

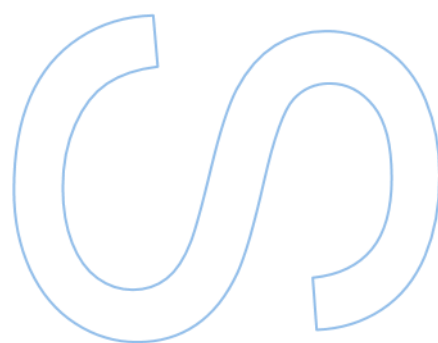
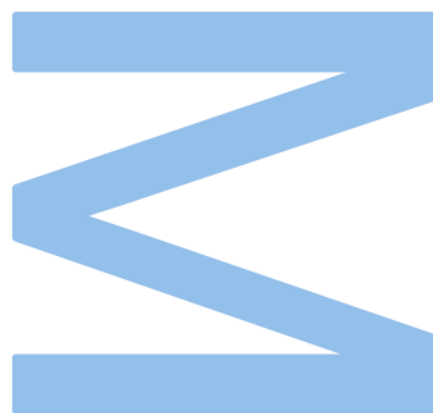
Uma interface para análise exploratória da base de dados MIMIC-III

Nuno Miguel Oliveira Tomás

Mestrado em Engenharia de Redes e Sistemas Informáticos
Departamento de Ciências de Computadores
2022

Orientador

Inês de Castro Dutra
Professor Auxiliar
Faculdade de Ciências da Universidade do Porto



N

S

O

Declaração de Honra

Eu, Nuno Miguel Oliveira Tomás inscrito no Mestrado em Engenharia de Redes e Sistemas Informáticos da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente dissertação reflete as perspetivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega.

Ao entregar esta dissertação, declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente dissertação quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor.

Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Nuno Miguel Oliveira Tomás

29 de setembro de 2022

Agradecimentos

À professora Inês Dutra pelo apoio durante este ano inteiro, disponibilidade e conselhos. A sua capacidade de me fazer ver sempre o lado positivo revelou-se essencial para superar esta etapa da minha vida. Sem a sua ajuda e paciência este trabalho não teria sido possível.

Ao meu pai, irmã e namorada que durante este ano foram incansáveis e deram-me o apoio e ânimo necessário para fazer este trabalho

Agradeço ainda aos meus amigos mais próximos, com quem desabafei imensas vezes e me aconselharam

E, finalmente, em memória da minha mãe, sem a qual este trabalho não existiria.

Resumo

MIMIC-III é a terceira edição de uma base de dados médica de acesso gratuito, contendo registos variados de mais quarenta e seis mil pacientes. Esta é utilizada mundialmente por vários cientistas e investigadores como base dos seus trabalhos e estudos. No entanto, esta não se faz acompanhar de qualquer tipo de interface para ajudar os seus utilizadores a melhor conhecê-la ou aos dados em si contida, contendo dezenas de gigabytes de tamanho, o primeiro contacto com a mesma pode ser bastante intimidante e requererá, obrigatoriamente, ao ser utilizador capacidades de programação para, no mínimo, conseguir extrair informação geral da mesma devido ao seu enorme tamanho.

O objetivo deste trabalho é criar uma interface gráfica, utilizando R e R Shiny, que coleta dados relativos aos pacientes desta base de dados, relaciona as várias tabelas presentes na mesma e apresenta informação relevante de uma forma simples, mas sem descuidar funcionalidades. Através de gráficos, tabelas e texto gerados em tempo real, o utilizador é capaz de ficar a conhecer bastantes detalhes sobre a base de dados sem necessitar de saber qualquer tipo de linguagem de programação.

Na página inicial é oferecida uma visão geral da demografia dos pacientes da MIMIC-III. O restante está dividido em 4 menus principais: “Patients”, “Admissions”, “Diagnoses” e “ICU”, cada uma delas fornece informação específica desse mesmo menu em submenus.

Palavras-chave: MIMIC-III, ferramenta de visualização de dados, R, Shiny, Dashboard.

Abstract

MIMIC-III is the third edition of a freely accessible medical database, containing varied records of over forty-six thousand patients. Used worldwide by many scientists and researchers as a basis for their work and studies. However, it is not accompanied by any kind of interface to help its users get to know it better or the data it contains, which with dozens of gigabytes in size. The first contact with it can be quite intimidating and will require, necessarily, its user programming skills to, at least, extract general information from it due to its enormous size.

The objective of this work is to create a graphical interface, with R and R Shiny, that collects data regarding patients from this database, relates the various tables present in it, and presents relevant information in a simple way, but without neglecting functionality.

Through graphs, tables and text generated in real time, the user can get to know a lot of details about the database without needing to know any kind of programming language.

An overview of MIMIC-III patient demographics is provided on the home page. The remaining information is divided into 4 main menus: “Patients”, “Admissions”, “Diagnoses” and “ICU”, each of which provides specific information from that same menu in submenus.

Keywords: MIMIC-III, data visualization tool, R, Shiny, dashboard

Índice

Declaração de Honra	ii
Agradecimentos	iii
Resumo	iv
Abstract	v
Índice	vi
Lista de Tabelas	viii
Lista de Figuras	ix
Lista de Abreviaturas	xiii
1. Introdução	1
1.1 Motivação	1
1.2 Objetivo	1
1.3 Sumário	2
2. Conceitos fundamentais	3
2.1 Base de dados relacional	3
2.2 Códigos ICD	4
2.2.1 Evolução dos códigos ICD	5
2.2.2 Formato do código ICD-9	5
2.3 MIMIC. Visão e história	7
2.4 MIMIC-III - Breve descrição e acesso à base de dados	8
2.4.1 Métodos usados para recolha de dados na MIMIC-III	9
2.4.2 Atualização	10
2.4.3 Dados Clínicos	10
2.4.4 Dados Fisiológicos para MIMIC-III Waveform	12
2.4.5 Anonimização dos dados da MIMIC-III	13
2.4.6 Merge dos dados	14
2.4.7 Organização da MIMIC-III	14

