	155
Desmame ventilatório	Ashutosh, Lee et al	1992	Sim	Sim	=	25
Pós-RCP	Ebell	1993	Não	Não	+	218
Tempo internamento na UCI						
> 15 dias	Buchman, Kubos et al	1994	Sim	Sim	+	491
Cirurgia cardíaca	Tu e Guerriere	1993	Sim	Não	++	1409
Diagnóstico						
EAM	Selker, Griffith et al	1995	Sim	Sim	+	5773
Hipomagnesemia	Verive, Irazuzta et al	2000	Sim	Sim	+	463
Níveis aminoglicosídeos	Yamamura, Nishizawa et al	1998	Sim	Sim	++	30

Referência: primeiro autor; Valida.: utilização grupo de validação; Estat.: comparação com técnica estatística; CS: comparação com score; Result.: resultado global; =: similar; +: rede neuronal ligeiramente superior; ++: melhor rede neuronal; N.º: número de doentes considerados no trabalho; TCE: traumatismo cranioencefálico; *Bypass Ao Cor*: *Bypass* aorto-coronário; RCP: reanimação cárdio-pulmonar; EAM: enfarte agudo miocárdio.

Os modelos neuronais artificiais têm demonstrado serem capazes de prever o tempo de internamento na UCI (Buchman et al., 1994, Frize et al., 1995, Tu & Guerriere, 1993) e estiveram entre os primeiros a utilizar redes neuronais em bases de dados

clínicas. Avaliaram a relação entre variáveis demográficas, fisiológicas e terapêuticas e o tempo de internamento na UCI. Demonstraram um poder preditivo superior a um modelo de regressão logística na previsão da probabilidade de doentes cirúrgicos terem um tempo de internamento na UCI igual ou superior a sete dias.

Tu et al (Tu & Guerriere, 1993) desenvolveram um modelo neuronal artificial para estratificar os doentes em três grupos de risco para internamento prolongado na UCI após cirurgia cardíaca (Fraser & Turney, 1990).

Outros trabalhos, utilizando modelos neuronais artificiais para a previsão de mortalidade na UCI têm demonstrado que estas têm um desempenho igual ou mesmo melhor que os modelos baseados na regressão logística (Doig et al., 1993, Dybowski et al., 1996, Frize et al., 1995).

Clermont et al (Clermont et al., 2001) demonstraram que os modelos neuronais artificiais tinham um desempenho comparável ao modelo APACHE III, para a previsão da mortalidade hospitalar (áreas sob a curva ROC = 0,836 e 0,839, respectivamente; *goodness-of-fit* = 12,6 e 15,3, respectivamente; $p = 0,13$ e $0,05$, respectivamente). Os modelos neuronais artificiais foram também capazes de ter um desempenho semelhante ao APACHE III original, mesmo quando não foram designadas variáveis preditivas específicas à priori (área sob a curva ROC = 0,8; *goodness-of-fit* = 14,4; $p = 0,07$).

Ashish Nimgaonkar et al (Nimgaonkar et al., 2004) compararam o desempenho entre o APACHE II e modelos neuronais artificiais numa UCI universitária indiana. Os modelos neuronais artificiais mostraram um melhor desempenho do que o modelo APACHE II para a previsão da mortalidade hospitalar (áreas sob a curva ROC = 0,87 versus 0,77; *goodness-of-fit* = 22,4 versus 123,5). A análise do ganho de informação por cada uma das vinte e duas variáveis revelou que os modelos neuronais artificiais conseguiam prever a mortalidade usando apenas quinze variáveis (área sob a curva ROC = 0,88; *goodness-of-fit* = 27,7).

Outra aplicação comum dos modelos neuronais artificiais na UCI é a análise em tempo real da configuração dos traçados de electrocardiogramas e electroencefalogramas. Um algoritmo baseado num modelo neuronal artificial identificava em tempo real a isquemia cardíaca, com elevada sensibilidade, baseando-se na análise do segmento ST (Maglaveras et al., 1998). Os modelos neuronais artificiais têm também sido utilizados na análise de padrões electroencefalográficos em crianças (Si et al., 1998) e em adultos sob sedação (Veselis et al., 1991).

Por último, os modelos neuronais artificiais foram também aplicados na análise de padrões hemodinâmicos nos doentes graves (Hanson et al., 1998, Spencer et al., 1997) e na embolia pulmonar (Tourassi et al., 1993).

4. Vantagens e limitações das redes neuronais artificiais

Em termos gerais, os modelos neuronais artificiais requerem menos conhecimento estatístico para o seu desenvolvimento, em oposição aos modelos de regressão logística, em que é necessário um treino e conhecimento aprofundado de vários conceitos estatísticos (Tu 1996).

As redes neuronais artificiais têm a capacidade de detectar implicitamente qualquer relação não linear e complexa entre as variáveis preditivas e as variáveis dependentes (Guerriere & Detsky 1991, Hinton, 1992).

Se existir uma quantidade significativa de não linearidade entre as variáveis preditivas e os *outcomes*, na amostra de treino, então a rede neuronal artificial automaticamente ajusta os pesos das conexões na sua estrutura para evidenciar essas não linearidades (Tu 1996).

A camada interna das redes neuronais tem a capacidade de detectar interações ou inter-relações entre todas as variáveis dos nodos de entrada (Lautentiis & Ravdin, 1994). Os modelos neuronais artificiais são uma “caixa negra” e têm uma capacidade limitada de explicitamente identificar possíveis relações de causalidade. Não são muito adequados quando é importante a interpretação por parte do utilizador das relações capturadas, daí a designação de “caixa negra”, o que os pode tornar menos atractivos para o entendimento clínico (Lucas & Abu-Hanna, 1999).

Na regressão logística é possível identificar quais as variáveis que mais influenciam o *outcome* de interesse, de acordo com o valor dos coeficientes β e dos *odds ratios* associados.

O modelo neuronal artificial tem ainda questões metodológicas por resolver (Smith, 1993).

É difícil concluir que foi desenvolvido o melhor modelo possível para uma determinada aplicação, face à existência de múltiplos algoritmos de treino e à natureza empírica do desenvolvimento do modelo. No presente, existem poucas linhas de orientação para a predeterminação da estrutura da rede e do algoritmo de treino que terão o melhor desempenho para um determinado problema (Tu, 1996).

Ao contrário dos métodos estatísticos, as redes neuronais artificiais são adaptativas. Os dados são apresentados iterativamente, à medida que a rede treina. A seguir, esta reavalia as classificações e previsões efectuadas. Durante as iterações a rede é treinada para reconhecer padrões nos dados (Lipmann, 1987).

As redes neuronais artificiais são resistentes ao ruído presentes nos dados e são capazes de expressar interações complexas entre as variáveis, sendo mais flexíveis

do que a regressão onde o tipo particular de função a ser aprendida tem de ser conhecido a priori (Lucas & Abu-Hanna, 1999).

A abordagem epidemiológica e estatística tradicional de análise de dados usa testes hipotético-dedutivos para o teste de hipóteses, através de métodos paramétricos e não-paramétricos como o teste do qui-quadrado, a correlação e a regressão (Altman, 1991, Bland, 2000). Contudo, estes métodos têm limitações. Embora disponibilizem uma medida do significado estatístico, não indicam necessariamente a importância clínica (Last et al., 1999). As técnicas de *Data Mining* não oferecem vantagens claras sobre a análise univariada ou bivariada, podem contudo aumentar a capacidade da análise multivariada, particularmente quando existem interações complexas entre as variáveis (Altman, 1991, Biganzoli et al., 2002, Bland, 2000, Dusseldorp & J., 2001). No âmbito da Medicina, o desempenho das redes neuronais artificiais tem tido como termo de comparação o das técnicas estatísticas, particularmente entre a rede neuronal *Perceptron* multi-camada e a regressão logística múltipla (Tu, 1996, Tu & Guerriere, 1993).

Dependendo do problema em estudo, esta comparação tem conduzido a uma alternância entre o optimismo (Duh et al., 1998, Eldar et al., 2002) e o pessimismo (Borque et al., 2001, Sinha et al., 2001), na utilização das redes neuronais artificiais na área da medicina.

Numa revisão de vinte e oito aplicações distintas, Sargent conclui que as redes neuronais são, no pior dos casos, equivalentes ou ligeiramente superiores à regressão logística múltipla por não terem que depender das exigências rígidas da independência das variáveis e dos pressupostos inerentes ao modelo logístico (Sargent, 2001).

Nos dados clínicos ocorrem frequentemente interações entre as variáveis independentes e a avaliação destas interações e os seus efeitos na variável dependente é difícil (Dusseldorp & J., 2001).

A regressão logística é importante na geração de estimativas de sobrevida em populações e para a identificação de factores de risco, contudo não são recomendadas para a previsão do doente individualizado (Anand et al., 1999, Bottaci et al., 1997, Suter et al., 1994).

Como sumariado no Quadro 3, as redes neuronais oferecem vantagens e limitações quando comparadas com a regressão logística para a previsão de *outcomes* clínicos.

Quadro 3 - Vantagens e limitações do uso de redes neuronais artificiais para a previsão de outcomes clínicos. Adaptado de (Tu, 1996).

Vantagens dos modelos neuronais artificiais	Limitações dos modelos neuronais artificiais
Exigem menos treino estatístico para o desenvolvimento de modelos.	São uma “caixa negra” e têm capacidade limitada de identificar explicitamente relações de causalidade possíveis.
Conseguem detectar implicitamente relações complexas, não-lineares entre as variáveis dependentes e independentes.	Podem ser de utilização mais difícil no terreno.
Têm a capacidade de detectar todas as interações possíveis entre as variáveis preditivas.	Exigem maiores recursos computacionais.
Os modelos podem ser desenvolvidos usando múltiplos algoritmos.	São susceptíveis à sobre-especialização.
	O desenvolvimento do modelo é empírico e algumas questões metodológicas estão por resolver.

O processo de Descoberta do Conhecimento em Bases de Dados pode contribuir para ultrapassar estes problemas através da identificação de padrões previamente inaparentes, ou através da aprendizagem a partir dos dados para o suporte à tomada de decisão, previsão, prognóstico ou diagnóstico (Downs et al., 1996).

5.Potencialidades da Descoberta de Conhecimento em Bases de Dados na Medicina Intensiva

A divulgação da utilização dos sistemas de informação nas UCIs e o crescimento das bases de dados clínicas provenientes de diversas fontes requerem o recurso a ferramentas especializadas no armazenamento, na integração e análise dos dados para o seu uso efectivo (Coiera, 1997, Peña-Reyes & Sipper, 2000, Shortliffe & Blois, 2001).

O simples registo, codificação e armazenamento da informação gerada diariamente nas UCIs, embora possam ser utilizados para análise retrospectiva, restringem a sua utilização como ferramenta de suporte à tomada de decisão clínica (à cabeceira do doente), uma vez que o volume de informação está para além da capacidade de integração do clínico (Peña-Reyes & Sipper, 2000, Shortliffe & Barnett, 2001).

A recolha, gestão, análise e interpretação da informação são fundamentais para a boa prática clínica, particularmente em decisões relativas ao prognóstico e ao processo de decisão clínica (Shortliffe & Barnett, 2001).

Na área da Medicina Intensiva, os dados clínicos colocam vários desafios à utilização do processo de Descoberta de Conhecimento em Bases de Dados, nomeadamente:

- subjectividade e heterogeneidade dos dados com origem em múltiplas fontes; alguma informação, como os dados demográficos e os antecedentes clínicos pré-admissão à UCI, permanece constante; outros dados são registados durante a estadia na UCI, tais como a monitorização de parâmetros, os resultados dos exames complementares de diagnóstico, a utilização de técnicas de suporte de função de órgão, a nutrição, a ocorrência de infeções nosocomiais e a terapêutica administrada;
- características individuais dos doentes; é muitas vezes difícil comparar valores absolutos de parâmetros entre doentes; por exemplo, os valores de muitos parâmetros são considerados anormais se modificam significativamente dos valores prévios para um doente em particular, e não se desviam de um valor universal; outro exemplo, os doentes podem responder mais lentamente ou mais rapidamente a determinadas intervenções terapêuticas; por isso, para além da aprendizagem "global" a partir dos doentes, será útil um outro nível "local" de aprendizagem das características e do doente individualizado;
- as bases de dados clínicas são muitas vezes caracterizadas pela imperfeição (valores de parâmetros omissos), incorrecção (ruído sistemático ou aleatório nos dados), escassez (poucos registos disponíveis ou não representativos) e inexactidão (selecção inadequada de parâmetros); as características destes dados devem ser tomadas em consideração no desenvolvimento de modelos de previsão, alarmes inteligentes e instrumentos de apoio à decisão.

Todas estas características revelam como o processo de Descoberta de Conhecimento em Bases Dados é potencialmente útil na Medicina Intensiva. Este processo permite extrair informação útil e disponibilizar inferências a partir dos dados disponíveis, de forma semi-automática, sob a forma simbólica ou sub-simbólica, graças à sua capacidade de aprendizagem e tolerância a erros e dados omissos ou com grande quantidade de ruído (Bajo & Ballesteros, 2002).

6. Bibliografia

- Abu-Hanna, A. & de Keizer, N. (2003) Integrating classification trees with local logistic regression in Intensive Care prognosis, *Artificial Intelligence in Medicine*, 29, pp. 5–23.
- Abu-Hanna, A. & Lucas, P.J.F. (2001) Prognostic models in medicine AI and statistical approaches, *Special issue of Methods Inf Med*, 40, pp. 1-5.
- Aitchison, T.C., Sire, I.J.M., Watt, D.C. & MacKie, R.M. (1995) Prognostic trees to aid prognosis in patients with cutaneous malignant melanoma, *BMJ* 311, pp. 1536–1539.
- Altman, D.G. (1991) *Practical statistics for medical research* (London, Chapman Hall/CRC).
- Anand, S.S., Smith, A.E., W., H.F., Anand, J.S., Hughes, J.G. & Bartels, P.H. (1999) An evaluation of intelligent prognostic systems for colorectal cancer, *Artificial Intelligence in Medicine*, 15(2), pp. 193-214.
- Andreassen, S., Woldbye, M., Falck, B. & Andersen, S.K. (1987) MUNIN—A causal probabilistic network for interpretation of electromyographic findings in: J. McDermott (Ed) *Tenth International Joint Conference on Artificial Intelligence* (Los Altos, CA; Morgan Kaufmann).
- Bajo, N.S. & Ballesteros, M.Á. (2002) Redes neuronales: concepto, aplicaciones y utilidad en medicina, *Aten Primaria*, 30 (2), pp. 119-120.
- Bath, P. & Philp, I. (1998) A hierarchical classification of dependency amongst older people using artificial neural networks, *Health Care in Later Life*, 3(11), pp. 59-69.
- Bath, P.A., Morgan, K., Pendleton, N., Clague, J., Horan, M. & Lucas, S. (2000) A new approach to risk determination: Prediction of new falls among community-dwelling older people using a genetic algorithm neural network *Journal of Gerontology (medical science)*, 55A, pp. 7-21.
- Bellazzi, R. & Zupan, B. (2006) Predictive data mining in clinical medicine: Current issues and guidelines, *Int. J. Med. Inform.*, pp. doi:10.1016/j.ijmedinf.2006.1011.1006
- Benoit, G. (2002) Data mining, *Annual Review of Information Science and Technology*, 36, pp. 265-310.
- Berry, M. & Linoff, G. (1997) *Data Mining Techniques – For Marketing, Sales and Customer Support* John Wiley & Sons).

- Berson, A., Smith, S. & Thearling, K. (2000) *Building data mining applications for CRM* (McGraw-Hill.).
- Biganzoli, E., Boracchi, P. & Marubini, E. (2002) A general framework for neural network models on censored survival data, *Neural Networks*, 15(2).
- Bigus, J.P. (1996) *Data Mining with Neural Network – Solving Business Problems from Application Development to Decision Support* (McGraw-Hill).
- Bland, M. (2000) *An introduction to medical statistics* (Oxford, U.K., Oxford Medical Publications).
- Borque, A., Sanz, G., Allepuz, C., Plaza, L. & Rioja, L.A. (2001) The use of neural networks and logistic regression analysis for predicting pathological stage in men undergoing radical prostatectomy: a population based study., *J Urol*, 166, pp. 1672-1678.
- Bottaci, L., Drew, P.J., Hartley, J.E., B., H.M., Farouk, R., Lee, P.W.R., Macintyre, I.M.C., Duthie, G.S. & Monson, J.R.T. (1997) Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions., *Lancet*, 350(9076), pp. 469-472.
- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and Regression Trees* (Belmont, CA, Wadsworth International).
- Buchman, T.G., Kubos, K.L., Seidler, A.J. & Siegforth, M.J. (1994) A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit, *Crit Care Med* 22, pp. 750-762.
- Chester, M. (1993) *Neural Network - A tutorial* (Prentice-Hall).
- Clermont, G., Angus, D.C., DiRusso, S.M., Griffin, M. & Linde-Zwirble, W.T. (2001) Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models, *Crit Care Med* 29, pp. 291-296.
- Coiera, E. (1997) *Guide to medical informatics, the Internet and telemedicine*. (London, Arnold).
- Cross, B.S., Harrison, R.F. & Kennedy, R.L. (1995) Introduction to neural networks, *Lancet*, 346, pp. 1075-1079.
- Dietterich, T.G. (1997) Limitations on Inductive Learning (Extended Abstract) (Oregon State University).
- Doig, G.S., Inman, K.J. & Sibbald, W.J. (1993) Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression, *Proc Annu Symp Comput Appl Med Care*, pp. 361-365.

- Downs, J., Harrison, R.F., Kennedy, R.L. & Cross, S.S. (1996) Application of the fuzzy ARTMAP neural network model to medical pattern classification tasks, *Artificial Intelligence in Medicine*, 8(4), pp. 403-428.
- Duh, M.S., Walker, A.M., Pagano, M. & Kronlund, K. (1998) Prediction and cross-validation of neural networks versus logistic regression: using hepatic disorders as an example, *Am J Epidemiol* 147, pp. 407-413.
- Dusseldorp, E. & J., M.J. (2001) Prediction in medicine by integrating regression trees into regression analysis with optimal scaling, *Methods of Information in Medicine*, 40, pp. 403-409.
- Dybowski, R., Weller, P., Chang, R. & Gant, V. (1996) Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm, *Lancet*, 347, pp. 1146-1150.
- Eldar, S., Siegelmann, H.T., Buzaglo, D., Matter, I., Cohen, A. & Sabo, E. (2002) Conversion of laparoscopic cholecystectomy to open cholecystectomy in acute cholecystitis: artificial neural networks improve the prediction of conversion, *World J Surg*, 26, pp. 79-85.
- Ennis, M., Hinton, G., Naylor, D., Revow, M. & Tibshirani, R. (1998) A comparison of statistical learning methods on the GUSTO database, *Statistics in Medicine*, 17, pp. 2501-2508.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996a) Data mining and knowledge discovery in databases, *Commun. ACM*, 39, pp. 24-26.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996b) The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, 39(11), pp. 27-34.
- Fayyad, U.M. (1991) On the Induction of Decision Trees for Multiple Concept Learning (EECS Dept. Univ. Michigan).
- Flexer, A. (1996) Statistical evaluation of neural networks experiments: minimum requirements and current practice in: R. Trappl (Ed) *13th European Meeting on Cybernetics and Systems Research*, (Vienna, Austria,
- Fraser, R.B. & Turney, S.Z. (1990) An expert system for the nutritional management of the critically ill, *Comput Methods Programs Biomed* 33, pp. 175-180.
- Frawley, W.J., Piatetsky-Shapiro, G. & Matheus, C.J. (1991) Knowledge Discovery in Databases, An Overview, in: G. Piatetsky-Shapiro & B. Frawley (Eds) *Knowledge Discovery in Databases* AAAI-MIT Press).
- Freitas, A. (2000) Understanding the Crucial Differences between Classification and Discovery of Association Rules – A Position Paper, *ACM SIGKDD Explorations*, 2(1).

- Frize, M., Ennett, C.M., Stevenson, M. & Trigg, H.C. (2001) Clinical decision support systems for intensive care units: using artificial neural networks., *Med. Eng. Phys*, 23(3), pp. 217–225.
- Frize, M., Solven, F.G. & Stevenson, M. (1995) Computer- assisted decision support systems for patient management in an intensive care unit, *Medinfo* 8, pp. 1009–1012.
- Giudici, P. (2003) *Applied Data Mining Statistical Methods for Business and Industry* Wiley & Sons).
- Goebel, M. & Gruenwald, L.A. (1999) Survey of Data Mining and Knowledge Discovery Software Tools, *ACM SIGKDD Explorations* 11(1), pp. 20-23.
- Guerriere, M.R. & Detsky, A.S. (1991) Neural networks: what are they? , *Ann Intern Med* 115, pp. 906-907.
- Hagan, M., Demuth, H. & Beale, M. (1996) *Neural Network Design* PWS Publishing Company).
- Halpern, N.A. & Pastores, S.M. (1999) Technology introduction in critical care: just knowing the price is not enough! , *Chest*, 116, pp. 1092–1099.
- Hanley, J.A. & Mc Neil, B.J. (1982) The meaning and use of the area under a receiver-operating characteristic (ROC) curve, *Radiology* 143, pp. 29–36.
- Hanson, C.W., Weiss, Y. & Frasch, F. (1998) Neurofuzzy analysis of hemodynamic data, *Abstr. Anesthesiology* 89(suppl), pp. 474.
- Hanson, C.W.r. & Marshall, B.E. (2001) Artificial intelligence applications in the intensive care unit, *Crit Care Med* 29, pp. 427–435.
- Harrison, T.H. (1998) *Intranet Data Warehouse* Berkeley Brasil).
- Haykin, S.S. (1999) *Neural networks : A comprehensive foundation*. (Upper Saddle River, NJ, Prentice Hall International).
- Hinton, G.E. (1992) How neural networks learn from experience, *Sci Am* 267, pp. 145-151.
- Jodouin, J.-F. (1994) *Les Réseaux de neurones : Principes et définitions* (Paris, Editions Hermès).
- Kass, G.V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics*, 29, pp. 119-127.
- Kayaalp, M., Cooper, G.F. & Clermont, G. (2000) Predicting ICU mortality: a comparison of stationary and nonstationary temporal models, *Proc. AMIA Symp*, 2000, pp. 418–422.
- Kibler, D. & Langley, P. (1988) Machine Learning as an Experimental Science, *Machine Learning*, 3 (1), pp. 5–8.
- Kohavi, R. & Provost, F. (1998) Glossary of terms, *Machine Learn* 30, pp. 271—274.

- Kohn, L.T., Corrigan, J.M. & Donaldson, M.S. (2000) *To err is human: building a safer health system* (Washington DC, Committee on Quality of HealthCare in America, Institute of Medicine).
- Kohonen, T. (1982) Self-Organized formation of Topologically Correct Feature Maps, *Biological Cybernetics*, 43, pp. 32-48.
- Kohonen, T. (1989) *Self Organization and Associative Memory* (Berlin, Springer-Verlag).
- Kohonen, T. (1995) *Self-organizing maps* (Berlin, Springer Verlag).
- Kuo, W.J., Chang, R.F., Chen, D.R. & Lee, C.C. (2001) Data mining with decision trees for diagnosis of breast tumour in medical ultrasonic images, *Breast Cancer Research and Treatment*, 66(51-57).
- Last, M., Schenker, A. & Kandel, A. (1999) Applying fuzzy hypothesis testing to medical data, in: A.S.S.O. N. Zhong (Ed) *New directions in rough sets, data mining, and granular-soft computing* (Berlin, Springer).
- Lautentiis, M.D. & Ravdin, P.M. (1994) Survival analysis of censored data: Neural network analysis detection of complex interactions between variables, *Breast Cancer Res Treat*, 32, pp. 113-118.
- Lavrac, N. (1999) Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, 16, pp. 3-23.
- Lee, S.J. & Siau, K. (2001) A review of data mining techniques, *Industrial Management & Data Systems*, pp. 41-46.
- Levine, R.F. (2001) Conference concluding remarks, *Cancer*, 91, pp. 1696-1697.
- Lipmann, R.P. (1987) An introduction to computing with neural nets, *IEEE ASSP Magazine*, 4, pp. 4-22.
- Lucas, P. (2004) Baysean analysis, pattern analysis, and data mining in health care, *Current opinion in Critical Care* 10, pp. 399-403.
- Lucas, P.J., de Bruijn, N.C. & Schurink, K. (2000) A probabilistic and decision theoretic approach to the management of infectious disease at the ICU., *Artif Intell Med* 19, pp. 251-279.
- Lucas, P.J.F. & Abu-Hanna, A. (1999) Prognostic methods in medicine, *Artif Intell Med* 15(2), pp. 105-119.
- Maglaveras, N., Stamkopoulos, T. & Pappas, C. (1998) An adaptive backpropagation neural network for real-time ischemia episodes detection: development and performance analysis using the European ST-T database, *IEEE Trans Biomed Eng*, 45, pp. 805-813.
- McCulloch, W.S. & Pitts, W. (1946) A Logical Calculus of the Ideas Imminent in Nervous Activity, *Bulletin of Mathematical Biophysics*, 5, pp. 115-133.

- Michalski, R.S., Bratko, I. & Kubat, M. (1997) *Machine learning and data mining: Methods and applications* (New York, John Wiley).
- Nimgaonkar, A., Karnad, Dilip R. , Sudarshan, S., Ohno-Machado, L. & Kohane, I. (2004) Prediction of mortality in an Indian intensive care unit. Comparison between APACHE II and artificial neural networks, *Intensive Care Med* 30, pp. 248–253.
- Ohno-Machado, L. (1996) Sequential use of neural networks for survival prediction in AIDS, *Proc. AMIA Annu. Fall. Symp*, 1996, pp. 170–174.
- Ohno-Machado, L. & Musen, M.A. (1997) Sequential versus standard neural networks for pattern recognition: an example using the domain of coronary heart disease, *Comput. Biol. Med*, 27(4), pp. 267–281.
- Ohno-Machado, L., Resnic, F.S. & Mathne, M.E. (2006) Prognosis in critical care, *Annu Ver. Biomed Eng.*, 8, pp. 567-599.
- Patterson, D. (1996) *Artificial Neural Networks – Theory and Applications* (Singapore, Prentice Hall).
- Patterson, D.W. (1990) *Introduction to Artificial Intelligence and Expert Systems* Prentice Hall, Englewood Cliffs).
- Peña-Reyes, C.A. & Sipper, M. (2000) Evolutionary computation in medicine: An overview, *Artificial Intelligence in Medicine*, 19, pp. 1-23.
- Pendharkar, P.C., Rodger, J.A., Yaverbaum, G.J., Herman, N. & Benner, M. (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns, *Expert Systems with Applications*, 17, pp. 223-232.
- Penny, W. & Frost, D. (1996) Neural networks in clinical medicine, *Med Decis Making* 16, pp. 386-398.
- Quinlan, J.R. (1987) Simplifying decision trees, *International Journal of Man-Machine Studies*.
- Quinlan, J.R. (1988) *Decision trees and multi-valued attributes* Machine Intelligence).
- Quinlan, J.R. (1993) *C4.5: programs for machine learning* (San Mateo (CA), Morgan Kaufman).
- Quinlan, J.R. (1997) *C5.0 Data Mining Tool*
- Rezende, S.O. (2003) *Sistemas Inteligentes – Fundamentos e Aplicações* (Brazil, Editora Manole).
- Richards, G., Rayward-Smith, V.J., Sonksen, P.H., Carey, S. & Weng, C. (2001) Data mining for indicators of early mortality in a database of clinical records, *Artificial Intelligence in Medicine*, 22, pp. 215-231.
- Rosenblatt, R. (1959) *Principles of Neurodynamics* (New York, Spartan Books).

- Rosenfeld, B.A., Dorman, T. & Breslow, M.J. (2000) Intensive care unit telemedicine: alternate paradigm for providing continuous intensivist care, *Crit Care Med* 28, pp. 3925–3931.
- Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning representations by backpropagating errors, *Nature* 323, pp. 533-536.
- Sargent, D.J. (2001) Comparison of artificial neural networks with other statistical approaches. Results from medical data sets, *Cancer*, 91, pp. 1636-1642.
- Sarle, W.S. (1997) Neural network FAQ, part 1 of 7: introduction, periodic posting to the Usenet newsgroup comp.ai.neural-nets
- Schaffer, C. (1994) A Conservation Law for Generalization Performance in: W.W.C.H. Hirsh (Ed) *Eleventh International Conference on Machine Learning* (New Jersey, New Brunswick).
- Schwabacher, M., Hirsh, H. & Ellman, T. (1995) Inductive Learning from Engineering Design Optimization, in: D.A.P. Riddle (Ed) *ICML Workshop on Applying Machine Learning in Practice*
- Schwarzer, G., Vach, W. & Schumacher, M. (2000) On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology, *Statist Med* 19, pp. 541-561.
- Shortliffe, E.H. & Barnett, G.O. (2001) Medical data: Their acquisition, storage and use, in: E.H.P. Shortliffe, L. E. (Ed) *Medical informatics computer applications in health care and biomedicine* (New York, Springer).
- Shortliffe, E.H. & Blois, M.S. (2001) The computer meets biology and medicine: Emergence of a discipline, in: E.H.S.L.E. Perreault (Ed) *Medical informatics computer applications in health care and biomedicine* (New York, Springer.).
- Si, Y., Gotman, J. & Pasupathy, A. (1998) An expert system for EEG monitoring in the pediatric intensive care unit, *Electroencephalogr Clin Neurophysiol* 106, pp. 488–500.
- Sinha, M., Kennedy, C.S. & Ramundo, M.L. (2001) Artificial neural network predicts CT scan abnormalities in paediatric patients with closed head injury, *J Trauma* 50, pp. 308-312.
- Smith, M. (1993) *Neural Networks for Statistical Modeling* (New York, Van Nostrand Reinhold).
- Spencer, R.G., Lessard, C.S. & Davila, F.S.d., recognition and prediction of haemodynamic patterns in the intensive care unit (1997) Selforganising discovery, recognition and prediction of haemodynamic patterns in the intensive care unit, *Med Biol Eng Comput*, 35, pp. 117–123.

- Suter, P., Argmandis, A. & Beaufils, F. (1994) Predicting outcome in ICU patients. Consensus Conference organized by the ESICM and SRLF., *Intensive Care Med*, 20, pp. 390-397.
- Tourassi, G.D., Floyd, C.E. & Sostman, H.D. (1993) Acute pulmonary embolism: Artificial neural network approach for diagnosis, *Radiology*, 189, pp. 555-558.
- Trujillano Cabello, J., Badía Castelló, M., March Llanes, J., Rodríguez Pozo, Á., Serviá, G., L. & Sorribas Tello, Y.A. (2005) Redes neuronales artificiales en Medicina Intensiva. Ejemplo de aplicación con las variables del MPM II, *Med Intensiva*, 29 (1), pp. 13-20.
- Trujillano, J., March, J., Badia, M., Rodríguez, A. & Sorribas, A. (2003) Aplicación de las redes neuronales artificiales para la estratificación de riesgo de mortalidad hospitalaria, *Gac Sanit* 17(6), pp. 504-511.
- Trujillano, J., March, J. & Sorribas, A. (2004) Aproximación metodológica al uso de redes neuronales artificiales para la predicción de resultados en medicina, *Med Clin (Barc)* 122 (Supl 1), pp. 59-67.
- Tu, J.V. (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, *J Clin Epidemiol*, 49(11), pp. 1225-1231.
- Tu, J.V. & Guerriere, M.R.J. (1993) Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery, *Comput Biomed Res* 26, pp. 220-229.
- Veselis, R.A., Reinsel, R. & Sommer, S. (1991) Use of neural network analysis to classify electroencephalographic patterns against depth of midazolam sedation in intensive care unit patients, *J Clin Monit* 7, pp. 259-267.
- Wei, J.T., Zhang, Z. & Barnhill, S.D. (1998) Understanding artificial neural networks and exploring their potential applications for the practicing urologist, *Urology*, 52, pp. 161-172.
- Weiss, S. & Indurkha, N. (1998) *Predictive Data Mining – a practical guide* Morgan Kaufmann).
- Widrow, B. & Lehr, M. (1990) 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and BackPropagation *Proceedings of the IEEE* (New York,
- Zhong, N. & Dong, J. (2002) Mining interesting rules in meningitis data by cooperatively using GDT-RS and RSBR, *Lecture Notes in Artificial Intelligence*, 2336, pp. 405-416.

CAPÍTULO IV

Indução de modelos de previsão de mortalidade e da falência de órgãos e sistemas no doente grave: contribuição pessoal

Capítulo IV- Indução de modelos de previsão de mortalidade e da falência de órgãos e sistemas no doente grave: contribuição pessoal

São apresentados, sob a forma de artigos publicados em revistas, actas de conferências internacionais e de capítulo de livro, os resultados do trabalho de investigação desenvolvido.

1. Introdução

Este capítulo inclui os resultados do trabalho de investigação apresentados sob a forma de artigos originais publicados em revistas, capítulo de livro e actas de conferências internacionais. Precedendo cada trabalho de investigação, inclui-se um preâmbulo que tem por objectivo expor o enquadramento de cada trabalho, complementar a descrição de materiais e métodos e apresentar alguns dos resultados mais relevantes.

As contribuições pessoais podem ser enquadradas em duas vertentes de acordo com os objectivos definidos nesta dissertação. A saber, a indução de modelos de previsão da mortalidade intra-hospitalar e os modelos de previsão de falência de órgãos e sistemas no doente grave admitido na UCI.

Para além dos objectivos citados, procurou-se também viabilizar o desenvolvimento de um sistema de apoio à decisão inteligente em Medicina Intensiva.

O percurso da investigação e a elaboração dos trabalhos apresentados neste capítulo seguiu as seguintes fases:

- dar formação aos elementos do Serviço de Cuidados Intensivos do Hospital Geral de Santo António, para participarem no estudo *European ICU Study (EURICUS) II - The effect of harmonizing and standardizing the nursing tasks on the Intensive Care Units of the European Community* financiado pela *Commission of the European Communities* (bolsa: BMH4-CT96-0817) e pela *Foundation for Research on Intensive Care in Europe (FRICE)* (Miranda 1999) ;
- registo, recolha e verificação dos dados necessários para referido estudo (a nível local);
- consulta e análise das bases de dados do estudo EURICUS II, de acordo com os objectivos propostos pela *Foundation for Research on Intensive Care in Europe* com

sede no (*FRICE*) *Health Services Research Unit* do Hospital Universitário de Groningen (Holanda).

Nesse âmbito, foi realizado um estágio no *Health Services Research Unit* do Hospital Universitário de Groningen durante doze meses, a convite e sob a orientação do Dr. Reis Miranda, onde procedi ao tratamento dos dados necessários para a elaboração das bases de dados necessárias para a presente dissertação.

A continuidade do processo de investigação decorreu no Serviço de Cuidados Intensivos do Hospital Geral de Santo António no Porto e na Escola de Engenharia da Universidade do Minho, do qual resultou este trabalho de dissertação sob a orientação do Professor Doutor J. Lopes Gomes e a co-orientação dos Professores. Doutores Maia Neves e Filipe Santos.

2. População e métodos

Para o desenvolvimento desta investigação foram utilizados dados derivados do estudo EURICUS II cuja metodologia e resultados foram já publicados (Miranda 1999; Fidler, Nap et al. 2004).

O estudo EURICUS II decorreu entre 1998 e 1999 (durante dez meses, de Novembro a Agosto) em quarenta e duas UCIs de nove países da União Europeia: Dinamarca, Holanda, França, Luxemburgo, Reino Unido, Itália, Espanha, Noruega e Portugal. Em cada país, a selecção das UCIs englobou diferentes tipos de hospitais (universitários e não universitários). A participação e selecção das UCIs foram feitas a convite da FRICE, sem o propósito de representar o país envolvido (Miranda 1999).

A dimensão mediana dos hospitais participantes foi de 500 camas [amplitude inter-quartis (AIQ) 350 – 850] e o número mediano de camas por UCI foi de 8 (AIQ 8 – 10). Trinta e sete UCIs eram médico-cirúrgicas, três cirúrgicas e duas médicas. Vinte e duas UCIs pertenciam a hospitais não-universitários. (Fidler, Nap et al. 2004)

Todos os doentes consecutivamente admitidos nas UCIs participantes foram incluídos no estudo EURICUS II .

Para cada doente, os dados foram registados manualmente de forma padronizada (codificados por país, UCI) pelos intensivistas e enfermeiros, de forma prospectiva (durante dez meses). Foram consideradas as seguintes variáveis:

- idade (em anos) à data de admissão na UCI;
- principal categoria diagnóstica de admissão, usando uma lista de setenta e nove categorias possíveis (Knaus, Wagner et al. 1991);
- variáveis necessárias para o cálculo do SAPS II score (Le Gall, Lemeshow et al. 1993) (às 24 horas após a admissão à UCI);
- variáveis necessárias para cálculo do SOFA score diário por órgão e sistema (durante um período de 4 meses)(Vincent, de Mendonça et al. 1998);
- tipo de admissão: cirúrgica não programada (1); cirúrgica programada (2) e médica(3);
- local de origem da admissão à UCI: bloco operatório (1), sala de recobro (2), sala de emergência (3), enfermaria geral (4), outra UCI (5), outro hospital (6) e outros locais (7);
- duração do tempo de internamento na UCI (em dias);
- estado vital do doente (à data da alta hospitalar).

Os enfermeiros registaram hora a hora durante toda a estadia do doente na UCI os eventos clínicos adversos.

Os eventos clínicos adversos (Quadro 1 e 2) foram definidos como a frequência, gravidade e duração dos valores fora dos limites de normalidade preestabelecidos, para quatro parâmetros monitorizados:

- frequência cardíaca (**HR**);
- tensão arterial sistólica (**BP**);
- saturação arterial de O₂ (**SaO₂**);
- débito urinário horário (**UR**).

Os eventos clínicos adversos foram classificados em eventos (Ev) ou eventos críticos (EvCr), dependendo da gravidade e do período de tempo em que os valores dos quatro parâmetros estiveram fora dos intervalos considerados normais.

Quadro 1 – Eventos clínicos adversos .

Evento	Intervalo de normalidade	Continuamente fora do intervalo de normalidade	Intermitentemente fora do intervalo de normalidade
BP (mmHg)	90-180	≥ 10'	≥ 10' em 30'
SaO ₂ (%)	≥ 90	≥ 10'	≥ 10' em 30'
HR (ppm)	60-120	≥ 10'	≥ 10' em 30'
UR (ml/hora)	≥ 30	≥ 1 hora	

Quadro 2 – Eventos críticos adversos.

Evento Crítico	Intervalo de normalidade	Continuamente fora do intervalo de normalidade	Intermitentemente fora do intervalo de normalidade	Evento em qualquer altura
BP (mmHg)	90-180	≥ 60'	≥ 60' em 120'	BP < 60
SaO ₂ (%)	≥ 90	≥ 60'	≥ 60' em 120'	SpO ₂ < 80
HR (ppm)	60-120	≥ 60'	≥ 60' em 120'	HR<30 ou HR>180
UR (ml/hora)	≥ 30	≥ 2 horas		≤ 10

Os eventos clínicos adversos são definidos como variáveis binárias (0 ou 1, se dentro ou fora do intervalo de normalidade). Caso os seus valores se encontrem fora dos limites de normalidade, é necessário registar a duração em minutos para a classificação em evento ou evento crítico.

Assim, considera-se um evento quando o valor do parâmetro analisado se mantém fora dos limites de normalidade por um período tempo contínuo ≥ a 10 minutos (1 hora no caso de UR) e menor que 60 minutos (2 horas no caso de UR). Considera-se ainda como um evento quando num período de 30 minutos, se verificam valores fora dos limites, cuja soma durante o período em questão é ≥ a 10 minutos.

Considera-se um evento crítico, quando o valor do parâmetro se mantém fora dos limites de normalidade por um período contínuo de tempo ≥ a 60 minutos (2 horas no caso do UR).

Considera-se ainda como evento crítico quando, num período de tempo de duas horas, se verificam valores fora dos limites que não ultrapassem os 60 minutos de duração, mas cuja soma durante o período em questão, seja igual ou superior a 60 minutos. Além disso, também se considera como evento crítico quando o parâmetro analisado se situa em valores considerados graves (BP <60mmHg, SaO₂ <60%, HR < 30 ou > 180 ppm e UR ≤ 10 ml/hora).

O processo de ensino e instrução formal para o registo das variáveis nos impressos foi desenvolvido de forma faseada, envolvendo os coordenadores (nacionais e locais) e posteriormente todos os intensivistas e enfermeiros de cada UCI.

A metodologia do controlo da qualidade destes registos está descrita no manual de procedimentos do EURICUS II (Miranda 1999) e ocorreu a três níveis: coordenador da UCI, coordenador nacional e no centro de coordenação em Groningen (*Health Services Research Unit* do Hospital Universitário de Groningen), envolvendo a verificação e dupla validação do preenchimento dos impressos normalizados.