2.4.8 As tabelas de dados da MIMIC-III	16
2.5 Linguagem R.....	19
2.5.1 R, Linguagem Funcional	19
2.5.2 R, Linguagem Orientada a Objetos	21
3. Estado da arte	23
4. Uma ferramenta de visualização dos dados do MIMIC-III	37
4.1 R e motivação para a sua utilização	38
4.2 Dados.....	38
4.2.1 Tabelas do MIMIC-III utilizadas	39
4.2.2 Tabelas derivadas.....	43
4.3 Desenho – R Shiny dashboard Framework	45
4.3.1 Requisitos de sistema	47
4.3.2 Navegação e menu lateral	48
4.3.3 Ecrãs do dashboard e porquê	50
5. Resultados.....	87
5.1 Sistema de visualização simples, direto, objetivo	87
5.2 Tipos de visualizações possíveis.....	91
5.2.1 Dashboard	91
5.2.2 Patients	95
5.2.3 ADMISSION.....	100
5.2.4 DIAGNOSES	102
5.3 Erros que a ferramenta pode evidenciar	110
5.4 Patient Mortality.....	113
6. Conclusão.....	115
6.1 Limitações	116
6.2 Trabalhos futuros	116
7. Bibliografia.....	117

Lista de Tabelas

Tabela 1- Exemplo de tabela	3
Tabela 2 – Exemplo de Tabela de Compras	4
Tabela 3- Resumo de todas as tabelas da MIMIC-III v1.4. Adaptado de (Johnson et al., 2016).	18
Tabela 4 - Exemplo de um ficheiro de dados em formato CSV	20

Lista de Figuras

Figura 1- Estrutura de um código ICD-9	6
Figura 2- Fonte de dados da MIMIC-II, adaptado de Secondary Analysis of Electronic Health Records pag 45 (MIT Critical Data, 2016).....	12
Figura 3- Tabela ADMISSIONS	15
Figura 4- Tabela ICUSTAYS	15
Figura 5 - Exemplo do funcionamento dos parâmetros únicos de pacientes, hospitalizações e internações. Os IDS são meramente exemplificativos.	18
Figura 6- Origem dos dados da MIMIC-III retirado de (Johnson et al., 2016).....	24
Figura 7- MIMIC-III Diagrama de entidade de relacionamento da MIMIC-III, extraído de (Miles, 2017).	26
Figura 8- Interface gráfica de acesso à base de dados. Devido a falta de credenciais é impossível utilizá-la. Extraído de (Miles,2017).....	27
Figura 9- Interface de acesso à base de dados, com duas queries cujo output é apresentado simultaneamente. Retirado de (Miles, 2017).....	28
Figura 10- Output de dados da ferramenta, suporta mouse-over highlighting. Retirado de (Miles,2017).....	28
Figura 11- PhysioBank Automated Teller Machine ,retirado de ("PhysioBank ATM", 2022).	29
Figura 12- PhysioBank Automated Teller Machine, signal output. Retirado de ("PhysioBank ATM", 2022).....	30
Figura 13- LightWAVE, seleção de input. Retirado de ("LightWAVE 0.71", 2022).....	30
Figura 14- Output de sinal vital pelo LightWAVE. Retirado de ("LightWAVE 0.71", 2022).	31
Figura 15- Ferramenta permite a seleção de múltiplos subjects de uma só vez. Retirado de (Festag & Spreckelsen, 2019).....	31
Figura 16- Output da ferramenta, retirado de (Festag & Spreckelsen, 2019).....	32
Figura 17- MIMIC-II Data Visualization Tool, Introdução retirado de (Lee et al, 2015).....	34
Figura 18-MIMIC-II Data Visualization Tool, Output do menu explore, retirado de (Lee et al, 2015)	34

Figura 19- MIMIC-II Data Visualization Tool, menu Compare, retirado de (Lee et al, 2015).....	35
Figura 20- Diagrama do Dashboard.....	50
Figura 21- Dashboard, página inicial.....	52
Figura 22- Estrutura do submenu patient search by ID.....	54
Figura 23- General Patient Info, antes de selecionar opções no filtro as boxes estão vazias.....	57
Figura 24- Output de summary() dos dados filtrados do submenu General Patients Info .	58
Figura 25- Output dos dados filtrados apresentado numa tabela do submenu General Patients Info.....	58
Figura 26- Output dos dados filtrados apresentados em gráficos, submenu General Patients Info.....	59
Figura 28 -General Admissions Info.....	62
Figura 29- Dataframe para realização de histograma.....	63
Figura 30- Histogramas com mouse-over do submenu Admissions Info.....	63
Figura 31 - Submenu Search by Admission ID.....	65
Figura 32- General Diagnoses Info.....	66
Figura 33- Dataframe DiagnosesPlot.....	68
Figura 34- Mouse-over histograma.....	69
Figura 35- Submenu Search Specific ICD9 group details.....	69
Figura 36- BoxPlot de LOS.....	71
Figura 37 -BoxPlot de LOS, informação em mouse-over.....	71
Figura 38- Diagnoses Search by Specific ICD.....	72
Figura 39- Dataframe resultante de left join de DIAGNOSES com D_ICD_DIAGNOSES	73
Figura 40- Submenu First Diagnoses.....	76
Figura 41- Dataframe DiagnosesPlot2.....	77
Figura 42 Submenu Diagoses Search by patient ID.....	78
Figura 42- Dataframe ADMISSIONS, paciente 305 com 2 admissões.....	80

Figura 43 - Dataframe ADMISSIONS, contendo todas as admissões do paciente 305.....	80
Figura 44- Submenu general ICU stays info.....	82
Figura 45- Submenu Patient Mortality, no days selected.....	85
Figura 46 - Gráficos do submenu Patient Mortality.....	86
Figura 47- Menu Lateral, organizado em menus e submenus.....	88
Figura 48- Menu lateral exposto.....	88
Figura 49 - Menu lateral oculto.....	89
Figura 50 - Submenu General Patients Info, localização do menu lateral de input na esquerda.....	90
Figura 51 - Patient Search by ID, menu de input na esquerda.....	90
Figura 52- Submenu Specific ICD9 group details, menu de input na esquerda.....	91
Figura 53 -Dashboard, página inicial.....	92
Figura 54 - Histograma de idades, mouse-over na idade 51 anos.....	93
Figura 55- Histograma de LOS.....	94
Figura 56-Zoom in no histograma de LOS.....	94
Figura 57- Histograma de LOS, mouse-over.....	95
Figura 58 - Submenu general patient info, seleção de género Masculino, etnia branca e seguro"selfpay".....	96
Figura 59- Tabela dos pacientes incluídos no filtro.....	97
Figura 60 - Pesquisa na tabela pelo ano 2187. Download destes dados é possível.....	98
Figura 61- Pesquisa na tabela pelo ano 2187.....	98
Figura 62- Pesquisa pelo SUBJECT_ID 1382 no submenu Patient Search y ID.....	99
Figura 63- Pesquisa pela admissão HADM_ID 183420.....	100
Figura 64- 5 separadores do submenu general admissions info.....	100
Figura 65- Submenu General Admissions Info, separador Admissions Type.....	101
Figura 65- Mouse-over sob barras do histograma apresenta informação.....	101
Figura 67 - Histograma da frequência dos grupos de ICD.....	102

Figura 68- Mouse-over sobre a barra 390-459 do histograma.....	103
Figura 69 - Comparação de boxplots do gráfico com mouse-over sob a barra 520-579.....	104
Figura 70- Zoom com mouse-over na barra 001-139.....	104
Figura 71- Zoom com mouse-over na barra 140-239.....	105
Figura 72 - Seleção de input no submenu general diagnoses info.....	105
Figura 73 - Doenças do sistema circulatório selecionada.....	106
Figura 74 - Seleção de doenças do sistema circulatório.....	107
Figura 75 - Seleção de doenças mentais.....	107
Figura 76 - Sugestão de opções durante inserção do código ICD.....	108
Figura 77 - Seleção da patologia com doença com ICD 4019.....	108
Figura 78- Submenu First Diagnoses.....	109
Figura 79 - Tabela comparando código ICD9 com Diagnosis.....	110
Figura 80- Paciente 124 com perfuração do intestino.....	110
Figura 81- Filtragem da tabela pelo HADM_ID 138376.....	111
Figura 82- Data de admissão do paciente 124 no submenu Patient Search by ICD.....	112
Figura 83- Admissão 138376 do paciente 24.....	112
Figura 84- Estudo de mortalidade 28 dias	114

Lista de Abreviaturas

ATM	AUTOMATED TELLER MACHINE
BIDMC	BETH ISRAEL DEACONESS MEDICAL CENTER
CSV	COMMA SEPARATED VALUE
DMF	DEATH MASTER FILE
HIPAA	HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY CAT
ICD	INTERNATIONAL CLASSIFICATION OF DISEASES
ICU	INTENSIVE CARE UNIT
LOS	LENGTH OF STAY
MIMIC	MEDICAL INFORMATION MART FOR INTENSIVE CARE
SSA	SOCIAL SECURITY ADMINISTRATION
UI	USER INTERFACE

1.Introdução

1.1 Motivação

MIMIC-III é a terceira edição de uma base de dados médica de acesso gratuito, contendo registos variados de mais quarenta e seis mil pacientes distintos do hospital norte-americano Beth Israel Deaconess Medical Center, entre os anos 2001 e 2012. Esta base de dados é utilizada mundialmente por vários cientistas e investigadores como base dos seus trabalhos e estudos.

No entanto, esta não se faz acompanhar de qualquer tipo de ferramenta introdutória que facilite o contacto com a mesma, sendo composta por dezenas de tabelas, algumas com várias dezenas de *gigabytes*.

Atualmente, não existe qualquer tipo de ferramenta que permita ao utilizador interagir com a base de dados e ficar a conhecer os dados presentes na mesma sem possuir conhecimentos de programação. Existem trabalhos que implementam interfaces de visualização de séries temporais da MIMIC-III (Festag & Spreckelsen, 2019), ("PhysioBank ATM", 2022) ou ("LightWAVE 0.71", 2022), mas ferramentas que permitam conhecer dados demográficos dos pacientes e outros dados, apenas (Lee et al, 2015) implementou uma interface gráfica para o fazer, que já não se encontra disponível.

1.2 Objetivo

Assim, o objetivo do nosso trabalho é criar uma ferramenta que ajude os seus utilizadores a conhecer detalhes sobre a MIMIC-III e os seus pacientes, sem necessitar de saber qualquer tipo de linguagem de programação. Uma ferramenta com uma interface visualmente apelativa e bem organizada, rica em funcionalidades que ajude não só novos utilizadores da MIMIC-III a conhecer detalhes da base de dados, mas também utilizadores mais avançados da mesma. O facto desta ferramenta permitir consultar dados que requeriam *queries* longas de SQL de várias tabelas da MIMIC, por exemplo, permite-lhes

poupar tempo na sua investigação. Além disso, a possibilidade de descarregar a informação apresentada para a sua própria máquina é uma mais-valia, permitindo que a ferramenta seja utilizada para estudar a viabilidade de um estudo na MIMIC-III, por exemplo.

Esta ferramenta não se limita apenas a filtrar dados de tabelas já existentes do MIMIC-III: relaciona-as e apresenta informações importantes que, de outra forma, estariam espalhadas por diferentes tabelas; apresenta dados estatísticos relevantes em formato de texto, gráficos interativos e tabelas e ainda possui submenus que podem ajudar a iniciar estudos, tais como a comparação da triagem feita á entrada no hospital com a real enfermidade diagnosticada pelo médico ou a mortalidade dos pacientes após receberem alta do hospital.

1.3 Sumário

Nesta tese começamos por primeiro introduzir informação relevante no capítulo 2, abordando conceitos necessários para compreender a nossa ferramenta e fornecendo, simultaneamente, contexto para o nosso trabalho.

No capítulo 3 é detalhado o estado da arte, tendo sido realizada uma pesquisa de artigos relacionados com a MIMIC-III e de ferramentas já existentes para visualização de dados da mesma. O capítulo 4 aborda em detalhe, não só a metodologia utilizada na construção da ferramenta, mas também as técnicas utilizadas para resolver os problemas encontrados. Fornece ainda uma explicação do porquê de algumas funções serem implementadas, descrevendo, de seguida, como o foram.

Exemplos de utilização dos diferentes menus da ferramenta são apresentados no capítulo 5, realçando a usabilidade da ferramenta, que tipo de informações da base de dados revela direta e indiretamente e de que forma é realizada a visualização dos dados.

O capítulo 6 retira ilações, sendo feita uma crítica e comparação a trabalhos já existentes, apontando limitações e pontos fortes do nosso trabalho e ainda que trabalhos futuros seriam interessantes.

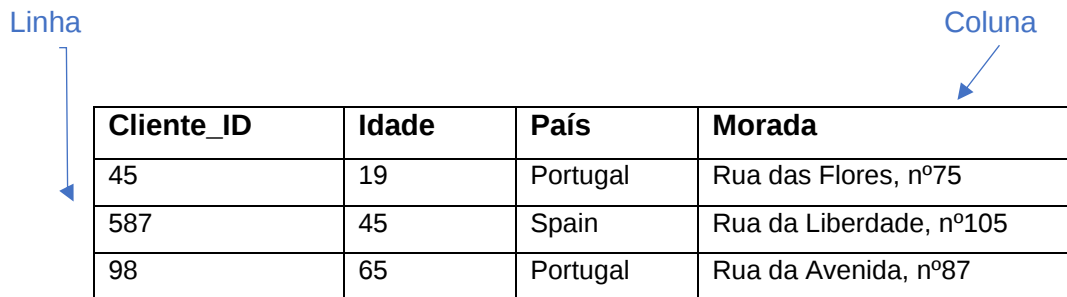
2. Conceitos fundamentais

Este capítulo cobre conhecimento base dos conceitos necessários para entender a ferramenta de visualização de dados MIMIC-III, abordando um diferente a cada subsecção deste capítulo.