A partir dos dados do estudo EURICUS II, foram criadas duas bases de dados distintas, utilizadas para a investigação desenvolvida nesta dissertação:

A. Uma com todas as variáveis descritas, à excepção do SOFA score diário por órgão e sistema, tendo ainda sido incluídas novas variáveis derivadas das originais, que incluiu 13 164 doentes;

B. Outra, com a inclusão do SOFA score diário por órgão e sistema, durante quatro meses, e que incluiu 4864 doentes.

Para a presente dissertação, foram excluídos da análise os doentes coronários agudos, os pós-operatórios de cirurgia coronária, os queimados, os com idade inferior a dezoito anos, os doentes com tempo de internamento nas UCIs inferior a oito horas e os doentes em morte cerebral (admitidos para manutenção de dador de órgãos). No caso dos doentes readmitidos na UCI no mesmo hospital de admissão, foi analisada apenas a primeira admissão.

A avaliação estatística foi efectuada recorrendo aos parâmetros descritivos habituais (média, desvio padrão, mediana, proporção). As variáveis quantitativas foram comparadas por provas paramétricas (*t-Student*, ANOVA) ou equivalentes não paramétricas (Wilcoxon, Kruskal-Wallis, Mann-Whitney); as proporções foram comparadas pela prova do qui-quadrado ou, quando apropriado, pela prova exacta de Fisher.

Para a indução de modelos de previsão foram utilizadas redes neuronais artificiais do tipo *perceptron* multi-camada com treino realizado por retropropagação do erro; árvores de decisão e indução de regras decisão. Para a indução de modelos de segmentação foram aplicadas redes neuronais de Kohonen.

O processo de treino das redes neuronais foi desenvolvido através de múltiplas iterações de forma a excluir resultados aleatórios e assegurar a individualização e generalização.

Para a avaliação dos modelos gerados de forma semi-automática foi usado o método de Houldout (Flexer 1996), em que a população é dividida em duas partes mutuamente exclusivas, sendo uma usada para treino e a outra para validação dos modelos.

Para evitar a sobre-especialização dos modelos foi definido um número máximo de épocas (número de ajustes dos pesos das interligações entre os nodos) de treino das redes neuronais artificiais. Outra condição de paragem do treino foi o da variação do erro, isto é, quando a variação do erro se aproximava do zero, o treino era interrompido.

O desempenho dos modelos gerados foi avaliado através do cálculo da precisão, sensibilidade, especificidade e capacidade discriminativa [(avaliada pela área sob a curva *receiver-operating characteristic* (área sob a curva ROC)] (Hanley and Mc Neil

1982) e respectivos intervalos de confiança a 95%. Para além dos métodos descritos, foi efectuada a comparação com os resultados obtidos pelos modelos obtidos por regressão logística.

3. Indução de modelos de previsão da mortalidade intra-hospitalar

Nesta secção é apresentado o trabalho *Mortality assessment in intensive care units via adverse events using artificial neural networks* publicado na revista Artificial Intelligence in Medicine (2006), Elsevier 36, 223-234. ISSN: 0933-3657.

Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, José Neves.

3.1. Preâmbulo

A monitorização de vários parâmetros é uma das actividades centrais numa UCI. O principal objectivo é o de avaliar as variações dos valores medidos em relação a um intervalo de "normalidade". O pressuposto subjacente é que o *outcome* final (estado vital à data da alta hospitalar) do doente está dependente da intensidade e frequência da ocorrência destes eventos.

Uma das características dos modelos gerais de previsão do risco de mortalidade mais utilizados nas UCIs é o de se basearem nos dados do doente obtidos nas primeiras 24 horas após a admissão à UCI. Toda a informação após a admissão é excluída. Estes modelos não incorporam os dados da monitorização contínua do doente, isto é, não são habitualmente actualizados com durante a estadia do doente na UCI. Este facto limita a sua utilização como ferramenta de apoio ao processo de decisão clínico em tempo útil, no doente individualizado.

Este trabalho apresenta uma direcção alternativa para a indução de modelos de previsão da mortalidade intra-hospitalar do doente grave admitido na UCI. Neste estudo foram desenvolvidos modelos de previsão da mortalidade intra-hospitalar, após o primeiro dia de internamento, baseados nos dados estáticos (ex.: SAPS II score, idade, tipo e local de origem da admissão) e nos dados dinâmicos resultantes da monitorização horária de quatro parâmetros (tensão arterial sistólica, saturação arterial de O₂, frequência cardíaca e débito urinário horário), registados por rotina em todas as UCIs e designados por eventos clínicos adversos nesta dissertação (definidos na secção 2).

Após o pré-processamento de dados (selecção, limpeza de dados omissos/incorrectos e transformação de variáveis) dos 15027 doentes, foram incluídos para análise 13164 doentes (87,6%).

As principais características da população estudada estão ilustradas no Quadro 3.

Quadro 3 - Características gerais da população estudada

População, (n)	13164
Idade (anos), mediana (AIQ) ^a	65 (46 – 76)
SAPS II , mediana (AIQ) ^a	30 (18 – 43)
Tempo de internamento na UCI (dias), mediana (IQ-range)	2 (1 – 5)
Tipo de admissão, [n, (%)]	
Cirúrgica – não programada	3411 (25,9%)
Cirúrgica - programada	3311 (25,2%)
Médica	6442 (48,9%)
Local de origem da admissão, [n, (%)]	
Bloco operatório	4974 (37,8%)
Sala de recobro	510 (3,9%)
Sala de emergência	3721 (28,3%)
Enfermaria	2431 (18,5%)
Outra UCI	433 (3,3%)
Outro hospital	823 (6,3%)
Outros	272 (2,1%)
Diagnóstico [n, (%)]	
Pós-cirúrgico	7342 (55,8%)
Médico	5822 (44,2%)
Mortalidade intra-hospitalar, [n, (%)]	2484 (19,6%)

^a Amplitude inter-quartis

A idade mediana dos doentes foi de 65 anos, com uma diferença significativa entre os valores médios das idades entre os sobreviventes e falecidos (50 ± 21 vs 59 ± 22 anos, $p = 0,001$). O valor da mediana do SAPS II score foi de 30, com uma diferença significativa entre sobreviventes e falecidos (29 ± 16 vs 50 ± 21 , $p = 0,001$). O tempo mediano de internamento na UCI foi de 2 dias, com uma diferença significativa entre os valores médios dos sobreviventes e falecidos (4 ± 6 vs 8 ± 10 , $p = 0,001$) (Quadro 4). O tipo de admissão mais frequente foi o pós-operatório. A mortalidade intra-hospitalar foi de 2484 doentes (19,6%).

Quadro 4 - Características da população estudada de acordo com o estado vital à data da alta hospitalar.

	População total (n = 13164)	Sobreviventes (n = 10680)	Falecidos (n = 2484)	p
Idade (anos)	59.5 ± 21.6	50 ± 21	59 ± 22	0.001
SAPS score	31.9 ± 18.3	29 ± 16	50 ± 21	0.001
Tempo de internamento (dias)	5.3 ± 9	5 ± 9	8 ± 11	0.000

Valores expressos em média ± desvio padrão.

Em 26,7% dos doentes não ocorreram eventos clínicos adversos. Este subgrupo de doentes teve um valor mediano do SAPS score de 19 (amplitude inter-quartis 9 a 28), e um tempo mediano de internamento de um dia (amplitude inter-quartis 1 a 2 dias).

Registou-se um número total de 294884 eventos clínicos adversos, sendo os mais frequentes os eventos urinários e os em tensão arterial sistólica (34% e 31%, respectivamente (Quadro 5).

Quadro 5 - Frequência total de eventos clínicos adversos

Eventos Clínicos Adversos			
	Eventos	Eventos críticos	Total
Tensão arterial sistólica (BP)	66583 (77%)	23826 (27%)	90409 (31%)
Frequência cardíaca (HR)	50993 (59%)	17582 (20%)	68575 (23%)
Saturação arterial de O ₂ (SaO ₂)	26506 (31%)	8284 (10%)	34790 (12%)
Débito urinário (ml/h) [UR]	72935 (84%)	28175 (32%)	101110 (34%)
Total	217017 (100%)	77867 (100%)	294884 (100%)

Analizou-se a frequência e duração dos eventos clínicos adversos entre as características dos doentes: idade, tipo de admissão, local de origem da admissão para a UCI e categoria diagnóstica. (Quadro 6)

A frequência e a duração dos eventos clínicos adversos variaram significativamente ($p = 0,001$) entre as variáveis categóricas consideradas. Os doentes com mais de 75 anos têm quase duas vezes mais eventos clínicos adversos do que os com menos de 46 anos de idade. Os doentes com o tipo de admissão médica têm maior frequência e

duração de eventos clínicos adversos dos que os admitidos após cirurgia não programada ou programada.

Quadro 6 - Case mix e ocorrência de eventos clínicos adversos (número total de doentes analisados 13164).

Variáveis categóricas	Ev/dia	EvCr/dia	% tempo em EvCr/dia
Idade (anos)			
≤ 46	0 (0 – 1,9)	0 (0 – 0,8)	0 (0 – 2,1)
47 – 65	1 (0 – 2,5)	0,3 (0 – 1)	0,7 (0 – 3,1)
66 – 75	1,2 (0,1 – 3,0)	0,5 (0 – 1,3)	1,2 (0 – 4,2)
≥ 76	1,5 (0,3 – 3,6)	0,6 (0 – 1,5)	1,5 (0 – 4,5)
Tipo de admissão			
Cirúrgica não programada	1 (0 – 2,8)	0,4 (0 – 1,1)	0,8 (0 – 3,6)
Cirúrgica programada	0,9 (0 – 2,5)	0,4 (0 – 1)	0,7 (0 – 2,8)
Médica	1 (0 – 3)	0,5 (0 – 1,3)	1 (0 – 3,9)
Local de origem pré-admissão à UCI			
Bloco operatório	1 (0 – 2,7)	0,5 (0 – 1)	0,9 (0 – 3,3)
Sala de recobro	0,9 (0 – 3)	0,3 (0 – 1,1)	0,6 (0 – 3,1)
Sala de emergência	0,6 (0 – 2,2)	0,2 (0 – 1)	0,3 (0 – 2,5)
Enfermaria	1,3 (0,1 – 3,4)	1,3 (0,1 – 3,4)	1,3 (0,1 – 3,4)
Outra UCI/Hospital	1,2 (0 – 3,4)	0,6 (0,1 – 1,4)	1,4 (0,2 – 4,6)
Categoria diagnóstica : não pós-operatório			
Cardiovascular/vascular	1,2 (0 – 3,3)	0,5 (0 – 1,4)	1 (0 – 4,2)
Respiratório	1,3 (0,3 – 3,2)	0,5 (0 – 1,3)	1,2 (0 – 4)
Gastrointestinal	0,5 (0 – 2,3)	0,3 (0 – 1)	0,4 (0 – 3)
Neurológico	0,9 (0 – 2,4)	0,3 (0 – 1)	0,6 (0 – 3,1)
Sepsis	2 (0,5 – 5)	1 (0,3 – 2,3)	3,6 (0,6 – 10,4)
Trauma	0,5 (0 – 1,7)	0,1 (0 – 0,8)	0,2 (0 – 2,1)
Outras, ex.: Metabólica, Hematológica, Renal	0,8 (0,1 – 2,5)	0,3 (0 – 1,2)	0,8 (0 – 4,6)
Categoria diagnóstica : pós-operatório			
Cardiovascular/vascular	1,5 (0,3 – 3,5)	0,7 (0 – 1,4)	1,6 (0 – 4,2)
Gastrointestinal	1 (0 – 3)	0,5 (0 – 1,2)	1 (0 – 3,7)
Neurológico	0,5 (0 – 1,8)	0,2 (0 – 0,8)	0,3 (0 – 1,6)
Trauma	1,1 (0,3 – 2,9)	0,4 (0 – 1,2)	1 (0 – 3,7)
Outras, ex.: Renal, Ginecológico e Ortopédico	0,6 (0 – 2,2)	0,3 (0 – 0,9)	0,5 (0 – 1,9)

Valores expressos em mediana e amplitude inter-quartis

p = 0.001, dentro de cada uma das cinco variáveis categóricas

Para os doentes sobreviventes, a incidência de eventos/dia foi de $1,6 \pm 2,7$ e de eventos críticos/dia de $0,6 \pm 0,8$. Para a população de falecidos, a incidência de eventos/dia foi de $5,4 \pm 6$, de eventos críticos/dia de $2,1 \pm 1,5$ e o tempo de $9,6 \pm 7,1$ horas (Quadro 7).

Quadro 7 - Frequência e duração dos eventos clínicos adversos

	Sobreviventes	Falecidos	p
Eventos/dia	1,6 ± 2,7	5,4 ± 6	0,001
Eventos críticos/dia	0,6 ± 0,8	2,1 ± 1,5	0,001
Tempo em eventos críticos/dia (horas)	1,8 ± 3,2	9,6 ± 7,1	0.001

Valores expressos em média ± desvio padrão

Foi desenvolvida uma aproximação baseada nos eventos clínicos adversos e na aplicação de modelos gerados por redes neuronais artificiais, materializando um dos principais objectivos desta dissertação: a indução de modelos de previsão de mortalidade intra-hospitalar. Os modelos gerados pelos métodos automáticos são comparados com os resultados obtidos pelo modelo SAPS II e com um outro modelo obtido por regressão logística que inclui as mesmas variáveis que os modelos desenvolvidos pela rede neuronal artificial.

Foi avaliada a importância relativa das variáveis independentes sobre o *outcome* final (mortalidade intra-hospitalar).

No desenvolvimento dos vários modelos foram consideradas as seguintes variáveis:

- o SAPS II score; (Le Gall, Lemeshow et al. 1993), às 24 horas após a admissão à UCI);
- a idade (em anos);
- o tipo de admissão;
- o local de origem da admissão à UCI
- os eventos clínicos adversos.

As dozes variáveis relativas aos eventos clínicos adversos foram transformadas em número médio de eventos e eventos críticos por dia de internamento do doente na UCI, (quadro 3), designadas por:

- **NBP** e **NCRBP** – número de eventos e eventos críticos de tensão arterial sistólica;
- **NHR** e **NCRH** – número de eventos e eventos críticos de frequência cardíaca;
- **NO2** e **NCRO2** - número de eventos e eventos críticos de saturação arterial de O2;
- **NUR** e **NCRUR** - número de eventos e eventos críticos de débito urinário horário.

Para os eventos críticos inclui-se também a duração média diária), designados por:

- **TCRBP** - tempo em eventos críticos de tensão arterial sistólica;

- **TCRHR** – tempo em eventos críticos de frequência cardíaca;
- **TCRO2** – tempo em eventos críticos de tensão arterial de O₂;
- **TCRUR** - tempo em eventos críticos de débito urinário horário.

Devido ao tipo de distribuição da ocorrência de eventos clínicos adversos, as variáveis correspondentes foram submetidas a transformação logarítmica²¹.

Foram induzidos seis modelos baseados em redes neuronais artificiais (ANN) e regressão logística (LR):

- **ANN ALL** – incluindo todas as variáveis, excepto a idade;
- **ANN Case Outcomes** - com todas as variáveis excepto o SAPS II score;
- **ANN Outcomes** - incluindo apenas os eventos clínicos adversos;
- **LR ALL** - incluindo todas as variáveis, excepto a idade;
- **LR Case Outcomes** - com todas as variáveis excepto o SAPS II score;
- **LR Outcomes** - que inclui apenas os eventos clínicos adversos;

Foram ainda considerados dois modelos baseados nos SAPS II:

- **LR SAPS II** - modelo original do SAPS II;
- **LR SAPS II B**: equivalente ao modelo original, excepto que os parâmetros internos da equação logística (β_0 , β_1 e β_2) foram ajustados com base numa sub-população de treino balanceado relativamente ao outcome de interesse.

Para o desenvolvimento dos modelos automáticos utilizou-se uma rede neuronal artificial unidireccional, do tipo *Perceptron* multi-camada. O processo de treino da rede neuronal foi supervisionado, tendo sido aplicado uma variante do algoritmo de retropropagação (o RPROP).

O método de Houldout foi usado para evitar a sobre-especialização da rede neuronal artificial. Os modelos foram treinados e avaliados, ao longo de trinta iterações.

Os resultados dos modelos de previsão (redes neuronais e regressão logística) são apresentados como valores médios da precisão, sensibilidade, especificidade e da capacidade discriminativa (área sob a curva ROC) e respectivos intervalos de confiança a 95%. Para a comparação da capacidade discriminativa entre os modelos obtidos por redes neuronais e os por regressão logística foi utilizado o teste t-Student, aceitando-se como significativo se $p < 0,05$.

Os modelos obtidos com base só nos eventos clínicos adversos têm uma capacidade discriminativa significativamente melhor do que a do SAPS II ($p < 0,05$), com valores

²¹ Transformação de uma variável conseguida através do seu logaritmo. Muitas vezes utilizada quando a distribuição da frequência da variável sugere um distribuição enviesada, tendo como objectivo de conseguir uma distribuição próxima da normalidade Everitt, B. (2006). Medical Statistics from A to Z, Cambridge University Press.

das curvas ROC entre 83,9 e 87,1% (redes neuronais artificiais), 82,6 – 85,2% (regressão logística) e 80% (SAPS score).

Quando utilizadas as mesmas variáveis, os modelos desenvolvidos pelas redes neuronais apresentam melhor desempenho que o por regressão logística (incremento de 1,3 – 2%)

A análise da importância relativa à mortalidade intra-hospitalar das variáveis incluídas nos diferentes modelos revela que, para os modelos por redes neuronais artificiais (que incluem todas as variáveis), o somatório da importância relativa dos eventos clínicos adversos é de 80,2%. Os eventos em débito urinário horário e os em saturação de O₂ são os mais relevantes para a mortalidade hospitalar, com importâncias relativas de 19,2% a 45,5% e 18,6 a 30,7%.

3.2 Mortality assessment in intensive care units via adverse events using artificial neural networks

Apresenta-se o artigo integral publicado:

Mortality assessment in intensive care units via adverse events using artificial neural networks.

Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, José Neves.

Publicado na revista *Artificial Intelligence in Medicine* (2006), Elsevier, 36, 223-234. ISSN: 0933-3657.

Summary

Objective: This work presents a novel approach for the prediction of mortality in intensive care units (ICUs) based on the use of adverse events, which are defined from four bedside alarms, and artificial neural networks (ANNs). This approach is compared with two logistic regression (LR) models: the prognostic model used in most of the European ICUs, based on the simplified acute physiology score (SAPS II), and a LR that uses the same input variables of the ANN model.

Materials and Methods: A large dataset was considered, encompassing forty two ICUs of nine European countries. The recorded features of each patient include the final outcome, the case mix (e.g. age) and the intermediate outcomes, defined as the daily averages of the out of range values of four biometrics (e.g. heart rate). The SAPS II score requires seventeen static variables (e.g. serum sodium), which are collected

within the first day of the patient's admission. A nonlinear least squares method was used to calibrate the LR models while the ANNs are made up of multilayer perceptrons trained by the RPROP algorithm. A total of 13164 adult patients were randomly divided into training (66%) and test (33%) sets. The two methods were evaluated in terms of receiver operator characteristic (ROC) curves.

Results: The event based models predicted the outcome more accurately than the currently used SAPS II model ($P < 0.05$), with ROC areas within the ranges 83.9–87.1% (ANN) and 82.6–85.2% (LR) vs 80% (LR SAPS II). When using the same inputs, the ANNs outperform the LR (improvement of 1.3 – 2%).

Conclusion: Better prognostic models can be achieved by adopting low cost and real-time intermediate outcomes rather than static data.

Keywords: Artificial neural networks; Classification; Data mining; Intensive care; Logistic regression.

1. Introduction

In the last decades, there has been an increasing development in intensive care medicine, where the goal is to provide the best outcome for critically ill patients.

Indeed, a worldwide expansion occurred in the number of intensive care units (ICUs) [1]. Moreover, scoring the severity of illness has become a daily practice, with several metrics available, such as the acute physiology and chronic health evaluation system (APACHE II), the simplified acute physiology score (SAPS II) or mortality probability model (MPM), just to name a few [2].

The intensive care improvement comes with a price, being ICUs responsible for an increasing percentage of the health care budget. Resource availability limitations force them to make sure that intensive care is applied only to those who are likely to benefit from it. Critical decisions include interrupting life-support treatments and writing do-not-resuscitate orders when intensive care is considered futile. Under this context, mortality assessment is a crucial task, being used not only to predict the final clinical outcome but also to evaluate the ICU effectiveness. The prevalent prognostic models are built using a logistic regression over a static score (e.g. SAPS II); i.e., computed with data collected only within the first twenty four hours of the patient's admission. This limits the impact of clinical decision making, since the scores are usually not updated during the patients' length of stay.

On the other hand, the use of data mining in medicine is a rapidly growing field, which aims at discovering some structure in large clinical heterogeneous data [3]. This interest arose due to the rapid emergence of electronic data management methods, holding valuable and complex information. Human experts are limited and may overlook

important details, while automated discovery tools can analyze the raw data and extract high-level information for the decision-maker [4].

The artificial neural networks (ANNs) are one of the most successful data mining techniques, denoting a set of connectionist models inspired by the behaviour of the human brain and presenting useful capabilities for medicine such as nonlinear learning, multi-dimensional mapping and noise tolerance [5]. The interest in ANNs was stimulated by the advent of the back propagation algorithm in 1986. Since then, the number of ANN publications in Medicine has spawned from two in 1990 to five hundred in 1998 [6] and the search term "neural network computer" in the MEDLINE database displays more than two thousand articles from 1999 to 2004.

In the past, there has been work comparing ANNs and logistic regression models for ICU mortality prediction, reporting either better [7] [8] or similar [9] [10] [11] performances.

Yet, in all these studies, the ANNs were trained with the static variables used by the APACHE II score. This work follows an alternative direction, the use of data collected after the first twenty four hours of a patients' admission. A similar approach has been proposed by Kayaalp and his collaborators [12] [13] where they adopted a time series prediction point of view, using twenty three temporal fields, such as the daily sequential organ failure assessment (SOFA) score, which takes time and costs to be obtained. However, in previous work [14], it has been suggested that the SOFA could be replaced by real-time and less costly outcomes, known as events, which are automatically measured as out of range values of four commonly bedside monitored physiological parameters. Hence, this article presents a novel approach for ICU mortality prediction, based on the use of daily intermediate events.

A final remark will be given to the relation between prognostic models and treatments.

A prognostic scoring system should be independent of treatment, providing a means of measuring disease or health status of a patient, where usually a higher score corresponds to greater severity. Different elements contribute to this total score, including physiological variables. Although therapeutic approaches may influence the final patient outcome, their effect will be reflected upon the average number of events per day, the alarms signs. By augmenting a prognostic model with utility assessments of potential outcomes and indicating particular variables for decision support, optimal decisions for a group or individual patients can be determined [15].

The paper is organized as follows: first, the ICU clinical data is presented and the prognostic models are introduced (Section 2); next, a description of the performed experiments is given, being the results analyzed and discussed (Section 3); finally, closing conclusions are drawn (Section 4).

2. Materials and methods

2.1 Clinical data

This work adopted part of data collected during the EURICUS II project [16], which involved forty two ICUs of nine European Union countries, from November/98 to August/99. The patient's data was manually collected and registered by the nursing staff. In every hour, the monitored bedside parameters were introduced into daily patient records. The whole data was gathered at the Health Services Research Unit of the Groningen University Hospital, the Netherlands. The final database presented one entry (or example) per patient. The patients with age lower than eighteen burned or with bypass surgery were discarded. It should be noticed that these last two classes of patients are often not treated in ICUs but in specialized areas (e.g. burn or coronary units). In addition, four entries were removed due to the presence of missing values [17], remaining a total of 13164 records.

The main features of the clinical data are described in Table 1.

Table 1: The attributes of the intensive care data.

Attribute	Description	Domain Values
SAPS II	SAPS II score	{0.1, ..., 163}
age	Patients' age	{18, 19, ..., 100}
admtyp	Admission type	{1, 2, 3} ^a
admfro	Admission origin	{1, 2, ..., 7} ^b
NBP	Average number of blood pressure events	[0.0, ..., 33.0]
NCRBP	Average number of critical blood pressure events	[0.0, ..., 6.0]
NHR	Average number of heart rate events	[0.0, ..., 42.0]
NCRHR	Average number of critical heart rate events	[0.0, ..., 6.0]
NO2	Average number of oxygen events	[0.0, ..., 28.0]
NCRO2	Average number of critical oxygen events	[0.0, ..., 6.0]
NUR	Average number of urine events	[0.0, ..., 38.0]
NCRUR	Average number of critical urine events	[0.0, ..., 8.0]
TCRBP	Average time (min.) of blood pressure critical events	[0.0, ..., 36.0]
TCRHR	Average time (min.) of heart rate critical events	[0.0, ..., 33.0]
TCRO2	Average time (min.) of oxygen critical events	[0.0, ..., 33.0]
TCRUR	Average time (min.) of urine critical events	[0.0, ..., 40.0]
death	The occurrence of death	{0, 1} ^c

^a 1- unscheduled surgery, 2- scheduled surgery, 3-medical.

^b 1- operating theatre, 2- recovery room, 3- emergency room, 4- general ward, 5-other ICU, 6- other hospital, other sources

^c 0- no death, 1- death

The case mix appears in the first four rows, an information that remains unchanged during the patient's admission. The frequency distributions (or histograms) related to these variables are plotted in Figure 1, which also includes the patient's length of stay

and the final outcome. Next, there are twelve variables related to the intermediate outcomes, which are defined from four monitored biometrics: the systolic blood pressure (BP), the heart rate (HR), the arterial oxygen saturation (O2) and the urine output (UR). Finally, the last attribute denotes the patients' final outcome.

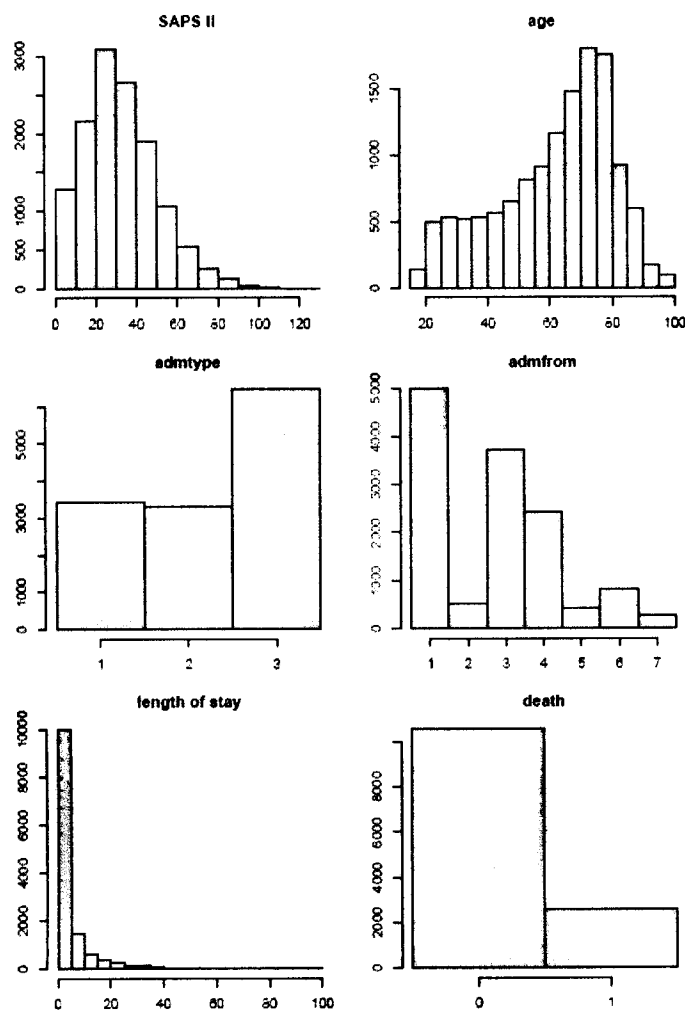


Figure 1: The histograms for the case mix (SAPS II, age, admtype and admfrom) variables, the length of stay and the final outcome (death).

In a UCI, there are several events that may affect the patient's condition (e.g. shock, extemporary extubation or hypoxia). For the selection of an event, it is important that its occurrence and duration can be registered by physiological changes (e.g. shock and not pneumonia). Moreover, such physiological variables should be commonly registered in the ICUs, at regular intervals. These are main reasons for the choice of the four biometrics. A panel of seven EURICUS II experts elaborated protocol that defines the normal ranges for these parameters (see Table 2). When an out of range value occurs for a given interval, an alarm is triggered, defining an event. A critical event is a more serious event and it is classified by a longer event or a more extreme

out of range measurement. Thus, each event or critical event is defined as a binary variable, which will be set to 0 (false), if the physiologic value lies within the advised range; or 1 (true) else, according to the time criterion.

The first eight outcomes of Table 1 denote the number of events/critical events, while the latter ones denote the total time in minutes considering only the critical 3 events. The whole twelve outcomes were analyzed as daily averages using the patients' records, since this allows prediction at early stages of the patients' admission.

During the patients' length of stay in the ICU, intermediate outcomes of the first day can feed into a real-time prognostic model; in next day, an accumulative knowledge is built by averaging current events with the ones obtained from the previous day; and so on.

Table 2: The protocol for the out of range measurements.

	BP	O2	HR	UR
Normal Range	90 – 180mmHg	≥ 90%	60 – 120bpm	≥ 30ml/h
Event ^a	≥ 10min.	≥ 10min.	≥ 10min.	≥ 1h
Event ^b	≥ 10min. in 30min.	≥ 10min. in 30min.	≥ 10min. in 30min.	–
Critical Event ^a	≥ 1h	≥ 1h	≥ 1h	≥ 2h
Critical Event ^b	≥ 1h in 2h	≥ 1h in 2h	≥ 1h in 2h	–
Critical Event ^c	< 60mmHg	< 80%	< 30bpm ∨ > 180bpm	≤ 10ml/h

^a Defined when continuously out of range.

^b Defined when intermittently out of range.

^c Defined anytime.

2.2 Logistic regression

The logistic regression is a statistical method that is commonly used within the Medicine field to model the probability of binary events (e.g. the proportion of patients that respond to a given therapy). It operates a smooth non linear transformation, where the probability of the event k is given by [4]:

$$P_k = \frac{1}{1 + \exp(\beta_0 + \sum_{i=1}^N \beta_i x_i)} \quad (1)$$

where $\beta_0, \dots, \beta_N$ denote the parameters of the model and $x_1, \dots, x_N$ the dependent variables.

On the other hand, the SAPS II is the most widely European score for mortality assessment, ranging from 0 to 163, where a high value indicates a high death probability [18]. The model encompasses a total of seventeen variables (Table 3), such as age, previous health status or diagnosis, which are collected within the first twenty four hours of the patient's admission. Some of these variables are associated to clinical tests (e.g. blood samples) that are expensive.

Table 3: The SAPS II variables (adapted from [18]).

Attribute	Description
age	Patients' age (in years)
heart rate	Worst value in 24 hours (either low or high)
systolic blood pressure	Worst value in 24 hours (either low or high)
body temperature	Highest temperature in degrees
ventilation	Lowest PaO ₂ /FiO ₂ ratio if ventilated
urinary output	Total urinary output in 24 hours
serum urea level	Highest value in mmol/L, g/L or mg/dL
WBC count	Worst WBC count from the scoring sheet
serum potassium level	Worst value in mmol/L
serum sodium level	Worst value in mmol/L
serum bicarbonate level	Lowest value in mEq/L
bilirubin	Highest value in μ mol/L or mg/dL
Glasgow coma score	Lowest value
admission	Unscheduled surgical, scheduled surgical or medical
AIDS	Yes, if HIV-positive with illness
hematologic malignancy	Yes, if lymphoma, acute leukemia or multiple myeloma
metastatic cancer	Yes, if proven by surgery or any other method

The prognostic model, known as predictive death rate (pdrk) is given by the logistic regression:

$$\begin{aligned}
 \text{logit}_k &= B_0 + B_1 \times \text{SAPSII}_k + B_2 \times \ln(\text{SAPSII}_k + 1) \\
 \text{pdr}_k &= \exp(\text{logit}_k) / (1 + \exp(\text{logit}_k))
 \end{aligned}
 \tag{2}$$

where SAPSIIk denotes the score for the patient k, being B0, B1 and B2 internal parameters estimated by calibration procedures. The majority of the European ICUs use the values B0 = -7.7631, B1 = 0.0737 and B2 = 0.9971, which were optimized in the study of Le Gall and his collaborators [18] by applying a Hosmer-Lemeshow goodness-of-fit test to an international database with a total of 13152 examples collected from September/1991 to February/1992. The predicted class (Pk) for the k patient is given by the nearest class value to a decision threshold D within the range [0.0, 1.0]:

$$P_k = \begin{cases} 0, & \text{if } \text{pdr}_k < D \\ 1, & \text{else} \end{cases}
 \tag{3}$$

where the class values Pk = 0 denotes no death and Pk = 1 death.

2.3 Artificial neural networks

The multilayer perceptron is one of the most popular ANN architectures. It consists of a feedforward network, where processing neurons are grouped into layers and connected by weighted links [5]. Supervised learning is achieved by an iterative adjustment of the network connection weights, called the training procedure, in order to minimize an error function, which is computed over a set of training examples (E). Each $p \in E$ example maps an input vector $(x_{p1}, \dots, x_{pN})$ with a vector of target values $(t_{p1}, \dots, t_{pM})$, where N and M denote the number of the input and output neurons. Typically, the sum squared error (SSE) is used as the cost function [19]:

$$SSE = \sum_{p \in E} \sum_{i \in M} (t_i^p - s_i^p)^2 \quad (4)$$

where s_i^p is the ANN output value.

The network is activated by feeding the input layer with the input vector and then propagating the activations in a feed forward fashion, via the weighted connections, through the entire network. For a given input $(x_1, \dots, x_N)$, the state of a neuron (S_i) is computed by:

$$s_i = f(w_{i,0} + \sum_{j \in I} w_{i,j} \times s_j) \quad (5)$$

where I represents the set of nodes reaching node i ; f the activation function, possibly of nonlinear nature; $w_{i,j}$ the weight of the connection between nodes j and i ; and $s_1 = x_1, \dots, s_N = x_N$. The $w_{i,0}$ connections, called bias, work as a constant term, since their input is always 1.0. Bias connections are normally added to multilayer perceptrons, since they increase the network learning flexibility; i.e., the hyperplanes defined by the neurons are not constrained to pass through the origin [20].

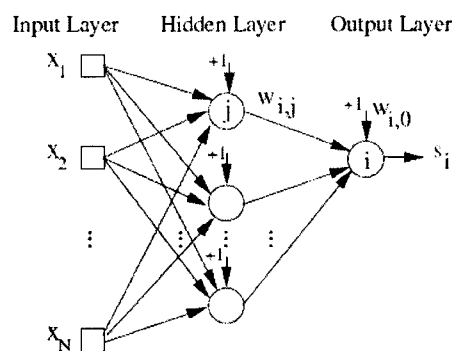


Figure 2: A fully connected network one hidden layer, 1 output node and bias connections.

Multilayer perceptrons with bias connections, one hidden layer with a fixed number of hidden nodes, one output node and logistic activation functions $f(x) = 1/(1+\exp(-x))$ were adopted (Figure 2). The predicted class (P_k) for the k example is given by:

$$P_k = \begin{cases} 0, & \text{if } ANN_k < D \\ 1, & \text{else} \end{cases} \quad (6)$$

where ANN_k denotes the neural output value for the k patient.

Before training the ANNs, the data was preprocessed: the input values were standardized into the range $[-1.0, 1.0]$ and an 1-of-C encoding, which uses one binary variable per each class, was applied to the nominal (non ordered) attributes with few categories (admtyp and admfrom). For example, the admtyp variable is fed to 3 input nodes, according to the scheme: $1 \rightarrow (1.0 \ 1.0 \ 1.0)$; $2 \rightarrow (-1.0 \ 1.0 \ -1.0)$; and $3 \rightarrow (-1.0 \ -1.0 \ 1.0)$.

At the beginning of the training process, the network weights are randomly set within the range $[-1.0, 1.0]$. Then, the resilient propagation (RPROP) algorithm is applied during the training, with the aim of minimizing the SSE. The RPROP is an enhanced version of the back propagation algorithm, performing a local adaptation of the weight-updates based on behaviour of the error function, given by the local gradient information. A full description of the RPROP details can be found in [19]. This algorithm contains two parameters, which were set to the RPROP advised values:

$\Delta 0 = 0.1$ and $\Delta \max = 50.0$. Nevertheless, the choice of these parameters is rather uncritical, since the error convergence is usually insensitive to the $\Delta 0$ and $\Delta \max$ values. Furthermore, when compared with other algorithms, such as the standard back propagation, the RPROP presents a faster training, requiring less computational effort [19] [21]. Finally, the training is stopped when the error slope is approaching zero or after a maximum of 100 epochs [22].

When compared to other data mining techniques, such as decision trees, ANNs often present a higher predictive accuracy [23]. However, in data mining applications, besides obtaining a high predictive performance, it is also important to provide explanatory knowledge; i.e., what has the model learned? With ANNs this can be given after the training process, by measuring the importance of the inputs and extracting a set of rules.

In the past, several approaches have been proposed to measure the importance of inputs, such as the weight-elimination algorithm [24], which is based in the network smallest weights. However, a huge (tiny) input-to-hidden weight does not necessarily mean that the input has a huge (tiny) effect on the output, since the weight effect is

limited by the squashing functions of the hidden neurons [25]. A better alternative is to use sensitivity analysis [26], which has outperformed other input selection techniques, such as forward selection and genetic algorithms. It can be measured as the variance (V_a) produced in the output (y) when the input attribute (a) is moved through its entire range:

$$\begin{aligned} V_a &= \sum_{i=1}^L (y_i - \bar{y}) / (L - 1) \\ R_a &= V_a / \sum_{j=1}^A V_j \end{aligned} \quad (7)$$

where A denotes the number of input attributes and R_a the relative importance of the a attribute. The y_i output is obtained by holding all input variables at their average values. The exception is x_a , which varies through its range with L levels. In this work, L was set to two for the binary attributes ($x_a \in \{-1.0, 1.0\}$) and eleven for the continuous inputs ($x_a \in \{-1.0, -0.8, -0.6, \dots, 0.8, 1.0\}$). The extraction of knowledge from ANNs is still an active research area [23]. Currently, there are two main approaches [27]: the use of decompositional and pedagogical techniques. The former models the ANN at the minimum level of granularity: first, rules are extracted from each individual neuron (hidden and output); then, the subsets of rules are aggregated to form a global relationship. The latter algorithms extract the direct relationships between the inputs and outputs of the 8 ANN. By using a black-box point of view, less computation effort is required and a simpler set of rules may be achieved.

2.4 Statistics

Thirty runs were applied in all experiments. The results are shown in terms of the mean and t-student 95% confidence intervals. The accuracy estimates are achieved using the Holdout method [28]: in each simulation, the available data is divided into two mutually exclusive partitions, using stratified sampling. The training set is used during the modeling phase, while the test set is adopted after training, in order to compute the accuracy estimates.

A common tool for classification analysis is the confusion matrix which matches the test results (predicted) and patient's real condition (actual) values [29]. From the matrix, three measures of performance can be defined [30]: the sensitivity (Sen), also known as recall and type II error; the specificity (Spe), also known as precision and type I error; and the accuracy (Acc), also called correct classification rate, which gives an overall evaluation. The receiver operating characteristic (ROC) curve shows the performance of a two class classifier across the range of possible threshold (D) values, plotting one minus the specificity (x-axis) versus the sensitivity (y-axis) [31]. The global

accuracy is given by the area under the curve ($AUC = \int_0^1 ROCdD$). A random classifier will have an AUC of 0.5, while the ideal value should be close to 1.0.

3 Results

3.1 Training setup

All experiments reported in this work related to the ANNs, including the RPROP algorithm, were conducted using an object oriented programming environment developed in JAVA [32]. On the other hand, the logistic regression models and the statistics, including the Acc, Spe, Sen, AUC and t-student tests; were computed using the R statistical environment [33].

The commonly used 2/3 and 1/3 partitions were adopted for the definition of the training and test set sizes [28]. A model is said to over fit when it correctly handles the training data but fails to generalize. Usually, over fitting is a critical issue in ANN modeling. As pointed out by Sarle [34], the best way to avoid over fitting is to use a large and diverse dataset. Since this is such a case, generalization loss is unlikely to occur, especially if small ANNs are adopted. To accomplish this and also to reduce the computational searching space, the number of hidden nodes was set to round ($N/2$), where N denotes the number of input nodes. It should be stressed that this rule does not constitute the optimal solution. However, in data mining applications, it is often more important to select the training data carefully than decisions regarding learning algorithms [35]. In effect, previous experiments [14] have shown that: the number of hidden nodes has a minor impact, when compared with filtering or feature selection setups; and with the round ($N/2$) setup over fitting is not an important issue.