2.1 Base de dados relacional

MIMIC-III é uma coleção de dados de pacientes organizados numa base de dados relacional. Para compreender a forma como a MIMIC-III está estruturada é necessário primeiro compreender o que é uma base de dados relacional e a forma mais fácil de o explicar é comparando-a com uma base de dados não relacional.

Normalmente, numa base de dados relacional os dados são organizados em tabelas que contêm informação sobre uma entidade e representam categorias pré-definidas através de linhas e colunas. Essas tabelas contêm os seus dados organizados em colunas (atributos) e linhas (registos). Os atributos especificam um tipo de dados e cada registo contém o valor desse tipo em específico. Um exemplo de uma tabela desse tipo é a Tabela 1.



Cliente_ID	Idade	País	Morada
45	19	Portugal	Rua das Flores, nº75
587	45	Spain	Rua da Liberdade, nº105
98	65	Portugal	Rua da Avenida, nº87

Tabela 1- Exemplo de tabela

Todas as tabelas de uma base de dados relacional têm um atributo ou coluna conhecido como chave primária, que é um identificador único de uma linha. No caso da tabela X, a chave primária é Customer_ID.

Cada linha pode ser utilizada para criar uma relação entre tabelas diferentes usando uma chave externa - uma referência para uma chave primária de outra tabela.

Cliente_ID	Compra_ID	Idade	País	Morada
45	515654	19	Portugal	Rua das Flores, nº75
587	651215	45	Spain	Rua da Liberdade, nº105
98	165152	65	Portugal	Rua da Avenida, nº87

Tabela 2 – Exemplo de Tabela de Compras

Suponhamos o seguinte exemplo. Consideremos duas tabelas, Tabela 1 e Tabela 2 numa base de dados relacional. Na primeira, o Cliente_ID é uma chave primária que identifica um único comprador na base de dados. O seu ID é único e mais nenhum pode ter o mesmo.

A Tabela 2 contém informações sobre transações, onde Compra_ID é uma chave primária pois não podem existir duas compras com o mesmo ID, é único. Nesta tabela, Cliente_ID é uma chave externa. Podemos assim ligar um cliente com uma compra usando a chave externa (Cliente_ID).

Como podemos observar, as duas tabelas estão agora relacionadas através do Cliente_ID. Uma base de dados desenhada desta forma permite a utilização de *queries* declarativas, que consiste em definir o que queremos extrair do sistema sem expressar como o sistema precisa de calcular o resultado. É esta a grande vantagem de utilizar bases de dados relacionais.

2.2 Códigos ICD

ICD é um acrónimo para *Internacional Classification of Disease* e oferece um método de classificação de doenças, lesões e causas de morte.

Os códigos ICD, publicados pela *World Health Organization* (WHO), foram criados com o objetivo de criar um standard nos métodos de registo e catalogação de doenças diagnosticadas em todo o mundo, facilitando a investigação das suas causas, o seu rastreamento e respetivos tratamentos (National Cancer Institute, 2018).

2.2.1 Evolução dos códigos ICD

Existem várias versões destes códigos desde o início da sua publicação em 1948 (Hirsch et al., 2016).. As primeiras cinco versões dos códigos ICD foram incluídas num único volume que incluía um *index* em ordem alfabética e uma lista tabular. A sexta geração, ICD-6, requeria já dois volumes, incluía não só designações de mortalidade, mas também de morbilidade e foi a primeira a incluir doenças do fórum mental. A versão seguinte, ICD-7, desenvolvida pela *United States Public Health Service*, foi ao encontro das crescentes necessidades dos hospitais, passando a incluir materiais necessários para categorizar necessidades hospitalares (Hirsch et al., 2016). Esta versão partiu de uma modificação da versão anterior, passando a indexar registos hospitalares e a classificar procedimentos cirúrgicos. A *Public Health Service* publicou ainda uma oitava versão, ICDA-8, baseada na ICD-8 e especialmente focada nas necessidades dos Estados Unidos da América em 1968. Esta versão teve uma ênfase especial em comunicação de morbilidades e mortalidade. (Benzon et al., 2008)

2.2.2 Formato do código ICD-9

Esta versão, a nona do *International Classification of Disease*, é a utilizada na base de dados MIMIC-III.

Publicada em 1977 pelo *Department of Knowledge Management and Sharing of the World Health Organization* (Cartwright, 2013), foi uma importante transição que aumentou a granularidade com categorias de 4 dígitos e subdivisões de 5 dígitos (Centers for Medicare & Medicaid Services, 2021).

Um código ICD-9 consiste de um carácter numérico ou alfanumérico no primeiro dígito, sendo os restantes numéricos. Pode conter um mínimo de três dígitos e um máximo de cinco e é decimal após o terceiro dígito (à exceção dos códigos começados por V onde o ponto decimal se encontra entre o segundo e terceiro dígito). A estrutura de um código ICD-9 está detalhada na Figura 1(Cartwright, 2013).

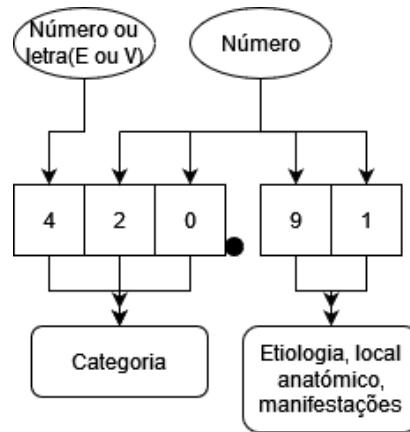


Figura 1 - Estrutura de um código ICD-9

Os primeiros três dígitos ditam a categoria, por exemplo: entre 001 e 139 encontra-se a categoria “Infeções e doenças de parasitas”, entre 140 e 239 “Neoplasmas”.

Existem 17 divisões desde 001-139 a 800-999 (Lesões e Envenenamento). Os códigos começados por V e E correspondem a “Fatores que influenciam cuidados de saúde” e “Causas externas de ferimento”, respetivamente. Os códigos devem também ser sempre atribuídos ao nível mais alto de especificidade e não são adicionados zeros para tornar o tamanho dos códigos uniforme, por exemplo 421.0(0).