In the preliminary experiments, all inputs were considered. The exception is the age attribute, which was not included since it is already used by the SAPS II metric. The decision threshold was set to the middle of its range ($D = 0.5$). The table also presents the AUC values, which were computed using a 0.005 grid search for the D value, with a total of 201 decision points. To prevent *tuning*, the adjustment of the models' parameters to the test performances, the training configuration will be improved using only training errors (Table 4).

Table 4: The training set performances (in percentage).

Setup	Acc	Sen	Spe	AUC
Constant Predictor	80.45±0.00	0.00±0.00	100.00±0.00	50.00±0.00
Normal	85.75±0.10	41.78±0.34	96.43±0.08	84.02±0.14
Balanced	78.22±0.20	76.45±0.37	79.99±0.36	86.42±0.19
Log Balanced	79.87±0.19	78.99±0.22	80.76±0.26	88.15±0.19

The constant predictor is defined as the most prevalent outcome. This naive classifier provides an accuracy of 80.45%, corresponding to the proportion of no death cases. The first neural simulation (Normal) presents a better accuracy (85.75%) when compared with the former setup. Yet, from the sensitivity point of view, the first two configurations are poor classifiers. Since there is a higher number of a false condition in the training set, the learning process will tend to favour such cases.

Although popular within the data mining community, the accuracy measure is not sufficient in Medicine, where a test should report both high sensitivity and specificity values [30]. To solve this handicap, a common solution is to balance the training data [17], by using a filtering method that selects an equal number of true and false examples. Therefore, a balanced setup was devised by adopting a under sampling procedure where false (*no death*) examples were randomly deleted from the training set. This setup obtained better results, with a double increase in the sensitivity (34.7%) when compared to the decrease in the specificity (16.4%). Finally, since all outcomes presented high positive skewed distributions, with mean values close to zero, a third configuration was tested (Log Balanced), where the intermediate outcomes suffered the logarithm transform $\ln(x + 1)$. Since this last setup gave the best results, it will be the selected training technique. It should be noted that paired t-tests confirmed statistical differences ($P < 0.05$) between the setups for all measures.

3.2 Test set performances

The results reported on this section were computed over the test sets, which contain the 1/3 samples that were not used in any part of the training process. Each test set also keeps the original proportions of death cases, which is 19.55%. The comparison will be performed between artificial neural network (ANN) and logistic regression (LR) models. In total, eight distinct models will be evaluated (Table 5):

ANN/LR ALL - similar to the Log Balanced setup, where all inputs are used except the age;

ANN/LR Case Outcomes - with all inputs except the SAPS II;

ANN/LR Outcomes - this only uses the intermediate outcomes;

LR SAPS II - the most used European prognostic model [18], defined by Equation 2, with the fixed parameters $B_0 = -7.7631$, $B_1 = 0.0737$ and $B_2 = 0.9971$; and

LR SAPS II B - equivalent to the previous model, except that in each run, the parameters (B_0 , B_1 and B_2) are calibrated to a balanced training.

The LR models (except LR SAPS II) were optimized by a nonlinear least squares algorithm that minimizes the SSE, under the R environment [33].

The results are shown in Table 5, while the correspondent ROCs is plotted in Figure 3.

Paired *t*-tests were also performed, confirming statistical differences ($P < 0.05$) for the

AUC metric between all models, except the pairs: LR SAPS II/LR SAPS II B ($P = 0.27$); ANN Outcomes/LR Case Outcomes ($P = 0.40$); and ANN Case Outcomes/LR All ($P = 0.13$).

Table 5: The test set performances (in percentage).

Setup	Acc	Sen	Spe	AUC
ANN All	79.21±0.24	78.11±0.51	79.48±0.35	87.12±0.21
ANN Case Outcomes	78.22±0.26	75.78±0.66	78.82±0.36	85.52±0.20
ANN Outcomes	77.60±0.31	70.00±0.59	79.45±0.48	83.88±0.23
LR All	75.97±0.29	77.06±0.68	75.71±0.40	85.17±0.27
LR Case Outcomes	77.15±0.26	70.63±0.67	78.73±0.41	83.78±0.24
LR Outcomes	77.57±0.15	65.91±0.62	80.41±0.21	82.59±0.23
LR SAPS II	82.60±0.14	42.57±0.50	92.33±0.12	79.84±0.26
LR SAPS II B	69.37±0.32	77.57±0.64	67.37±0.48	80.04±0.25

First, the last two setups will be analyzed. Due to the balanced training, the second logistic setup presents higher sensitivity values than the first one. Yet, in terms of the ROC analysis, and despite of using quite different learning procedures, the two logistic models are statistically equivalent. Furthermore, both setups are outperformed by all models that use the intermediate events, with a difference ranging from 2.19% to 6.72%. This is an important result, since it backs the main claim of this work: the selected daily events do provide useful information for predicting ICU mortality.

When using the same inputs, the ANNs outperform the LR models, with improvements of 1.95% (**All**), 1.74% (**Case Outcomes**) and 1.29% (**Outcomes**). Regarding the feature selection, the best option, in terms of predictive performance, is to use all attributes. When replacing the SAPS II attribute by the age (setup **Case Outcomes**), there is a decay of 1.60% (ANN) and 1.39% (LR). A similar effect occurs when only the intermediate outcomes are used, with a 1.64% (ANN) and 1.9% (LR) decrease in terms

of the AUC values. It is also important to stress that if the AUC is above 80% then the discriminator is considered excellent [18].

Since the first six setups are above this limit, this means that high quality results were achieved.

3.3 Explanatory knowledge

Table 6 shows the relative importance of the inputs for the event based models when applying the sensitivity analysis procedure (Eq. 7). To improve clarity, the importance of the attributes related to each biometric were summoned into a single value (last four columns). The obtained results are consistent for both ANN and LR models. For the All selection, the event based variables denote a strong influence, summing a total of 80.2% (ANN) and 76.5% (LR).

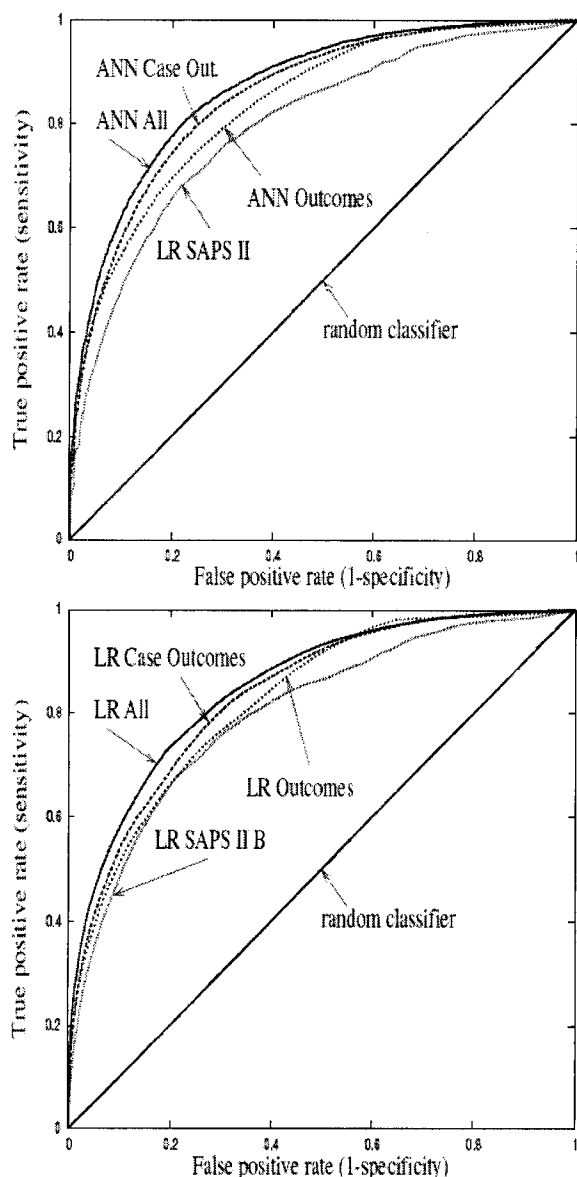


Figure 3: The receiver operating characteristic curves.

Furthermore, the table suggests that the case mix is not needed in this configuration, a scenario that changes when the SAPS II is excluded (setup **Case Outcomes**). When comparing the bedside parameters, the importance order changes from model to model. The oxygen saturation is the most relevant biometric in the first two configurations. The urine output comes in second place for **ANN/LR All** setups and presents the highest influence when only event based data is used.

Table 6: The relative importance of the input variables (in percentage).

Setup	SAPS II	age	admttype	admfrom	BP*	HR*	O2*	UR*
ANN All	16.8	—	1.0	2.0	15.9	14.4	30.7	19.2
ANN Case Outcomes	—	14.3	5.9	11.7	13.1	16.7	22.5	15.8
ANN Outcomes	—	—	—	—	16.9	15.8	21.8	45.5
LR All	19.3	—	2.8	1.4	17.9	16.2	23.1	19.3
LR Case Outcomes	—	19.2	3.6	13.4	15.3	14.0	18.6	15.9
LR Outcomes	—	—	—	—	24.1	18.6	26.5	30.8

* All attributes related to the variable where included: number of events/critical events and the time.

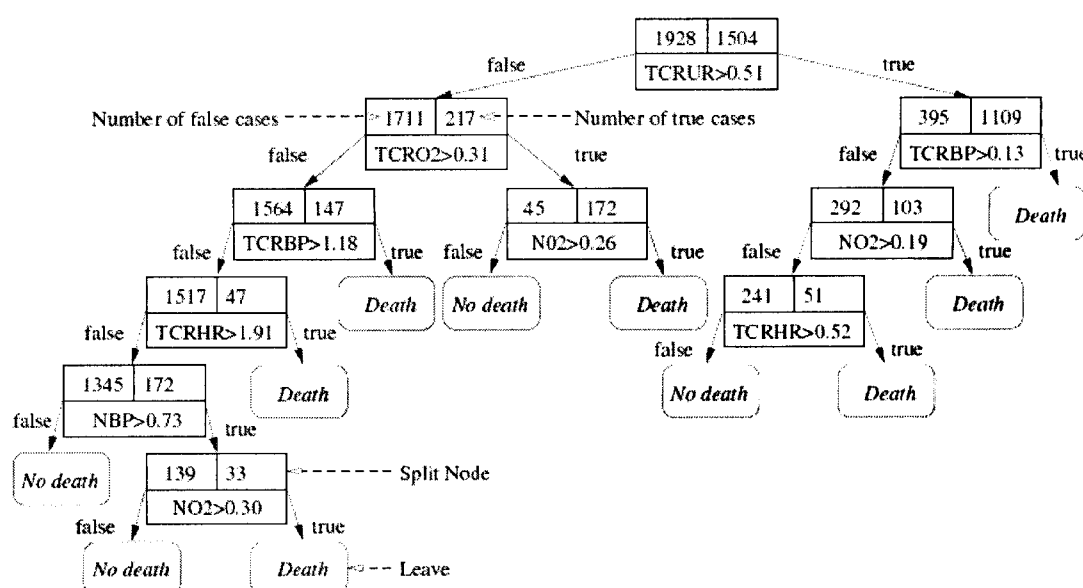


Figure 4: Example of a decision tree extracted from a trained neural network. Variables denote the logarithm transform $\ln(x + 1)$ of the intermediate outcomes from Table 1.

For demonstrative purposes, a simple pedagogical rule extraction technique will be applied to the ANN setup that only includes the intermediate outcomes. The intention is to provide a descriptive model of what a trained ANN has learned. The rules will be presented in terms of a decision tree [36], a branching structure based on 13 split nodes, that test a given feature, and leaves, which assign a class label. First, a new dataset was constructed, by considering the inputs of a given balanced training set, in a total of 3432 cases; and the corresponding ANN output class labels. Then, a decision tree algorithm was applied to this dataset by using the rpart library of the R environment [33]. The obtained decision tree managed to explain the ANN behaviour with an accuracy of 90%. To simplify the visualization, the tree was pruned to a maximum of

ten splits (Figure 4). As an example, the following rule corresponds to the top right path of the tree, explaining a total of 1109 cases:

*if $\ln(TCRUR + 1) > 0.51 \wedge \ln(TCRBP + 1) > 0.13$ then *Death**

It is interesting to notice that the most important attribute is the time of urine critical events (TCRUR), which appears at the root node. At the next levels, the relevant features are the oxygen saturation (TCRO2) and blood pressure (TCRBP), while the heart rate outcome only appears at the fourth level. Hence, the decision tree corroborates the sensitivity analysis procedure for input relevance (third row of Table 6).

4. Conclusion

The surge of data mining techniques, such as artificial neural networks (ANN), has created new exciting possibilities in medicine. In this work, these techniques were applied for the prediction of death in intensive care units (ICUs) by using daily intermediate outcomes, which are defined from four commonly bedside monitored variables. In contrast, the current ICU prognostic models, based on a logistic regression (LR), use data collected only in the first day of the patient's admission.

Both approaches were tested in a large database with a total of 13164 records taken from forty two ICUs of nine European countries. Experiments were drawn to test several training configurations, being the best performances obtained by the use of a balanced training procedure and a logarithm transform on the intermediate outcomes. The test set results clearly favour the event based models, when compared with current ICU models. Furthermore, when using the same input variables, the ANNs presented better performances than the LR methods, with improvements within the range 1.29-1.95%. Moreover, several input selection combinations were also considered, being the best result obtained by the use of all input variables except the patients' age. However, since the SAPS II score requires time and costs to be obtained, it can be replaced by the age, provided that a decrease of 1.39-1.60% in performance is acceptable.

The presented approach aims at improving clinical decision-making by providing the best estimate of the patient's condition. It is not intended for evaluating ICUs, since a bad performing ICU will produce more critical events and consequently the poor performance will not be identified. The results have been analyzed under the receiver operator characteristic (ROC) curves, which plots one minus the specificity versus the sensitivity,

when varying a decision threshold ($D \in [0.0, 1.0]$). Globally, high quality results were achieved, with area under the curve (AUC) values higher than 80%. In a real scenario, the decision threshold could be optimized for distinct purposes. High sensitivity models would allow an early identification of deteriorating patients, while high specificity ones would be essential for interrupting life-support treatments.

Nowadays, the majority of the European ICUs adopted the SAPS II prognostic model, which is based on a logistic regression. Yet, this model was outperformed by all event based models (ANN and LR). Furthermore, the proposed approach (setup **Case Outcomes**) requires fewer variables. The SAPS II index was developed in the last decade and currently there is an intensive research for the development of its successor, the SAPS III [38]. This work shows that a different direction should be pursued: the use of intermediate information rather than static data.

Regarding the comparison between the logistic regression and ANNs, although the latter present a higher accuracy, the former method is easier to interpret.

Yet, it is possible to extract human friendly rules from trained ANNs, as shown in the previous section. Nevertheless, if the physicians are not comfortable with the ANNs, then the LR **Case Outcomes** setup is advised.

This study was based on an *off-line* learning, since the data mining techniques were applied after the conclusion of the EURICUS II project [16]. However, the proposed approach opens room for the development of tools for clinical decision support, which are expected to enhance the physician response and proactiveness.

Although the data was manually collected, it could be acquired with a low cost and in real-time, since the events can be automatically fired from the four monitored variables. However, in a real environment there are several phenomena which may cause bad readings (e.g. spurious heart rates may be caused by loose leads). In such scenarios, data quality and cleansing procedures should be addressed [37] and/or the data could be validated by the ICU staff. It is also interesting to notice that such system could potentially provide more updated predictions (e.g. every hour).

Another important aspect is related to the concept of *lead time*: how long before the patient died was the system able to predict the death? Since the full daily records were not available for this study, this question will be addressed in future work.

Indeed, it is intended to extend this approach to an on-line learning environment, by attaching computer systems with friendly human interfaces into ICUs, with the capability to learn and respond in real-time. This will allow us to obtain, after some time, a

valuable feedback from the physicians, in terms of trust and acceptance of this alternative

solution. Currently, a pilot project, called *INTCare*, is being developed under this perspective at the ICU of the Hospital Geral de Santo António, Oporto, Portugal.

Acknowledgments

We thank FRICE and the BIOMED project BMH4-CT96-0817 for the provision of part of the EURICUS II data and support for this study, which is integrated in a PhD program, developed at Instituto de Ciências Biomédicas Abel Salazar from University of Porto and the Departments of Computer Science/Information Systems from the University of Minho. We also would like to thanks the anonymous reviewers for their helpful comments.

References

- [1] C. Hanson and B. Marshall, Artificial intelligence applications in the intensive care unit, *Crit. Care Med.*, 29(2) (2001) 1–9.
- [2] D. Teres and P. Pekow, Assessment data elements in a severity scoring system (Editorial), *Intensive Care Med.* 26 (2000) 263–264.
- [3] K. Cios and G. Moore, Uniqueness of Medical Data Mining, *Artificial Intelligence in Med.* 26 (2002) 1–24.
- [4] D. Hand, H. Mannila and P. Smyth, *Principles of Data Mining* (MIT Press, Cambridge MA, USA, 2001).
- [5] S. Haykin, *Neural Networks - A Comprehensive Foundation* (Prentice-Hall, New Jersey, USA, 2nd edition, 1999).
- [6] R. Dybowski, Neural Computation in Medicine: Perspectives and Prospects, in: H. Malmgreen, M. Borga and L. Niklasson, Eds, *Artificial Neural Networks in Medicine and Biology* (Springer, Heidelberg, Germany, 2000) 26–36.
- [7] R. Dybowski, P. Weller, R. Chang and V. Gant, Prediction of outcome in critically ill patients using artificial neural network synthesized by genetic algorithm. *Lancet*, 347(9009) (1996) 1146–1150.
- [8] A. Nimgaonkar and S. Sudarshan, Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Intensive Care Med.*, 30 (2004) 248–253.
- [9] G. Doig, K Inman, W. Sibbald, M. Martin and J. Robertson, Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-

learning neural network with multivariate logistic regression. in: C. Safran, ed, proceedings of Annual Symposium on Computer Applications in Medical Care (McGraw-Hill, New York, USA, 1993) 361–365.

[10] L. Wong and J. Young, A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. *Anaesthesia*, 54 (1999) 1048–54.

[11] G. Clermont, D. Angus, S. DiRusso, M Griffin and W. Linde-Zwirble, Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit. Care Med.*, 29(2) (2001) 291–296.

[12] M. Kayaalp, G. Cooper and G. Clermont, Predicting ICU mortality: A comparison of stationary and nonstationary temporal models. In: J. Overhage, ed, *Converging Information, Technology, and Health Care*, proceedings of AMIA Symposium (AMIA, Los Angeles CA, USA, 2000) 418–422.

[13] M. Kayaalp, G. Cooper and G. Clermont, Predicting with Variables Constructed from Temporal Sequences. In: proceedings of the Eight International Workshop on Artificial Intelligence and Statistics, (Key West FL, USA, 2001) 220–225.

[14] Á. Silva, P. Cortez, M. Santos, L. Gomes and J. Neves, Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks, in: I. Seruca, J. Cordeiro, S. Hammoudi and J. Filipe, eds, *Enterprise Information Systems VI* (Springer, Germany, 2005).

[15] A. Abu-Hanna and P. Lucas, Prognostic Models in Medicine: AI and Statistical approaches. *Methods Inf. Med.*, 40 (2001) 1–5.

[16] R. Miranda, The effect of harmonizing and standardising the nursing tasks on Intensive Care Units (ICU) of the European Community, Second European Intensive Care Unit Study (EURICUS-II). Technical Report BMH4-CT96-0817, Health Services Research Unit - Groningen University Hospital, The Netherlands, 1999.

[17] D. Pyle, *Data Preparation for Data Mining* (Morgan Kaufmann, S. Francisco CA, USA, 1999).

[18] J. Le Gall, S. Lemeshow and F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European / North American multicenter study, *JAMA* 270 (1993) 2957–2963.

[19] M. Riedmiller, Supervised Learning in Multilayer Perceptrons – from Backpropagation to Adaptive Learning Techniques, *Computer Standards and Interfaces* (1994).

[20] K. Hornik, Some new results on neural network approximation. *Neural Networks*, (1993) 1069–1072.

- [21] R. Mendes, P. Cortez, M. Rocha and J. Neves, Particle Swarms for Feedforward Neural Network Training, in: proceedings of The 2002 Int. Joint Conference on Neural Networks, 2 (IEEE, Honolulu, Havai, USA, 2002) 1895–1899.
- [22] L. Prechelt, Early Stopping – but when?, in: G. Orr and K. Muller, eds, Neural Networks: Tricks of the trade (Springer, Heidelberg, Germany, 1998).
- [23] R. Setiono, Techniques for Extracting Classification and Regression Rules from Artificial Neural Networks, in: Computational Intelligence: The Experts Speak, D. Fogel and C. Robinson, Eds, (IEEE, Piscataway NJ, USA, 2003) 99–114.
- [24] C. Ennett and. Frize, Weight-elimination neural networks applied to coronary surgery mortality prediction, IEEE Trans. Inf. Technol. Biomed. 7(2) (2003) 86–92.
- [25] N. Cardell, W. Joerding and Y. Li, Why Some Feedforward Networks Cannot Learn Some Polynomials, Neural Computation 6 (1994) 761–766.
- [26] R. Kewley, M. Embrechts and C. Breneman, Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks, IEEE Trans. on Neural Networks 11(2000) 668–679.
- [27] A. Tickle, R. Andrews, M. Golea and J. Diederich, The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks, IEEE Trans. on Neural Networks 9(6) (1998) 1057–1068.
- [28] A. Flexer, Statistical evaluation of neural networks experiments: Minimum requirements and current practice, in: R. Trappl, Ed, proceedings of the 13th European Meeting on Cybernetics and Systems Research 2 (Vienna, Austria, 1996) 1005–1008.
- [29] R. Kohavi and F. Provost, Glossary of Terms, Machine Learning 30 (1998) 271–274.
- [30] D. Essex-Sorlie, Medical Biostatistics & Epidemiology: Examination & Board Review (McGraw-Hill/Appleton & Lange, Norwalk CT, USA, Int edition, 1995).
- [31] J. Hanley J. and B. McNeil, The meaning and use of the area under the Receiver Operating Characteristics (ROC) curve, Radiology 143 (1982) 29–36.
- [32] M. Rocha, P. Cortez and J. Neves, Simultaneous Evolution of Neural Network Topologies and Weights for Classification and Regression, Lecture Notes in Computer Science 2902 (2005) 59–66.
- [33] R Development Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2004).
- [34] W. Sarle, Stopped Training and Other Remedies for Overfitting. In: proceedings of the 27th Symposium on the Interface of Computer Science and Statistics, (Pittsburgh PA, USA, 1995) 352–360.

- [35] N. Lavrac, H. Motoda, T. Fawcett, P. Langley and P. Adriaans, Introduction: Lessons learned from data mining applications and collaborative problem solving. *Machine Learning*, 57(1-2) (2004) 13–32.
- [36] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees* (Wadsworth, Belmont CA, USA, 1984).
- [37] E. Turban, J. Aronson and T. Liang, *Decision Support Systems and Intelligent Systems* (Prentice-Hall, UK, 2004).
- [38] The SAPS III Outcomes Research Group, From the Evaluation of the Individual Patient to the Evaluation of the ICU. Available at <http://www.saps3.org> (2005).

4. Indução de modelos de previsão de falência de órgãos e sistemas

Nesta secção é apresentado o trabalho *Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks* (2006) publicado em I. Seruca, J. Cordeiro, S. Hammoudi and J. Filipe (Eds.), *Enterprise Information Systems VI*, Springer, The Netherlands, pp. 127-134, ISBN: 1-4020-3674-4.

Á. Silva, P. Cortez, M.F. Santos, L. Gomes and J. Neves.

4.1. Preâmbulo

A síndrome de falência múltipla de órgãos continua a estar associada a um mau prognóstico e constitui uma das causas mais frequentes de mortalidade na UCI (Beal and Cerra 1994).

As taxas de mortalidade referidas na literatura variam entre 20% a 100%, de acordo com o número, gravidade, duração, tipo, combinação e definição de falência de órgão (Knaus, Draper et al. 1985; Fagon, Chastre et al. 1993; Hebert, Drummond et al. 1993; Tran, Miguel et al. 1993; Zimmerman, Knaus et al. 1996).

Os vários índices para a avaliação da falência múltipla de órgãos diferem substancialmente. Além disso, as variáveis que descrevem a falência de órgãos e sistemas não são tão específicas quanto deveriam. A maioria destes índices utiliza uma associação de variáveis fisiológicas com parâmetros analíticos (creatinina, bilirrubina).

A motivação subjacente a estes índices foi a necessidade de encontrar uma forma objectiva para descrever a falência de órgãos no doente individualizado. Contudo, algumas questões não estão ainda resolvidas, tais como: os valores analíticos obtidos são realmente os piores desse dia? Será que o intervalo de tempo do registo destes

Índices reflecte de facto um processo dinâmico? O conhecimento acumulado sobre o impacto da disfunção fisiológica e a subsequente falência de órgãos é limitado.

Uma abordagem baseada nos parâmetros clínicos registados hora a hora, ao nível do doente, provavelmente reflectirá melhor a relação entre a disfunção fisiológica e a subsequente falência de órgãos e sistemas. Para além disso, este processo será mais precocemente avaliado dado que não estará dependente de exames complementares. Para adquirir uma melhor compreensão desta questão, torna-se necessário conhecer a relação entre a disfunção fisiológica e a falência de órgãos e sistemas.

A avaliação da morbilidade atempada do doente grave (ex.:disfunção/falência de órgãos e sistemas) através de parâmetros objectivos é crucial para a selecção e implementação duma estratégia terapêutica precoce e adequada, no âmbito da UCI.

Existem vários índices de quantificação de falência de órgãos e sistemas entre os quais o *Sequential Organ Failure Assessment (SOFA score)* (Vincent, de Mendonça et al. 1998).

O *SOFA score* quantifica a disfunção/falência de seis órgãos e sistemas: respiratório, hematológico, hepático, cardiovascular, neurológico e renal, pontuados de 0 - função normal, 1 a 2 - disfunção e 3 a 4 – falência. A sua pontuação varia entre 0 e 24. O *SOFA score* regista os piores valores dos parâmetros definidos, em intervalos regulares de vinte e quatro horas.

O *SOFA score* permite a avaliação sequencial da falência de órgãos e sistemas, contudo não possibilita previsão da falência de órgãos ou sistemas do doente grave durante a estadia na UCI.

Com o presente trabalho pretende-se concretizar um dos objectivos desta dissertação, a indução de modelos de previsão da falência de órgãos e sistemas no doente grave.

Este estudo explora a relação entre os eventos clínicos adversos (frequência diária de eventos e eventos críticos) e o desenvolvimento subsequente da falência de órgãos (definido com *SOFA score* ≥ 3 , por órgão e sistema) no doente admitido na UCI.

Foram analisados os dados de 4864 doentes, que correspondem a um total de 30570 registos diários. Após pré-processamento dos dados foram excluídos 4864 registos referentes ao último dia de internamento de cada doente na UCI e 397 registos por dados omissos ou errados. As características gerais da população estão descritas no Quadro 1 e 2:

Quadro 1 - Características gerais da população

População, (n)	4864
Idade (anos), mediana (AIQ)^a	67 (51 – 77)
SAPS II, mediana (AIQ)^a	31 (21 – 44)
Tempo de internamento na UCI (dias), mediana (IQ-range)	2 (1 – 7)
Tipo de admissão, [n, (%)]	
Cirúrgica – não programada	1185 (24,4)
Cirúrgica – programada	1136 (23,4)
Médica	2543 (52,3%)
Local de origem da admissão, [n, (%)]	
Bloco operatório	1643 (33,8)
Sala de recobro	201 (4,3)
Sala de emergência	1335 (27,4)
Enfermaria	1003 (20,6)
Outra UCI	198 (4,1)
Outro hospital	332 (6,8)
Outros	152 (3,1)
Mortalidade na UCI, [n, (%)]	652 (13,4%)

^a amplitude inter-quartis

Quadro 2 - Características gerais da população de acordo com o estado vital à data da alta hospital

	População total	Sobreviventes	Mortos	p
Idade (anos),	66,2 ± 17,3	63,6 ± 17,5	67,3 ± 16	0,001
SAPS II score	36,4 ± 16,3	34 ± 14,8	50,3 ± 18	0,001
SOFA score (total, máximo)	8 ± 5	6 ± 5	14 ± 5	0,001

Valores expressos como média ± desvio padrão

A frequência dos eventos clínicos adversos foi significativamente ($p = 0,001$) nos doentes falecidos. (Quadro 2)

Quadro 3 - Frequência dos eventos clínicos adversos

	Sobreviventes	Falecidos	p
Eventos/dia	1,9 ± 2,7	5,6 ± 6	0,001
Eventos críticos/dia	0,8 ± 0,5	3,3 ± 2,5	0,001

Valores expressos como média ± desvio padrão

Para o desenvolvimento de modelos de previsão do SOFA score por órgão do dia seguinte aplicaram-se técnicas do *Data Mining*, em concreto modelos de redes neuronais artificiais, usando como variáveis os eventos clínicos adversos, o SAPS II score, o SOFA score por órgão, a idade, o tipo e origem da admissão à UCI.

Foi gerada uma nova variável binária para cada órgão e sistema tendo sido transformados todos os valores do SOFA score por órgão a prever (do dia seguinte) numa variável binária (0 - se pontuação <3, e 1 se pontuação ≥ 3).

Como o valor diário do SOFA score reflecte a pontuação obtida das últimas 24 horas, o primeiro passo consistiu em criar seis novas variáveis, por doente, a saber: o SOFA_{d-1} (como variável binária: 0 ou 1) por órgão e sistemas. Estas novas variáveis foram incorporadas nos registos imediatamente anteriores, tendo sido eliminado os registos do último dia de internamento.

Foram criados de forma automática vinte e quatro modelos, para a previsão do valor de SOFA score por órgão do dia (SOFA_d) a partir dos dados registados nesse dia.

Os vinte quatro modelos resultam do somatório do desenvolvimento de modelos de previsão por órgão a partir de quatro subconjunto de variáveis, e designados por:

A – inclui apenas o valor registado do SOFA score do dia anterior (uma variável);

B – inclui todas as variáveis: idade, SAPS II score, tipo e local de origem de admissão, frequência de eventos clínicos adversos do dia anterior e o SOFA score do dia anterior;

C - inclui todas as variáveis (idade, SAPS II score, tipo e local de origem de admissão e frequência de eventos clínicos adversos do dia anterior;

D – inclui apenas a frequência de eventos clínicos adversos do dia anterior.

Para a modelação foram utilizadas as redes neuronais artificiais unidireccionais do tipo *Perceptron* multi-camada, com aprendizagem supervisionada. Para o processo de treino da rede foi utilizado o algoritmo de retropropagação, na sua variante RPROP. O treino da rede e respectiva validação foram conseguidos através de trinta iterações. Em cada iteração os dados disponíveis foram divididos em duas amostras mutuamente exclusivas, amostra de treino e de validação. Os modelos foram criados após o processo de treino na amostra treino e depois aplicados na amostra de validação, para a obtenção dos resultados.

O desempenho dos modelos (pré e pós balanceamento dos dados) foi avaliado pelos valores médios da sensibilidade, especificidade e precisão (percentagem do número de acertos).

Os resultados do desempenho dos modelos são apresentados nos quadros 5, 6 e 7 do artigo. Neste último, comparam-se os resultados obtidos por outros modelos [*Naive Bayes*, JRIP (indução de regras de decisão)]

Dos resultados obtidos, após balanceamento dos dados, salienta-se que os modelos que não incorporam o SOFA_{d-1} têm menor precisão. Os modelos D (que incluem

apenas a frequência de eventos clínicos adversos do dia anterior) após balanceamento da amostra de treino apresentam uma precisão média de cerca de 70% associada a uma especificidade para previsão de falência de órgão entre 68% (hepático) e 77% para o renal.

Apresenta-se o trabalho em versão integral publicado sob a forma de capítulo de livro:

4.2 Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks

Multiple organ failure diagnosis using adverse events and neural networks

Álvaro Silva, Paulo Cortez, Manuel Santos, Lopes Gomes, José Neves.

In I. Seruca, J. Cordeiro, S. Hammoudi and J. Filipe (Eds.), *Enterprise Information Systems VI*, pp. 127-134, Springer, The Netherlands, 2006. ISBN: 1-4020-3674-4.

Abstract: In the past years, the *Clinical Data Mining* arena has suffered a remarkable development, where *intelligent data analysis* tools, such as *Neural Networks*, have been successfully applied in the design of medical systems. In this work, *Neural Networks* are applied to the prediction of organ dysfunction in *Intensive Care Units*. The novelty of this approach comes from the use of *adverse events*, which are triggered from four bedside alarms, being achieved an overall predictive accuracy of 70%.

Key words: Intensive Care Medicine, Classification, Clinical Data Mining, Multilayer Perceptrons.

1 Introduction

Scoring the severity of illness has become a daily routine practice in *Intensive Care Units (ICUs)*, with several metrics available, such as the *Acute Physiology and Chronic Health Evaluation System (APACHE II)* or the *Acute Physiology Score (SAPS II)*, just to name a few (Teres and Pekow, 2000). Yet, most of these prognostic models (given by *Logistic Regression*) are static, being computed with data collected within the first 24 hours of a patient's admission. This will produce a limited impact in clinical decision making, since there is a lack of accuracy of the patient's condition, with no intermediate information being used.

On the other hand, the *Clinical Data Mining* is a rapidly growing field, which aims at discovering patterns in large clinical heterogeneous data (Cios and Moore, 2002). In particular, an increasing attention has been set of the use of *Neural Networks* (connectionist models that mimic the central nervous system) in *Medicine*, with the

number of publications growing from two in 1990 to five hundred in 1998 (Dybowski, 2000).

The interest in *Data Mining* arose due to the rapid emergence of electronic data management methods, holding valuable and complex information (Hand et al., 2001). However, human experts are limited and may overlook important details, while techniques such as *neural networks* have the potential to solve some of these hurdles, due to capabilities such as nonlinear learning, multi-dimensional mapping and noise tolerance (Bishop, 1995).

In *ICUs*, *organ failure diagnosis* in real time is a critical task. Its rapid detection may allow physicians to respond quickly with therapy. Moreover, multiple organ dysfunctions will highly increase the probability of the patient's death.

The usual approach to detect organ failure is based in the *Sequential Organ Failure Assessment (SOFA)*, a diary index, ranging from 0 to 4, where an organ is considered to fail when its *SOFA* score is equal or higher than 3 (Vincent et al., 1996). However, this index takes some effort to be obtained (in terms of time and costs).

This work is motivated by the success of previous applications of *Data Mining* techniques in *ICUs*, such as *predicting hospital mortality* (Santos et al., 2002). The aim is to study the use of *Neural Networks* for organ failure prediction (identified by high *SOFA* values) of six systems: *respiratory*, *coagulation*, *hepatic*, *cardiovascular*, *central nervous* and *renal*. Several approaches will be tested, using different *feature selection*, *training* and *modelling* configurations. A particular focus will be given to the use of daily intermediate adverse events, which can be automatically obtained from four hourly bedside measurements, with fewer costs when compared to the *SOFA* score.

The paper is organized as follows: first, a the *ICU* clinical data is presented, being transformed into a classification task; then, the neural models for organ failure diagnosis are introduced; next, a description of the different experiments performed is given, being the results analyzed and discussed; finally, closing conclusions are drawn.

2. Materials and Methods

2.1 Clinical Data

In this work, a part of the *EURICUS II* database (www.frice.nl) was adopted which contains data related to 5355 patients from 42 *ICUs* and 9 European Union countries, during a period of 4 months, from 1998 to 1999. The database has one *entry* (or *example*) per each day (with a total of 30570), being its main features described in Table 1:

_ the first six rows denote the *SOFA* values (one for each organ) of the patient's condition in the previous day. In terms of notation, these will be denoted by $SOFA_{d-1}$, where d represents the current day;

_ the *case mix* appears in the next four rows, an information that remains unchanged during the patient's internment, containing: the *admission type* (1 - Unscheduled surgery, 2 - Scheduled surgery, 3 - Medical), the *admission origin* (1 - Operating theatre, 2 - Recovery room, 3 - Emergency room, 4 - General ward, 5 - Other ICU, 6 - Other hospital, 7 - Other sources), the *SAPSII* score and the patient's age; figure 1 shows the frequency distributions of these attributes;

Finally, the last four rows denote the intermediate outcomes, which are triggered from four monitored biometrics: the *systolic Blood Pressure (BP)*, the *Heart Rate (HR)*, the *Oxygen saturation (O2)* and the *URine Output (UR)*. A panel of *EURICUS II* experts defined the normal ranges for these four variables (Tables 2 and 3). Each *event* (or *critical event*) is defined as a binary variable, which will be set to 0 (false), if the physiologic value lies within the advised range; or 1 (true) else, according to the time criterion.

Before attempting modelling, the data was pre-processed, in order to set the desired classification outputs.

First, six new attributes were created, by sliding the $SOFA_{d-1}$ values into each previous example, since the intention is to predict the patient's condition ($SOFA_d$) with the available data at day d ($SOFA_{d-1}$ *case mix* and adverse events). Then, the last day of the patient's admission entries were discarded (remaining a total of 25309), since in this cases, no $SOFA_d$ information is available. Finally, the new attributes were transformed into binary variables, according to the expression:

$$\begin{aligned} &0 ; \text{if } SOFA_d < 3 \text{ (false, no organ failure)} \\ &1 ; \text{else (true, organ dysfunction)} \end{aligned} \quad (1)$$

2.2 Neural Networks

In *MultiLayer Perceptrons*, one of the most popular *Neural Network* architectures, *neurons* are grouped into *layers* and only *forward connections* exist (Bishop, 1995). Supervised learning is achieved by an iterative adjustment of the network *connection weights* (the *training procedure*), in order to minimize an error function, computed over the training examples (*cases*).

The state of a neuron (s_i) is given by: (Haykin, 1999):

$$s_i = f(w_{i,0} + \sum_{j \in I} w_{i,j} \times s_j) \quad (2)$$

where I represents the set of nodes reaching node i , f activation function (possibly of nonlinear nature), $w_{i,j}$ the weight of the connection between nodes j and i (when $j = 0$, it is called *bias*); and $s_1 = x_1, \dots, s_n = x_n$, being $x_1, \dots, x_n$ the input vector values for a network with n inputs.

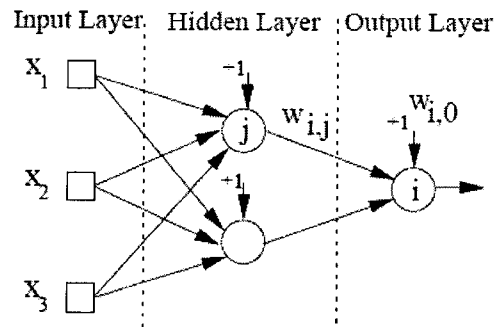


Figure 2: A fully connected network with 3 inputs, 2 hidden nodes, 1 output and *bias* connections.

All experiments reported in this work will be conducted using a neural network object oriented programming environment, developed in *JAVA*. Fully connected *Multilayer Perceptrons* with *bias* connections, one hidden layer (with a fixed number of hidden

$$(f(x) = \frac{1}{1+e^{-x}})$$

nodes) and logistic activation functions were adopted for the organ failure classification (Figure 2). Only one output node is used, since each organ system will be modeled by a different network. This splitting is expected to facilitate the *Neural Network* learning process. Therefore, nearest class value:

$$P_k = \begin{cases} 0, & \text{if } s_{k,o} < 0.50 \\ 1, & \text{else} \end{cases} \quad (3)$$

where $s_{k,o}$ denotes the output value for the o output node and the k input example.

Before feeding the *Neural Networks*, the data was preprocessed: the input values were standardized into the range $[-1, 1]$ and a *1-of-C* encoding (one binary variable per class) was applied to the nominal attributes (non ordered) with few categories (*SOFA_{d-1}*, *admtype* and *admfrom*). For example, the *admtype* variable is fed into 3 input nodes, according to the scheme:

$$\begin{array}{lcl} 1 & \rightarrow & -1 \quad -1 \quad 1 \\ 2 & \rightarrow & -1 \quad 1 \quad -1 \\ 3 & \rightarrow & -1 \quad -1 \quad 1 \end{array}$$

At the beginning of the training process, the network weights are randomly set within thrange $[-1,1]$.

Then, the *RPROP* algorithm (Riedmiller, 1994) is applied, due to its faster convergence and stability, being stopped when the training error slope is approaching zero or after a maximum of E epochs (Prechelt, 1998).

2.2.1 Statistics

To insure statistical significance, 30 runs were applied in all tests, being the accuracy estimates achieved using the *Holdout* method (Flexer, 1996). In each simulation, the available data is divided into two mutually exclusive partitions, using stratified sampling: the *training set*, used during the modeling phase; and the *test set*, being used after training, in order to compute the accuracy estimates. A common tool for classification analysis is the *confusion matrix* (Kohavi and Provost, 1998), a matrix of size $L \times L$, where L denotes the number of possible classes (*domain*). This matrix is created by matching the *predicted (test result)* and *actual (patients real condition)* values. When $L = 2$ and there are four possibilities (Table 4): the number of *True Negative (TN)*, *False Positive (FP)*, *False Negative (FN)* and *True Positive (TP)* classifications.

Table 4: The 2×2 confusion matrix.

↓ actual \ predicted →	negative	positive
negative	<i>TN</i>	<i>FP</i>
positive	<i>FN</i>	<i>TP</i>

From the matrix, three accuracy measures can be defined (Essex-Sorlie, 1995): the *Sensitivity* (also known as *recall* and *Type II Error*); the *Specificity* (also known as *precision* and *Type I Error*); and the *Accuracy*, which gives an overall evaluation. These metrics can be computed using the following equations:

$$\begin{aligned}
 \text{Sensitivity} &= \frac{TP}{FN+TP} \times 100 (\%) \\
 \text{Specificity} &= \frac{TN}{TN+FP} \times 100 (\%) \\
 \text{Accuracy} &= \frac{TN+TP}{TN+FP+FN+TP} \times 100 (\%)
 \end{aligned}
 \tag{4}$$

3. Results

3.1 Feature Selection

Four different feature selection configurations will be tested, in order to evaluate the input attribute importance:

A - which uses only the $SOFA_{d-1}$ values (1 variable).

B - where all available input information is used ($SOFA_{d-1}$ of the corresponding organ system, the *case mix* and the *adverse events*, in a total of 9 attributes);

C - in this case, the $SOFA_{d-1}$ is omitted (8 variables); and

D - which uses only the four *adverse outcomes*.

Since the *SOFA* score takes costs and time to obtain, in this study, a special attention will be given to the last two settings.

In the initial experiments, it was considered more important to approach feature selection than model selection. Due to time constrains, the number of hidden nodes was set to round ($N/2$), where N denotes the number of input nodes ($N = 5$, $N = 21$, $N = 16$ and $N = 4$, for the **A**, **B**, **C** and **D** setups); and $round(x)$ gives nearest integer to the x value.

The commonly used 2=3 and 1=3 partitions were adopted for the *training* and *test* sets (Flexer, 1996), while the maximum number of training epochs was set to $E = 100$. Each input configuration was tested for all organ systems, being the accuracy measures given in terms of the mean of thirty runs (Table 5).

The **A** selection manages to achieve a high performance, with an *Accuracy* ranging from 86% to 97%, even surpassing the **B** configuration. This is not surprising, since it is a well established fact that the *SOFA* is a adequate score for organ dysfunction.

Therefore, the results suggest that there is a high correlation between $SOFA_{d-1}$ and $SOFA_d$.

When the *SOFA* index is omitted (**C** and **D**), the *Accuracy* values only decay slightly. However, this measure (which is popular within *Data Mining* community) is not sufficient in *Medicine*. Ideally, a test should report both high *Sensitivity* and *Specificity* values, which suggest a high level of confidence (Essex- Sorlie, 1995). In fact, there seems to be a tradeoff between these two characteristics, since when the *SOFA* values are not present (Table 5), the *Sensitivity* values suffer a huge loss, while the *Specificity* values increase.

3.2 Balanced Training

Why do the **A/B** selections lead to high *Accuracy / Specificity* values and low *Sensitivity* ones? The answer may be due to the biased nature of the organ dysfunction distributions; i.e., there is a much higher number of *false* (0) than *true* (1) conditions (Figure3).

One solution to solve this handicap, is to *balance* the training data; i.e., to use an equal number of true and false learning examples. Therefore, another set of experiments was

devised (Table 6), using random sampling training sets, which contained 2=3 of the true examples, plus an equal number of false examples. The test set was composed of the other 1=3 positive entries. In order to achieve a fair comparison with the previous results, the negative test examples were randomly selected from the remaining ones, with a distribution identical to the one found in the original dataset (as given by Figure 3). The obtained results show a clear improvement in the *Sensitivity* values, specially for the **C** configuration, stressing the importance of the *case mix* attributes.

Yet, the overall results are still far from the ones given by the **A** selection.

3.3 Improving Learning

Until now, the main focus was over selecting the correct training data. Since the obtained results are still not satisfactory, the attention will move towards better *Neural Network* modelling. This will be achieved by changing two parameters: the *number of hidden nodes* and the *maximum number of training epochs*.

Due to computational power restrictions, these factors were kept fixed in the previous experiments. However, the adoption of *balanced training* leads to a considerable reduction of the number of training cases, thus reducing the required training time.

Several experimental trials were conducted, using different combinations of hidden nodes ($H = 4, 8, 16$ and 32) and maximum number of epochs ($E = 100, 500$ and 1000), being selected the configuration which gave the lowest training errors ($H = 16$ and $E = 1000$). These setup lead to better results for all organ systems and accuracy measures (Table 6). To evaluate the obtained results, a comparison with other *Machine Learning* classifiers was performed (Table 7), using two classical methods from the *WEKA Machine Learning* software (Witten and Frank, 2000): *Naive Bayes* - a statistical algorithm based on probability estimation; and *JRIP* - a learner based on "IF-THEN" rules.

Although presenting a better *Accuracy*, the *Naïve Bayes* tends to emphasize the *Specificity* values, giving poor *Sensitivity* results. A better behavior is given by the *JRIP* method, with similar *Sensitivity* and *Specificity* values. Nevertheless, the *Neural Networks* still exhibit the best overall performances.

4 Conclusions

The surge of novel bio-inspired tools, such as *Neural Networks*, has created new exciting possibilities for the field of *Clinical Data Mining*. In this work, these techniques were applied for *organ failure diagnosis* of *ICU* patients.

Preliminary experiments were drawn to test several feature selection configurations, being the best results obtained by the solely use of the *SOFA* value, measured in the previous day. However, this score takes much more time and costs to be obtained, when compared with the physiologic adverse events. Therefore, another set of experiments were conducted, in order to improve the use of the latter outcomes. First, the training sets were balanced to contain similar proportions of positive and negative examples. Then, the number of hidden nodes and training epochs was increased. As the result of these changes, an improved performance was gained, especially in terms of *sensitivity*. A final comparison between the *SOFA* score and the proposed solution (the **C** improved setup), still favours the former, although the *Sensitivity* values are close (being even higher for the **C** configuration in the *coagulation* and *liver systems*). Nevertheless, it is important to stress the main goal of this work: to show that is it possible to diagnose organ failure by using cheap and fast intermediate outcomes (within our knowledge this is done for the first time). The results so far obtained give an overall accuracy of 70%, which although not authoritative, still back this claim. In addition, the proposed approach opens room for the development of automatic tools for clinical decision support, which are expected to enhance the physician response. In future research it is intended to improve the performances, by exploring different *Neural Network* types, such as *Radial Basis Functions* (Bishop, 1995).

Another interesting direction is based in the use of alternative *Neural Network* training algorithms, which can optimize other learning functions, such as *Evolutionary Algorithms* (Rocha et al., 2003), since the gradient-based methods (e.g., the *RPROP* (Riedmiller, 1994)) work by minimizing the *Sum Squared Error*, a target which does not necessarily correspond to maximizing the *Sensitivity* and *Specificity* rates. Finally, it is intended to enlarge the experiments to other *ICU* applications (e.g., predicting *life expectancy*).

REFERENCES

- Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.
- Cios, K. and Moore, G. (2002). Uniqueness of Medical Data Mining. *Artificial Intelligence in Medicine*, 26(1- 2):1–24.
- Dybowski, R. (2000). Neural Computation in Medicine: Perspectives and Prospects. In et al., H. M., editor, *Proceedings of the ANNIMAB-1 Conference (Artificial Neural Networks in Medicine and Biology)*, pages 26–36. Springer.
- Essex-Sorlie, D. (1995). *Medical Biostatistics & Epidemiology: Examination & Board Review*. McGraw-Hill/Appleton & Lange, International edition.

- Flexer, A. (1996). Statistical evaluation of neural networks experiments: Minimum requirements and current practice. In *Proceedings of the 13th European Meeting on Cybernetics and Systems Research*, volume 2, pages 1005–1008, Vienna, Austria.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press, Cambridge, MA.
- Haykin, S. (1999). *Neural Networks - A Comprehensive Foundation*. Prentice-Hall, New Jersey, 2nd edition. 7
- Kohavi, R. and Provost, F. (1998). Glossary of Terms. *Machine Learning*, 30(2/3):271–274.
- Miranda, R. (2003). *Foundation for Research on Intensive Care in Europe*. Available from <http://www.frice.nl/>.
- Prechelt, L. (1998). *Early Stopping – but when?* In: *Neural Networks: Tricks of the trade*, Springer Verlag, Heidelberg.
- Riedmiller, M. (1994). Supervised Learning in Multilayer Perceptrons - from Backpropagation to Adaptive Learning Techniques. *Computer Standards and Interfaces*, 16.
- Rocha, M., Cortez, P., and Neves, J. (2003). Evolutionary Neural Network Learning. In Pires, F. and Abreu, S., editors, *Progress in Artificial Intelligence, EPIA 2003 Proceedings, LNAI 2902*, pages 24–28, Beja, Portugal. Springer.
- Santos, M., Neves, J., Abelha, A., Silva, Á., and Rua, F. (2002). Augmented Data Mining Over Clinical Databases Using Learning Classifier System. In *Proceedings of the 4th Int. Conf. on Enterprise Information Systems - ICEIS 2002*, pages 512–516, Ciudad Real, Spain.
- Teres, D. and Pekow, P. (2000). Assessment data elements in a severity scoring system (Editorial). *Intensive Care Med*, 26:263–264.
- Vincent, J., Moreno, R., Takala, J., Willatss, S., Mendonça, A. D., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction / failure. *Intensive Care Med*, 22:707–710.
- Witten, I. and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA. 8

Na próxima secção é apresentado o trabalho ***Rating organ failure via adverse events using data mining in intensive care medicine*** submetido à revista Artificial intelligence in Medicine em 2 de Maio de 2007.

Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, José Neves

4.3 Preâmbulo

Os eventos clínicos adversos (*outcomes* intermédios) são expressão do grau de disfunção fisiológica e da evolução da gravidade da doença do doente admitido na UCI. Para além da função de alarme, a sua integração na compreensão do processo fisiopatológico do desenvolvimento da falência de órgãos ainda não foi cabalmente estudada.

A função alarme dos monitores de cabeceira constitui um princípio básico da Medicina Intensiva, enquanto elemento indicador de disfunção fisiológica a que se associa uma intervenção e uma análise do processo fisiopatológico desencadeador subjacente. A demonstração quantitativa e qualitativa destas relações não lineares e multivariada não é uma tarefa fácil.

Os resultados obtidos do estudo *Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks* a partir de modelos induzidos automaticamente para a previsão diária da falência de órgão e sistemas, incorporando os eventos clínicos adversos revelaram valores de precisão interessantes. Este facto estimulou o aprofundamento do tema em questão: a previsão da falência de órgãos através de diferentes técnicas de modelação.

O objectivo deste trabalho foi a previsão para o dia seguinte do estado funcional dos órgãos e sistemas: normal, disfunção e falência, ao longo do internamento na UCI. Como objectivo secundário, surgiu a comparação do desempenho dos modelos obtidos por redes neuronais artificiais, árvores de decisão e regressão logística.

Para o presente estudo foram incluídos 4864 doentes, a que corresponde um total de 30570 registos diários, cujas características foram descritas na secção anterior.

Foram consideradas como variáveis independentes a idade, o SAPS II score, o tipo e local de origem de admissão à UCI, bem como os eventos clínicos adversos

A variável dependente, SOFA score por órgão do dia seguinte foi transformada em três variáveis binárias, indicando função normal (SOFA score = 0), disfunção (SOFA score = 1 ou 2) e falência (SOFA score 3 ou 4).

Para a previsão da função de cada órgão ou sistema foram desenvolvidos seis modelos para cada um dos três *outcomes* de interesse através de redes neuronais artificiais, árvores de decisão e regressão logística, isto é, foram testados cinquenta e quatro modelos diferentes de previsão.

A aplicação informática utilizada foi o R, com minimização do quadrado do erro.

Os resultados de cada modelo são apresentados como um valor entre 0 e 1, interpretáveis como a probabilidade de ocorrência do *outcome* de interesse.

A comparação da capacidade discriminativa dos modelos foi feita através do cálculo da área sob a curva ROC.

Os valores apresentados no quadro 4 do artigo, são os valores médios das áreas sob a curva ROC e respectivos intervalos de confiança a 95%, obtidos após trinta iterações do processo de treino e avaliação. A análise da significância estatística destes valores médios foi feita através do teste t-Student aceitando-se como significativo se $p < 0,05$.

No quadro 4 do estudo estão ilustrados os resultados para os três *outcomes* de interesse referidos (normalidade, disfunção e falência).

Os modelos obtidos pelas árvores de decisão evidenciaram uma menor capacidade discriminativa (61,4; 54,92 e 54,95% área sob a curva ROC) que os desenvolvidos por regressão logística ou por redes neuronais artificiais. Os melhores resultados foram obtidos pelas redes neuronais artificiais (70,29; 64,44 e 74,50% área sob a curva ROC) para os três *outcomes* de interesse (normal, disfunção ou falência de órgão).

Os três modelos gerados para previsão de disfunção de órgão apresentam capacidades discriminativas inferiores a 70% avaliadas pela área sob a curva ROC.

Para a previsão de falência, as redes neuronais e a regressão logística apresentam valores médios da área sob a curva ROC superiores a 70% (no caso das redes neuronais, o valor médio é de 74,5%).

As diferenças entre as capacidades discriminativas dos modelos desenvolvidos são significativas ($p < 0,05$), excepto para a previsão de falência do sistema hematológico.

A importância relativa de cada variável independente, atribuída pelo melhor modelo de previsão de falência de cada órgão ou sistema é apresentada no quadro 5. O peso do conjunto das variáveis relacionadas com eventos clínicos adversos varia entre 50% para o sistema nervoso central e 99% para o hepático.

A apresentação simbólica através de uma árvore de decisão dos melhores modelos (redes neuronais para todos os órgãos, excepto a coagulação) obtidos para previsão de falência de cada órgão é apresentada na figura 4.

As árvores de decisão apresentadas devem ser interpretadas como exemplo ilustrativo do “funcionamento” das redes neuronais.

4.4 Rating organ failure via adverse events using data mining in intensive care medicine

Apresenta-se o trabalho integral submetido à revista Artificial Intelligence in Medicine:

Rating organ failure via adverse events using data mining in intensive care medicine

Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, José Neves

Abstract

Objective: The sequential organ failure assessment (SOFA) is an expert-driven score that is widely used in European intensive care units (ICUs) to access the degree of organ failure. Since some SOFA variables require costs and time to be obtained (due to clinical tests), an alternative data-driven score is proposed in this study based in the use of adverse events, which are defined as the out of range measurements of four routinely collected physiological variables.

Materials and Methods: An analysis of a large European prospectively collected database was considered, with a total of 30570 daily records taken from 4864 patients and involving forty two European ICUs. The input variables include the case mix (i.e. age, diagnosis, admission type and admission from) and adverse events defined from four biometrics (i.e. systolic blood pressure, heart rate, pulse oximeter oxygen saturation and urine output). The output target is the organ status (i.e. normal, dysfunction or failure) of six organ systems (respiration, coagulation, liver, cardiovascular, neurological and renal), as measured by the SOFA score. Three different data mining (DM) methods were tested: decision trees (DTs), logistic regression (LR) and neural networks (NNs). These methods were programmed in the R statistical environment and optimised using least squares algorithms. The original dataset was randomly divided into training (66%) and test (33%) sets and thirty runs were applied to each model. The receiver operator characteristic (ROC) curve was used for accuracy assessment (discriminative power). Finally, input variable relevance and rule extraction procedures were used to get descriptive knowledge from the best predictive models.

Results: In general, the best performance was obtained by the NNs, with improvements of 1.3-2.0% over the LR (second best method). The global predicted ROC areas for the normal, dysfunction and failure conditions were respectively 70.3%, 64.4% and 74.5% (78.0% for the neurological system). Regarding the failure status, the adverse events denoted a high importance, which ranged from 50% (neurological) to 99% (liver).

Conclusion: By exploring data mining techniques and event variables, a new practical daily bedside score can be developed for early organ failure detection. This approach constitutes an alternative to other subjective scores.

Keywords: Adverse event; Artificial neural networks; Critical care; Data mining; Decision trees; Logistic regression; Organ failure assessment.

Keywords: Adverse event; Artificial neural networks; Critical care; Data mining; Decision trees; Logistic regression; Organ failure assessment.

1. Introduction

The severity of patients illness and organ dysfunction cannot be easily accessed by physicians, thus since the early 1980s clinical scores have been developed to rate conditions more objectively in the intensive care unit (ICU) setting [1]. Indeed, in the context of intensive medicine, scores are very useful to [2]: i) improve the quality of intensive care; ii) provide a risk stratification for patient selection in clinical trials; iii) help physicians and families in end-of-life decisions; and iv) guide local planning of resources.

The majority of these scores use are static, since they use data collected only on the first ICU day, such as the acute physiology and chronic health evaluation system (APACHE), the simplified acute physiology score (SAPS) or mortality probability model (MPM) [3]. Yet, these static scores fail to recognize several factors that can influence the patient outcome after the first 24 hours (e.g. the therapeutics strategy and the patients' response).

More recently, dynamic (or sequential) scores have been designed, where the data and scores are updated on a daily basis. The most used scores include [4]: the sequential organ failure assessment (SOFA), multiple organs dysfunction score (MODS) and logistic organ dysfunction (LOD). Our focus is on the SOFA score that was first proposed to evaluate morbidity (degree of organ failure) [5] and latter it has been shown to be related with mortality [3].

The SOFA scores six organ systems (respiratory, coagulation, hepatic, cardiovascular, neurological and renal) on a scale ranging from 0 to 4, according to the degree of failure. This is a subjective score, in the sense that it was developed by a panel of experts who choose a set of variables and rules based on their personal opinions [4].

Although the SOFA is widely used in European ICUs, there are some issues regarding the collection of its 12 input variables (Table 3). For instance, the measurements of platelets, bilirubin and creatinine levels are time and resources consuming. Also, for

some variables, the comparison is made using the worst value obtained in the last 24 hours and it is not clear how many daily times those variables should be evaluated.

Finally, the Glasgow coma score value is an imprecise and subjective score. [6]. A potential alternative is to use adverse events which are more precise, routinely registered or can be automatically collected as the out of range measurements of four commonly monitored ICU biometrics (i.e. systolic blood pressure, heart rate, pulse oximeter oxygen saturation and urine output). In previous work [7], it has been shown that these events can be successfully used to predict mortality. Although therapeutic strategies may influence the patient outcome, such information will not be directly used in this work, since any therapeutic effect will be reflected upon the number and severity of events.

One of the most promising recent developments in intensive care consists in the use of artificial intelligence/data mining techniques [1] [2]. The fast growing amount of data collected had led to vast and complex databases that exceeded the human capability for comprehension without using computational resources. The goal of data mining (DM) is to discover interesting knowledge from the raw data by using automatic discovery tools [8].

There are several DM techniques, each one with its own purposes and advantages. The majority of the severity scores use statistical methods such as the logistic regression (LR), which is easy to interpret. Yet, such classical statistics may not be suitable for the complex non-linear relationships often found in biomedical data [1]. Neural networks (NNs) are connectionist models inspired by the behaviour of the human brain [9]. In ICUs, NNs are the most popular artificial intelligence method, due to advantages of non-linear learning and high flexibility. Indeed, NNs have been applied to predict mortality and length of stay [1] [2]. Decision Trees (DTs) are another useful DM technique, generating hierarchical rule models that are easy to understand by humans [10]. In intensive care, DTs have been applied in mortality prognosis [11].

In the past, the majority of studies using DM methods in ICU environments were focused in mortality assessment [2], while the application of DM to organ failure is rather scarce. Matis et al. [12] used 15 variables (e.g. age, bilirubin, creatinine) to train a NN in order to predict liver failure after transplantation. The obtained accuracy ranged from 70% (using data prior to the operation) to 88% (5 days after the transplantation). A NN was also successfully used to access the cardiac failure of 58 patients, using 20 variables (e.g. heart rate, blood pressure) [13]. In previous work [14], NNs have outperformed DTs for multiple organ failure prediction, obtaining an overall classification accuracy of 70%. A kernel logistic regression was used by Pearcea et al. in order to predict acute pancreatitis. The model included 8 variables (e.g. age,

respiratory rate, creatinine) and outperformed a daily updated APACHE II prognostic model.

Motivated by the results obtained in [14], a novel approach is presented in this work for ICU organ dysfunction/failure prediction, where several DM techniques (e.g. LR, NNs and DTs) are evaluated. The aim is to show if it is possible to predict the condition (i.e. normal, dysfunction or failure), as measured by the SOFA score, of six organ systems (respiratory, coagulation, hepatic, cardiovascular, neurological and renal) by using adverse event data as the input variables. The proposed approach will be tested on a large database, which includes daily records of 4864 patients taken from forty two European ICUs.

The paper is organized as follows: firstly, the ICU clinical data and data mining models are presented (Section 2). Next, a description of the performed experiments is given, being the results analysed and discussed (Section 3). Finally, closing conclusions are drawn (Section 4).

2. Materials and methods

2.1 Intensive care data

The database used in the present study adopted part of the data of from the EURICUS II study. The EURICUS II project was conducted from November/98 to August/99 and encompassed forty-two ICUs from nine European Union countries (see [16] for more details). The bedside physiological variables were manually and routinely registered by the nursing staff (every hour) and the full database was later gathered at the Groningen University Hospital, the Netherlands. The study steering committee decided to consider all patients except those with age lower than eighteen years old, burned or with bypass surgery, since these subgroups of patients are mainly treated in specialized areas (e.g. burn or coronary units). The final database contains a total of 30570 daily records taken from 4864 critically ill patients.

Two main criteria were used for the event definition. First, its occurrence and duration should be registered by physiological changes (e.g. shock and not pneumonia). Second, the related physiological variables should be routinely registered at regular intervals. Four biometrics filled these requirements: the systolic blood pressure (BP), the heart rate (HR), the pulse oximeter oxygen saturation (SpO₂) and the hourly urine output (UR). The normal ranges for these parameters (see Table 1) were set by a panel of seven experts. An alarm is triggered if there is an out of range value during a given time, defining an event. For each biometric, the daily number of events was stored.

When a longer event occurs or a more extreme measurement is found, it is called a critical event. For this last case, the database includes daily entries with the number of critical events and its duration (in minutes). Table 2 shows a synopsis of the ICU variables considered. The first four attributes (the case mix) are static, being collected during the patient's admission. The next twelve variables are related to the adverse events.

Table 1: The protocol for the out of range measurements

	BP	SpO ₂	HR	UR
Normal Range	90 – 180mmHg	≥ 90%	60 – 120bpm	≥ 30ml/h
Event ^a	≥ 10min.	≥ 10min.	≥ 10min.	≥ 1h
Event ^b	≥ 10min. in 30min.	≥ 10min. in 30min.	≥ 10min. in 30min.	–
Critical Event ^a	≥ 1h	≥ 1h	≥ 1h	≥ 2h
Critical Event ^b	≥ 1h in 2h	≥ 1h in 2h	≥ 1h in 2h	–
Critical Event ^c	< 60mmHg	< 80%	< 30bpm ∨ > 180bpm	≤ 10ml/h

BP - blood pressure, HR - heart rate, SpO₂ - pulse oximeter oxygen saturation, UR

- urine output.

^a Defined when continuously out of range.

^b Defined when intermittently out of range.

^c Defined anytime.

Table 2: The intensive care variables

Attribute	Description	Min	Max	Mean ^a
admttype	admission type	Categorical ^b		
admfrom	admission origin	Categorical ^c		
SAPS II	SAPS II score	0	118	39.7±16.9
age	age of the patient	18	100	62.5±18.2
NBP	number of blood pressure events	0	24	0.7±1.9
NCRBP	number of critical blood pressure events	0	10	0.3±0.7
NHR	number of heart rate events	0	24	0.6±2.2
NCRHR	number of critical heart rate events	0	10	0.2±0.6
NSpO₂	number of oxygen events	0	24	0.4±1.8
NCRSpO₂	number of critical oxygen events	0	7	0.1±0.4
NUR	number of urine events	0	24	1.0±2.8
NCRUR	number of critical urine events	0	7	0.4±0.8
TCRBP	time of blood pressure critical events (min)	0	29.3	1.0±2.9
TCRHR	time of heart rate critical events (min)	0	29.3	1.0±3.5
TCRSpO₂	time of oxygen critical events (min)	0	29.3	0.4±2.4
TCRUR	time of urine critical events (min)	0	29.3	1.7±4.7

^a mean and sample standard deviation

^b 1- unscheduled surgery, 2- scheduled surgery, 3-medical.

^c 1- operating theatre, 2- recovery room, 3- emergency room, 4- general ward, 5-other ICU, 6- other hospital, other sources

In a daily basis, the SOFA score was computed by using the collection of raw data presented in Table 3. A total of twelve variables (e.g. platelets and bilirubin) are used to access the degree of failure for six organ systems (respiration, coagulation, liver, cardiovascular, neurological and renal). The SOFA values range from 0 to 4, with the following interpretation: 0 – normal; 1 or 2 – dysfunction; 3 or 4 – failure.

Table 3: The SOFA variables and scoring rules (adapted from [3])

Organ/ Variable	SOFA Score				
	0	1	2	3	4
respiration					
PaO ₂ /FIO ₂ (mmHg)	>400	≤400	≤ 300	≤ 200 ^a	≤ 100 ^a
coagulation					
platelets×10 ³ /mm ³	>150	≤150	≤ 100	≤ 50	≤ 20
liver					
bilirubin (μmol/l)	>20	<32	< 101	< 204	> 204
cardiovascular					
hypotension ^b	None	MAP< 70 mmHg	dop.≤5 or dobutamine (any dose)	dop.<5 or epi.≤0.1 or norepi.≤0.1	dop.>15 or epi. > 0.1 or norepi.>0.1
neurological					
Glasgow coma score	15	13-14	10-12	6-9	<6
renal					
creatinine (μmol/l)	<110	≥110	≥ 171	≥ 300	≥ 440
or urine output				<500mL/day	<200ml/day

PaO₂ - arterial oxygen tension, FIO₂ - fractional inspired oxygen.

MAP - mean arterial pressure, dop. - dopamine, epi. - epinephrine,

norepi. - norepinephrine.

^a - with respiratory support.

^b - agents administered for at least 1 hour (doses in μg/kg per min).

Figure 1 plots the respective histograms, which denote skewed distributions

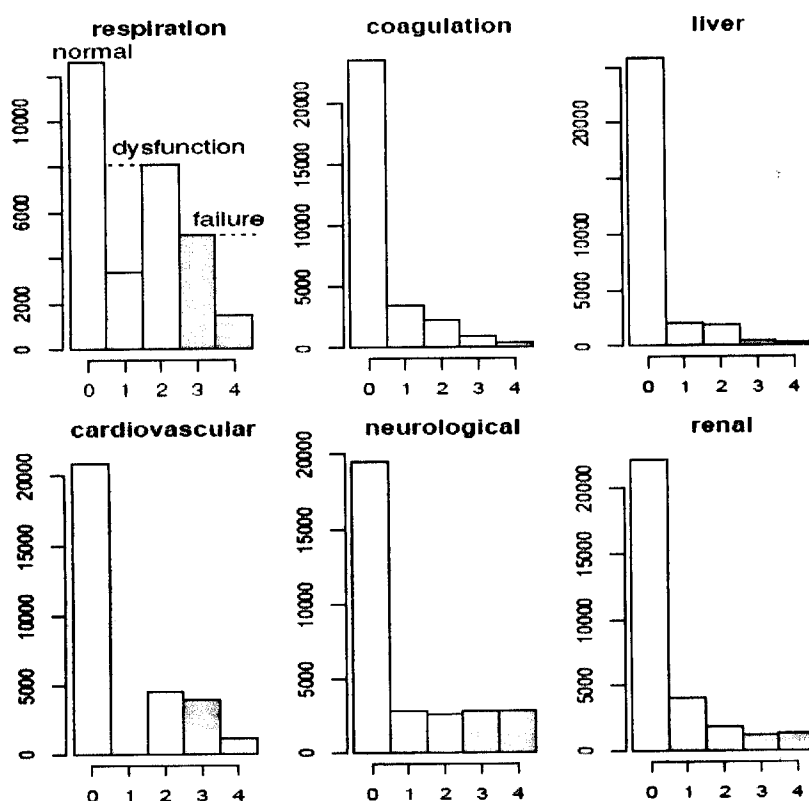


Figure 1:

2.2 Data mining methods

Data mining (DM) is a new emerging area that lies at the intersection of statistics, artificial intelligence and data management. DM tasks can be classified into two categories [8]: descriptive, where the intention is to characterize the properties of the data; and predictive, to forecast the unknown value of an output target given known values of other variables (the inputs). Predictive tasks can be further divided into classification, when the output domain is discrete, and regression, when the dependent variable is continuous.

In intensive care, the area under curve (AUC) of the receiver operating characteristic (ROC) graph is one of the most common metric for prognostic scores [2]. It measures the degree of discrimination that can be obtained from a given model. The ideal method should present an AUC of 1.0, while an AUC of 0.5 denotes a random classifier. In the medical literature, values of AUC above 0.7 are often considered acceptable [1] [2].

This study will consider the 16 input variables from Table 2. In order to perform a ROC analysis, each SOFA variable was transformed into 3 binary (0/1) outputs, one for each organ condition: normal, dysfunction and failure. The binary variables are true (1) if the

corresponding status is achieved (e.g. failure= 1 if SOFA= 3 V 4). Thus, each DM method will model the probability for a given condition and the output is interpreted as positive (or true) if $P > D$, where P denotes the probability predicted by the method and $D \in [0.0, 1.0]$ is a decision threshold. The ROC curve shows the performance of a two class classifier across the range of possible threshold values (D), plotting one minus the specificity (x-axis) versus the sensitivity (y-axis) [17].

The logistic regression (LR) is a common statistical method, where the probability of the event k is given by [8]:

$$P_k = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^N \beta_i x_i)}}$$

where $\beta_0, \dots, \beta_N$ denote the parameters of the model and $x_1, \dots, x_N$ the dependent variables.

The multilayer perceptron is a popular Neural Network (NN), where processing neurons are grouped into layers and connected by weighted links [9]. The NN is activated by feeding the input layer with the input variables and then propagating the activations in a feedforward fashion, via the weighted connections, through the entire network. The most common activation function is the logistic function used in the LR method. Hence, a NN can be seen as a generalization of the LR model (Figure 2). In contrast with the LR, NNs are capable of modelling complex discrimination functions [2]. The response of a neuron (y_i) is computed by:

$$y_i = \frac{1}{1 + \exp^{-(b_i + \sum_{j \in I} w_{i,j} \times y_j)}}$$

where I represents the set of nodes reaching node i ; $w_{i,j}$ the weight of the connection between nodes j and i ; b_i is the a constant (called bias); and $y_1 = x_1, \dots, y_N = x_N$. A fully connected network, with one hidden layer of H hidden nodes, will be adopted in this work. The initial NN weights will be randomly set within the range $[-0.7, +0.7]$ [18]. Next, the training algorithm is applied and stopped when the error slope approaches zero or after a maximum of E epochs. Since the NN cost function is nonconvex (with multiple minima), 5 runs were applied to each neural configuration, being selected the NN with the lowest error. Under these settings, the NN performance will depend on the number of hidden nodes (H). The simplest NN, with $H = 0$ and equivalent to the LR model, will present limited learning capabilities, while a large network will induce generalization loss (i.e., overfitting).

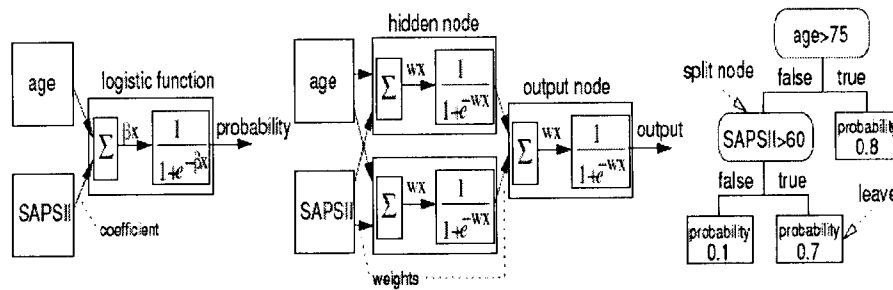


Figure 2:

The decision tree (DT) [10], also termed classification or regression tree (depending on the type of task), is a branching structure that represents a set of rules, distinguishing classes or values in a hierarchical form. This representation can be translated into a set of IF-THEN rules, which are easy to understand by humans. The source data is splitted into subsets, based on the attribute test value and the process is repeated in a recursive way. Graphically, they present a tree structure (Figure 2) and are formed by split nodes, which test a given feature, and leaves, which assign a class label (classification) or a value (regression). Before training the LR and NNs, the data will be pre-processed [18]: the input values will be scaled into a zero mean and one standard deviation range and an *1-of-C* encoding, which uses one binary variable per each class, was applied to the nominal attributes with few categories (admttype and admfrom). For example, the admttype variable is fed to 3 input nodes, according to the scheme: 1 $\rightarrow$ (-1.0 - 1.0 1.0); 2 $\rightarrow$ (-1.0 1.0 - 1.0); and 3 $\rightarrow$ (-1.0 - 1.0 1.0).

Turning to the issue of descriptive knowledge, DTs and LR models are easier to interpret than NNs. Nevertheless, it is possible to gather knowledge about what the NN has learned by measuring the relative importance of the inputs and extracting rules. Regarding the former, the sensitivity analysis [19] is a simple method that has outperformed other input selection techniques (e.g. forward selection and genetic algorithms). The relative importance (R_a) can be measured as the variance (V_a) produced in the output (y) when the input attribute (a) is moved through its entire range:

$$V_a = \frac{\sum_{i=1}^L (y_i - \bar{y})^2}{(L - 1)}$$

$$R_a = V_a / \sum_{j=1}^A V_j$$

where A denotes the number of input attributes. The y_i output is obtained by holding all input variables at their average values. The exception is x_a , which varies through its

range with L levels. In this work, L was set to two for the binary attributes ($x_a \in \{-1.0, 1.0\}$) and 6 for the continuous inputs ($x_a \in \{-1.0, -0.6, 1.0\}$).

The extraction of knowledge from NNs is still an active research domain [20].

In this work, the pedagogical technique presented in [7] will be adopted, where the direct relationships between the inputs and outputs of the NN are extracted by using a classification tree.

2.3 Statistics

In all experiments presented in this work, thirty runs were applied. The results will be shown in terms of the mean and respective t-student 95% confidence intervals.

Paired t-tests will be used to confirm statistical differences.

The predictive estimates are achieved using the Holdout method [21], where in each run the available data is divided into two mutually exclusive partitions, using stratified sampling. The training set (with 2/3 of the data) is only used during the learning phase, while the test set (1/3) is adopted after training, in order to access the performances of each model. Regarding the NNs, the number of hidden nodes (H) will be optimized using a grid search, where $H \in \{H_L, H_L+g, H_L+2g, H_U\}$, H_L and H_U denote the lower and upper bounds; and g is a small constant value. To avoid overfitting, the training data is further divided into training (44% of the original dataset) and validation sets (22%). After selecting the best NN (i.e. with the lowest error in the validation set), the NN is retrained with all training data.

3. Results

3.1 Predictive Performance

All experiments reported in this study were conducted using the R statistical environment [22], an open source and high-level matrix programming language that provides a powerful suite of tools for statistical and data analysis. The algorithm 2 and 3 presented in [17] were used to compute the ROC curves and AUC values.

The DT node split was adjusted for the reduction of the sum of squares, the LR parameters were optimized using a least squares algorithm [18] and the NNs were trained using $E = 100$ epochs of the efficient BFGS algorithm [23], from the family of quasi-Newton methods. After preliminary experiments, the number of hidden nodes tested in the grid search was set to $H \in \{2, 4, 6, 8, 10, 12\}$.

Table 4: The discrimination power (AUC values, in percentage) for each organ, condition and method (values of AUC>70% are in bold)

Condition	Organ	Method		
		DT	LR	NN
normal	respiration	63.50±0.25	69.42±0.17	71.27 ± 0.23 (8)*
	coagulation	53.13±0.13	64.05±0.18	65.36 ± 0.33 (8)
	liver	53.98±0.93	65.71±0.24	66.89 ± 0.43 (8)
	cardiovascular	65.50±0.17	69.61±0.17	72.80 ± 0.26 (10)
	neurological	68.53±0.23	71.80±0.16	73.73 ± 0.26 (10)
	renal	63.77±0.32	70.37±0.14	71.70 ± 0.23 (9)
	Mean	61.40	68.49	70.29
dysfunction	respiration	53.82±0.28	59.37 ± 0.16	62.48 ± 0.23 (8)
	coagulation	50±0	60.68 ± 0.20	61.45 ± 0.36 (8)
	liver	50.13±0.26	63.37 ± 0.27	64.87 ± 0.37 (8)
	cardiovascular	58.89±0.23	64.52 ± 0.23	66.99 ± 0.30 (8)
	neurological	56.27±0.66	58.78 ± 0.23	63.57 ± 0.38 (10)
	renal	60.41±0.4	66.78 ± 0.17	67.25 ± 0.30† (7)
	Mean	54.92	62.25	64.44
failure	respiration	53.76±0.24	68.69 ± 0.15	71.32 ± 0.21 (8)
	coagulation	50.00±0.00	73.18 ± 0.41	72.66 ± 0.51† (6)
	liver	50.37±0.42	73.45 ± 0.47	76.77 ± 0.82 (8)
	cardiovascular	58.71±0.24	69.49 ± 0.21	72.39 ± 0.36 (10)
	neurological	56.73±0.18	76.40 ± 0.18	78.09 ± 0.27 (10)
	renal	60.14±0.33	74.02 ± 0.39	75.77 ± 0.48 (8)
	Mean	54.95	72.54	74.50

* - Median number of hidden nodes (*H*).

† - The differences are not significant when compared with the LR (*p* > 0.05).

The results are shown in Table 4. A total of 3 (conditions) × 6 (organs) × 3 (methods) = 54 different models were tested. The DTs obtained the worst performances and were outperformed by both the LR and NN methods. This is not considered a surprising

result, since the majority of the inputs are numeric. Also, the DT probability estimate is a piecewise constant. In contrast, the LR and NN techniques can learn more smooth functions. Indeed, the LR shows an overall improvement from 7% to 17%, when compared with the DTs. Nevertheless, the best results are obtained by the NNs, which outperform the LR with a margin of 1.3% for the normal condition, a value that increases to 1.8% and 2.0% for the dysfunction and failure status. The NN/LR differences are significant ($p < 0.05$) in all cases except for the renal dysfunction and coagulation failure predictions. For the NNs, the table also presents the median number of the selected hidden nodes, which ranged from 7 to 10.

When analyzing the organ condition discrimination, the dysfunction condition is more difficult to predict. In effect, none of the presented models has acceptable values (AUC higher than 70%). The normal status shows a higher discrimination, with 2 LR and 4 NN acceptable models. Finally, the failure condition presents the most accurate predictions. The LR models are acceptable for the coagulation, liver, neurological and renal, while the NNs obtain good performances for all organs. In particular, the NN central nervous system failure (neurological) model shows a high quality AUC, close to 80%. For demonstrative purposes the ROC curves are plotted for the failure condition (Figure 3). In the graphs, the NN curves are above the other methods, except for the LR in the coagulation system.

3.2 Descriptive knowledge

This section will provide explanatory knowledge that can be useful for the intensive care domain. The goal not to infer about the predictive capabilities of each model, as measured in the previous section, but to give a simple description that summarizes the DM models. Thus, the whole dataset will be used in the descriptive experiments. Furthermore, the focus will be given to the failure condition, which achieved the best predictive performances.

Table 5 presents the relative relevance (in percentage) of each input variable for accessing organ failure. The best predictive models from Table 4 were adopted: NN for all organ systems except coagulation, which uses the LR. The procedure described in Section 2.2 was used for the NN optimisation. The original data was divided into training (2/3) and validation sets (1/3). Then, the selected NN was retrained with the full dataset.