Vejamos como funciona esta identificação com o seguinte código ICD-9 de exemplo: 420.91. A primeira parte, 420, encontra-se no grupo entre 390-459, logo é uma doença do sistema circulatório. Dentro deste grupo há subgrupos, por exemplo, 390-392 corresponde a “febre reumática aguda”. Neste caso 420 está no subgrupo “Outras formas de doenças do coração”, onde 420 corresponde a “pericardite aguda”. A seguir temos o grupo correspondente a 420.9, correspondente a “pericardites agudas não especificadas e outros” que contém 420.91. 420.91 identifica finalmente a doença: Pericardite idiopática aguda.

É importante notar que embora os códigos ICD-9 contenham um ponto decimal, na MIMIC-III todos os códigos ICD não o contém. O exemplo anterior, está presente como 42091 na base dados.

Finalmente, os códigos ICD-9 na MIMIC-III são atribuídos no final da estadia do doente e são usados pelo hospital para efeitos de cobrança.

2.3 MIMIC. Visão e história

No início do ano 2000, um grupo de investigadores pertencente ao *Massachusetts Institute of Technology* (MIT) do *Laboratory of Computational Physiology*, reconhecendo o detalhe e importância dos dados recolhidos em ambiente de ICU (*Intensive Care Unit*) por parte dos seus sistemas de monitorização constante, abriu a possibilidade de criar uma nova geração desses mesmos sistemas de monitorização. O objetivo primordial era melhorar o possível desfecho dos pacientes hospitalizados, acompanhando o seu estado fisiológico e recorrendo a modernas técnicas de processamento de sinal, reconhecimento de padrões, modelos computacionais e decisão clínica baseada em conhecimento previamente adquirido.

A sua visão era não só recolher estes dados, processá-los e pô-los à disposição do pessoal médico, mas também esse novo sistema de monitorização ser capaz de formular hipóteses fisiopatológicas que conseguissem melhor explicar os dados observados. Assim, tornar possível a deteção prematura de doenças e ajudar a melhorar as *guidelines* de intervenções terapêuticas (MIT Critical Data, 2016).

Propuseram-se assim, a desenvolver e tornar pública uma base de dados “substancial e representativa” obtida de hospitais e pacientes de ICU. (MIT Critical Data, 2016).

Em 2003 este grupo de investigadores do MIT juntamente com indústria (*Philips Medical Systems*) e o hospital *Beth Israel Deaconess Medical Center* (BIDMC), receberam fundos do *National Reasoning in Critical Care* para lançar o projeto “*Integrating Signals, Models and Reasoning in Critical Care*”, cujo objetivo major era criar uma enorme base de dados de investigação de cuidados intensivos. Este estudo foi aprovado pelo *Institutional Review Boards of BIDMC* (Boston, MA) e pelo MIT (Cambridge, MA).

Este estudo não precisou da autorização individual de cada paciente porque o estudo não impactava os cuidados clínicos prestados e todas as informações sensíveis foram de-identificadas/anonimizadas.

Iniciando o estudo, começaram a ser recolhidos extensos dados quer fisiológicos, quer clínicos de todos os pacientes admitidos ao ICU do hospital BIDMC. O processo de obtenção dos dados foi invisível e continuo para o *staff* médico, sendo cada registo de

paciente iniciado com admissão na ICU e terminado com a sua alta médica final. Assim nasceu a MIMIC-II.

Foram capturados três tipos de dados: dados clínicos, dados fisiológicos de alta-definição (*Waveforms* e séries temporais de sinais vitais, juntamente com alarmes capturados nos monitores instalados nas camas de internamento) e ainda informação de óbitos obtidos do Social Security Administration Death Master Files (Cambridge, MA).

Existem três versões desta base de dados:

- MIMIC-II, contém dados obtidos entre 2001 e 2008, e os seus dados foram capturados unicamente através do sistema de monitorização *Philips CareVue*. Descontinuada ("MIMIC II Clinical Database - Demo", 2016);
- MIMIC-III, contém dados de 2001 a 2012. Inclui dados capturados pela MIMIC-II com o sistema *CareVue*, e foram acrescentados dados extra do sistema *MetaVision* que entrou em funcionamento em 2008 no hospital;
- MIMIC-IV, cujos dados foram capturados em 2008-2019, e a sua fonte é inteiramente *MetaVision*.

2.4 MIMIC-III - Breve descrição e acesso à base de dados

MIMIC-III, um acrónimo para *Medical Information Mart for Intensive Care*, (previamente denominada de *Multiparameter Intelligent Monitoring in Intensive Care*, na sua segunda versão, (Johnson et al., 2016) é uma grande base de dados cujo acesso é gratuito, com dados anonimizados recolhidos relativos a doentes hospitalizados entre os anos 2001 e 2012 que ficaram internados na unidade de cuidados intensivos do hospital Beth Israel Deaconess Medical Center, situado em Boston, Massachusetts. Esta base de dados serve atualmente como suporte para vários estudos epidemiológicos, aperfeiçoamento de decisões clínicas e desenvolvimento de ferramentas eletrónicas.

MIMIC-III inclui variadas informações dos pacientes: informação demográfica, resultados de testes laboratoriais, procedimentos clínicos, medicamentos, sinais vitais com a precisão de ~1 *data point* por hora, notas escritas pelo pessoal médico, relatórios de imagem e mortalidade (tanto no hospital como fora), (Medical Information Mart for Intensive Care, 2021). Para além de incorporar uma população diversa e variada de pacientes de *ICU*,

diferentes idades, etnias e estratos sociais também possui dados muito granulares, incluindo sinais vitais, testes laboratoriais e medicação.

Pelo facto de esta base de dados conter informação tão detalhada e sensível sobre o histórico clínico dos seus pacientes, o seu acesso não é imediato e não basta fazer o seu download o website PhysioNetWorks, que pertence a PhysioNet.

Para assegurar que a MIMIC-III é tratada com o respeito e cuidado que merece é preciso seguir duas etapas que se encontram detalhadas no website da mesma, as quais tivemos de seguir para poder realizar este trabalho (Johnson et al., 2016):

1. Completar um curso online para assegurar a proteção de participantes humanos durante a investigação, que inclui os requerimentos exigidos pela HIPAA (*Health Insurance Portability and Accountability Act*);
2. Assinar um acordo de uso de dados, que baliza qual o uso apropriado desta base de dados, aspetos relativos a segurança e proíbe esforços para tentar identificar pacientes individuais.

Após fornecer os dados académicos e concluir estas etapas de forma cuidadosa, tivemos de aguardar cerca de uma semana para, então, sermos autorizados a descarregar e utilizar a MIMIC-III.