The case mix variables (age, admfrom, admtype and SAPS II) denote an influence around 30%, except for the liver (1%) and renal organs (14%). In particular, the severity of illness index (SAPS II) denotes a high relevance in case of the coagulation and

neurological systems. In general, the age of the patient presents a low impact (except for the neurological system). However, it should be noted that age is already used in the calculation of the SAPS II index. Finally, the admission variables (type and origin) have a moderate influence in the respiration, cardiovascular and neurological failures. Regarding the event variables, as expected they present the highest impact: 50% neurological, 65% – respiration, 69% – coagulation, 74% – cardiovascular, 86% – renal and 99% – liver. The urine output (UR) events are highly relevant for the renal (as anticipated) and liver failure. The systolic blood pressure (BP) variables are the most relevant for the respiration and cardiovascular organs. The oxygen events (S_pO_2) denote a high importance for the respiration and coagulation systems, while the heart rate (HR) seems to be the least relevant biometric.

The input importance analysis will be complemented with a knowledge extraction procedure. Since the intention is to provide a compact description, only the most relevant variables from Table 5 will be selected for each model. Six variables (with the highest coefficients) were selected for the LR and coagulation system (Figure 4). The retrained LR is much simpler (Figure 4) and it is capable of replicating the 16 input variables LR with only a 0.1% error. Regarding the NNs, it will be shown how a simple set of classification rules can be obtained. As an example, the decision threshold (D) was first maximized to get the best sensitivity $\times$ specificity value. Then, the D value is used to compute the NN output classes (e.g. true/false) for each patient. Next, a DT is used to learn the NN behaviour [7], where the training set is composed by the NN inputs and outputs. Finally, the DTs were pruned to a maximum depth of four levels. The obtained trees (Figure 4) managed to mimic the NN behaviour with low classification errors: 5% (respiration), 9.8% (liver), 3.2% (cardiovascular), 0.7% (neurological) and 1.7% (renal). For demonstrative purposes, the next rule corresponds to the top right path of the renal tree: IF SAPS II >47.5 $\wedge$ TTUR > 6.99 THEN *true*.

4. Conclusion

The prediction of organ failure severity is crucial in intensive care units (ICUs). Indeed, several expert-driven scores have been developed for such purpose, such as the sequential organ failure assessment (SOFA), which is widely used in Europe. The SOFA score can be used to access the degree of organ failure, in terms of normal, dysfunction and failure conditions. However, there are drawbacks with several SOFA variables, which require time and costs to be obtained (e.g. creatinine or bilirubin) or are imprecise (e.g. Glasgow coma score).

In this study, an alternative data-driven (bedside monitoring) approach was tested for rating the failure of six organ systems (i.e. respiration, coagulation, liver, cardiovascular, neurological and renal). A large database was considered, with 30750 daily records taken from forty-two European ICUs. The tested input variables included the case mix and adverse events, which were measured as the out of range values of four routinely collected physiological variables. Moreover, several data mining (DM) techniques were explored, including decision trees (DTs), logistic regression (LR) and neural networks (NNs).

The experiments were conducted using a receiver operator characteristic (ROC) analysis. The results show that the NNs are the best learning models, outperforming the LR (second best method) with improvements ranging from 1.3% to 2.0% and presenting more acceptable models (ROC area higher than 70%). The only exception is the coagulation failure, where the LR obtained a non significant higher performance. The average performance for the organ status shows that failure condition is more feasible to predict (74.5%) than the other conditions (70.3% for normal and 64.4% for dysfunction). Particularly, impressive results (ROC area higher than 75%) were achieved for the prediction of failure in three organ systems (liver, renal and neurological).

Finally, two procedures (input relevance and rule extraction) were applied to simplify the best models of the failure condition. This descriptive analysis is particularly useful for the NNs, since they are less understandable when compared to the LR and DT methods. The experiments confirmed the importance of the adverse events, with an overall impact that ranges from 50% to 99%.

It should be noted that the data was manually collected in the EURICUS II project [16]. Nevertheless, there is a potential for the automatic collection of the adverse events, by using a real-time computerized data acquisition system from bedside monitors and applying quality procedures (e.g. data validated by the ICU staff), as shown in [24]. Hence, the proposed approach could be applied using precise, low cost and real-time variables. Moreover, such system could give more updated predictions (e.g. every hour), a scenario that is not possible with the SOFA score (due to its definition). Another interesting research issue would be the inclusion of the event variables to the SOFA scoring system, yielding an expanded score. None of the tested models obtained satisfactory ROC curves for the dysfunction status, suggesting that this condition is very difficult to access. However, it should be noted that organ failure development is a continuous process where the borders for each stage are necessarily fuzzy. To our knowledge this is the first attempt where bedside monitored data is modelled for such purpose and further exploratory research is needed. In particular, we believe that a

temporal analysis, such as presented in [25], where the evolution of each organ during the patient length of stay is modelled, is the most promising direction. In the future, we intent to test all these possibilities in a real environment, such as the INTCare pilot project [24], where a friendly decision support system is being developed at the ICU of the Hospital Geral de Santo António, Oporto, Portugal.

Acknowledgments

The authors wish to thank FRICE (www.frice.nl) and all the anonymous participants of the EURICUS II project.

References

- [1] A. Rosenberg. Recent innovations in intensive care unit risk-prediction models. *Current Opinion in Critical Care*, 8:321–330, 2002.
- [2] L. Ohno-Machado, F. Resnic, and M. Matheny. Prognosis in Critical Care. *Annual Rev Biomed Eng*, 8:567–599, 2006.
- [3] R. Moreno, J. Vincent, R. Matos, A. Mendonça, F. Cantraine, L. Thijs, J. Takala, C. Sprung, M. Antonelli, H. Buining, and S. Willatts. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. *Intensive Care Med*, 25:696–696, 1999.
- [4] J. Le Gall. The use of severity scores in the intensive care unit. *Intensive Care Med*, 31:1618–1623, 2005.
- [5] J. Vincent, R. Moreno, J. Takala, S. Willatss, A. Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction / failure. *Intensive Care Med*, 22:707–710, 1996.
- [6] J. Paetz. Finding optimal decision scores by evolutionary strategies. *Artif Intell Med*, 32:85–95, 2004.
- [7] Á. Silva, P. Cortez, M. F. Santos, L. Gomes, and J. Neves. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artif Intell Med*, 36:223–234, 2006.
- [8] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.
- [9] S. Haykin. *Neural Networks - A Comprehensive Foundation*. Prentice-Hall, New Jersey, 2nd edition, 1999.
- [10] L. Breiman, J. Friedman, R. Ohlsen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Monterey, CA, 1984.

- [11] A. Abu-Hanna and N. Keizer. Integrating classification trees with local logistic regression in intensive care prognosis. *Artif Intell Med*, 29:5–23, 2003.
- [12] S. Matis, H. Doyle, I. Marino, R. Mural, and E. Uberbacher. Use of neural networks for prediction of graft failure following liver transplantation. In *Proceedings of the 8th IEEE Symposium on Computer-Based Medical Systems*, pages 133–140, Washington, DC, USA, 1995. IEEE.
- [13] M. Gils, H. Jansen, K. Nieminen, R. Summers, and P. Weller. Using artificial neural networks for classifying ICU patient states. *Engineering in Medicine and Biology Magazine*, 16:41–47, 1997.
- [14] Á. Silva, P. Cortez, M. Santos, L. Gomes, and J. Neves. Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks. In I. Seruca, J. Cordeiro, S. Hammoudi, and J. Filipe, editors, *Enterprise Information Systems VI*, The Netherlands, 2006. Springer.
- [15] C. Pearcea, S. Gunn, A. Ahmeda, and C. Johnson. Machine Learning Can Improve Prediction of Severity in Acute Pancreatitis Using Admission Values of APACHE II Score and C-Reactive Protein. *Pancreatology*, 6:123–131, 2006.
- [16] R. Miranda. The effect of harmonising and standardising the nursing tasks on Intensive Care Units (ICU) of the European Community, Second European Intensive Care Unit Study (EURICUS-II). Technical Report BMH4-CT96-0817, Groningen University Hospital, The Netherlands, 1999.
- [17] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical report, HP Laboratories Palo Alto, Kluwer, March 2004.
- [18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY, USA, 2001.
- [19] R. Kewley, M. Embrechts, and C. Breneman. Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks. *IEEE Trans Neural Networks*, 11(3):668–679, May 2000.
- [20] R. Setiono. Techniques for Extracting Classification and Regression Rules from Artificial Neural Networks. In D. Fogel and C. Robinson, editors, *Computational Intelligence: The Experts Speak*, pages 99–114. Piscataway, NY, USA, IEEE, 2003.
- [21] A. Flexer. Statistical evaluation of neural networks experiments: Minimum requirements and current practice. In *Proceedings of the 13th European Meeting on Cybernetics and Systems Research*, volume 2, pages 1005–1008, Vienna, Austria, 1996.
- [22] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3, <http://www.R-project.org>, 2006.

- [23] M. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.
- [24] P. Gago, M.F. Santos, A. Silva, P. Cortez, J. Neves, and L. Gomes. INTCare: A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine. *Journal of Decision Systems*, 14(3):241–259, 2005.
- [25] M. Kayaalp, G. Cooper, and G. Clermont. Predicting with Variables Constructed from Temporal Sequences. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, USA, pages 220–225, 2001.

Na próxima secção é apresentado o trabalho ***Organ Failure Prediction Based On Clinical Adverse Events: A Cluster Model Approach*** publicado em: *Artificial Intelligence and Applications (AIA2003)*, 695-700, ACTA Press, 2003.

Álvaro Silva, João Pereira, Manuel Santos, Lopes Gomes, José Neves.

4.5 Preâmbulo

Neste estudo são desenvolvidos modelos de previsão de falência de órgãos e sistemas, no contexto do suporte à tomada de decisão clínica.

O objectivo deste trabalho é a geração de modelos de previsão da falência de órgãos para o dia n (*dia seguinte*) até ao quinto dia de internamento na UCI, baseado nas variáveis do dia $n-1$ e anteriores.

Para atingir o objectivo enunciado, foi realizado um trabalho de segmentação e aplicação de indução de regras de decisão, árvores de decisão e redes neuronais artificiais de forma a gerar um conjunto de modelos de previsão de falência de órgãos e sistemas, visualizáveis através de uma matriz tridimensional.

Para além dos critérios de exclusão e de inclusão dos doentes para análise referidos, foi associado mais um critério de inclusão: tempo de internamento na UCI $\geq$ dois dias, o que englobou 4425 doentes.

As variáveis independentes utilizadas para modelação da previsão foram: o SAPS II score, o tipo e local de origem da admissão à UCI, o diagnóstico de admissão (pós-operatório e médico) e frequência diária de eventos clínicos adversos. As referidas variáveis são utilizadas para a previsão da falência de órgão e sistema para o segundo dia de internamento na UCI. A partir do terceiro dia são usados apenas os eventos clínicos adversos, até ao quinto dia de internamento na UCI. Para a construção dos

modelos, não são considerados os valores de SOFA score de qualquer dia de internamento.

Foram criadas seis novas variáveis (binárias) com base no SOFA score por órgão e sistema, de acordo com a seguinte expressão: 0 se SOFA score <3 e 1 se SOFA score $\geq$ 3 (falência de órgão), considerados como os *outcomes* de interesse.

Com o intuito de se criar modelos de previsão de falência para cada um dos órgãos e sistemas de forma independente, foi adoptada uma abordagem por segmentos. A população em estudo foi segmentada em três subgrupos (designados por segmentos 00, 10 e 20) com recurso à aplicação da rede neuronal artificial de Kohonen. A identificação dos critérios de segmentação foi conseguida através da utilização do algoritmo C5. Para a geração de modelos de previsão em cada segmento, foi uma vez mais utilizado o algoritmo C5 de forma a construir as respectivas árvores de decisão.

Estes procedimentos foram efectuados para o segundo, terceiro, quarto e quintos dias de internamento na UCI. Os modelos gerados exploram a previsão da falência de órgão e sistema em cada segmento.

Foi construída uma estrutura de suporte (*framework*) tridimensional para a identificação e aplicação do modelo gerado para o segmento em que um doente se enquadra, em função das suas características, dia de internamento e órgão ou sistema em questão.

Como resultado final das partições efectuadas, foi possível obter 72 segmentos. Para a compreensão do modelo de segmentação, foram induzidas árvores de decisão através do algoritmo C5.

A título de exemplo, apresenta-se de seguida a árvore de decisão criada para a segmentação no segundo dia de internamento para o sistema hematológico:

Tipo de admissão $\leq$ 2

Origem da admissão $\leq$ 1 -> segmento 20

Origem da admissão > 1

Tipo de admissão $\leq$ 1 -> segmento 20

Tipo de admissão > 1 -> segmento 10

Tipo de admissão > 2

Origem da admissão $\leq$ 1

SAPS II score $\leq$ 49 -> segmento 10

SAPS II score > 49 -> segmento 00

Origem da admissão > 1 segmento -> 00

A partir da árvore apresentada é possível criar as seguintes regras que contêm os critérios de segmentação:

Regras para 00:

Regra #1 para 00:

Se tipo de admissão > 2
e origem de admissão $\leq$ 1
e SAPS II score > 49
então -> 00

Regra #2 para 00:

Se tipo de admissão > 2
e origem de admissão > 1
então -> 00

Regras para 10:

Regra #1 for 10:

Se tipo de admissão > 1
e tipo de admissão $\leq$ 2
e origem de admissão > 1
então -> 10 (155, 1.0)

Regra #2 for 10:

se tipo de admissão > 2
e origem de admissão $\leq$ 1
e SAPS II score $\leq$ 49
então -> 10

Regras para 20:

Regra #1 for 20:

se tipo de admissão $\leq$ 2
e origem de admissão $\leq$ 1
então -> 20

Regra #2 for 20:

se tipo de admissão $\leq$ 1
e origem de admissão > 1
então -> 20

Se nenhuma regra se aplicar: -> 00

A título ilustrativo, é apresentada uma parte da árvore de decisão gerada para o segmento 00 do sistema respiratório, para o terceiro dia de internamento:

SAPS II score $\leq$ 41

nrofo2cr1 $\leq$ 0

nrofo2ev2 $\leq$ 1

nrofhrcr2 $\leq$ 1

SAPS II score ≤ 17

nrofurcr1 $\leq 0 \rightarrow 0$

...

Não existirá falência do sistema respiratório no terceiro dia de internamento se:

- o doente tem um SAPS II score ≤ 17 e;
- não tem nenhum evento crítico de saturação arterial de O₂ no primeiro dia e;
- tem menos de dois eventos de saturação arterial de O₂ no segundo dia e;
- tem menos de dois eventos críticos de frequência cardíaca no segundo dia e;
- não tem nenhum evento crítico de débito urinário no primeiro dia .

Para a validação dos modelos foram utilizados os seguintes métodos: a validação cruzada, onde foi efectuada uma divisão da amostra em 10 conjuntos (*10-fold cross validation*) e a matriz de confusão.

Neste trabalho, foi considerada uma estrutura tri-dimensional de suporte designada por *framework* de segmentação, que permite gerir e caracterizar o conjunto de modelos de previsão criados (para o segundo, terceiro, quarto e quinto dias de internamento).

A utilização desta ferramenta possibilita a identificação e aplicação do modelo de previsão gerado para cada segmento, tendo em consideração o dia e sistema de órgão em questão. Para tal, é necessário que em primeira instância, se identifique qual o segmento em que o doente ficará enquadrado. Esta associação é efectuada em função dos critérios de segmentação para cada dia de internamento e sistema de órgão em questão. A identificação e aplicação de um modelo a um doente em particular têm por base três vértices: dia, órgão e segmento, baseados em métricas de avaliação como a precisão, especificidade e sensibilidade.

A utilização da *framework* é descrita da seguinte forma:

- em função do dia de internamento, órgão e segmento, é identificado o respectivo modelo de previsão de falência de órgão e sistema;
- é aplicado o modelo e verificada a previsão;
- é repetido o procedimento para cada órgão e sistema por dia de internamento.

Numa das arestas (Figura 2 do estudo) constam os quatro dias de internamento considerados para efeitos de previsão; noutra aresta, podemos visualizar os seis órgãos e sistemas; enquanto que numa terceira, são observados os três segmentos em análise.

Como se pode verificar, o mesmo doente poderá pertencer a segmentos diferentes no mesmo dia de internamento, bem como mudar de segmento ao longo da sua estadia na UCI.



Os resultados da precisão dos modelos de previsão para o *outcome* de interesse ao quinto dia de internamento na UCI gerados são apresentados no quadro 5 do artigo. Verificámos valores de precisão elevados (74,2 a 98,6%).

Os valores relativos à precisão, especificidade e sensibilidade obtidos para todos os modelos, são apresentados na Figura 1 da presente secção, que agrega através de escalas de tons de cinzento, toda esta informação para cada um dos dias, órgãos e segmentos.

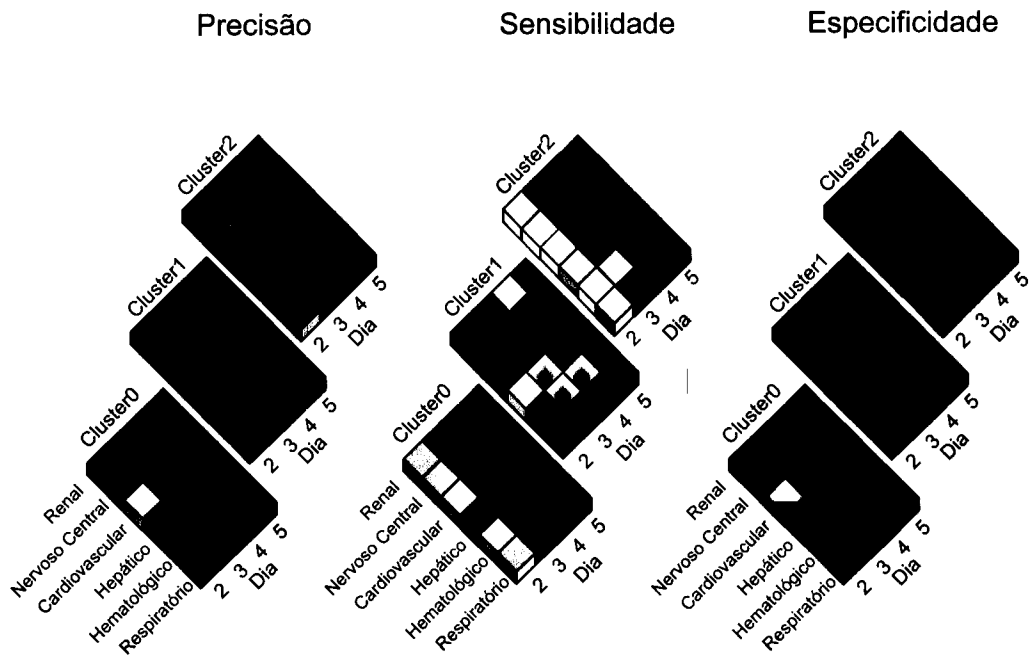
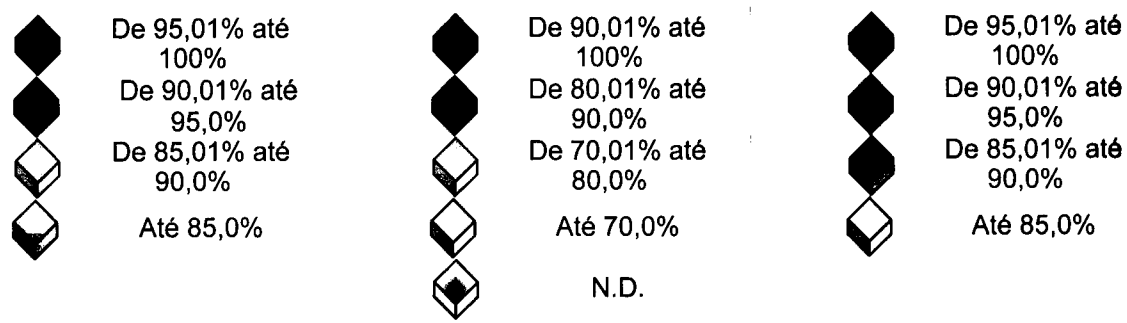


Figura 1 – Precisão, sensibilidade e especificidade dos modelos de previsão.



4.6 Organ Failure Prediction Based On Clinical Adverse Events: A Cluster Model Approach

Apresenta-se o artigo integral:

Organ failure prediction based on clinical adverse events: A cluster model approach

Álvaro Silva, João Pereira, Manuel Santos, Lopes Gomes, José Neves

Abstract

Medical prognosis has played an increasing role in health, namely in the critical care medicine. These factors have induced the medical community to take a more active interest in developing models for mortality prediction based on Artificial Intelligence (AI) techniques [1], that make possible the physicians pro-active action. In this context, the existence of large Databases (DB) containing Intensive Care Units (ICU) clinical information, motivate and enable the application of Data Mining (DM) techniques, in a Knowledge Discovery Database process (KDD), to induce prediction models of organ failure in a much more efficient way than other approaches (e.g., Logistic Regression)[2].

In this study we used a clustering approach as a strategy to predict the organs failures, based on the occurrence of Clinical Adverse Events (CAE) during the patient's stay in the ICU. It follows other studies made in this area (Santos e Silva [3, 4]), where DM techniques were used to predict hospital mortality.

The clinical database used was created from part of the database originated from the EURICUS II [5] study, after being subjected to processes elaborated with SPSS and SPSS Clementine tools [6].

Various models were created by applying the C5 Algorithm [7] in order to generate decision rules for easy interpretation, of each organ, for each day of stay, and for each cluster defined. This approach enables a better accuracy on some models, a homogeneous segmentation of patient's (for further medical analysis) and a tracking model of the patients' evolution during his stay. In such a way it is possible to identify the cluster where he fits and, consequently, to make use of the adjusted model to predict the status of a certain organ on the following day.

Key Words

Data Mining, Knowledge Discovery from Databases, Artificial Neural Networks, Organ Failure, Mortality Predicting Models based on Intermediate Outcomes, Intensive Care Units.

1. Introduction

Decision Support Systems (DSS) are present on different areas of activity where various solutions for different problems origins are needed. Health care is not an exception. So we shouldn't be surprised that realities like hospital ICUs take advantage of these DSS.

The Sequential Organ Failure Assessment (SOFA) [8, 9] scores the degree of failure of six organs. This score is calculated on a daily basis.

Considering the admission data and other variables taken on first admission day, as well as CAEs taken during the patient's stay in ICU, it is possible to predict the failure of each organ for the day following the last day of collecting data (time series). This is the assumption this study is going to validate.

This document shows a study on the influence of the CAEs to the dysfunction or failure of human organs occurring during the patients' stay at the ICU.

To generate predicting models for each organ, and for each day, we followed a cluster model approach. This allowed the conception of a clustering framework, representing the patient evolution along his stay. The paper is organized as follows: in the second and third chapters present the clinical data and some definitions about events and critical events; the fourth and fifth chapters introduce the process of data preparation, transformation and model generation; the last two chapters, before the eighth that concludes the article, are dedicated to the results presenting the achieved accuracies and the framework to organize the models into clusters.

2. Clinical Data

In this study a database was created based on EURICUS II, a study made in 42 ICUs on 9 UE countries, between 1997 and 1999. For a period of 4 months every admission to the ICU was included. This database includes the characteristics that define case mix [10], namely the Age, the Type of Admission (unscheduled surgery, scheduled surgery and medical), the Admission Source (Operating theatre, Recovery room, Emergency room, Ward, other ICU, other Hospital, other sources), Diagnosis, SAPS II [11], SOFA of each Organ System (Respiratory, Coagulation, Hepatic, Cardiovascular, Central Nervous and Renal), Mortality in the ICU and in the Hospital; Number of CAEs for each of the parameters monitored continuously, Length of Stay and Admission Day. In this study, from the 5355 patients admitted to the ICUs only 4425 (82.63%) stayed for two or more days, 3105 (57.98%) stayed three or more days and 2329 (43.49%)

four days or over. For the data concerning the fifth day of stay, only 1845 (34.35%) patients were considered.

3. Clinical Adverse Events

We considered as Events (Ev) or Critical Events (CrEv) the occurrences of values out of the established limits for four physiologic variables that are monitored continuously. These four variables are Heart Rate (HR), Systolic Blood Pressure (BP), Oxygen Saturation (SaO₂) and Urine Output (Diur). A group of clinical specialists determined the intervals considered normal for each one of these parameters. In the Table 1 it is presented the information related with the intervals considered as normal, as well as the situations in that they are considered as out of the limits.

Table 1 – Clinical Adverse Events

Events				
Event	Suggested Range	Continuously out-of-range	Intermittently out-of-range	
BP(mmHg)	90-180	≥ 10'	≥ 10' in 30'	
SaO ₂ (%)	≥ 90	≥ 10'	≥ 10' in 30'	
HR(bpm)	60-120	≥ 10'	≥ 10' in 30'	
Diur(ml/hour)	≥ 30	≥ 1 hour		

Critical Events				
Critical Event	Suggested Range	Continuously out-of-range	Intermittently out-of-range	Event Anytime
BP(mmHg)	90-180	≥ 60'	≥ 60' in 120'	BP < 60
SaO ₂ (%)	≥ 90	≥ 60'	≥ 60' in 120'	SpO ₂ < 80
HR(bpm)	60-120	≥ 60'	≥ 60' in 120'	HR < 30 or HR > 180
Diur(ml/hour)	≥ 30	≥ 2 hours		≤ 10

Adverse events were defined as binary variables, whose values correspond to one of two situations, in that the variable is within or not of the established limits (if yes, by how long). We considered as an Event when the value of the analyzed parameter maintains out of the limits, for a same or a period superior of continuous time to 10 min. (1 h. in the case of Diur.) and less than 60 min. (2 h in the case of Diur).

It is still considered an Event when, in a discontinuous way, values are verified out of the limits, but that are inferior to 10 min. and in a period of time of 30 min. maximum, since the sum of those is equal or greater to 10 min.

The definition of Critical Event is similar to the Event, but with different values. The times of 10 min. referred in the definition of Events, should be replaced by 1 hour, the 30 min. for 2 hours and 1 hour for 2 hours respectively.

A Critical Event can also be defined in some special situations, when the value of the analyzed parameter places among certain values that we could see in the Table 1.

We only can consider a new event, after a recovery period of 30 min. or more for BP, SaO2 and HR, and of 2 hours or more for Diur, with values inside of the intervals. In Critical Events, it should be considered a period greater than 2 hours for Diur and greater than 60 minutes for the remaining ones.

4. Data preparation

In this work we followed the Crisp-DM Methodology [12] and in the data preparation phase has been necessary to treat the wrong or omitted data. Besides, not all the variables were considered to generate the prediction models, as it is the case of the age, once it is already considered within SAPSII score. The Table 2 shows the variables that were considered in this study and their description.

Table 2 - Variables Description

Variable	Description	
ID	Patient number	*
Respirat	Respiratory System	**
Coagulat	Coagulation System	**
Liver	Hepatic System	**
Cardiova	Cardiovascular System	**
Cns	Central Nervous System	**
Renal	Renal System	**
Nrbpevnt	Number of BP Events/day	
Nrbpcriv	Number of BP Critical	
Nrofhrev	Number of HR Events/day	
Nrofhrcr	Number of HR Critical	
Nrofo2ev	Number of O2 Events/day	
Nrofo2cr	Number of O2 Critical	
Nrofurev	Number of Diur Events/day	
Nrofurcr	Number of Diur Critical	
Admfrom	Admission From	***
Admtype	Admission Type	***
Sapsiil	Simplified Acute Physiology	***
Diagn01	Diagnostic	***

* Not considered into the prediction models.

** Dependent variables.

*** Variables just considered in the first day.

Once we intend to predict an organ failure in a certain day, based in the data of previous days, it was necessary to transform the database structure, in order to capture

a temporary sequence of the variables (time series). The Tables 3 and 4 represent these two states.

Table 3 – Database before the transformation

ID	AdmFrom	AdmType	...	Nrofo2cr	Nrofurev	Nrofurcr
03	1	1		0	0	0
03	1	1		0	1	1
03	1	1		0	1	1
...	...	...	...	...	...	...

Table 4 – Database transformed

ID	Adm From	Adm Type	...	Nrof o2c1	Nrof ure1	Nrof urc1	...	Nrof o2c3	Nrof ure3	Nrof urc3
03	1	1		0	0	0		0	1	1

There are variables that are only related to the first (AdmFrom, AdmType, SAPSII and Diagn01) and are not considered to the remaining days.

For instance, if we considered the prediction of respiratory organ failure in the third day of stay at the ICU, we can verify that the SOFAs obtained for the first and second days are not considered. However the number of Ev and CrEv obtained for these two days are considered for the prediction in cause.

We just considered a temporary horizon of five days, because, in medical terms, the fifth day of stay in an ICU is considered a critical point in terms of the evolution of the patient's clinical state. The first day was not considered for prediction effects, once the goal is to predict organ failure based on the data collected in the previous days. In any way, the data collected in the first day is used as input for all the other ones.

5. Models Definitions and Construction

Making use of SPSS Clementine tool, we submitted the database to a Kohonen Network [13], to segment it in three distinct groups. Later, it was fallen back upon the C5 algorithm in way to generate a model of decision rules to understand each one of those clusters. This way, we obtained 3 models for each one of the dependent variables, and for each one of the days of stay in the ICU, totalling 18 models for each day.

The necessary procedures for mining and generation of models for each one of the clusters can be visualized in the Figure 1, which presents a stream of Clementine, elaborated for predicting central nervous system failure, in the fourth day of stay.

The validation of those models was made through the cross validation method (10 fold) and, at the end we could see the achieved results through the application of the models generated in each one of the clusters using a 2 x 2 matrix, which compares the results obtained with the real ones.

Table 5 – Accuracies obtained for the 5th day of stay

		Cluster 1	Cluster 2	Cluster 3
Respiratory System	Accuracy	74.20	75.00	75.80
	Std.Error	0.70	4.80	0.90
	N.Cases	1132	64	649
Coagulation System	Accuracy	92.60	98.60	94.50
	Std.Error	0.40	1.40	0.20
	N.Cases	649	65	1131
Hepatic System	Accuracy	97.90	98.60	97.20
	Std.Error	0.10	1.40	0.20
	N.Cases	1132	63	650
Cardiovascular System	Accuracy	80.60	89.80	81.50
	Std.Error	0.50	3.80	0.50
	N.Cases	1132	64	649
Central Nervous System	Accuracy	82.20	91.00	79.80
	Std.Error	0.50	2.50	1.00
	N.Cases	1132	64	649
Renal System	Accuracy	89.10	92.10	91.10
	Std.Error	0.50	3.60	0.50
	N.Cases	1134	62	649

7. Clustering Framework

The organ failure prediction models are denoted by $m(d_j, o_k, c_i)$, they were generated for each of the 6 different organs o_k where $k \in \{Renal, Central\ Nervous, Cardiovascular, Liver, Coagulation, Respiratory\}$, for each of the first 5 days of the stay d_i , where $j \in \{2,3,4,5\}$ and were organized in three distinct clusters c_i , where $i \in \{0,1,2\}$ (the cases in each of these clusters are closer).

These models are organized in a cube framework that makes possible the graphical presentation of the patient course along the stay in ICU as we can see in the Figure 2. For a given patient we have a prediction model for each one of the 6 organs (o_k) indexed to the day of stay (d_j) and the correspondent cluster (c_i).

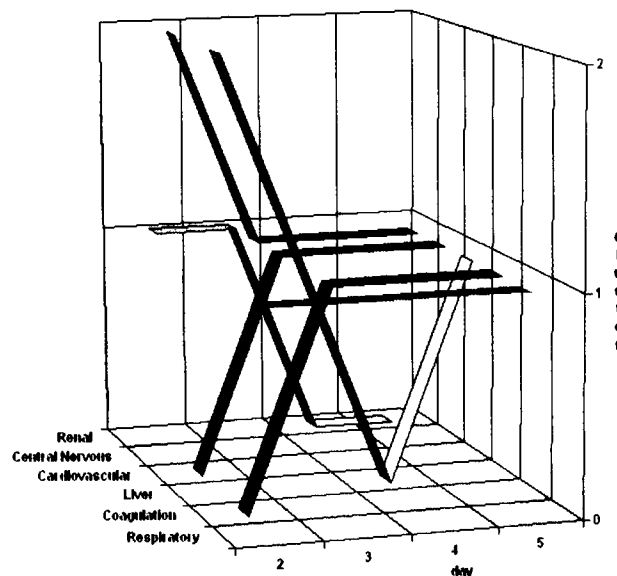


Figure 2 – Clustering Framework.

Be noticed that, in the same day, the correspondent models of a particular patient may belong to different clusters. In the example presented in the Figure 2, the prediction models for day 2 are given by:

$m(2, Renal, 1);$
 $m(2, Central Nervous, 2);$
 $m(2, Cardiovascular, 0);$
 $m(2, Liver, 2);$
 $m(2, Coagulation, 0);$
 $m(2, Respiratory, 1).$

For the same patient, for the fourth day, the models are given by:

$m(4, Renal, 0);$
 $m(4, Central Nervous, 1);$
 $m(4, Cardiovascular, 1);$
 $m(4, Liver, 0);$
 $m(4, Coagulation, 1);$
 $m(4, Respiratory, 1).$

As we can see, the prediction model for the *Central Nervous* system of this patient changed from the cluster 2 to the first cluster.

The database segmentation criteria, for the third day of stay related with the *Renal* system failure, can be visualized under the form of rules as following:

Rules for 0:

Rule #1 for 0:

if admfrom1 > 1

```
and admtype1 > 2
then -> 0 (1764, 0.999)
```

Rules for 1:

Rule #1 for 1:

```
if admfrom1 > 1
and admtype1 > 1
and admtype1 <= 2
then -> 1 (98, 0.99)
```

Rule #2 for 1:

```
if admfrom1 <= 1
and admtype1 > 2
then -> 1 (12, 0.929)
```

Rules for 2:

Rule #1 for 2:

```
if admtype1 <= 1
then -> 2 (774, 0.999)
```

Rule #2 for 2:

```
if admfrom1 <= 1
and admtype1 <= 2
then -> 2 (922, 0.999)
```

Default : -> 0

In this case, the variables that determined the classification in three clusters, were the Admission Type and Admission From. The values presented between parentheses stand for the support level and the confidence level, respectively.

As we can see, there is only one rule that respect to cluster 0, and two rules for each one of the clusters 2 and 3. If the admission from is medical, and the admission type is not the Operating Bloc or Recovery Room, the patient will be in the cluster 0. This rule was applied to 1764 cases.

The first day of stay was not considered, once it doesn't make sense to predict organ failure for this day, because the only data we have was collected in the same day.

It can be said for instance that, a patient with SAPSII 37, with the type of admission scheduled surgery, coming from the infirmary, with diagnosis pos-operative and with an event of O2 in the first day; it's framed in the cluster 1.

Each of the 72 models referred in the framework correspond to rules generated by the C5 algorithm. Consider for example the rules that predict the *Central Nervous* system for the fifth day:

Rules for 0:

Rule #1 for 0:

*if sapsii1 <= 54
then -> 0*

Rule #2 for 0:

*if nrofurev1 <= 5
and nrofhrev2 > 2
and nrofhrev4 <= 3
then -> 0*

Rules for 1:

Rule #1 for 1:

*if sapsii1 > 54
and nrbpevnt3 <= 5
and nrofhrev4 > 3
then -> 1*

Rule #2 for 1:

*if admfrom1 <= 3
and sapsii1 > 54
and diagn01 <= 0
and nrofurev4 > 2
then -> 1*

Rule #3 for 1:

*if nrofurev1 > 1
and sapsii1 > 54
and diagn01 > 0
and nrbpevnt2 <= 0
and nrofhrcr2 <= 0
then -> 1*

.....

.....

Rule #11 for 1:

*if nrbpevnt1 <= 0
and admfrom1 <= 3
and sapsii1 > 54
and nrofurcr2 <= 1
and nrbpcriv3 <= 0*

then -> 1

Default : -> 0

8. Conclusions and Further Work

As demonstrated in this study, the Clinical Adverse Events (CAE) that happen during the stay period in the ICU are important factors in the process of organ failure prediction. These CAEs, case mix and SAPSII were the base of the variables considered for organ failure prediction, and the interesting results obtained, are result of all this, as well of the clustering approach adopted.

Besides, this study presented a clustering framework, with the purpose of identifying and applying the model generated for the cluster in which a patient frames to, according to his characteristics. The majority of the models revealed high accuracies; witch is very useful for decision support.

Further work will include the optimization of each generated model and the deep medical characterization of the patients of each model (in the cube).

References

- [1] Hanson William C., Marshall Bryan E.: Artificial Intelligence applications in the intensive care unit, *Critical Care Medicine, Volume 29, Number 2*, 2001, 427:435.
- [2] Gilles Clermont, Derek C. Angus, Stephen M. DiRusso, et al: Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models, *Critical Care Medicine, Volume 29, Number 2*, 2001, 291:296.
- [3] Santos Manuel, Silva Álvaro, Abelha António, et al: Predicting Outcome in Critically Ill Patients using Artificial Intelligence Models, *Artificial Intelligence and Applications*, Málaga, Spain, 2002, 47-51.
- [4] Santos Manuel, Neves José, Abelha António, et al: Augmented Data Mining over Clinical Databases: Using Learning Classifier Systems, *Fourth International Conference on Enterprise Information Systems*, Ciudad Real, Spain, 2002, 512-516.
- [5] <http://www.frice.nl>, Foundation for Research on Intensive Care in Europe, visited in March 2003.
- [6] <http://www.spss.com>, SPSS Inc., visited in March 2003.
- [7] <http://www.rulequest.com>, Quinlan J.R. C5.0 Data Mining Tool, visited in March 2003.

- [8] Vincent J-L, De Mendonça A, Cantraine F. et al., Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study, *Critical Care Medicine*, Volume 26, 1998, 1793-1800.
- [9] Moreno R., Vincent J-L., Matos R. et al: The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study, 1999, *Intensive Care Med* 25, 686-696.
- [10] Fetter RB, Shin Y, Freeman JL, et al. Case mix definition by diagnosis-related groups, 1980, *Medical Care* 18, 1-53.
- [11] Le Gall JR, Lemeshow S, Saulnier F., A new simplified acute physiology score (SAPS II) base on an European/North American multicenter study, *JAMA* 1993;270:2957-2963.
- [12] Chapman, P., et al., CRISP 1.0 Process and User Guide. 2000, CRISP-DM Consortium (<http://www.crisp-dm.org>).
- [13] T. Kohonen, "Self-Organizing Maps," Springer-Verlag, 1995.

5.Viabilização de sistemas de suporte à decisão inteligente em Medicina Intensiva

Na próxima secção é apresentado o trabalho *INTCARE: A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine*, publicado em: Journal of Decision Systems,(2005) vol.14, n.º 3, GUPTA, Jatinder N.D., FORGIONNE, Guisseppi, GARRIDO, Leonardo, MORA Manuel (Eds.), Hermes science publications, Lavoisier, pp.241-259, ISBN: 2-7462-1405-9.

Pedro Gago, Manuel Filipe Santos, Álvaro Silva, Paulo Cortez, José Neves, Lopes Gomes

5.1. Preâmbulo

A ideia de conceber sistemas de suporte à decisão baseados na descoberta do conhecimento em bases de dados e no paradigma dos agentes, como uma forma resolver problemas complexos e dinâmicos, não é nova (Fayyad et al, 1996; Weiss 1999; Santos 1999). Todavia, uma grande parte destes conceitos e de arquitecturas necessitam de ser corroborados através de aplicações reais, a fim de se ter uma percepção da sua utilidade. Os sistemas de suporte à decisão são sistemas computacionais que suportam uma ou várias fases do processo de tomada de decisão individual, em equipa, e a nível organizacional ou inter-organizacional (Forgionne et al, 2000). A área dos sistemas de suporte à decisão constituiu uma das áreas de pesquisa mais comuns em sistemas de informação no passado, no entanto, um olhar mais próximo parece indicar que o interesse na área está em declínio (Claver et al, 2000) ou não está a aumentar à velocidade expectável (Vahidov et al, 2004). Devido às exigências colocadas aos instrumentos de apoio de decisão pelos ambientes dinâmicos e complexos (como é o caso das UCIs), a sua utilidade impõe que sejam situados e activos. Um sistema de suporte à decisão situado é similar a um sistema aberto na Inteligência Artificial, uma perspectiva que tem vindo a ganhar terreno desde os anos oitenta (Suchman 1987). Um sistema situado focaliza-se na interação com seu ambiente (Lueg et al, 1997), o seu comportamento resulta da combinação da finalidade para a qual foi concebido e de interacção com o ambiente (Rosenschein et al, 1995).

Segundo alguns autores, o futuro da pesquisa na área dos sistemas de suporte à decisão deverá passar pela adopção de métodos inteligentes como forma de possibilitar um suporte à decisão mais eficaz (Shim et al, 2002). A incorporação de técnicas da Inteligência Artificial, abre um novo espaço para um novo tipo,

potencialmente melhor sucedido de sistemas de suporte à decisão, chamado sistemas de suporte à decisão inteligentes

Neste trabalho introduz-se o INTCare, protótipo de um sistema de suporte à decisão inteligente no ambiente das UCIs. Em detalhe, o sistema pretende apoiar processo da tomada de decisão dos intensivistas:

- detectando condições que alertem para a necessidade de uma acção, através da análise dos dados automaticamente adquiridos e aplicando o modelo “mais relevante” para previsão da falência de seis órgãos e sistemas (hepático, respiratório, cardiovascular, hematológico, sistema nervoso central e renal), para o dia seguinte, durante a estadia na UCI;
- mantendo actualizada a previsão do risco de mortalidade intra-hospitalar actualizado;
- disponibilizando cenários da evolução clínica do doente grave.

Alguns dos modelos de classificação e previsão de *outcome* final (mortalidade intra-hospitalar) e de falência de órgão ou sistema, apresentados neste capítulo, foram integrados no INTCare com o objectivo da avaliação deste sistema, em “ambiente real”.

Neste protótipo a colheita dos dados é automática e contínua, sendo as fontes:

- monitores de cabeceira (frequência cardíaca, saturação arterial de O₂, e tensão arterial) à excepção do parâmetro débito urinário horário (registo manual), o que permite o cálculo automático dos eventos clínicos adversos de acordo com a definição estabelecida;
- aquisição de dados do processo clínico electrónico:
 - a idade (em anos);
 - o tipo de admissão (cirúrgica não programada; cirúrgica programada e médica)
 - o local de origem da admissão à UCI (bloco operatório, sala de recobro sala de emergência, enfermaria geral, outra UCI, outro hospital e outros locais;
 - o estado vital do doente (à data da alta hospitalar)

As variáveis seguintes, foram registadas manualmente:

- SAPS II score (às 24 horas após a admissão à UCI);
- SOFA score diário, por órgão e sistema.

A estrutura do sistema INTCare e o fluxo da informação estão ilustrados na Figura 2 do estudo. A partir dos dados colhidos manualmente e adquiridos automaticamente dos monitores de cabeceira, os dados são pré-processados pelo sistema e armazenados. A partir de modelos previamente gerados (resultantes dos trabalho apresentados neste

capítulo), é criada uma base de conhecimento que será utilizada para disponibilizar as previsões. Para a obtenção de uma previsão o sistema utiliza os dados do doente individualizado (ponto 2 da Figura 2), aos quais são aplicados os modelos de interesse (mortalidade ou falência de órgão) presentes na base de conhecimento (ponto 3 da figura). Para a avaliação de um cenário clínico é possível simular alterações dos parâmetros registados (ponto 4 da Figura 2), sendo a esses novos dados aplicados os modelos de interesse referidos (ponto 5 da Figura 2).

A actualização do sistema INTCare é automática, sendo iniciada na sequência de avaliação do desempenho dos modelos por comparação com os resultados observados. Na discordância pode ser eliminado o modelo em questão e substituído por um outro treinado com novos dados (ponto 1 da Figura 2).

Na Figura 4 visualiza-se a evolução SOFA score por órgão (observados e estimados pelo INTCare) bem como a probabilidade de mortalidade estimada pelo sistema à data da previsão.

Para uma amostra de 23 doentes admitidos consecutivamente no Serviço de Cuidados Intensivos do Hospital Geral de Santo António, foram avaliados os resultados dos modelos de interesse através do cálculo da precisão, especificidade e sensibilidade (Quadro 2).

Os resultados interessantes traduzidos pelos valores da precisão para a previsão de falência de órgão e mortalidade intra-hospitalar, por terem sido obtidos partir de uma população com um número reduzido de doentes, não são generalizáveis, servindo apenas como indicador das potencialidades do sistema.

5.2 INTCARE: A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine

Apresenta-se o trabalho em versão integral:

INTCare: A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine

Pedro Gago, Manuel Filipe Santos, Álvaro Silva, Paulo Cortez, José Neves, Lopes Gomes

ABSTRACT. This paper introduces the INTCare system, an intelligent information system based on a completely automated Knowledge Discovery process and on the Agents paradigm. The system was designed to work in Hospital Intensive Care Units,

supporting the physicians' decisions by means of prognostic Data Mining models. In particular, these techniques were used to predict organ failure and mortality assessment. The main intention is to change the current reactive behaviour to a proactive one, enhancing the quality of service. Current applications and experimentations, the functional and structural aspects, and technological options are presented.

RÉSUMÉ. Cet article présente INTCare, un système d'information intelligent basé sur un procédé complètement automatisé de découverte de la connaissance et sur le paradigme d'agents. Le système a été conçu pour fonctionner dans les services hospitaliers de réanimation, soutenant les décisions des médecins au moyen de modèles d'extraction de données pronostiques. En particulier, ces techniques ont été employées pour prévoir l'échec d'organes et l'évaluation de la mortalité. La principale intention est de changer le comportement réactif habituel en proactif augmentant la qualité du service. Des applications et les expérimentations courantes, les aspects fonctionnels et structuraux, et des options technologiques sont présentés.

KEY WORDS: intelligent decision support, knowledge discovery, intensive care

MOTS-CLÉS: appui de décision intelligent, découverte de la connaissance, soin intensif

1. Introduction

The idea of conceiving Decision Making Support Systems (DMSS) making use of the Knowledge Discovery from Databases (KDD) and Agent-Based Systems paradigms, as a way to solve complex and dynamic problems, is not new (Fayyad et al., 1996; Weiss 1999; Santos 1999). However, a great part of these concepts (and architectures) need to be corroborated by real-world applications, in order to obtain a valuable feedback of its effective use. The intensive care medicine, due to its high volume and complexity of data, is a rich field to test this thesis.

DMSSs are computer-based systems that support one or several phases of the individual, team organizational or inter-organizational decision-making process (Forgionne et al., 2000). While DMSS constituted one of the most popular areas of research in information systems in the past, a closer look seems to indicate that the interest in DMSS is declining (Claver et al., 2000) or is not increasing at the rate that was expected (Vahidov et al., 2004). This is due to the requirements for decision support tools posed by new dynamic and complex environments (like the one in an ICU), demanding for situated and active DMSSs. A situated DMSS is similar to an open system in Artificial Intelligence (AI), a "view" that has been gaining popularity since the

late 1980s (Suchman 1987). In a situated system the focus is on the interaction with its environment (Lueg et al., 1997), the behaviour is the combined result of its purpose and its interaction with the environment (Rosenschein et al., 1995).

Other concerns come from the extended Keen's agenda for DMSS research to the year 2007 (Shim et al., 2002), which postulates that researchers and developers should be more prescriptive about effective decision making by using intelligent systems and methods. In effect, the incorporation of Artificial Intelligence techniques, opened room for a new and more successful kind of DMSS, called intelligent decision making support systems (i-DMSS).

This paper introduces the INTCare, an i-DMSS based on the KDD and the Agent-Based paradigms, to support intensive care medical activities. In particular, the system assists the physicians' decision making by: (i) detecting action demanding conditions by continuously scanning automatically acquired data and applying the relevant model to predict next day failure of six systems (liver, respiratory, cardiovascular, coagulation, central nervous and renal); (ii) maintaining an up-to-date in-hospital death probability value used in end-of-life decision making and (iii) evaluating scenarios for the evolution of the patient's condition, allowing physicians to compare the consequences of different medical procedures. The INTCare system is being tested in a real environment, the Intensive Care Unit (ICU) of Hospital Geral de Santo António (HGSA), Oporto, north of Portugal.

The following sections will introduce the problem definition and the state of the art, followed by the INTCare system formal description in terms of the Agent-Based paradigm. Then, the obtained results will be presented by means of their usability and efficiency. Next, a discussion will be set over the critical aspects of the INTCare system. Finally, the paper will be concluded, showing also perspectives of future work.

2. Theoretical Backgrounds and Related Work

2.1 *Decision Support Systems and Intensive Care*

In the beginning of the 80s, several expert systems were developed for medicine, such as MYCIN (Buchanan et al., 1984), CASNET (Kulikowski et al., 1982) and CADUCEUS (Pople 1985), just to name a few. A decade later, in the 90s, the common knowledge was that these expert systems, based exclusively on information retrieved from experts, were not sufficient for the complex real-world problems (Lavrac et al., 1997). A shift was

induced, where the main emphasis was the gathering of knowledge directly from the data, using intelligent data analysis. At the end of the 90s, this approach gained an increasing interest, especially in the medicine area, due to the high volume and complexity of the clinical data. In terms of intensive care medicine, the application of DM techniques (e.g. Artificial Neural Networks) is relatively new, although possessing a huge potential (Hanson et al., 2001). However, it should be stressed that the number of studies where these techniques are applied in real environments is very limited.

Everyday, intensivists have to make decisions that seriously affect their patients. First, they have to reach a diagnosis and then conceive therapeutic plans (e.g. prescribe treatments) in order to improve the patients' condition. However, ICU physicians have to make decisions that are even more challenging, such as the ones related to life-support treatments. Resource availability limitations force them to make sure that intensive care is applied only to those who are likely to benefit from it. Critical decisions include interrupting life-support treatments and writing do-not-resuscitate orders when intensive care is considered futile. In addition, the condition of some ICU patients is such that physicians are not able to completely assess it and will thus benefit from having extra quality information available.

A synopsis of the current decision making process in the ICU of the HGSA is presented in Figure 1. This process can be seen as unfolding in four phases (Simon 1977; Mora et al., 2003; Turban et al., 2004):

Intelligence - The medical and nursing staff collect physiological data regarding the patient's condition. Then, based on this information, the physicians evaluate the illness state to determine the pertinence of immediate action;

Design - Physicians conceive the possible therapeutic scenarios. Two main categories of scenarios can be considered: immediate action assumption (where there is a pressing need for intervention) and delayed group decision (guidelines for discharge or long term treatments). In the former case, if needed, the physician can consult internal or external colleagues. Delayed group decision choices include: patient's discharge, suspension of life-supporting treatment, writing do-not-resuscitate orders and applying a specific treatment;

Choice – After the identification of the alternatives and entwined with the previous stage, a final decision is taken (either individually or group based); and

Implementation – The corresponding therapeutic procedure is applied.

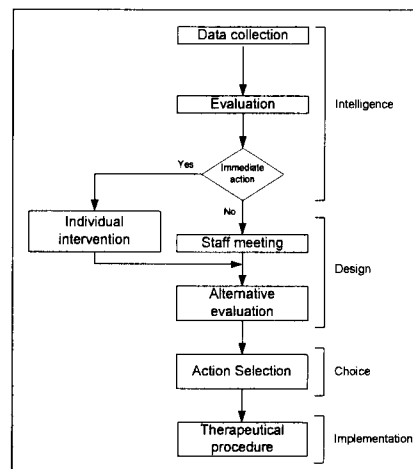


Figure 1. *The HGSA ICU decision-making process*

At the current stage, physicians rely on their experience and on the available clinical data (e.g. patients chart sheets) to determine if there is a need for action. The same information is used when taking decisions about the course of action to be followed. No use is made of the vast amounts of data, which are automatically collected through bed-side monitors. To the authors' knowledge, although several prognostic scores have been developed (e.g. SAPS II (Le Gall et al., 1993) or APACHE II (Knaus et al., 1991)), there are no intelligent DSS systems completely developed and based on the KDD paradigm in use in ICUs. Yet, both patient and physicians will benefit if high-level, reliable and timely information is available. Hence, the INTCare system aims at filling this gap. The latest version of the system supports the intelligence phase by continuously scanning (sensing the environment) the available information in order to uncover action-demanding situations. In addition, the design phase will be supported by evaluation of the consequences of the available therapeutic procedures. With the information gathered in the previous step, physicians can choose the procedure to follow. This is a particularly sensitive area as strong physician resistance is expected (Barnato et al., 2004). Finally, whenever allowed, INTCare can act on the environment controlling the equipment (e.g. infusion pumps).

2.2. Knowledge Discovery from Databases and Data Mining

The interest in Knowledge Discovery from Databases (KDD) and Data Mining (DM) arose due to the rapid emergence of electronic data management methods. In 1997, the Gartner Group suggested that DM is one of the top five key technologies that will have a major impact in the industry within the next years (Lee et al., 2001). In effect, these techniques are now widespread and applications can be found in diverse areas,

such as: marketing (Kim et al., 2004), banking (Hsieh 2004), manufacturing and production (Musaev 2004), brokerage and securities trading (Lam 2004), and health care (Li et al., 2004). References for other applications can be found in various publications, including (Turban et al., 2004; Santos et al., 2005) and (Liao 2003).

Within the Medicine arena, huge databases, with large, complex and multi-source information (e.g. text, images or numerical data), are commonplace. However, human experts are limited and may overlook important details. Furthermore, the classical data analysis (e.g. logistic regression) breaks down when such vast amounts of data are present. Hence, an alternative is to use automated discovery tools to analyse the raw data and extract high level information for the decision-maker (Hand et al., 2001).

An important step of the KDD process is related to the choice of the DM function, which will affect the selection of the DM algorithm (Fayyad et al., 1996). In an ICU setting we consider the most useful DM goals to be:

- *classification* - labels a given set of attributes into one of several predefined classes (e.g. diagnosing a disease according to the patient's symptoms);
- *regression* - maps a data item into a real-value variable (e.g. estimation of the patient's heart rate); and
- *clustering* - searches for natural groupings of objects based on similarity measures (e.g. segmenting patients into clusters according to similar profiles).

The above goals may involve the application of Machine Learning algorithms such as Decision Trees (DT) (Quinlan 1986; Lavrac et al., 1997), Learning Classifier Systems (LCS) (Lanzi et al., 2000; Santos 1999) and Artificial Neural Networks (ANN) (Haykin 1999; Dybowski 2000). More recently, there has been an emergent DM research area that involves the use of ensembles for supervised learning, where a set of classification/regression models are combined in some way to produce an answer (Dietterich 2001).

2.3. Agent Based Systems

The term agent is a metaphor allowing various definitions, interpretations and taxonomies. Actually, no one of them is universally accepted, despite this some positions are considered referential (Jennings et al., 1998; Nwama 1998; Weiss 1999; Ferber 1999). One of the most comprehensive definitions of agent was proffered by (Jennings et al., 1998) and is based on a strong and weak view. Applications of agents

are widespread and can be found in travel planning (Yim et al., 2004), e-commerce (Garcia-Serrano et al., 2004) and even as components of DMSSs (Hess et al., 2000).

In the context of this work, the AIMA definition prosecuted by (Russel et al., 2003) was adopted, stating that an agent is an entity capable of perceiving the environment and actuating on that environment. From a software engineering point of view, an agent is an abstraction that allows the construction of more complex systems designated by Agent-Based Systems or Agencies.

The requisites inherent to a situated and active system, in addition to the previous experience accumulated in the conception of agencies such as the DICE system (Santos 1999), and the AIDA agency (Abelha et al., 2003), dictated the adoption of the agent technology in the development of INTCare.

The various types of agents integrating the INTCare system are presented and discussed in the section 3.

2.4. Previous Work

Artificial Intelligence techniques have been used in order to enhance DMSSs. One can find examples of ANN and genetic algorithms (Kim et al., 2004), DT (Tsang et al., 2004), fuzzy logic (Lin et al., 2004) and DM (Michalowski et al., 2003) embedded in decision support systems. Likewise, agents are becoming commonplace as they provide a useful abstraction that facilitates systems conceptualisation. An example of use of agents in the medical area is the AMPLIA system (Vicari et al., 2003). In an ICU setting the Guardian system (Hayes-Roth et al., 1998) monitors and makes diagnoses of intensive-care patients. Another system, the Rihad ICU Program (Chang et al., 1988) was used to predict death. In our previous work, several DM models were applied to ICU data (Santos et al., 2002; Santos et al., 2005b; Silva et al., 2005).

These studies used off-line learning, where all data was stored first and then accessed repeatedly by the DM algorithms. The clinical data was collected during the EURICUS II research programme (Miranda et al., 2003), which involved a massive study in 42 ICUs from 9 countries during a period of 10 months, from 1998 to 1999. The database included thousands of daily records related to bedside measurements of critical ill patients, including features such as: the case mix - an information that remains unchanged during the patient's stay in the ICU (e.g. age or admission origin); the intermediate outcomes - being triggered from four monitored biometrics (e.g. the systolic blood pressure or urine output); and the patient's state - based on daily organ

failure scores (e.g. SOFA index (Vincent et al., 1998)) and the final outcome (death/no death). Intermediate outcomes are represented by the number of Events and Critical Events per day – the number of daily occurrences of values out of the established limits for four physiologic variables that are monitored continuously. These four variables are the Heart Rate, the Systolic Blood Pressure, the Oxygen Saturation and the Urine Output. A group of clinical specialists determined the intervals considered normal for each one of these parameters (Silva et al., 2005).

This vast amount of information was modelled by:

- A LCS to predict the length of stay (short or prolonged) and the outcome (death/no death) of patients from the HGSA ICU, attaining an accuracy of 72% (Santos et al., 2002);
- A clustering framework to organ dysfunction diagnosis (Santos et al., 2005b), where patients were segmented into clusters (using a Self-Organizing Map) and then each cluster was processed with the C.5 DT algorithm (the classification accuracies of the clusters ranged from 74% to 98%); and
- An ANN approach (Silva et al., 2005), where easily acquired daily inputs (the intermediate outcomes) were fed into Multilayer Perceptrons, in order to predict the failure of six organs and systems (e.g. hepatic or respiratory), achieving an overall accuracy of 70%.

The outputs of the organ failure prediction models are measured by the SOFA (Sequential Organ Failure Assessment) score, which provides values within the range (Hospital Statistics) (normal function) {1, 2} (dysfunction) and {3, 4} (failure). As for mortality assessment, the value one indicates hospital mortality while zero represents survival. Intermediate values can be read as probabilities of occurrence.

The results obtained proved the usefulness of the DM approach but, at the same time, suggested the need for more work in order to gain:

- real-time data collection from bed-side monitors;
- real -time outcome and intermediate outcomes prediction;
- scenario evaluation;
- increased predictive accuracy over time;
- explanation capabilities;
- simulation of therapeutic procedures consequences;
- control of effectors (e.g. infusion pumps).

To achieve these objectives the INTCare project was started. Due to its dimension and the particularities of the medical environment an incremental development was judged to be adequate. At the present time, the three first points are completed.

3. INTCare System

The INTCare is an agent-based system, composed by several semi-autonomous agents in charge for the functionalities inherent to the system. Conceptually, it can be viewed as set of four subsystems (Figure 2): Data Entry, Knowledge Management, Inference and Interface. Formally, the INTCare system is defined as a tuple $\Xi \equiv \langle C_{INTCare}, \Delta_{INTCare}, a_{pp}, a_{cde}, a_{dm}, a_{pf}, a_{mi}, a_{dr}, a_{pd}, a_{sc}, a_{int}, a_{ic} \rangle$, where:

$C_{INTCare}$ is the context and corresponds to a logical theory, represented as a triple $\langle Lg, Ax, \Delta \rangle$, where Lg stands for an extension to the language of programming logic, Ax is a set of axioms over Lg , and Δ is a set of inference rules;

$\Delta_{INTCare}$ is the set of bridge rules defining the interaction among the systems' components (the agents);

$a_{pp}, \dots, a_{ic}$ are the system's agents.

This formalism corresponds to a logical framework, suitable to specify agent-oriented systems based on the notion of context logic, and some properties of object-oriented design such abstraction, encapsulation, modularity and hierarchy (Santos 1999). In this work, the agents are represented as logical theories with a specific context (different agents may involve different contexts). Several agents (i.e. contexts) can be put together and be able to reason about the behaviour of the entire system as a (heterogeneous) logical theory. A set of special rules called *bridge-rules* is applied to provide the interface among agents and systems of agents. These rules describe the agents' reactions to events occurring in their environment. The agents include a set of event types, and a set of time points. Next, the overall system is described making use of this formalism, explaining it in some detail.

3.1. Description of the Agents

The INTCare agents follow the subordinated architecture (Weiss 1999), they are social computational entities, semi-autonomous, reactive, with internal-state, and pre-defined goals incited by the system creators, whose activity contributes to the goal of the overall system. In a technical perspective the INTCare agents are high granularity objects



Model Language (PMML), which is a XML based model description language (URL1 2004);

Admission Data				SAPS Data		SOFA Data		Discharge Data		Urine Output	
AIDS	☐	Temperature	38.2	°C	Glasgow	12					
Metastatic cancer	☐	WBC	18.7	1000 / ml							
Hematologic malignancy	☐	Bilirubin	0.8	mg / dl							
Heart rate	115	# / min		Urine output	2.45	l / day					
Systolic BP	130	mmHg		Serum Urea or BUN	65	mg / dl					
Mechanic Ventilation / CPAV	☐	Potassium	3.1	mmol / l							
FiO2		%		Sodium	131	mmol / l					
PaO2		mmHg		HCO3-	25	mmol / l					

Figure 3. Clinical Data Entry agent User Interface

Performance (a_{pf}) agent continually consults, in a proactive way, the Data Warehouse for updates that allow statistics collection (e.g. discharge data that may or may not confirm a prediction made), as a base to calculate a set of assessment parameters maintained in the Performance Database. The evaluation metrics include classification accuracy, sensitivity and specificity values (Silva et al 2005). These statistics are updated every time new relevant information is collected. Whenever the collected statistics show that the performance has fallen below a predefined threshold (a configuration parameter) a new model is requested in order to allow the replacement of the poor performing one ($message_dm$);

Model Initialization (a_{mi}) agent populates the Knowledge Base with the models obtained from off-line training. This agent is currently used only when first starting INTCare. As external models may be represented in a PMML version that is not used in INTCare, this agent must perform a conversion before loading the models into the Knowledge Base;

Data Retrieval (a_{dr}) agent is an information agent, whose only objective is that of retrieving, from the Data Warehouse, the information requested by the interface agent (get_data) and returning it ($send_data$);

Prediction (a_{pd}) agent answers user questions by applying the adequate models (get_models , $predict$) contained in the Knowledge Base to the data stored in the Data

Warehouse. Next, results are sent back to the calling agent (*send_data*) and if the calling agent is the interface agent, the prediction made is recorded into the Data Warehouse (*update_dw*);

Scenario Evaluation (a_{sc}) agent makes it possible to the physician to create and evaluate what-if scenarios. After receiving the data from the interface agent, the scenario evaluation agent requests a forecast from the prediction agent (*re_prediction*), the scenario is then stored in the Scenarios Database (*store_scenario*) and the result is sent back to the interface agent (*send_result*);

Interface (a_{int}) agent allows (web-based) interaction with the system by providing an easy way for physicians to request prognostics and evaluate scenarios. Whenever new data is needed, this agent messages the data retrieval agent (*message_dr*). In Figure 4 a screen shot of the prediction window (*re_prediction*) is presented, where it is possible to consult the prognostics for a single patient and at the same time compare the past SOFA predictions with the real values. A holistic view of the patient is supported by presenting in the same chart the information regarding the diverse organs and systems. Also, the values predicted for the next day are shown, providing information that may suggest the need for immediate action. For the past days, the actual SOFA score is shown, together with the value predicted for the next day. It is also possible to consult the confidence levels associated with the prediction values (confidence button);

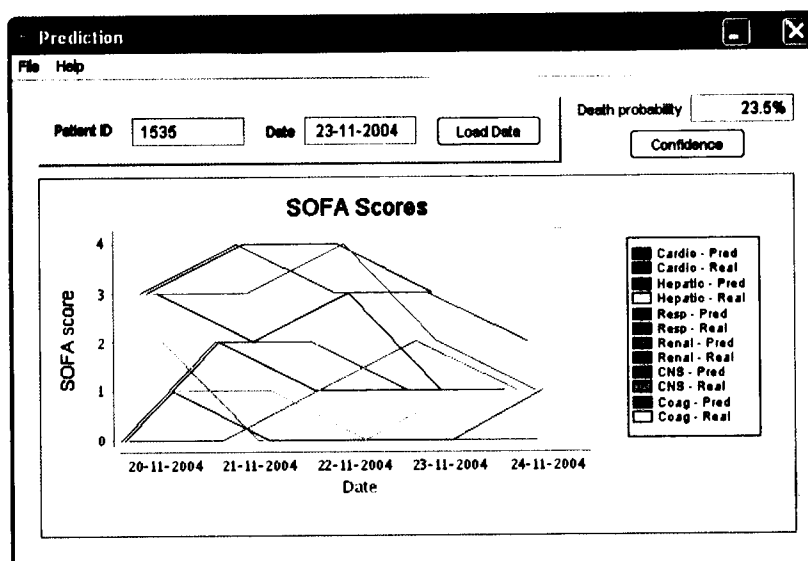


Figure 4. INTCare User Interface

Ensemble agent intended to enhance predictive performance by combining several models (e.g. using a majority voting scheme) in order to produce an efficient answer (Dietterich 2001);

Connection Agent (a_{ic}) is in charge for the knowledge interchange among the various instances of the INTCare system running in the ICUs.

The interaction among the INTCare's agents is represented in terms of the bridge rules $\Delta_{INTCare}$ described in Table 1 (where t stands for the system's time cycle). The numbers posted in the left of the rules were used in Figure 1 to label the arrows representing the interactions.

	Bridge-Rule	Description
1	$a_{pf}: occurs(message_dm, t)$ $a_{dm}: [occurs(read_dw, t) \wedge$ $occurs(train, t) \wedge$ $occurs(update_models, t)]$	When the <i>Performance</i> agent detects the need for the replacement of a model, the <i>Data Mining</i> agent is messaged, reads the available data in the Data Warehouse, trains a new model and finally saves it into the Knowledge Base.
2	$a_{int}: occurs(message_dr, t)$ $a_{dr}: [occurs(get_data, t) \wedge$ $occurs(send_data, t)]$	When triggered by the <i>Interface</i> agent, the <i>Data Retrieval</i> agent gets the patient's data from the Data Warehouse and sends it back to the <i>Interface</i> agent.
3	$a_{int}: occurs(re_prediction, t)$ $a_{pd}: [occurs(get_models, t) \wedge$ $occurs(predict, t) \wedge$ $occurs(send_data, t) \wedge$ $occurs(update_dw, t)]$	When the <i>Interface</i> agent requests a forecast, the <i>Prediction</i> agent gets the adequate model from the Knowledge Base applies it and finally sends the results back to the <i>Interface</i> agent and updates the Data Warehouse.
4	$a_{int}: occurs(re_scenario, t)$ $a_{sce}: [occurs(re_prediction, t) \wedge$ $occurs(store_scenario, t) \wedge$ $occurs(send_result, t)]$	When the <i>Interface</i> agent requests a scenario evaluation it sends the necessary data to the <i>Scenario Evaluation</i> agent that then requests the forecast from the <i>Prediction</i> agent, stores the scenario in the Scenarios Database and finally returns the scenario results to the <i>Interface</i> agent.
5	$a_{sce}: occurs(re_prediction, t)$ $a_{pd}: [occurs(get_models, t) \wedge$ $occurs(predict, t) \wedge$ $occurs(send_data, t)]$	When the <i>Scenario Evaluation</i> agent requests a forecast, the <i>Prediction</i> agent gets the adequate model(s) from the Knowledge Base applies it and finally sends the results back to the requesting agent.

Table 1. INTCare's bridge-rules

4. Deployments and Experimentation

INTCare was programmed using the Visual Basic .Net²² environment, PMML (to represent the models in the Knowledge Base), the Xelopes library (prudsys 2004) and Jbuilder. The system is currently under use in the ICU of the HGSA in Oporto, Portugal. This unit contains ten monitored beds, treating around 400 patients each year. Regarding the deployment, it was planned to occur in three separate steps. First, the system started to be used with limited data (acquired from a single bed-side monitor). This allowed for a less invasive introduction in the operating environment, as only two physicians were involved at this time. The next step extends data acquisition to the ten available monitors and widespread use of the system by the physicians working in this ICU. Finally, INTCare will be extended to two other ICUs within the same region of influence.

The current configuration includes one bed-side monitor (Datex Ohmeda S5) and two personal computers running under the MS-Windows operating system. One of the PCs is dedicated to data collection having installed the Collect 5 software from Datex Ohmeda; the other is where INTCare is available for use by the physicians. A TCP/IP network allows communication between the diverse components. All data acquisition software runs under Windows.

There is a model for each of the six organic systems considered and for each model nine input variables were used: admission type, admission origin, age of the patient, SAPS II score, number of adverse events (number of Events + number of Critical Events for each of the four continuously monitored variables) and the SOFA value for the previous day. Only five variables are used in the outcome prediction model: SAPS II score and the number of adverse events in the first 24 hours of ICU stay. The back-propagation algorithm (Rumelhart et al., 1986) was used for training all networks based on the data collected during the EURICOS II study.

In order to assess the quality of the predictions the confusion matrix (Kohavi et al., 1998) was used. This is a matrix of size $L \times L$, where L denotes the number of possible classes (*domain*). This matrix is created by matching the *predicted (test result)* and *actual (patients real condition)* values. When $L = 2$, there are four possibilities: the number of *True Negative (TN)*, *False Positive (FP)*, *False Negative (FN)* and *True Positive (TP)* classifications. From the matrix, three accuracy measures can be defined (Essex-Sorlie

²² All other trademarks or registered trademarks referred in this document are property of the respective owners.

1995): the *Sensitivity* (also known as *recall* and *Type II Error*); the *Specificity* (also known as *precision* and *Type I Error*); and the *Accuracy*, which gives an overall evaluation. These metrics can be computed using the following equations:

$$\text{Sensitivity} = TP / (FN+TP) \times 100 (\%) \quad [1]$$

$$\text{Specificity} = TN / (TN+FP) \times 100 (\%) \quad [2]$$

$$\text{Accuracy} = (TN+TP) / (TN+FP+FN+TP) \times 100 (\%) \quad [3]$$

Experimentation has started as planned, with data acquisition from one bed-side monitor, therefore the information is available for a limited number of patients and consequently the results can not be considered statistically meaningful. The available results are given in Table 2.

Prediction	Accuracy	Sensitivity	Specificity
In-hospital mortality	80.8	60.0	85.7
Respiratory	60.5	54.5	68.8
Coagulation	94.7	80.0	97.0
Hepatic	86.8	70.0	100
Cardiovascular	94.7	0.0	97.3
Central Nervous System	100	n.a.	100
Renal	97.4	96.2	100

Table 2. Prediction results (values in percentage)

A good classifier should present both high Sensitivity and Specificity values. Although some of the results attained can be considered excellent (e.g. Coagulation and Renal systems), it should be stressed that they are not meaningful due to the low number of patients considered. This lack of data also prevented the retraining of the models. We could not find other SOFA prediction systems in order to compare our results.

Nevertheless, our in-hospital death prediction results are comparable to those of the Rihad ICU Program (RCIP) even though INTCare always gives a death/no-death result whereas RICP reports “death” or “outcome unknown”, thus reducing the number of false positive answers. Nevertheless, the feedback received from the physicians in relation to the system usage is positive, the predictions were considered “interesting”. In the future, the system is expected to play a significant role in the ICU decision procedures.

5. Discussion and Technological Impact

Currently, intensive care decision-making is based on the individual physician's previous empirical experience. In the majority of ICUs, patients are continuously monitored with several biometric sensors. Yet, this information is not fully taken into the decision process. This scenario changes with the introduction of the INTCare system, where all this data is stored and made available. Furthermore, the INTCare is enriched with DM prognostic models, which can provide forecasts of both intermediate and final outcomes. In addition, the physicians may also evaluate different scenarios; i.e. to test what would occur if a given condition is changed.

Previous results (Santos et al., 2005b; Silva et al., 2005) show the reliability of the DM prognostic models. Nevertheless, introduction of model-based outcome prediction has to overtake some problems, including: reliability, availability and physician resistance (Barnato et al., 2004). In INTCare, the availability problem is partly solved by using automatic data acquisition techniques. Reliability is expected to increase when the ensemble agent is completed. Physician resistance is expected to soften as the system is to be used as a support tool and its performance is expected to improve with the feedback taken from the physicians.

A computer-based system has other advantages such as not being subject to between-physician variability in prognostic skill or beliefs regarding the construct of “futility”. Furthermore model-based predictions may be more equitable because they do not incorporate value-based judgments regarding the “worth” of one life over another (Barnato et al., 2004). The usability (interaction with humans) and technology acceptance results are presented in Table 3:

Table 3. Expected impact of the system

Situation	Present	Expected
access to patient's records	paper records	digital records
comparing present and old patient's data	someone has to find the right file – manual search	computer aided
decision making	during a physicians meeting	INTCare is another expert in the physicians meeting
scenario evaluation	not available	supported

6. Conclusions and further work

This paper presented an Agent-Based System called INTCare, for ICU real time Decision Support on clinical actuation using a KDD/DM approach for automatic knowledge acquisition. The architecture and functionalities were defined in terms of a formal framework. The major contributions of this approach are:

- The system is being tested in a real-world environment which means that it is tailored to the physician needs;
- A better overall predictive performance is enabled by the competition and combination of heterogeneous DM algorithms;
- The system implements all the KDD steps in an automatic way, leading to an autonomous situated Intelligent Information System;
- The Data Mining process is enhanced by the construction of a Data Warehouse which is automatically filled by a Pre-Processing agent.

The INTCare system is currently being tested in the ICU of HGSA. The preliminary results even promising are not absolutely credible due the limited number of cases recorded. Further experiments will consider a greater volume of monitored data and cases analysing

its impact in terms of on-line and incremental learning. Efforts will be made towards a major performance through the system tuning, the ensemble agent exploitation, involving more ICUs (from the same hospital and from other hospitals in Portugal or even from other European countries) and the inclusion of new variables in the prediction models (e.g. data regarding patient ventilation). Finally we will pay attention to the therapeutic procedures, covering the choice and design phases of the decision making process by suggesting possible actions and predicting their consequences. It is also planned the control of medical equipment (e.g. infusion pumps) automating the implementation phase.

Acknowledgements

The authors would like to thank the Director of the HGSA ICU Dr. Fernando Rua for the unconditional support and for creating the conditions for the deployment of INTCare. We would also like to thank FRICE for the provision of part of the EURICUS II data.

References

- Abelha S., Machado J., Santos M., Allegro S., Rua F., Paiva M. and J. Neves, "Agency for Integration, Diffusion and Archive of Medical Information", in *proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, Benalmádena, Spain, September 2003.
- Barnato A. and Angus D., "Value and role of intensive care unit outcome prediction models in end-of-life decision making". *Crit Care Clin*, vol. 20, no. 3, 2004, p. 345-362.
- Buchanan B.G. and Shortliffe E.H. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley, 1984.
- Chang R., Jacobs S., Lee B., "Predicting outcome among intensive care unit patients using computerised trend analysis of daily Apache II scores corrected for organ system failure". *Intensive Care Med* vol. 14, no. 5, 1988, p. 558-566.
- Claver E., Gonzales R., Llopis J., "An analysis of research in information systems (1981-1997)", *Information and Management*, vol. 37, no. 4, 2000, p. 181-195.
- Dietterich T., "Ensemble methods in machine learning". In: Kittler and Roli (eds), *Multiple Classifier Systems*, 2001, LNCS 1857, Springer, pp 1-15

- Dybowski R., "Neural Computation in Medicine: Perspectives and Prospects". In: Malmgreen et al. (eds) *Proceedings of the ANNIMAB-1 Conference (Artificial Neural Networks in Medicine and Biology)*. 2000, Springer, Berlin Heidelberg New York, pp 26-36
- Essex-Sorlie D. *Medical Biostatistics & Epidemiology: Examination & Board Review*. McGraw- Hill/Appleton & Lange, International edition, 1995.
- Fayyad U., Piatetsky-Shapiro G., Smyth P., "From Data Mining to Knowledge Discovery: An Overview". In Fayyad et al. (eds) *Advances in Knowledge Discovery and Data Mining*. 1996, AAAI Press / The MIT Press, Cambridge MA, pp 471-493.
- Ferber J. *Multi-Agent Systems – An Introduction to Distributed Artificial Intelligence*, Addison-Wesley, 1999.
- Forgionne G.A., Mora M., Cervantes F., Kohli R., "Development of integrated decision making support systems: a practical approach". In *Proceedings of the AIS Conference*, 2000, Long Beach CA, pp 2132-2134.
- Garcia-Serrano A.M., Martinez P., Hernandez J.Z., "Using AI techniques to support advanced interaction capabilities in a virtual assistant for e-commerce". *Expert Systems with Applications*, vol. 26, no. 3, 2004, p. 413-426.
- Hand D., Mannila H., Smyth P., *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.
- Hanson C., Marshall B., "Artificial intelligence applications in the intensive care unit". *Crit Care Med*, vol. 29, no. 2, 2001, p. 427-435.
- Hayes-Roth B., Larsson J.E., "Guardian: an intelligent autonomous agent for medical monitoring and diagnosis". *IEEE Intelligent Syst* 1998(January/February):58-64
- Haykin S., *Neural Networks – A Comprehensive Foundation (2nd Edition)*. Prentice-Hall, New Jersey, 1999.
- Hess T.J., Rees, L.P., Rakes T.R., "Using autonomous software agents to create next generation of decision support systems". *Decision Sciences*, vol. 31, no. 1, 2000, p. 1-31.
- Hsieh N.C., "An integrated data mining and behavioural scoring model for analyzing bank customers". *Expert Systems with Applications*, vol. 27, no. 4, 2004, p. 623-633.
- Jennings R., Wooldridge J., *Agent Technology Foundations, Applications and Markets*. Springer, Berlin Heidelberg New York, 1998.
- Knaus W.A. et al., "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults". *Chest*, vol. 100, no. 6, 1991, p. 1619-1636.

- Kim Y. and Street W.N., "An intelligent system for customer targeting: a data mining approach". *Decision Support Systems*, vol. 37, no. 2, 2004, p. 215-228.
- Kohavi R., Provost F., "Glossary of terms". *Machine Learning*, vol. 30, 1998, p.271-274.
- Kulikowski C.A., Weis S.M., "Representation of expert knowledge for consultation: the CASNET and EXPERT projects", in: P. Szolovits (ed.), *Artificial Intelligence in Medicine*, 1982, Westview Press, Boulder, 21-56.
- Lam M., "Neural network techniques for financial performance prediction: integrating fundamental and technical analysis". *Decision Support Systems*, vol. 37, no. 4, 2004, p. 567-581.
- Lanzi P., Solzman W., Wilson S., "Learning Classifier Systems from Foundations to Applications". *Lecture Notes in Artificial Intelligence 1813*, 2000, Springer, Berlin Heidelberg New York.
- Lavrac N., Keravnou E., Zupan B., *Intelligent Data Analysis in Medicine and Pharmacology*. Kluwer, Boston, 1997.
- Le Gall J.R., Lemeshow S., Saulnier F., "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study". *JAMA*, vol. 270, no. 24, 1993, p. 2957-2963.
- Lee S. and Siau K., "A review of data mining techniques". *Industrial Management & Data Systems*, vol. 101, no. 1, 2001, p. 41-46.
- Li L., Tang H. , et al., "Data mining techniques for cancer detection using serum proteomic profiling". *Artificial Intelligence in Medicine*, vol. 32, no. 2, 2004, p. 71-83.
- Lin, C... and Hsieh P-J., "A fuzzy decision support system for strategic portfolio management." *Decision Support Systems*, vol. 38, no. 3, 2004, p. 383-398.
- Liao S-h., "Knowledge management technologies and applications-literature review from 1995 to 2002". *Expert Systems with Applications*, vol. 25, no. 2, 2003, p. 155-164.
- Lueg C., Pfeifer R., "Cognition, situatedness, and situated design". *Second International Conference on Cognitive Technology*, 1997, IEEE Computer Society, Aizu, Japan
- Michalowski W., Rubin S., et al., "Mobile clinical support system for paediatric emergencies". *Decision Support Systems*, vol. 36, no. 2, 2003, p. 161-176.
- Miranda D., Nat R., Nijk A., Schaufeling W., Lapichino G., "Nursing activities score". *Crit Care Med*, vol. 31, no. 2, 2003, p. 374-382.