2.4.1 Métodos usados para recolha de dados na MIMIC-III

Todos os dados recolhidos no MIMIC-III foram recolhidos em operações de rotina hospitalar, não interferindo com o fluxo de trabalho do pessoal médico (Goldberger et al., 2000), (Johnson et al., 2016).

Foram recolhidos de várias fontes (Figura 2), entre elas:

- Arquivos oriundos de dois sistemas de informação de cuidados críticos: *Philips CareVue Clinical Information System* entre os anos 2001 e 2008 (*modelos M2331A and M1215A*; *Philips Health-care, Andover, MA*) e *iMDsoft MetaVision ICU* (*iMDsoft, Needham, MA*) entre os anos 2008 e 2012;

- Dados adicionais foram recolhidos nas extensivas bases de dados eletrónicas do hospital;
- SSA-DMF (*Social Security Administration Death Master File*), (DNP, 2017).

O ficheiro DMF (*Death Master File*) oriundo da SSA (*Social Security Administration*) norte-americana é um “ficheiro” criado pela SSA com registo de cidadãos falecidos que possuíam número de segurança social e cujas mortes foram reportadas à SSA. Contém mais de 94 milhões de registos. (National Technical Information Card, 2017)

Os seguintes dados são incluídos: número de segurança social, nome, data de nascimento, e data do óbito. Este ficheiro foi usado para registar a data de óbito de pacientes que tiveram alta do hospital.

2.4.2 Atualização

Em 2008 o hospital decidiu atualizar o seu sistema informático de ICU e conseqüentemente os seus processos de documentação. O objetivo foi substituir o sistema *Philips CareVue* pelo *iMDsoft's MetaVision*. Isto permitiu adicionar aproximadamente 20 000 novas admissões em ICU do que a versão MIMIC anterior, a MIMIC-II.

Assim, foi necessário incorporar duas formas distintas de apresentar os novos dados da ICU na base de dados do MIMIC, e ainda fazer *merge* do novo formato de dados com o design da estrutura existente do MIMIC. O formato *MetaVision* incluía novos formatos de dados tais como notas de progresso médico e registos de medicação tanto oral como intravenosa (MIT Critical Data, 2016).

2.4.3 Dados Clínicos

Dados clínicos de internamento foram descarregados de arquivos dos dois sistemas: CareVue Clinical Information System e de MDsoft MetaVision ICU.

Estes são a origem de dados clínicos como (Figura 2):

- medições fisiológicas verificadas por enfermeiras com data e hora (por exemplo, documentação horária da frequência cardíaca, pressão arterial, ou frequência respiratória);
- notas de evolução clínica documentadas por pessoal médico, seja médicos ou enfermeiras;
- administração de medicação intravenosa contínua e fluídos.

Todos os dados recolhidos pelos dois sistemas de informação foram misturados e fundidos ao construir as tabelas de dados do MIMIC-III, (Alistair et al., 2016).

No entanto, a forma como é registada a administração de fluídos difere entre os dois sistemas de informação. Nestes casos, quando não é possível misturar as duas fontes de informação é criada uma tabela de dados com um sufixo a indicar a fonte da informação. Por exemplo, todos os *inputs* do sistema *CareVue* são armazenados em `INPUTEVENTS_CV`, enquanto os monitorizados com o *Metavision* são armazenados em `INPUTEVENTS_MV`.

Os restantes dados foram extraídos das extensas bases de dados do próprio hospital e de registos de laboratórios de saúde. Entre estes dados incluem-se (MIT Critical Data, 2016):

- resultados de testes laboratoriais: análises sanguíneas, testes microbiológicos, testes químicos, gasometrias;
- dados demográficos de pacientes e mortalidade no próprio hospital;
- dados fisiológicos: sinais vitais horários, códigos de severidade clínica, definições de ventilador
- informações administrativas do hospital: datas de admissão/alta/morte, *tracking* de quartos, etc;
- medicações: prescrições médicas, medicação intravenosa, etc;
- notas e relatórios: sumários de alta, notas de progresso e evolução clínica, relatórios de eletrocardiogramas, ecocardiogramas e de imagem;
- dados relacionados com faturação como códigos ICD-9 (International Classification of Disease 9th Edition), códigos DGR (Diagnosis Group codes) e códigos CPT (Current Procedural Terminology)

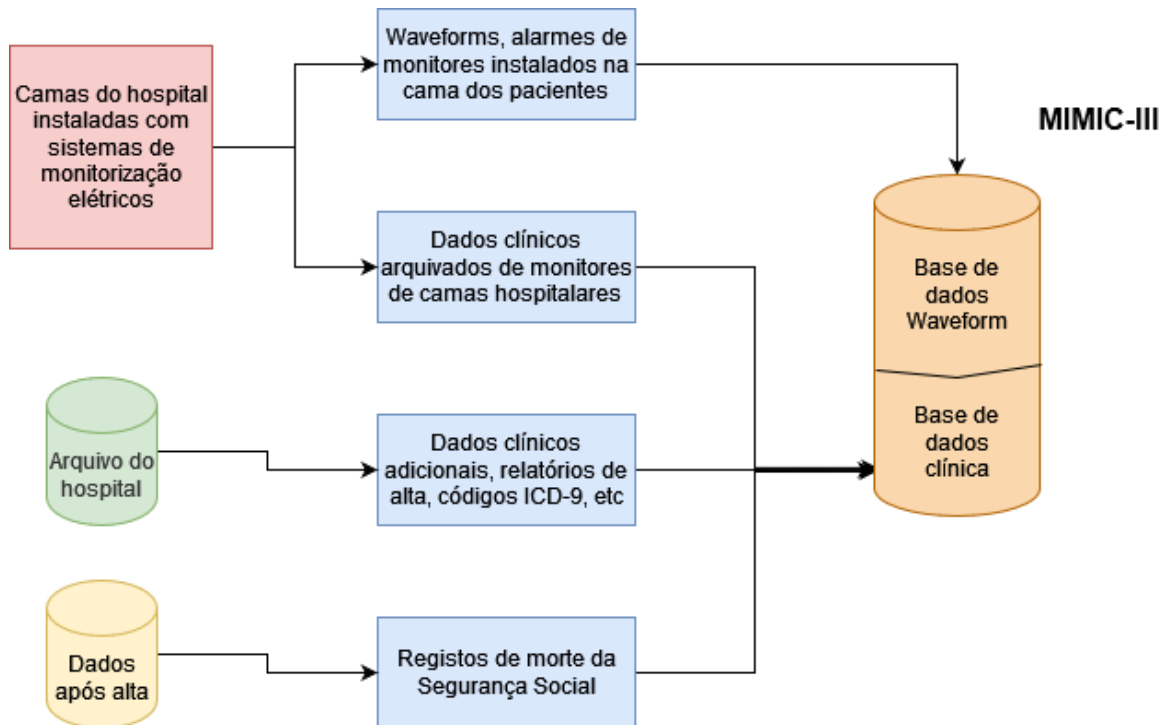


Figura 2- Fonte de dados da MIMIC-II, adaptado de *Secondary Analysis of Electronic Health Records* pag 45 (MIT Critical Data, 2016)