- Mora M., Forgionne, J. and Gupta J., *Decision Making Support Systems – Achievements and Challenges for the new decade*, Idea Group Publishing, 2003.
- Musaev A., "Analytic information technologies in oil refinery". *Expert Systems with Applications*, vol. 26, no.1, 2004, p. 81-85.
- Pople H.E., "Evolution of an Expert System: from INTERNIST to CADUCEUS". In: *Artificial Intelligence in Medicine* (ed. by I. De Lotto and M. Stefanelli), 1985, p. 179-208. Elsevier Science Publisher, Amsterdam.
- Prudsys AG., *Xeli's Intro - Introduction to XELOPES Version 1.2.3* (Xelopes Users Manual), 2004.
- Quinlan J., "Induction of Decision Trees". *Machine Learning*, vol. 1, no.1, 1986, p. 81-106.
- Rosenschein S.J., Kaelbling L.P., "A situated view of representation and control". *Artificial Intelligence*, vol. 73, no. 1-2, 1995, p. 149-173.
- Rumelhart D.E., Hinton G.E. and Williams R.J., "Learning internal representations by error propagation". In: Rumelhart, D.E. and McClelland, J.L., eds. *Parallel distributed processing, vol. 1: Foundations*. 1986, Cambridge MIT Press
- Russel S. and Norvig P., *Artificial Intelligence – A Modern Approach 2nd Edition*. Prentice Hall, New Jersey, 2003.
- Santos, M.F., *Sistemas de Classificação em Ambientes Distribuídos*. Ph.D. thesis, Universidade do Minho (in Portuguese), 1999.
- Santos M.F., et al., "Augmented Data Mining over Clinical Databases: Using Learning Classifier Systems". In: *Proceedings of the Fourth International Conference on Enterprise Information Systems (ICEIS 2002)*, INSTICC, p. 512-516.
- Santos M.F. and Azevedo C., *Data Mining – Descoberta de Conhecimento em Bases de Dados*, FCA Editora, Lisbon (in Portuguese), 2005.
- Santos M.F., Pereira J., Silva A., "A cluster framework for data mining models – an application to intensive medicine". In: Chen C-S. et al. (eds) *Proceedings of the seventh International Conference on Enterprise Information Systems (ICEIS 2005)*, INSTICC, p. 163-168.
- Shim J.P. et al., "Past, Present, and future of decision support technology". *Decision Support Systems*, vol. 33, no. 2, 2002, p. 111-126.

Silva A., Cortez P., Santos M.F., Gomes L., Neves J., "Multiple organ failure diagnosis using adverse events and neural networks". In: Seruca I. et al. (eds), *Enterprise Information Systems VI*, 2005, Springer.

Simon H., *The New Science of Management Decisions*, Rev. ed. Englewood Cliffs. Prentice Hall, New Jersey, 1977.

Suchman L., *Plans and Situated Actions: The Problem of Human -Machine Communication*, Cambridge Univ. Press, Cambridge, MA, 1987.

Tsang E., Yung P., et al., "EDDIE-Automation, a decision support tool for financial forecasting". *Decision Support Systems*, vol. 37, no. 4, 2004, p. 559-565.

Turban E., Aronson J., Liang T., *Decision Support Systems and Intelligent Systems*, Prentice Hall, New Jersey, 2004.

Vahidov R., Kersten G., "Decision station: situating decision support systems". *Decision Support Systems*, vol. 38, no. 2, 2004, p. 283-303.

Vicari R.M., Flores C.D., et al., "A multi-agent intelligent environment for medical knowledge". *Artificial Intelligence in Medicine*, vol. 27, no. 3, 2003, p. 335-366.

Vincent J-L., De Mendonça A., Cantraine F., et al., "Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study". *Crit Care Med*, vol. 26, no. 11, 1998, p. 1793-1800.

Weiss G., *Multiagent Systems A Modern Approach to Distributed Artificial Intelligence*, MIT Press, Cambridge MA, 1999.

Yim H.S., Ahn H.J., Kim J.W., Park S.J., "Agent-based adaptive travel planning system in peak seasons". *Expert Systems with Applications*, vol. 27, no. 2, 2004, p. 211-222.

URL1 http://soleunet.ijs.si/website/other/DMiner_and_PMML.pdf 2004.

6. Bibliografia

- Beal, A.L. & Cerra, F.B. (1994) Multiple organ failure syndrome in the 1990s: Systemic inflammatory response and organ dysfunction, *JAMA*, 271, pp. 226–233.
- Fagon, J.Y., Chastre, J. & Novara, A. (1993) Characterization of intensive care unit patients using a model based on the presence or absence of organ dysfunctions and/or infection: The ODIN model. , *Intensive Care Med*, 19, pp. 137–144.
- Flexer, A. (1996) Statistical evaluation of neural networks experiments: minimum requirements and current practice in: R. Trappl (Ed) *13th European Meeting on Cybernetics and Systems Research*, (Vienna, Austria)
- Hanley, J.A. & Mc Neil, B.J. (1982) The meaning and use of the area under a receiver-operating characteristic (ROC) curve, *Radiology* 143, pp. 29–36.
- Hebert, P.C., Drummond, A.J. & Singer, J. (1993) A simple multiple system organ failure scoring system predicts mortality of patients who have sepsis syndrome., *Chest*, 104, pp. 230–235.
- Knaus, W.A., Draper, E.A. & Wagner, D.P. (1985b) Prognosis in acute organ-system failure, *Ann Surg*, 202, pp. 685–693.
- Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., M., B., G., B.P., Sirio, C.A., J., M.D., Lotring, T. & Damiano, A. (1991) The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults, *Chest* 100, pp. 1619–1636.
- Le Gall, J.R., Lemeshow, S. & Saulnier, F. (1993) A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study,, *JAMA*, 270, pp. 2957–2963.
- Miranda, R. (1999). The effect of harmonising and standardising the nursing tasks on Intensive Care Units (ICU) of the European Community. Technical Report BMH4-CT96—0817. S. E. I. C. U. S. (EURICUS-II), Health Services Research Unit, Groningen University Hospital.
- Tran, D.D., Miguel, M.A. & van Leeuwen, P.A.M. (1993) Risk factors for multiple organ system failure and death in critically injured patients., *Surgery*, 114, pp. 21–30.
- Vincent, J.L., de Mendonça, A. & Cantraine, F. (1998) Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study, *Crit Care Med*, 26, pp. 1793–1800.
- Zimmerman, J.E., Knaus, W.A. & Wagner, D.P. (1996) A comparison of risks and outcomes for patients with organ system failures 1982–1990, *Crit Care Med*, 24, pp. 1633–1641.

CAPÍTULO V
Discussão global

CAPÍTULO V- Discussão global

É apresentada a discussão global dos trabalhos desenvolvidos, mediante um processo de análise dos resultados obtidos.

A falta de uma integração dos dados obtidos da monitorização de vários parâmetros do doente grave admitido na UCI pode comprometer a sua utilidade clínica. Na prática clínica, o intensivista analisa todos os dados com o propósito de prever uma alteração do estado do doente e seleccionar a terapêutica. Contudo, o crescimento exponencial de dados (Halpern & Pastores, 1999, Hanson & Marshall, 2001, Lucas et al., 2000, Rosenfeld et al., 2000) tem tornado evidente a necessidade do recurso a novas tecnologias e ferramentas para o seu processamento, no contexto do suporte à tomada de decisão (Cios & Moore, 2002, McKinley et al., 2001, Morris, 2000, Teres & Higgins, 2000).

Este facto tem induzido nos intensivistas um interesse no desenvolvimento de modelos de previsão baseando-se em técnicas oriundas da Inteligência Artificial (Hanson & Marshall, 2001).

Os modelos de previsão da mortalidade habitualmente utilizados são estáticos, não explorando a informação contida nos dados gerados pela monitorização de vários parâmetros registados por rotina em todas as UCIs (ex.: frequência cardíaca, saturação arterial de O₂, tensão arterial sistólica e débito urinário horário).

Com o trabalho *Mortality assessment in intensive care units via adverse events using artificial neural networks* propôs-se uma abordagem inovadora para a previsão da mortalidade intra-hospitalar, em que, para além dos dados relativos à admissão, são incorporados dados clínicos diários registados durante a estadia na UCI.

Explorou-se uma aproximação baseada nos eventos clínicos adversos e na aplicação de redes neuronais artificiais para a previsão da mortalidade intra-hospitalar.

Os desempenhos dos modelos gerados automaticamente foram comparados com os do SAPS II e um modelo obtido por regressão logística, após ajuste dos parâmetros internos.

Globalmente, foram conseguidos bons resultados, com valores da capacidade discriminativa superiores a 80%. Na prática clínica, os modelos com maior sensibilidade permitirão uma identificação precoce da deterioração do estado clínico do doente.

A capacidade discriminativa dos modelos que têm como variáveis apenas os eventos clínicos adversos foi significativamente superior ao modelo SAPS II (83,9 versus 79,8%, $p < 0,05$).

O modelo obtido por rede neuronal com a inclusão de todas as variáveis, excepto a idade, teve uma capacidade discriminativa de 87,1%.

A inclusão dos eventos clínicos adversos no modelo SAPS II fez aumentar a sua capacidade discriminativa de 79,8 para 85,2%. O valor da capacidade discriminativa do SAPS II na população de validação foi de 84,4% (Castella et al., 1995).

Os valores da capacidade discriminativa superiores a 80% são considerados muito interessantes (Randolph et al., 1998b).

Da análise da importância relativa dos eventos clínicos adversos, verificou-se que os eventos clínicos agregados têm um maior contributo para o *outcome* final, com uma precisão de 90%. Entre estes, os em débito urinário horário e em saturação arterial de O_2 tiveram uma importância relativa superior aos de tensão arterial sistólica e frequência cardíaca. De salientar que os modelos que apresentaram melhores resultados foram os que incluíam todas as variáveis disponíveis.

Constata-se que o tempo médio/dia em eventos críticos urinários, seguidos dos em eventos críticos em tensão arterial e saturação arterial de O_2 são os parâmetros que têm maior impacto no *outcome* final.

Estes resultados corroboram três premissas desta dissertação:

- a integração do grau de disfunção fisiológica durante a estadia (eventos clínicos adversos) na UCI tem um valor prognóstico mensurável e relevante;
- os modelos de previsão baseados nos eventos clínicos adversos têm um desempenho significativamente melhor ($p < 0,05$) do que o modelo SAPS II;
- as redes neuronais artificiais permitem melhorar o desempenho do SAPS II através da inclusão dos eventos clínicos adversos.

Para evitar a sobre-especialização (*overfitting*) dos modelos automáticos, os resultados foram obtidos numa amostra de teste (validação) e o número de nodos internos foi de metade do número de variáveis de entrada. (Silva et al., 2005.)

A previsão da mortalidade é numérica (valor entre 0 e 1) e os valores da sensibilidade, especificidade e precisão foram obtidos considerando que *outcomes* acima de 0,5 correspondem a 1 (morte).

Buchman et al (Buchman et al., 1994) estiveram entre os primeiros a aplicar rede neuronais artificiais em bases de dados clínicas. Demonstraram um poder preditivo superior de três tipos de redes neuronais face a um modelo por regressão logística, para a previsão da probabilidade de doentes cirúrgicos terem uma estadia na UCI de sete ou mais dias.

Outros trabalhos utilizando redes neuronais artificiais para a previsão da mortalidade em UCIs polivalentes têm demonstrado que estas têm um desempenho igual ou superior aos dos modelos por regressão logística. (Clermont et al., 2001, Doig et al., 1993, Dybowski et al., 1996)

Clermont et al (Clermont et al., 2001) demonstraram que o desempenho das redes neuronais era comparável ao do modelo APACHE III para a previsão da mortalidade hospitalar (área sob a curva ROC = 0,836 e 0,839, respectivamente). A rede neuronal foi também capaz de ter um desempenho muito semelhante ao do APACHE III, mesmo quando as variáveis independentes não foram pré-designadas (área sob a curva ROC = 0,8). Contudo, em todos estes estudos, as redes neuronais foram treinadas com variáveis estáticas utilizadas pelo modelo APACHE II. Este estudo segue uma direcção alternativa: a utilização de dados registados após as primeiras vinte e quatro horas da admissão do doente à UCI.

Como Marshal (Marshall 1999) afirmou, os modelos prognósticos têm dificuldade em descrever o que acontece aos doentes após a admissão na UCI.

Como demonstrado no passado por Wagner (Wagner et al., 1986), para outras variáveis, este estudo enfatiza objectivamente que os desvios dos limites da normalidade das variáveis fisiológicas monitorizadas ao longo do tempo, após a admissão na UCI, contribuem para uma previsão mais precisa do *outcome* final do doente grave.

Como referido no capítulo III, o desempenho do modelo SAPS II diminui quando é aplicado a populações diferentes daquelas em que foi desenvolvido (Le Gall et al., 1993). O modelo pode ser adaptado a uma população específica através do recurso a um processo de ajuste ou através da expansão do modelo (inclusão de novas variáveis).

Aergerter et al propuseram uma revisão do SAPS II e desenvolveram dois modelos de regressão logística baseados no SAPS II. O primeiro modelo derivou da utilização de uma estratégia de ajuste (usando apenas a pontuação do SAPS II como uma variável independente, e consequentemente mantendo todas as variáveis originais e os seus pesos), e um segundo modelo que reavaliou os itens originais do SAPS II e a integração de localização pré-admissão e comorbilidade crónica. Estes modelos demonstraram uma melhor calibração para a mortalidade intra-hospitalar do que o SAPS II original. Le Gall et al elaboraram outra proposta (Le Gall et al., 2005), adicionando seis variáveis de admissão ao SAPS II: idade, sexo, tempo de internamento hospitalar pré-admissão à UCI, localização do doente pré-admissão à UCI, categoria clínica e presença ou não de intoxicação por drogas ilícitas. As qualidades estatísticas do SAPS II expandido são melhores do que as do original, e mesmo do que do SAPS II ajustado (Le Gall, 2005).

A abordagem deste trabalho segue outra direcção: a inclusão de variáveis obtidas ao longo do internamento do doente na UCI associada à pontuação do SAPS II, que também mostrou melhorar o desempenho do SAPS II. Mais ainda, neste estudo, quando são usadas as mesmas variáveis, os modelos obtidos por redes neuronais apresentam um melhor desempenho que os por regressão logística, com incrementos entre 1,3 a 2%.

Kayaalp et al (Kayaalp et al., 2001, Kayaalp et al., 2000) adoptaram uma abordagem algo similar à do presente estudo, utilizando o SOFA score para a previsão diária do risco de mortalidade do doente durante a sua estadia na UCI. O estudo revelou que o SOFA score renal, respiratório e cardiovascular são as variáveis preditivas com maior peso.

Estes resultados corroboram de alguma forma a importância relativa no *outcome* final dos eventos clínicos adversos, revelada no presente estudo.

Uma das limitações do nosso estudo advém do facto dos modelos terem sido desenvolvidos a partir de uma base de dados estática, isto é, não vai recebendo dados em tempo real, o que pode limitar a capacidade de avaliação dos desempenhos dos modelos ao longo do tempo. Esta questão poderá ser contornada através da adopção de técnicas, também elas automáticas, de actualização dos modelos ao longo da sua utilização tendo como objectivo manter a sua validade acima de limites mínimos. Esta solução é abordada e explorada no sistema INTCare.

A aquisição automática dos parâmetros monitorizados tem demonstrado que influencia a qualidade dos dados quando comparada com o registo manual (Bosman et al., 1998, Suistomaa et al., 2000).

Hammond et al revelaram que quase 20% dos prestadores de cuidados de saúde registam incorrectamente os dados clínicos (Hammond et al., 1989).

Dois estudos evidenciam que existe a tendência dos enfermeiros registarem os valores normais e ignorarem os valores extremos, resultando numa variabilidade reduzida dos dados registados manualmente (Reich et al., 2000, Taylor & Whamond, 1975).

Na definição dos eventos clínicos, o requisito de duração foi de dez minutos antes do seu registo, com o intuito de eliminar tanto quanto possível erros de registo. Por outro lado, todos os impressos normalizados para o registo dos eventos clínicos adversos foram submetidos a várias verificações locais (na UCI) e centrais (no *Health Services Research Unit, Groningen University Hospital*) para avaliar a sua consistência lógica e a presença de erros ou dados omissos.

A não disponibilização dos registos diários de ocorrência dos eventos clínicos na base de dados é uma limitação presente no actual trabalho, que impossibilitou a análise

temporal da ocorrência dos eventos clínicos adversos. Para ultrapassar esta limitação, está a ser desenvolvido um sistema de aquisição dos dados em tempo real, a partir dos monitores de cabeceira do doente grave admitido na UCI.

A presente abordagem procura disponibilizar uma ferramenta de suporte à decisão clínica através da estimativa automática e dinâmica do estado clínico do doente, e não a geração de novos modelos de previsão do risco de mortalidade.

Os modelos gerados pelas redes neuronais podem ser actualizados diariamente a partir dos eventos clínicos adversos, disponibilizando previsões de mortalidade diárias, de acordo com a evolução do estado clínico do doente.

A previsão da falência de órgãos, disponível em tempo útil no doente grave, pode ter utilidade por várias razões. Em primeiro lugar, os dados da literatura sugerem que a identificação e intervenção precoce nos doentes com sepsis e choque séptico podem melhorar o prognóstico (Rivers et al., 2001). Em segundo, os atrasos no início da terapêutica podem promover a evolução adicional de disfunção de órgãos e o desenvolvimento adicional de disfunção de órgãos está associado a um risco aumentado de mortalidade hospitalar (Vincent et al., 1998).

Consequentemente, qualquer informação que possa orientar o processo de decisão clínico para decisões em tempo útil será de extrema mais-valia. A investigação apresentada para a previsão de falência de órgãos não tem por objectivo gerar novos índices de quantificação de falência de órgãos, mas antes, induzir modelos de previsão de falência de órgãos, através da aplicação de técnicas de *Data Mining*, incorporando o impacto potencial do grau de disfunção fisiológica aguda, ocorrida durante a estadia do doente na UCI (eventos clínicos adversos).

O estudo *Multiple organ failure diagnosis using adverse events and neural networks* tenta uma abordagem exploratória para a indução de modelos de previsão de falência de órgãos através de redes neuronais artificiais, independentemente da causa subjacente da falência de órgão ou sistema.

A falência de órgãos resulta de processos fisiopatológicos dinâmicos e complexos cujos mecanismos não são ainda completamente conhecidos.

Têm sido desenvolvidos vários sistemas para avaliar a gravidade da síndrome de falência múltipla de órgãos. Contudo, os vários índices para a avaliação da falência múltipla de órgãos diferem substancialmente. Além disso, as variáveis que descrevem a falência de órgãos e sistemas não são tão específicas quanto deveriam. A maioria destes índices utiliza uma associação de variáveis fisiológicas, com parâmetros analíticos (creatinina, bilirrubina).

Uma abordagem baseada nos dados clínicos monitorizados e registados de hora a hora, ao nível do doente, provavelmente reflectirá melhor a relação entre a disfunção

fisiológica (eventos clínicos adversos) e a subsequente falência de órgãos e sistemas. Para além disso, este processo será mais precocemente avaliado dado que não estará dependente de exames complementares, elemento crucial para a selecção e implementação duma estratégia terapêutica adequada.

A previsão em tempo útil da falência de órgãos, baseada na integração da informação presente nas variáveis tais como o índice de gravidade da doença (SAPS II), o índice de quantificação de falência de órgãos (SOFA score) e os eventos clínicos adversos limitaria potencialmente a deterioração funcional do doente grave. O SOFA score permite a quantificação da incidência de falência de órgãos e sistemas durante a estadia na UCI, no entanto, a previsão de falência de órgãos é ainda um tema em investigação.

Neste estudo, aplicamos uma nova abordagem para a indução de modelos previsão da falência de órgãos do dia seguinte, durante a estadia na UCI, explorando a informação presente nos eventos clínicos adversos (frequência diária agregada de eventos clínicos e eventos críticos em tensão arterial sistólica, frequência cardíaca, débito urinário e saturação arterial de O₂) .

A falência de órgãos a prever foi quantificada pelo SOFA score por órgão através da sua transformação numa variável binária (0 se SOFA score por órgão < 3 e 1 se SOFA score por órgão ≥ 3). O objectivo é a previsão de falência de órgão (SOFA score por órgão ≥ 3, por órgão) para o dia seguinte, através da aplicação de modelos automáticos, durante a estadia na UCI.

As variáveis de entrada incluídas foram: o SOFA score por órgão do dia, os eventos clínicos adversos, o tipo e origem da admissão, idade e o SAPS score.

Com o intuito de avaliar a importância das variáveis de entrada, foram testadas quatro tipo de configurações, designadas por:

SOFA - inclui apenas as pontuações por órgão SOFA $_{d-1}$;

ALL – onde são incluídas todas as variáveis;

NO SOFA – neste caso são omitidos as pontuações por órgão do SOFA $_{d-1}$;

ORM –usa apenas os eventos e os eventos críticos (frequência diária).

Dos resultados dos vários modelos desenvolvidos são de destacar os valores de previsões correctas de falência de órgão do dia seguinte, entre 87,6 a 94,6%, associados a bons valores de sensibilidade e especificidade.

Os resultados obtidos vêm corroborar uma das premissas desta dissertação: a integração de múltiplos dados e informação permite obter modelos de previsão com bons resultados (o modelo com todas as variáveis tem resultados semelhantes ao modelo só com as pontuações do SOFA score).

Os melhores resultados foram obtidos com o modelo que inclui as pontuações por órgão do SOFA avaliado no dia anterior. Contudo, estes resultados foram obtidos numa população não balanceada em termos de frequência de falência de órgãos, o que revela a capacidade do SOFA score de avaliar a incidência da disfunção de órgão, durante a estadia do doente grave na UCI. Mas o SOFA score requer uma revisão dos dados de um período de vinte e quatro horas para identificar o pior valor de cada variável, e a escolha do tempo ideal do registo do dados e o seu efeito na validade ainda não foram investigados (Zygun et al., 2005).

Os modelos de previsão com a inclusão única dos eventos clínicos adversos têm acuidades preditivas inferiores (70%), com bons valores de sensibilidade sendo as mais elevadas para o sistema hematológico e hepático, por isso não comparáveis com os resultados obtidos pelos restantes modelos.

A variação de resultados da previsão por órgãos observada pode resultar da diferente prevalência de disfunção de órgãos observada na população analisada (Essex-Sorlie, 1995).

Os resultados para o aparelho respiratório apresentam menor precisão, podendo depender de variáveis não incluídas nos modelos (ex.: práticas de desmame ventilatório, tipo de sedação utilizada) ou alguma limitação do SOFA score na classificação da falência respiratória (de acordo com a pontuação do SOFA respiratório, a falência pressupõe que o doente esteja sob ventilação mecânica para além do valor da razão $\text{PaO}_2 / \text{FIO}_2$).

Grande parte da literatura relativa falência do órgãos focaliza-se em mostrar uma correlação positiva entre o SOFA score e a mortalidade (Ferreira, 2001 #436) (Russell et al., 2000).

Outros trabalhos investigaram tendências temporais pré-especificadas, definidas como variações ao longo do tempo, tais como “subida” ou “descida” dos valores de SOFA score (Cabre et al., 2005., Ferreira et al., 2001, Levy et al., 2005) Estes trabalhos consideram um número reduzido de dias e têm por objectivo a previsão da mortalidade. A presente dissertação aponta para outra direcção.

Convém sublinhar que o principal objectivo deste estudo é explorar a possibilidade da previsão de falência de órgãos, utilizando apenas os eventos adversos clínicos. Os resultados obtidos, nomeadamente uma acuidade preditiva global de 70%, embora não afirmativos, têm o seu valor. Para além disso, a abordagem proposta abrirá espaço para investigações e estudos mais aprofundados do impacto da disfunção fisiológica aguda na falência de órgãos, durante a estadia do doente grave na UCI.

Prever esta dinâmica poderá disponibilizar um conhecimento mais preciso da evolução da falência múltipla de órgãos. Na prática clínica, é frequente calcular o número de

falência de órgãos para avaliar o prognóstico do doente à cabeceira do doente. O presente estudo revela o potencial da utilização de variáveis fisiológicas dinâmicas, disponíveis à cabeceira do doente, com uma capacidade preditiva da falência de órgãos interessante.

O principal objectivo do estudo *Rating organ failure via adverse events using data mining in intensive care medicine* foi o de induzir modelos de previsão diária do estado funcional de órgãos (aferido pelo SOFA score por órgão e sistema) com base nos eventos clínicos adversos, SAPS II score, idade e tipo e origem da admissão na UCI. Este estudo procura aprofundar os resultados do *Multiple organ failure diagnosis using adverse events and neural networks*.

No desenvolvimento dos modelos, foram aplicadas técnicas de *data mining*: redes neuronais, árvores de decisão e de regressão logística. Os três tipos de modelos de previsão induzidos tentam prever o estado funcional diário para cada órgão e sistema que pode ser classificado em: normalidade (SOFA score = 0), disfunção (SOFA score = 1 ou 2) e falência (SOFA score $\geq$ 3).

Salienta-se a não inclusão dos valores do SOFA score por órgão como variável de entrada, para estudar assim o impacto dos eventos clínicos adversos no estado funcional dos órgãos.

As previsões geradas pelas redes neuronais têm melhor capacidade discriminativa relativamente às outras técnicas utilizadas ($p < 0,05$), com incrementos de 1,3 a 2%, excepto no caso da falência sistema hematológico.

Para a previsão diária da falência de órgão os modelos obtidos por redes neuronais apresentam o um desempenho significativamente melhor ($p < 0,05$), avaliado em termos de capacidade discriminativa, com valores da área sob a curva ROC entre 78,1% para o sistema nervoso central e 71,3% para o sistema respiratório.

Para o estado de disfunção, todos os modelos apresentam valores de capacidade discriminativa inferiores a 70%. Para a previsão do estado funcional normal, os modelos apresentam um melhor desempenho médio (70,3%), particularmente para o sistema respiratório (71,3%), cardiovascular (72,8%), sistema nervoso central (73,4%) e renal (71,7%).

O presente estudo revela que em associação à idade, ao SAPS II score, ao tipo e local de origem da admissão à UCI, os eventos clínicos adversos têm um valor preditivo significativo ($p < 0,05$) para a previsão diária da falência de órgãos. Na realidade, nestes modelos a importância relativa dos eventos clínicos adversos é superior a 50% (variando entre 50,4% e 99,0%). O SAPS II score aquando da admissão tem uma menor importância relativa do que os eventos clínicos adversos agregados.

Nenhum dos modelos obteve boa capacidade discriminativa para a disfunção de órgão, provavelmente reflectindo que o desenvolvimento da falência de órgãos é um processo contínuo, em que classificação proposta pelo SOFA score foi definida subjectivamente (Russell et al., 2000).

Rivers et al (Rivers et al., 2001) demonstraram que a uma intervenção terapêutica atempada e orientada por objectivos, na sala de emergência, reduz a mortalidade intra-hospitalar, na sepsis. Esta diminuição de mortalidade, está associada à correcção e estabilização precoce do grau de disfunção fisiológica aguda, de forma a evitar o posterior desenvolvimento de falência múltipla de órgãos.

Como anteriormente demonstrado por Wagner et al (Wagner et al., 1986), para outras variáveis, este estudo enfatiza objectivamente que os desvios da normalidade dos valores dos quatro parâmetros fisiológicos têm um valor preditivo significativo no estado funcional dos órgãos, por outras palavras, a monitorização dos parâmetros em questão acrescenta informação relevante sobre o processo fisiopatológico da falência múltipla de órgãos.

A literatura sobre o desenvolvimento de modelos de previsão de falência de órgãos é escassa. Matis et al (Matis et al., 1995) utilizaram quinze variáveis (ex.: idade, bilirrubina, creatinina) para desenvolver o modelo de previsão de insuficiência hepática após transplante. Obtiveram uma acuidade entre 70% (usando os dados pré-transplante) e 88% (cinco dias após o transplante). Um modelo baseado numa rede neuronal artificial teve sucesso na previsão de insuficiência cardíaca em 58 doentes, usando vinte variáveis (ex.: frequência cardíaca, tensão arterial)(Gils et al., 1997).

A metodologia utilizada permitiu constatar que a inclusão dos eventos clínicos adversos permite a previsão da evolução do estado funcional dos órgãos e sistemas do doente grave, sem o recurso a dados laboratoriais, e em tempo útil. Além disso, estas variáveis são passíveis de recolha automática a partir dos monitores de cabeceira.

A valorização da monitorização de cabeceira destes quatro parâmetros é por vezes questionada por alguns investigadores (Chambrin et al., 1999, Cropp et al., 1994, Friesdorf & Buss, 1999, Friesdorf et al., 1994, Koski et al., 1990, Meredith & Edworthy, 1995).

Contudo, os resultados obtidos pelo presente estudo revelam a importância da valorização clínica da monitorização da disfunção fisiológica dos quatro parâmetros fisiológicos, não só para a acção terapêutica imediata mas também pelo seu impacto mensurável no decurso do processo fisiopatológico da falência múltipla de órgãos.

No trabalho *Organ failure prediction based on clinical adverse events: a cluster model approach*, aplicou-se uma abordagem baseada na segmentação propondo uma matriz tridimensional de suporte, considerando três dimensões:

1. o dia do internamento na UCI ($n \geq 2$ e $n \leq 5$);
2. o órgão e sistema (hepático, renal, respiratório, cardiovascular, hematológico e sistema nervoso central);
3. o segmento (0, 1 e 2).

Esta matriz acomoda setenta e dois modelos de previsão gerados para cada uma das combinações: dia x órgão x segmento. Para além disto, permite, conforme ilustrado no estudo (Figura 2), visualizar graficamente a sequência de modelos ajustados ao doente individualizado, ao longo da sua estadia na UCI.

A segmentação permite o agrupamento automático de doentes com características semelhantes, facilitando a modelação dentro de cada segmento.

Considerando os dados da admissão e os eventos clínicos adversos, é possível estimar a falência de órgão diária (avaliada pela transformação do SOFA score por órgão numa variável binária), durante o internamento, com a informação disponível até ao dia em análise.

Na população analisada foram identificados automaticamente, por aplicação de redes de Kohonen, três segmentos. Dentro de cada um e para cada órgão e dia de internamento (até quinto dia na UCI), foi aplicado um algoritmo de indução de árvores de decisão.

Em cada segmento, a previsão do estado funcional dos órgãos e sistemas revelou valores de precisão superiores a 74,2%, havendo segmentos com valores de precisão superiores a 95%. Estes resultados evidenciam padrões de evolução do processo de falência de órgãos e são referidos por vários investigadores (Deitch, 1992); (Moreno, 1999); (Ferreira, 2001); (Vincent, 1998).

A disfunção de órgãos é um processo dinâmico, daí que a sua avaliação sequencial e progressão durante a estadia na UCI é muito útil, uma vez que está correlacionado com o *outcome* final (Cook et al., 2001, Ferreira et al., 2001). O tipo de evolução da falência de órgãos pode ser útil para identificar os doentes que não respondem à terapêutica (Cabre et al., 2005).

A abordagem utilizada no presente estudo permite gerar modelos menos extensos e, consequentemente de compreensão e aplicação mais fácil. Trata-se de uma abordagem dinâmica, por oposição aos métodos habituais caracterizados pela análise única (à data da admissão à UCI ou sequencial de variáveis predefinidas), permitindo a incorporação de dados biométricos monitorizados por rotina.

O sistema INTCare: *A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine*, apresentado sob a forma de protótipo, constitui um dos

objectivos desta dissertação: a viabilização de um sistema de apoio à decisão inteligente em Medicina Intensiva.

O INTCare é uma concretização experimental do trabalho de investigação resultante da confluência do conhecimento da área da Medicina Intensiva com o conhecimento da área da Inteligência Artificial.

É um sistema inteligente de suporte à decisão, baseado no processo de Descoberta de Conhecimento em Bases de Dados e no paradigma dos agentes. Mais especificamente, este sistema efectua a previsão de disfunção de órgãos e sistemas e do risco de mortalidade hospitalar.

O INTCare tem por objectivo a automatização do processo de Descoberta de Conhecimento em Bases de Dados para o suporte à decisão clínica através de:

- da aquisição automática de dados dos monitores e aplicação de modelos de previsão da falência de órgãos e sistemas;
- actualização dos modelos de previsão do risco de mortalidade intra-hospitalar;
- avaliação de cenários sobre a evolução do estado clínico doente grave, permitindo a comparação das consequências de diferentes estratégias terapêuticas.

A utilização em ambiente real desta aplicação depende da sua autonomia, que é aumentada através de agentes responsáveis pela automatização das tarefas de aquisição de dados e manutenção da validade dos modelos.

A automatização do processo de Descoberta de Conhecimento em Bases de Dados recorre a uma arquitectura baseada em agentes que torna mais fácil o desenho da arquitectura ao mesmo tempo que simplifica as futuras tarefas de manutenção e introdução de novas funcionalidades. Como elementos constituintes, possui agentes dedicados ao tratamento de dados que possibilitam a aquisição em tempo real dos valores de diversos parâmetros biométricos normalmente monitorizados nas UCI. Estes agentes encarregam-se não só da recolha dos dados, mas também do controlo de qualidade dos mesmos. No centro do sistema encontram-se modelos, treinados com base em algoritmos de treino da Inteligência Artificial, (como redes neuronais, árvores de decisão ou algoritmos de segmentação).

O sistema é autónomo incorporando mecanismos que possibilitam a detecção de modelos com baixa utilidade e a substituição por outros.

A actualização do sistema é automática ou o conjunto de modelos pode ser substituído manualmente.

O sistema INTCare está actualmente a ser testado no Serviço de Cuidados Intensivos do Hospital Geral de Santo António. Os resultados preliminares, embora promissores,

são ainda embrionários. É necessário mais trabalho experimental com a inclusão de um volume maior de dados e casos para uma análise do seu impacto na prática clínica.

Alguns sistemas similares estão já implementados na área da Medicina Intensiva. Num estudo multicêntrico, Render et al (Render et al., 2000) desenvolveram um sistema computadorizado e automático para avaliação da gravidade da doença, utilizando o *Department of Veterans Affairs Integrated Health Systems Technology Architecture* (SISVistA). Este sistema regista automaticamente variáveis demográficas, laboratoriais e diagnósticas, para o desenvolvimento de um modelo semelhante ao APACHE II. Da análise de 4651 casos, o modelo SISVista revelou um boa discriminação e calibração para a previsão da mortalidade hospitalar (área sob a ROC = 0.86; testes de Hosmer-Lemeshow, $p > 0.2$). Este estudo preliminar demonstrou que é possível construir um modelo de prognóstico a partir de dados registados durante a estadia do doente na UCI. A inclusão destes últimos dados teve como impacto um incremento significativo da validade do modelo (Hadorn et al., 1993, Render et al., 2000).

O sistema Guardian (Hayes-Roth & J..E., 1998) baseia-se em agentes autónomos com uma arquitectura flexível nos quais vários algoritmos cooperam para a disponibilização de um diagnóstico em tempo real e para a monitorização e suporte ao diagnóstico de doentes submetidos a cirurgia coronária admitidos na UCI. Os resultados revelados mostram um bom desempenho.

O sistema Rihad ICU Program desenvolvido por Chang et al baseia-se na avaliação diária do APACHE II e na falência de órgãos para a previsão diária do risco de mortalidade (Chang et al., 1988, Chang, 1989). Este sistema foi posteriormente aplicado por vários investigadores (Holmes et al., 1994, Wauer, 2001) com resultados díspares (sensibilidade variando entre 2,2% a 51%).

1. Bibliografia

- Bosman, R.J., Oudemans van Straaten, H.M. & Zandstra, D.F. (1998) The use of intensive care information systems alters outcome prediction, *Intensive Care Medicine*, 24, pp. 953–958.
- Buchman, T.G., Kubos, K.L., Seidler, A.J. & Siegforth, M.J. (1994) A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit, *Crit Care Med* 22, pp. 750-762.
- Cabre, L., Mancebo, J., Solsona, J.F., Saura, P., Gich, I. & Blanch, L. (2005) Multicenter study of the multiple organ dysfunction syndrome in intensive care units: the usefulness of sequential organ failure assessment scores in decision making, *J Intensive Care Med*, 31(7), pp. 927–933.
- Castella, X., A., A. & Bion, J. (1995) A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter multinational study. The European/North American Severity Study Group. , *Crit Care Med*, 23, pp. 1327 - 1335.
- Chambrin, M.C., Ravoux , P., Calvelo-Aros, D., Jaborska, A., Chopin , A. & Boniface, B. (1999) Multicentric study of monitoring alarms in the adult intensive care unit (ICU): a descriptive analysis, *Intensive Care Med*, 25, pp. 1360-1366.
- Chang, R.W., Jacobs, S. & Lee, B. (1988) Predicting outcome among intensive care unit patients using computerized trend analysis of daily Apache II scores corrected for organ system failure, *Intensive Care Med*, 14, pp. 558–566.
- Chang, R.W.S. (1989) Individual outcome prediction models for intensive care units, *Lancet*, pp. 143–146.
- Cios, K. & Moore, G. (2002) Uniqueness of medical data mining, *Artif Intell Med*, 26, pp. 1-24.
- Clermont, G., Angus, D.C., DiRusso, S.M., Griffin, M. & Linde-Zwirble, W.T. (2001) Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models, *Crit Care Med* 29, pp. 291-296.
- Cook, R., Cook, D. & Tilley, J. (2001) Multiple organ dysfunction: baseline and serial component scores, *Crit. Care Med*, 29, pp. 2046–2050.
- Cropp, A.J., Woods, L.A., Raney, D. & Bredle, D.L. (1994) Name that tone. The proliferation of Alarms in the Intensive Care Unit, *Chest*, 105, pp. 1217-1220.

- Dybowski, R., Weller, P., Chang, R. & Gant, V. (1996) Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm, *Lancet*, 347, pp. 1146-1150.
- Doig, G.S., Inman, K.J. & Sibbald, W.J. (1993) Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression, *Proc Annu Symp Comput Appl Med Care*, pp. 361–365.
- Ferreira, F.L., Bota, D.P. & Bross, A. (2001) Serial Evaluation of the SOFA score to predict outcome in critically ill patients, *JAMA*, 286(1754-1758).
- Friesdorf, W. & Buss, B. (1999) Gobel M. Monitoring alarms – the key to patients safety in the ICU?, *Intensive Care Med*, 25, pp. 1350-1352
- Gils, M., Jansen, H., Nieminen, K., Summers, R. & Weller, P. (1997) Using artificial neural networks for classifying ICU patient states, *Engineering in Medicine and Biology Magazine*, 16, pp. 41–47.
- Hadorn, D.C., Keeler, E.B. & Rogers, W.H. (1993) *Assessing the performance of mortality prediction models* RAND/ UCLA/Harvard Center for Health Care Financing Policy Research).
- Halpern, N.A. & Pastores, S.M. (1999) Technology introduction in critical care: just knowing the price is not enough! , *Chest*, 116, pp. 1092–1099.
- Hammond, J., Ward, C.G., Johnson, M., Varas, R. & Marcial, E. (1989) The computerized burn unit: experience with a patient data management system, *Int J Clin Monit Comput*, 6, pp. 87–89
- Hanson, C.W.r. & Marshall, B.E. (2001) Artificial intelligence applications in the intensive care unit, *Crit Care Med* 29, pp. 427–435.
- Hayes-Roth, B. & J.E., L. (1998) Guardian: an intelligent autonomous agent for medical monitoring and diagnosis, *IEEE Intelligent Syst*, (January/February), pp. 58–64.
- Hebert, P.C., Drummond, A.J. & Singer, J. (1993) A simple multiple system organ failure scoring system predicts mortality of patients who have sepsis syndrome., *Chest*, 104, pp. 230–235.
- Holmes, L., Loughhead, K., Treasure, T. & Gallivan, S. (1994) Which patient will not benefit from further intensive care after cardiac surgery?, *Lancet*, 344, pp. 1200–1202.
- Kayaalp, M., Cooper, G. & Clermont, G. (2001) Predicting with variables constructed from temporal sequences *Eight International Workshop on Artificial Intelligence and Statistics* (Key West, FL, USA,

- Kayaalp, M., Cooper, G.F. & Clermont, G. (2000) Predicting ICU mortality: a comparison of stationary and nonstationary temporal models, *Proc. AMIA Symp*, 2000, pp. 418–422.
- Koski, E.M., Makivirta, A., Sukuvaara, T. & Kari, A. (1990) Frequency and reliability of alarms in the monitoring of cardiac postoperative patients, *Int J Clin Monit Comput* 7, pp. 129–133.
- Le Gall, J.R., Neumann, A., Hemery, F., Bleriot, J.P., Fulgencio, J.P., Garrigues, B., Gouzes, C., Lepage, E., Moine, P. & Villers, D. (2005) Mortality prediction using the SAPS II, an update for French ICUs, *Critical Care Med*, 9, pp. 645–652.
- Levy, M., Macias, W.L., Vincent, J.L., J.A., R., Silva, E. & Trzaskoma, B. (2005) Early changes in organ function predict eventual survival in severe sepsis, *Crit Care Med*, 33(10), pp. 20–2194.
- Lucas, P.J., de Bruijn, N.C. & Schurink, K. (2000) A probabilistic and decision theoretic approach to the management of infectious disease at the ICU., *Artif Intell Med* 19, pp. 251–279.
- Matis, S., Doyle, H., Marino, I., Mural, R. & Uberbacher, E. (1995) Use of neural networks for prediction of graft failure following liver transplantation in: IEEE (Ed) *8th IEEE Symposium on Computer-Based Medical Systems* (Washington, DC, USA,
- McKinley, B.A., Moore, F.A. & Sailors, R.M. (2001) Computerized decision support for mechanical ventilation of trauma induced ARDS: results of a randomized clinical trial, *J Trauma*, 50, pp. 415–424.
- Meredith, C. & Edworthy, J. (1995) Are there too many alarms in the intensive care unit? An overview of the problems, *J Adv Nurs* 2, pp. 15-20.
- Morris, A.H. (2000) Developing and implementing computerized protocols for standardization of clinical decisions., *Ann Intern Med*, 132, pp. 373–383.
- Randolph, A.G., Guyatt, G.H. & Richardson, W.S. (1998b) Prognosis in the intensive care unit: finding accurate and useful estimates for counseling patients, *Crit Care Med*, 26, pp. 767–772.
- Reich, D.L., Wood, R.K.J. & Mattar, R. (2000) Arterial blood pressure and heart rate discrepancies between handwritten and computerized anesthesia records., *Anesth Analg*, 91, pp. 612–616.
- Render, M.L., Welsh, D.E. & Kollef, M. (2000) Automated computerized intensive care unit severity of illness measure in the Department of Veterans Affairs: preliminary results.SISVista Investigators. Scrutiny of ICU Severity Veterans Health Systems Technology Architecture, *Crit Care Med*, 28, pp. 3540–3546.

- Rivers, E., Nguyen, B. & Havstad, S. (2001) Early goal directed therapy in the treatment of severe sepsis and septic shock., *N Engl J Med*, 345, pp. 1368–1377.
- Rosenfeld, B.A., Dorman, T. & Breslow, M.J. (2000) Intensive care unit telemedicine: alternate paradigm for providing continuous intensivists care, *Crit Care Med* 28, pp. 3925–3931.
- Russell, J.A., Singer, J. & Bernard, G.R. (2000) Changing pattern of organ dysfunction in early human sepsis is related to mortality, *Crit Care Med* 28, pp. 3405–3411.
- Silva, A., Cortez, P., Santos, M., Gomes, L. & Neves, J. (2005.) Multiple organ failure diagnosis using adverse events and neural networks, in: C.J. Seruca I, Hammoudi S, Filipe J, editors. (Ed) Enterprise information systems VI Springer).
- Taylor, D.E. & Whamond, J.S. (1975) Reliability of human and machine measurements in patient monitoring, *Eur J Intensive Care Med* 1, pp. 53–59.
- Teres, D. & Higgins, T.L. (2000) What if your hospital informatics department could provide a severity adjuster?, *Crit Care Med*, 28, pp. 3570–3571.
- Tran, D.D., Miguel, M.A. & van Leeuwen, P.A.M. (1993) Risk factors for multiple organ system failure and death in critically injured patients., *Surgery*, 114, pp. 21–30.
- Vincent, J.L., de Mendonça, A. & Cantraine, F. (1998) Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study, *Crit Care Med*, 26, pp. 1793–1800
- Wagner, D.P., Knaus, W.A., Elizabeth, A. & Draper, R.N. (1986) Physiologic Abnormalities From Acute Diseases. Evidence for a Predictable Relationship, *Arch Intern Med*, 146, pp. 1389-1396.
- Wauer, H. (2001) Unpublished data from the Charité University Hospital, Berlin National meeting about intensive care medicine in Bremen (Bremen).

CAPÍTULO VI

Conclusões

CAPÍTULO VI - Conclusões

O estudo *Mortality assessment in intensive care units via adverse events using artificial neural networks* explorou uma aproximação baseada nos eventos clínicos adversos e na aplicação de redes neuronais artificiais para a previsão da mortalidade intra-hospitalar a partir de uma base de dados com 13164 doentes. Este estudo confirmou três premissas desta dissertação:

- A integração nos modelos de previsão de mortalidade dos eventos clínicos adversos ocorridos durante a estadia do doente na UCI tem um valor prognóstico mensurável.
- Os modelos gerados por redes neuronais artificiais permitem melhorar o desempenho do modelo SAPS II através da incorporação dos eventos clínicos adversos.
- Os eventos clínicos adversos agregados têm a maior contribuição relativa para o outcome final do doente grave.

Note-se que os modelos de previsão da mortalidade que apresentam melhores resultados são os que incluem todas as variáveis, realçando a importância da integração dos dados.

O estudo *Multiple organ failure diagnosis using adverse events and neural networks* faz uma abordagem exploratória para a previsão de falência de órgãos e sistemas do dia seguinte, durante a estadia do doente na UCI, através da aplicação de modelos induzidos por redes neuronais artificiais. O estudo revelou que os eventos clínicos adversos permitem com razoável acuidade gerar previsões correctas de falência de órgão.

O estudo *Rating organ failure via adverse events using data mining in intensive care medicine* teve como principal objectivo induzir modelos de previsão diária do estado funcional de órgãos. Neste estudo, salienta-se o facto de não se usar como variável independente SOFA score por órgão, estudando assim o impacto dos eventos clínicos adversos na previsão do estado funcional dos órgãos. Os resultados obtidos confirmam a possibilidade de gerar modelos de previsão do estado funcional dos órgãos e sistemas, e que os eventos clínicos adversos têm uma importância relativa mais elevada.

A metodologia utilizada permitiu ainda constatar que a inclusão dos eventos clínicos adversos permite a previsão da evolução do estado funcional dos órgãos e sistemas do doente grave, sem o recurso a dados laboratoriais e em tempo útil.

No estudo *Organ failure prediction based on clinical adverse events: a cluster model approach*, aplicou-se uma abordagem baseada na segmentação propondo uma matriz tridimensional de suporte, permitindo a visualização da evolução estimada para o doente individualizado. Considerando os dados da admissão e os eventos clínicos adversos, é possível estimar com elevada precisão a falência de órgãos diária.

Os resultados obtidos motivaram um aprofundar da interdisciplinariedade entre a Medicina Intensiva e a área da Inteligência Artificial, que se traduziu no desenvolvimento de um protótipo de sistema de apoio à decisão inteligente em Medicina Intensiva, designado por INTCare. Este modelo tem por objectivo a disponibilização de mecanismos para a aquisição automática de dados a partir dos monitores de cabeceira e a actualização de modelos de prognóstico, no contexto do suporte à decisão clínica. Mais especificamente, este sistema efectua a previsão de disfunção de órgãos e sistemas e do risco de mortalidade hospitalar.

A participação num projecto com intervenientes que pensam, se exprimem, actuam e resolvem os problemas de formas distintas foi uma experiência exigente e simultaneamente gratificante. O trabalho que agora termina vem realçar novos aspectos e abrir possibilidades de investigação futura que anteriormente não estavam bem definidos.

Algumas pistas podem ser equacionadas em termos de trabalho futuro:

- análise sequencial dos eventos clínicos adversos;
- experimentação via INTCARE de um sistema de apoio à decisão em Medicina Intensiva e a utilização dos resultados para fazer evoluir os modelos desenvolvidos ;
- análise do impacto dos sistemas na prática clínica, na qualidade do processo de cuidados de saúde do doente grave admitido na UCI.

Resumo

Na Medicina Intensiva o prognóstico através avaliação do risco da mortalidade e da quantificação da falência de órgãos e sistemas, introduzidos no início da década de oitenta, assumiu desde sempre uma relevância específica.

Estes modelos foram desenvolvidos para quantificar o grau de disfunção fisiológica e comorbilidades sobre os quais é estimado o risco de mortalidade intra-hospitalar. A premissa fundamental da maioria destes modelos é a de que “as variáveis clínicas aquando ou próximo da admissão do doente à UCI estão associadas ao risco de mortalidade intra-hospitalar”. Os modelos mais utilizados são construídos usando a regressão logística sobre uma pontuação estática. A sua aplicação destes modelos na prática clínica, no contexto do suporte à tomada de decisão é ainda controverso e não aceite de forma universal.

Para além destes modelos, surgiram os índices de quantificação de falência de órgãos com o intuito de monitorizar a evolução da gravidade do doente durante a estadia na UCI. Contudo, estes índices não incorporam a informação em tempo útil nem integram a capacidade de previsão de falência de órgãos e sistemas, pelo que a sua utilidade como sistemas de suporte ao processo de tomada de decisão é de alguma forma limitada. Acresce que, a informação obtida por estes índices de quantificação de falência de órgãos não é incorporada pelos modelos gerais de prognóstico.

Simultaneamente assistimos à introdução de novas tecnologias de monitorização, suporte e terapêutica e à progressiva informatização das UCIs, via implementação de sistemas de informação clínica que têm contribuído para que a UCI seja um ambiente gerador de um volume exponencial de dados.

A monitorização de vários parâmetros é uma das actividades fundamentais na UCI. Entre os parâmetros monitorizados de hora a hora, por rotina em todas as UCIs estão a: frequência cardíaca, saturação arterial de O₂, tensão arterial e débito urinário horário. Os desvios dos limites da normalidade destes quatro parâmetros são designados nesta dissertação por eventos clínicos adversos.

Apesar da relevância atribuída à monitorização destes eventos clínicos adversos, a sua interpretação não tem resultado em práticas normalizadas e também não se tem traduzido em conhecimento “novo”.

Na prática clínica, a informação disponibilizada pela monitorização dos eventos clínicos adversos e pelos modelos gerais de prognóstico e índices de quantificação de falência

de órgãos é interpretada de forma desarticulada, não tirando proveito do conhecimento inerente às sinergias potencialmente existentes, consequentemente subaproveitada no processo de decisão clínica.

Esta realidade coloca novos desafios aos intensivistas para a integração e extracção de conhecimento ou informação potencialmente útil relativa ao prognóstico e à evolução da gravidade da(s) doença(s). A quantidade de informação acerca das determinantes prognósticas do doente grave é hoje extensa existe a necessidade de complementar os métodos estáticos com modelos dinâmicos, adaptáveis à medida que nova informação vai estando disponível.

A evolução corrente no processo de decisão clínica tem mostrado uma abertura crescente relativamente à introdução de técnicas de análise inteligente de dados para a extracção de conhecimento útil.

Um dos propósitos desta dissertação é o de explorar a incorporação de dados dinâmicos, obtidos dos eventos clínicos adversos, no modelo SAPS II e no SOFA score, utilizando métodos do processo da Descoberta do Conhecimento em Bases de Dados para a indução de modelos de previsão da mortalidade e falência de órgãos, como forma de viabilizar a construção de sistemas automáticos, situados de apoio à decisão clínica inteligente.

Para o desenvolvimento da investigação foi criada uma base de dados a partir de subconjunto de dados colhidos aquando do estudo EURICUS II englobando registos de 13164 doentes de quarenta e duas UCIs de nove países da União Europeia, seguindo a metodologia inerente ao processo de *Data Mining*.

O estudo *Mortality assessment in intensive care units via adverse events using artificial neural networks* explorou uma aproximação baseada nos eventos clínicos adversos e na aplicação de redes neuronais artificiais para a previsão da mortalidade intra-hospitalar. Os resultados dos modelos gerados automaticamente foram comparados com os do SAPS II e um modelo obtido por regressão logística. A capacidade discriminativa dos modelos que têm como variáveis apenas os eventos clínicos foi significativamente superior ao modelo SAPS II (83,9 versus 79,8%, $p > 0,05$). O modelo obtido por rede neuronal com a inclusão de todas as variáveis, excepto a idade teve uma capacidade discriminativa de 87,1%. Da análise da importância relativa dos eventos clínicos adversos verifica-se que eventos agregados têm o maior contributo para o *outcome* final. Destes, os em débito urinário horário e em saturação de O_2 têm uma importância relativa superior aos de tensão arterial e frequência cardíaca. Os modelos que apresentam melhores resultados são os que incluem todas as variáveis disponíveis.

O estudo *Multiple organ failure diagnosis using adverse events and neural networks* faz uma abordagem exploratória para a previsão de falência de órgãos. A falência de

órgãos a prever foi quantificada pelo SOFA score por órgão através da sua transformação numa variável binária (0 se SOFA score <3 e 1 se SOFA score ≥ 3). O objectivo é a previsão de falência de órgão para o dia seguinte. As variáveis de entrada foram: o SOFA score por órgão do dia, os eventos clínicos adversos, o tipo e origem da admissão, idade e o SAPS score. Os vários modelos desenvolvidos apresentam previsões correctas de falência de órgão entre 87,6 a 94,6%. Os modelos de previsão, com a inclusão única dos eventos clínicos adversos, têm acuidades preditivas mais baixas (70%).

No estudo *Rating organ failure via adverse events using data mining in intensive care medicine* o principal objectivo é o de induzir modelos de previsão diária do estado funcional de órgãos (aferido pelo SOFA score) com base nos eventos clínicos adversos, SAPS II score, idade e tipo e origem da admissão na UCI. Neste estudo foram aplicadas várias técnicas de data mining. Os três tipos modelos de previsão induzidos tentam prever para cada órgão e sistema: normalidade (SOFA score = 0), disfunção (SOFA score = 1 ou 2) e falência (SOFA score ≥ 3). As previsões geradas pelas redes neuronais têm melhor capacidade discriminativa relativamente às outras técnicas utilizadas ($p < 0,05$). Salienta-se o facto de não se usar como variável de entrada os valores do SOFA score por órgão. Os resultados obtidos revelam que em associação à idade, ao SAPS II score, ao tipo e local de origem da admissão à UCI, os eventos clínicos adversos têm um valor preditivo significativo ($p < 0,05$) para a previsão diária da falência de órgãos. Nestes modelos a importância relativa dos eventos clínicos adversos é superior a 50% (variando entre 50,4% e 99,0%). O SAPS II score aquando da admissão tem uma menor importância relativa do que os eventos clínicos adversos agregados. A metodologia utilizada revela que a inclusão dos eventos clínicos adversos permite a previsão da evolução do estado funcional dos órgãos e sistemas do doente grave, em tempo útil.

No trabalho *Organ failure prediction based on clinical adverse events: a cluster model approach*, aplicou-se uma abordagem baseada na segmentação propondo uma matriz tridimensional de suporte, permitindo a visualização da evolução estimada para o doente individualizado. Considerando os dados da admissão e os eventos clínicos adversos, é possível estimar a falência de órgão diária, durante o internamento, com a informação disponível até ao dia respectivo. Na população analisada foram identificados por aplicação de redes de Kohonen, três segmentos. Em cada segmento os modelos de previsão do estado funcional dos órgãos revelam valores de precisão superiores a 74,2%, havendo segmentos com valores de precisão superiores a 95%.

O sistema *INTCare* exposto sob a forma de protótipo constitui um dos objectivos desta dissertação, a viabilização de um sistema de apoio à decisão inteligente, e tem por objectivos disponibilizar mecanismos para a aquisição automática de dados e actualização de modelos de previsão. Especificamente este sistema efectua a previsão de falência de órgãos e do risco de mortalidade hospitalar. Como elementos constituintes possui agentes dedicados ao tratamento de dados que possibilitam a aquisição automática, em tempo real dos valores de diversos parâmetros biométricos monitorizados nas UCI. No centro do sistema encontram-se modelos, treinados com base em algoritmos de treino da Inteligência Artificial. O sistema é autónomo incorporando mecanismos que possibilitam a detecção de modelos com baixa utilidade e a substituição por outros. A actualização do sistema é automática ou o conjunto de modelos pode ser substituído manualmente.

Resumé

Dans la Médecine Intensive, le pronostic à partir de l'évaluation du risque de mortalité et de la quantification de la défaillance multi-viscérale, introduit au début des années 80, a gagné de l'importance parmi les investigateurs.

Malgré la différence de noms des modèles de pronostic, tous ont été développés pour quantifier le degré de dysfonction organique et comorbidités sur lesquels on estime le risque de mortalité à l'hôpital. La prémisse fondamentale de ces modèles est "les variables cliniques, au moment de l'admission du patient à l'unité de soins intensifs (USI), sont associées au risque de mortalité à l'hôpital".

Bien que l'utilisation de ces modèles et indices soit de plus en plus adoptée, leur application dans la pratique clinique pour le support à la prise de décision est encore controversée et n'est pas acceptée de façon universelle.

Au delà de ces modèles pronostiques, d'autres indices de quantification de défaillance organique sont apparus (ex.: SOFA score), pour monitoriser l'évolution de la gravité du malade pendant son séjour à l'USI. Cependant, ces indices n'incluent pas l'information en temps utile et n'intègrent pas la capacité de prévision de défaillance organique et systèmes, donc leur usage comme système de support à la prise de décision est limité. En plus, l'information obtenue à partir de ces indices de quantification de défaillance organique n'est pas considérée par les modèles généraux de pronostic.

En même temps, on assiste à une croissance de la dimension des ressources, à une complexité croissante des maladies, à l'introduction de nouvelles technologies de

monitoring, de support et thérapeutique du malade critique, et à la progressive informatisation des USIs à travers le développement de systèmes d'information clinique, ce qui contribue pour que l'USI soit une ambiance génératrice d'un énorme volume de données.

Le monitoring de plusieurs paramètres est une des activités fondamentales chez l'USI. Parmi les paramètres monitorisés heure par heure, par routine dans toutes les USIs, on trouve les suivants: fréquence cardiaque, saturation artérielle d'O₂, tension artérielle et débit urinaire horaire. Les déviations des limites de la normalité de ces quatre paramètres (évalués en termes de fréquence, durée et gravité) sont désignés dans ce travail par évènements cliniques adverses.

Malgré l'importance donnée au monitoring de ces évènements cliniques adverses, leur interprétation et leur valorisation n'ont pas comme résultat des pratiques normalisées et ne produisent pas de «nouvelles» connaissances. Le manque d'une intégration effective des évènements cliniques adverses et de leur représentation comme connaissance limite leur utilité dans l'appui/ support à la décision clinique.

Dans la pratique clinique, l'information mise à disposition par le monitoring des évènements cliniques adverses et par les modèles généraux de pronostic et indices de quantification de défaillance organique est interprétée de façon non-articulée, n'apportant pas de bénéfices à la connaissance inhérente aux synergies potentiellement existantes. Malgré la richesse des données disponibles au chevet du malade, en général, la synthèse et l'interprétation de cette information est faite de forme compartimentée et par conséquent, on n'en profite pas suffisamment dans le procès de décision clinique.

Cette réalité complexe (volume et diversité de données) pose de nouveaux défis aux intensivistes pour l'intégration et l'extraction de connaissance ou d'information utiles sur le pronostic et l'évolution de la gravité de la maladie.

La quantité d'information sur les déterminantes pronostiques du malade grave est aujourd'hui extensive et on doit compléter les méthodes actuelles statiques avec des modèles dynamiques, qui s'adaptent à chaque fois qu'il y a une nouvelle information.

L'évolution du procès de décision clinique a montré une ouverture croissante à l'introduction de techniques d'analyse intelligente de données pour l'extraction de connaissance utile.

Un des propos de ce travail a été d'explorer l'incorporation de données dynamiques, obtenues à partir du monitoring de quatre paramètres physiologiques surveillés par monitoring par routine dans toutes les USIs (modèles cliniques adverses) dans le modèle SAPS II et dans le score SOFA, utilisant méthodes du procès de la Découverte

de Connaissance en Bases de Données pour l'induction de modèles de prévision de la mortalité et défaillance organique, permettant la construction de systèmes automatiques et dynamiques d'appui à la décision intelligente.

Pour le développement de la recherche, on a créé une base de données à partir d'un sous-ensemble de données cueillies lors de l'étude multicentrique EURICUS Elle inclut des enregistrements de 13164 malades de quarante-deux USIs de neuf pays de l'Union Européenne, suivant la méthodologie inhérente au procès du *Data Mining*.

L'étude *Mortality assessment in intensive care units via adverse events using artificial neural networks* a exploré une approximation basée aux événements cliniques adverses et à l'application de réseaux neuronaux artificiels pour la prévision de la mortalité intra-hospitalaire dans une base de données avec 13164 malades. Les résultats des modèles gérés automatiquement ont été comparés avec ceux du SAPSII et d'un modèle obtenu par régression logistique. La capacité discriminative des modèles qui ont comme variables indépendantes seulement les événements cliniques a été significativement supérieure au modèle SAPS II (83,9 versus 79,8%, $p > 0,05$). Le modèle obtenu par réseau neuronal avec l'inclusion de toutes les variables, sauf l'âge, a eu une capacité discriminative de 87,1% (aire sous la courbe ROC).

D'après l'analyse de l'importance relative des événements cliniques adverses, on a vérifié que les événements agrégés avaient la plus grande contribution à l'*outcome* final. D'entre eux, ceux exprimés en débit urinaire horaire et saturation d'O₂ ont eu une importance relative supérieure à celle de la tension artérielle et de la fréquence cardiaque. Il faut souligner que les modèles qui ont présenté de meilleurs résultats ont été ceux qui incluait toutes les variables disponibles.

L'étude *Multiple organ failure diagnosis using adverse events and neural networks* fait une approche exploratoire pour la prévision de la défaillance organique à travers l'application de modèles induits par des réseaux neuronaux artificiels. La défaillance organique à prévoir a été quantifiée par le *SOFA score* par organe à travers sa transformation dans une variable binaire (0 si *SOFA score* < 3 et 1 si *SOFA score* ≥ 3). L'objectif est la prévision de défaillance organique pour le lendemain, à travers l'application de modèles automatiques. Les variables d'entrée ont été : le *SOFA score* par organe du jour, les événements cliniques adverses, le type et origine de l'admission, âge et le *SAPS score*.

Les différents modèles développés présentent des prévisions correctes de défaillance organique entre 87,6% et 94,6%. Ces résultats dépendent de la présence du *SOFA score* dans le modèle de prévision. Le modèle de prévision avec l'inclusion unique des événements cliniques adverses a des acuités prédictives plus basses (70%).

Dans l'étude *Rating organ failure via adverse events using data mining in intensive care medicine* l'objectif principal a été celui d'induire des modèles de prévision quotidienne de l'état fonctionnel d'organes (d'après le SOFA score) en s'appuyant sur les événements cliniques adverses, SAPS II score, âge, type et origine de l'admission dans l'USI. Les trois types de modèles de prévision induits essaient de prévoir pour chaque organe et système : normalité (SOFA score = 0), dysfonction (SOFA score = 1 ou 2) et défaillance (SOFA score ≥ 3). Les prévisions gérées par les réseaux neuronaux ont une meilleure capacité discriminative en comparaison avec celle d'autres techniques utilisées ($p < 0,05$), sauf dans le cas de défaillance du système hématologique.

Dans cette étude, on souligne le fait de ne pas utiliser comme variable d'entrée les valeurs du SOFA score par organe, étudiant de cette manière l'impacte du degré de dysfonction physiologique quotidien (événements cliniques adverses) dans l'état fonctionnel des organes.

Les résultats obtenus montrent que, en association avec l'âge, le SAPSII score, le type et le local d'origine de l'admission à l'USI, les événements cliniques adverses ont une valeur prédictive plus signifiante ($p < 0,05$) pour la prévision quotidienne de la défaillance organique. En vérité, dans ces modèles l'importance relative des événements cliniques adverses est supérieure à 50,4% (variant entre 50,4% et 99,0%).

Le SAPS II score au moment de l'admission a une moindre importance relative que celle des événements cliniques adverses agrégés.

La méthodologie utilisée a permis de constater que l'inclusion des événements cliniques adverses permet la prévision de l'évolution de l'état fonctionnel des organes et des systèmes du malade grave, en temps utile.

Dans le travail *Organ failure prediction based on clinical adverse events : a cluster model approach*, on a appliqué une approche basée sur la segmentation en proposant une matrice tridimensionnelle de support, en permettant la visualisation de l'évolution estimée pour le malade individualisé.

Si on considère les données de l'admission et les événements cliniques adverses, il est possible d'estimer la défaillance organique quotidienne (évaluée par la transformation du SOFA score par organe dans une variable binaire), pendant le séjour à l'hôpital, avec l'information disponible jusqu'au jour respectif.

Dans la population analysée on a identifié automatiquement, par l'application de réseaux de Kohonen, trois segments, dans chaque organe et pour chaque organe et jour de séjour à l'hôpital (jusqu'au cinquième jour dans l'USI) et on a appliqué un algorithme d'induction d'arbres de décision.

Dans chaque segment, les modèles de prévision de l'état fonctionnel des organes et systèmes ont révélé des valeurs de précision supérieures à 74,2%, et il y a même des segments avec des valeurs de précision supérieure à 95%.

Le système INTCare exposé sous la forme de prototype a constitué un des objectifs de cette étude, c'est-à-dire la viabilisation d'un système d'appui à la décision intelligente et met à la disposition des mécanismes pour l'acquisition automatique de données et l'actualisation de modèles de pronostic.

Spécifiquement, ce système effectue la prévision de la défaillance organique et systèmes et du risque de mortalité hospitalière.

Il a comme éléments constitutifs des agents destinés au traitement de données qui permettent l'acquisition automatique en temps réel des valeurs de différents paramètres biométriques normalement monitorisés chez les USIs. Dans le centre du système, on trouve des modèles entraînés basés sur des algorithmes d'entraînement de l'Intelligence Artificielle. Le système est autonome incorporant des mécanismes qui permettent la détection de modèles avec une faible utilité et leur remplacement par d'autres. L'actualisation du système est automatique ou l'ensemble des modèles peut être remplacé manuellement.

Abstract

In Intensive care Medicine, the prognosis using the mortality risk evaluation and the organs and systems failure quantification, introduced in the beginning of eighties, has assumed particular relevance.

These models were developed to quantify the degree of physiologic dysfunction and comorbidities that enable the prediction of hospital mortality. The main idea of these models is that "the several clinical variables which are present at the moment the patient is admitted to the ICU are associated to the risk of mortality in hospital. The models commonly in use are built through logistic regression applied to a static score. Their application on clinical practice as support decision tools is still controversial and unaccepted in a universal way.

Besides these models, the scores of organs failure quantification created to monitor the evolution of patient's severity at the ICU do not include the capacity of predicting the organs and systems failure. Thus, their utility as support decision tools is also limited and the information they obtain is not included by the general models of prognosis.

At the same time we assist to the introduction of new monitoring, support and therapeutic technologies and to the implementation of clinical information systems making of ICUs a source of an enormous quantity of data.

The monitoring of several parameters is one of the main activities in a ICU. These parameters which are hourly and routinely monitored are: heart rate, arterial oxygen saturation, blood pressure and hourly urine output. The deviations from the ranges of normality are designed in this thesis as clinical adverse events.

Although the monitoring of these clinical adverse events has been considered relevant, their interpretation has neither resulted in normalized practices nor in "new" knowledge.

There is a lack of effective integration of the clinical adverse events and of their representation as knowledge limiting their use as support clinical decision tools.

In clinical practice, the information available as result of the monitoring of clinical adverse events and the general prognosis models and the scores of organs failure quantification is interpreted in a discontinuous way taking no profit of their capacity as clinical decision support tools.

This is a challenge for intensivists to extract and integrate knowledge or information about prognosis and illnesses evolution severity which might be useful. The extensive quantity of information about the critical patient's prognosis requires the complementarity between static methods and dynamic ones, which may adapt themselves to the new information available.

The clinical support decision has increasingly adopted data intelligent analysis techniques for useful knowledge extraction.

One of the purposes of this work is to explore the incorporation of dynamic data obtained through the clinical adverse events into the models SAPSII and SOFA score, by using methods of Knowledge Discovery from Databases to induce mortality and organs failure prediction and to enable the creation of automatic systems for the intelligent clinical support decision.

For this purpose, it was developed a database using a data set obtained through the EURICUS II study considering the records of 13164 patients from forty-two ICUs of nine countries of the European Union, by using a methodology of Data Mining.

The study *Mortality assessment in intensive care units via adverse events using artificial neural network* was based on the use of adverse events and artificial neural networks for the prediction of mortality in intensive care units. The results were compared with the ones obtained from the prognostic model SAPS II and a logistic regression model.

The discriminative power of the models using only the adverse events as variables was superior to the one obtained by SAPS II (83,9 versus 79,8%, $p > 0,05$). The model

obtained by neural networks including all the variables except age has shown a discriminative power of 87,1%.

From the analysis of the relative importance of the clinical adverse events (aggregated) it was demonstrated that those had the highest impact on the final outcome. Among them those in hourly urinary output and arterial oxygen saturation had a relative importance superior to the ones in blood pressure and heart rate. The models which show better results include all the available variables.

The study *Multiple organ failure diagnosis using adverse events and neural networks* makes an exploratory approach for the organs failure prediction. The organs failure was quantified by SOFA score by organ, transforming it into a binary variable (0 if SOFA score <3 and 1 if SOFA score ≥ 3). The goal is the organs failure prediction for the following day. The input variables were the daily SOFA score by organ, the clinical adverse events, the type and source of admission, age and the SAPS score.

The different models which were developed reveal precise predictions of organs failure between 87,6 and 94,6%. The prediction models which only include clinical adverse events show lower acuity prediction (70%).

In the study *Rating organ failure via adverse events using data mining in intensive care medicine* the main objective is to induce daily prediction models of organs functional conditions (aferido by SOFA score) based on clinical adverse events, SAPS II score, age and type and source of admission into the ICU. In this study several techniques of data mining were applied. The three types of induced prediction models try to predict for each organ and system normality (SOFA score = 0), dysfunction (SOFA score = 1 or 2) and failure (SOFA score ≥ 3).

The predictions which result from neural networks show better discriminative power than other techniques ($p \leq 0,05$). It should be noticed that the values of SOFA score by organ were not used as input variable. The obtained results reveal that, in association to age, SAPS II score, type and source of admission to the ICU, the clinical adverse events has a significant predictable value ($p < 0,05$) for the daily organs failure prediction.

In these models the relative relevance of clinical adverse events is higher than 50% (from 50,4% to 99,0%). The SAPS II score at the moment the patient is admitted has lower relative relevance than the aggregated clinical adverse events. The methodology applied shows that the inclusion of clinical adverse events enables the prediction of critical patient organs and systems functional condition, in real time.

In the study *Organ failure prediction based on clinical adverse events: a cluster model approach* it has been used an approach based on segmentation by proposing a three-dimensional support matrix to enable the visualization of the predicted evolution of the

individual patient. Considering the admission data and the clinical adverse events, it is possible to predict the daily organs failure, during the hospital stay, using the available information till the respective day. They were identified in the considered population three segments by using the Kohonen networks. In each segment, the prediction models of functional condition of organs show acuity values higher than 74,2%, and in some segments, higher than 95%.

The INTCare system as a prototype is one of the goals of this work, that is, the realization of an intelligent decision support system to supply automatic data acquisition mechanisms and prediction models update. More specifically, this system achieves the organs failure and hospital mortality prediction. In terms of constituent elements it shows dedicated agents to the data treatment which enable the automatic acquisition in real time.

In the system's centre there are models which are trained from Artificial Intelligence's training algorithms. The system is autonomous and includes mechanisms that enable the detection of models that show low utility and their substitution by other ones. The system's update is automatic otherwise the set of models may be manually substituted.

ERRATA

Na página 182, devem constar, imediatamente antes do sub-capítulo “3.2 Balanced Training”, as tabelas:

Table 5: The feature selection performances (in percentage).

Organ	A			B			C			D		
	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.
respirat	86.3	72.4	90.2	86.2	70.0	90.8	77.9	4.4	98.8	77.6	1.8	99.4
coagulat	97.4	68.8	98.7	97.3	59.6	99.0	95.8	4.6	99.9	95.7	0.0	100
liver	98.3	68.6	99.1	98.3	60.2	99.4	97.3	7.6	99.9	97.3	0.0	100
cardiova	94.2	84.1	96.3	94.2	84.0	96.3	82.8	7.5	99.0	82.2	0.5	99.8
cns	95.7	92.7	96.4	95.7	92.3	96.4	83.5	23.4	97.1	81.6	0.4	99.9
renal	95.5	71.3	97.8	95.3	66.6	98.1	91.4	5.7	99.7	91.1	0.3	100
Mean	94.6	76.3	96.4	94.5	72.1	96.7	88.1	8.9	99.1	87.6	0.5	99.96

Acc. - Accuracy, Sen. - Sensitivity, Spe - Specificity.

Table 6: The balanced C, D and C improved performances (in percentage).

Organ	C			D			C (improved)		
	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.
respirat	61.3	66.4	59.8	67.1	41.1	74.5	63.3	70.4	61.3
coagulat	67.6	66.8	67.7	73.7	41.5	75.1	70.0	72.0	69.9
liver	70.0	71.6	70.0	66.9	36.5	67.8	72.5	77.3	72.4
cardiova	65.9	62.5	66.7	68.2	37.9	74.8	69.1	66.3	69.8
cns	73.6	63.9	75.7	66.8	36.3	73.7	75.2	72.2	75.8
renal	67.8	65.6	68.0	73.2	37.6	76.6	71.9	70.5	72.0
Mean	67.7	66.2	68.0	69.3	38.5	73.8	70.3	71.5	70.2

Acc. - Accuracy, Sen. - Sensitivity, Spe - Specificity.

Table 7: The classifiers performances for the C selection balanced data (in percentage).

Organ	Naive Bayes			JRIP			Neural Networks		
	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.
respirat	73.5	25.2	87.3	62.8	61.9	63.0	63.3	70.4	61.3
coagulat	83.3	24.8	85.8	67.8	62.4	68.0	70.0	72.0	69.9
liver	70.8	54.3	71.2	75.7	73.7	75.7	72.5	77.3	72.4
cardiova	73.4	33.4	82.0	66.6	70.3	65.8	69.1	66.3	69.8
cns	76.3	41.3	84.2	77.6	74.4	78.3	75.2	72.2	75.8
renal	76.8	45.6	79.9	69.1	68.5	69.2	71.9	70.5	72.0
Mean	75.7	37.4	81.7	69.9	68.5	70.15	70.3	71.5	70.2

Acc. - Accuracy, Sen. - Sensitivity, Spe - Specificity.

Na página 198, deve constar a tabela:

Table 4: The discrimination power (AUC values, in percentage) for each organ condition and method (values of AUC>70% are in bold)

Condition	Organ	Method		
		DT	LR	NN
normal	respiration	63.50±0.25	69.42±0.17	71.27 ± 0.23 (8)*
	coagulation	53.13±0.13	64.05±0.18	65.36 ± 0.33 (8)
	liver	53.98±0.93	65.71±0.24	66.89 ± 0.43 (8)
	cardiovascular	65.50±0.17	69.61±0.17	72.80 ± 0.26 (10)
	neurological	68.53±0.23	71.80±0.16	73.73 ± 0.26 (10)
	renal	63.77±0.32	70.37±0.14	71.70 ± 0.23 (9)
	Mean	61.40	68.49	70.29
dysfunction	respiration	53.82±0.28	59.37 ± 0.16	62.48 ± 0.23 (8)
	coagulation	50±0	60.68 ± 0.20	61.45 ± 0.36 (8)
	liver	50.13±0.26	63.37 ± 0.27	64.87 ± 0.37 (8)
	cardiovascular	58.89±0.23	64.52 ± 0.23	66.99 ± 0.30 (8)
	neurological	56.27±0.66	58.78 ± 0.23	63.57 ± 0.38 (10)
	renal	60.41±0.4	66.78 ± 0.17	67.25 ± 0.30 [†] (7)
	Mean	54.92	62.25	64.44
failure	respiration	53.76±0.24	68.69 ± 0.15	71.32 ± 0.21 (8)
	coagulation	50.00±0.00	73.18 ± 0.41	72.66 ± 0.51[†] (6)
	liver	50.37±0.42	73.45 ± 0.47	76.77 ± 0.82 (8)
	cardiovascular	58.71±0.24	69.49 ± 0.21	72.39 ± 0.36 (10)
	neurological	56.73±0.18	76.40 ± 0.18	78.09 ± 0.27 (10)
	renal	60.14±0.33	74.02 ± 0.39	75.77 ± 0.48 (8)
	Mean	54.95	72.54	74.50

★ - Median number of hidden nodes (H).

Na página 200, deve constar a tabela:

Table 5: The relative importance of the input variables for organ failure (R_a values, in percentage).

Organ (Model [†])	admtype	admfrom	SAPSI	age	BP*	HR*	SpO ₂ *	UR*
respiration (NN)	10.0	8.2	14.8	1.8	31	7.5	25.3	1.3
coagulation (LR)	0.1	0.4	26.4	4.1	8.5	7.5	38.7	14.3
liver (NN)	0.3	0.5	0.1	0.1	1.8	0.8	0.4	96.1
cardiovascular (NN)	5.3	6.5	8.9	5.1	29.5	7.1	9.9	27.7
neurological (NN)	7.8	3.7	25.8	12.3	15.5	9.2	9.7	16
renal (NN)	0.1	0.8	12.8	0.1	1.8	1.1	2.3	80.9

* – All attributes related to the variable where summed (number of events, critical events and the time).

† – Best predictive method (NN – neural network or LR – logistic regression)

Na página 201, deve constar a figura:

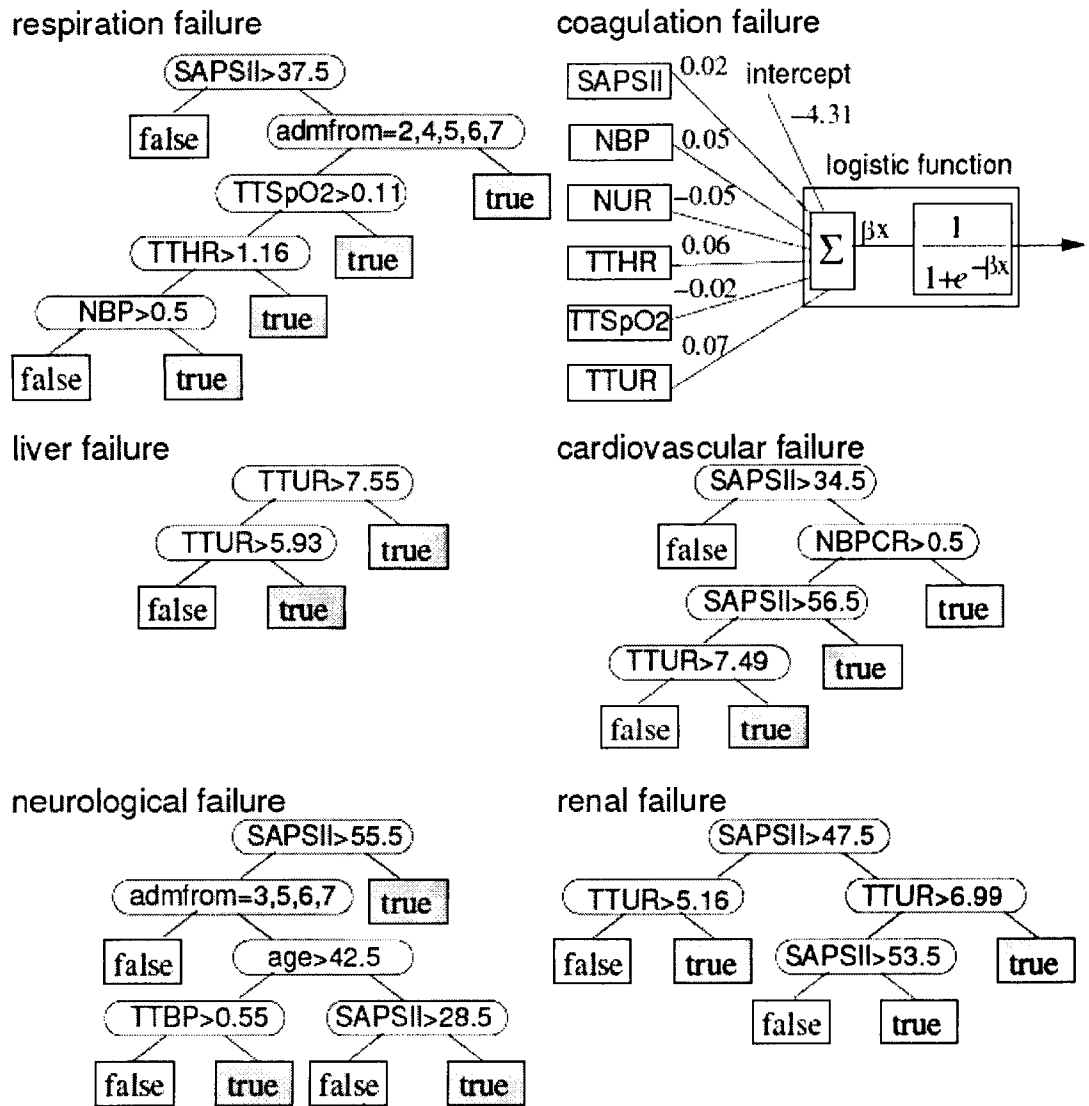


Figure 4:

2/4/08

BIBLIOTECA
DO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS
"ABEL SALAZAR"