2.4.4 Dados Fisiológicos para MIMIC-III Waveform

Para um grupo de pacientes, dados fisiológicos no formato *Waveform* foram recolhidos de monitores instalados nas suas camas de ICU. Entre estes dados incluem-se: eletrocardiogramas, dados de pressão sanguínea em formato *Waveform*, fotopletismogramas e pneumogramas.

Os monitores adquirem dados fisiológicos com vários parâmetros em formato *Waveform* e digitalizam-nos; processam sinais como frequência cardíaca, pressão sanguínea e saturação de oxigénio em séries temporais (*time-series*). Os dados em formato *Waveform*, tais como eletrocardiogramas, pressão sanguínea e outros, são *sampled* a 125Hz, e as séries temporais atualizadas com frequência de um minuto.

Todos estes dados são depois armazenados numa base de dados central e que serve de suporte às várias ICUs do hospital. A cada duas ou quatro semanas, foram transportados

do hospital para os laboratórios, anonimizados, convertidos para um formato *open-source* e incorporados na base de dados *MIMIC-III Waveform*.

2.4.5 Anonimização dos dados da MIMIC-III

A anonimização de dados pessoais é uma técnica de processamento de dados que consiste em remover ou modificar informações que possam identificar um indivíduo. Essa técnica resulta em dados anonimizados, que não podem ser associados a nenhum indivíduo específico, (Coimbra, 2022).

Os identificadores podem ser diretos ou indiretos. O identificador direto mais comum é o nome que assume relevo especial como sinal designativo que permite a individualização da pessoa. Contudo, nem sempre este identificador direto é bastante para identificar o titular. Por vezes, só o seu cruzamento com outros indicadores indiretos, através de combinações únicas, permite a identificação eficaz do titular. Incluem-se em identificadores indiretos: moradas, idades, números de telefone, etc.

A MIMIC-III teve também de seguir indicações precisas de anonimização dos seus dados, de acordo com as especificações *HIPAA*.

Dados sensíveis e protegidos foram removidos de campos *clear text*, tais como diagnósticos e notas médicas utilizando algoritmos complexos de de-identificação baseados em *pattern-matching* de expressões regulares e pesquisas de dicionário. (Neamatullah et al., 2008). Estes algoritmos demonstraram serem mais capazes em identificar nota médicas sensíveis e que deviam ser protegidas, do que os próprios médicos (MIT Critical Data, 2016).

Antes de os dados serem incorporados na base de dados foram removidos dezoito identificadores identificados pela *HIPAA*, tais como: nome do paciente, número de telefone, morada e datas. Todas as datas foram alteradas, para um tempo futuro, sendo adicionados *X* anos como *offset* aleatório para cada paciente de forma consistente, com o intuito de manter os intervalos temporais. No que diz respeito às horas, dias da semana e sazonalidade aproximada, estes parâmetros foram conservados. É importante salientar

que, de forma a cumprir os regulamentos impostos pela HIPAA, todas as datas de nascimento dos pacientes com mais de 89 anos foram modificadas de modo a ocultar a idade real dos mesmos. Assim, a idade destes doentes é apresentada como sendo 300 ou mais anos.

2.4.6 Merge dos dados

Os pacientes são identificados nas bases de dados do hospital pelo seu número único de doente (*Medical Record Number*) e pelo seu número fiscal. O número fiscal é o que permite identificar um mesmo paciente que teve várias admissões, permitindo fazer *merge* de várias fontes hospitalares. Os dados foram depois organizados numa base de dados relacional bem documentada, que contém registos médicos para cada paciente.

Como referido anteriormente, foi necessário converter o formato proprietária de dados em formato *Waveform* para um *open-source*, neste caso WFDB (*WaveForm DataBase*). Este consiste em um formato de dados *open-source* utilizado por bases de dados públicas no website *PhysioNet* (MIT Critical Data, 2016).

2.4.7 Organização da MIMIC-III

A MIMIC-III é uma base de dados relacional e está organizada em 26 tabelas, todas elas em formato CSV (*comma separated value*). Uma tabela consiste numa estrutura de dados semelhante a folha de cálculo, contendo colunas e linhas. Cada coluna contém informação consistente (tal como um número único de identificação do doente (*SUBJECT_ID*) ou de admissão no hospital Beth Israel Deaconess Medical Center (*HADM_ID*)), e cada linha contém uma instância dessa informação. Por exemplo, uma linha contendo o número inteiro 4 na coluna *SUBJECT_ID* significa que o número identificador do paciente é 4 (Figura 3).

ROW_ID	SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION
1	2	163353	2138-07-17 19:04:00	2138-07-21 15:48:00		NEWBORN	PHYS REFERRAL/NORMAL DELI
2	3	145834	2101-10-20 19:08:00	2101-10-31 13:58:00		EMERGENCY	EMERGENCY ROOM ADMIT
3	4	185777	2191-03-16 00:28:00	2191-03-23 18:41:00		EMERGENCY	EMERGENCY ROOM ADMIT
4	5	178980	2103-02-02 04:31:00	2103-02-04 12:15:00		NEWBORN	PHYS REFERRAL/NORMAL DELI
5	6	107064	2175-05-30 07:15:00	2175-06-15 16:00:00		ELECTIVE	PHYS REFERRAL/NORMAL DELI
6	7	118037	2121-05-23 15:05:00	2121-05-27 11:57:00		NEWBORN	PHYS REFERRAL/NORMAL DELI
7	8	159514	2117-11-20 10:22:00	2117-11-24 14:20:00		NEWBORN	PHYS REFERRAL/NORMAL DELI

Figura 3-Tabela ADMISSIONS

As tabelas são ligadas por identificadores que normalmente têm o sufixo “ID” (Figura 4). Por exemplo, um valor de *SUBJECT_ID* refere-se a um único paciente e é usado para o identificar em todas as tabelas, tal como *HADM_ID*, que por sua vez identifica uma única admissão no hospital Beth Israel Deaconess Medical Center. *ICUSTAY_ID* é outro exemplo, referindo-se a uma única admissão na unidade de cuidados intensivos (ICU).

ROW_ID	SUBJECT_ID	HADM_ID	ICUSTAY_ID	DBSOURCE	FIRST_CAREUNIT
1	2	163353	243653	carevue	NICU
2	3	145834	211552	carevue	MICU
3	4	185777	294638	carevue	MICU
4	5	178980	214757	carevue	NICU
5	6	107064	228232	carevue	SICU
6	7	118037	278444	carevue	NICU
7	7	118037	236754	carevue	NICU
8	8	159514	262299	carevue	NICU

Figura 4- Tabela ICUSTAYS

Uma exceção para esta regra é *ROW_ID*, que é apenas um identificador único nessa tabela.

Todas as tabelas cujo nome tem o prefixo “D_” são dicionários, e contém definições para identificadores. Por exemplo: *D_ICD_DIAGNOSES* contém a definição de cada um dos 14 567 códigos ICD9, com uma coluna com *short-title* e outra com *long-title*, onde, ao contrário da primeira, é descrito mais pormenorizadamente o código ICD9. Assim, fazendo *merge* da tabela *DIAGNOSES_ICD* (que apenas contém o número do código ICD9) e

D_DIAGNOSES_ICD, é possível identificar mais rapidamente qual o diagnóstico do paciente.

Para além disso, todos os eventos anotados por pessoal médico são guardados em tabelas de “eventos”. Um exemplo deste tipo de tabela é o seguinte: *OUTPUTEVENTS* que contém todas as medições de *output* de um paciente (quer seja através do sistema eletrónico *CareVue* ou *MetaVision*, ao contrário do que acontece com *INPUTEVENTS*) ou a tabela *LABEVENTS*, que contém os resultados laboratoriais por paciente.

2.4.8 As tabelas de dados da MIMIC-III

Nesta seção, aprofundamos um pouco mais quais são as tabelas da MIMIC-III, que dados contêm e como podem ser divididas de acordo com o seu conteúdo e consequentemente, a sua função.

Podemos dividir as tabelas em quatro grupos:

- tabelas usadas para definir e seguir pacientes (a azul mais claro na Tabela 3);
- tabelas que contêm informação recolhida em ICU (a azul-claro na Tabela 3);
- tabelas que contêm informação recolhida nos registos do próprio hospital (a cinzento mais claro na Tabela 3);
- tabelas que são dicionários, para cruzar referências dos códigos com as suas respetivas definições (a cinzento-claro na Tabela 3).

A documentação oficial da MIMIC-III refere que, ao contrário da versão anterior, MIMIC-II, na versão da MIMIC tabelas derivadas, não foram incluídas. Uma tabela derivada consiste na junção de duas tabelas ou mais numa só, por exemplo *D_ICD_DIAGNOSES* com *DIAGNOSES_ICD*, ambas com informação relativa a diagnósticos. Nesta versão é encorajada a criação de tabelas derivadas por parte dos seus utilizadores, e foi decidido não as incluir de raiz na MIMIC-III, mantendo assim a base de dados o mais *raw* possível. Uma breve descrição de todas as tabelas presentes na MIMIC-III (v1.4) encontra-se na Tabela 3.

Nome da tabela	Descrição/Função
ADMISSIONS	Cada hospitalização única para cada paciente. Define HADM_ID.
CALLOUT	Informação sobre quando um paciente foi aceite para alta médica e quando a recebeu realmente.
ICUSTAYS	Informação relativa a cada internamento em UCI. Define ICUSTAY_ID.
PATIENTS	Informação de cada paciente único. Define SUBJECT_ID.
SERVICES	Serviço clínico em que cada paciente é registado.
TRANSFERS	Informação relativa a movimentações entre camas dos pacientes dentro do hospital. Inclui entradas e saídas de ICU.
CAREGIVERS	Cada profissional de saúde que registou dados na base de dados. Define CGID.
CHARTEVENTS	Todas as observações manuscritas dos doentes.
DATETIMEEVENTS	Todas as datas de observações. Por exemplo: data de diálise ou inserção de linhas.
INPUTEVENTS_CV	Todos os <i>inputs</i> em ICU de pacientes monitorizados pelo Philips CareVue system.
INPUTEVENTS_MV	Todos os <i>inputs</i> em ICU de pacientes monitorizados pelo IMDsoft MetaVision.
NOTEEVENTS	Notas de-identificadas de enfermeiros e médicos, relatórios de ECG e imagem e sumários de alta.
OUTPUTEVENTS	Todos os <i>inputs</i> de pacientes em ICU.
PROCEDUREEVENTS_MV	Todos os procedimentos médicos para o grupo de pacientes monitorizados em ICU pelo sistema IMDsoft MetaVision.
CPTEVENTS	Procedimentos registados como códigos CPT (Current Procedural Codes).
DIAGNOSES_ICD	Códigos ICD9 atribuídos pelo hospital.
DRGCODES	Códigos DRG (Diagnosis Related Groups) usados para fins de faturação/cobrança.
LABEVENTS	Medições laboratoriais para pacientes no hospital e dentro/fora de clínica.
MICROBIOLOGYEVENTS	Medições microbiológicas e sensibilidades antibióticas
PRESCRIPTIONS	Medicações pedidas (usadas ou não) para um dado paciente.
PROCEDURES_ICD	Procedimentos efetuados nos pacientes seguindo os códigos ICD.
D_CPT	Dicionário de alto nível de códigos CPT.
D_ICD_DIAGNOSES	Dicionário de códigos ICD relativos a diagnósticos.
D_ICD_PROCEDURES	Dicionário de códigos ICD relativos a procedimentos.

D_ITEMS	Dicionário de ITEMID s, exceto os relacionados a testes laboratoriais.
D_LABITEMS	Dicionário de ITEMID s relacionados com testes laboratoriais.

Tabela 3- Resumo de todas as tabelas da MIMIC-III v1.4. Adaptado de (Johnson et al., 2016).

É importante notar que cada ICUSTAY_ID corresponde a uma única HADM_ID e a um único SUBJECT_ID. Da mesma forma, cada HADM_ID corresponde a um único SUBJECT_ID. Ou seja, cada internamento em ICU pertence a uma única hospitalização que por sua vez pertence a um só paciente.

No entanto, é importante notar que para além de um paciente poder ter múltiplas admissões, também pode ter múltiplas hospitalizações em ICU, quer seja na mesma hospitalização ou em várias. Assim, um único SUBJECT_ID pode corresponder a vários HADM_ID e múltiplos ICUSTAY_ID (Figura 5).

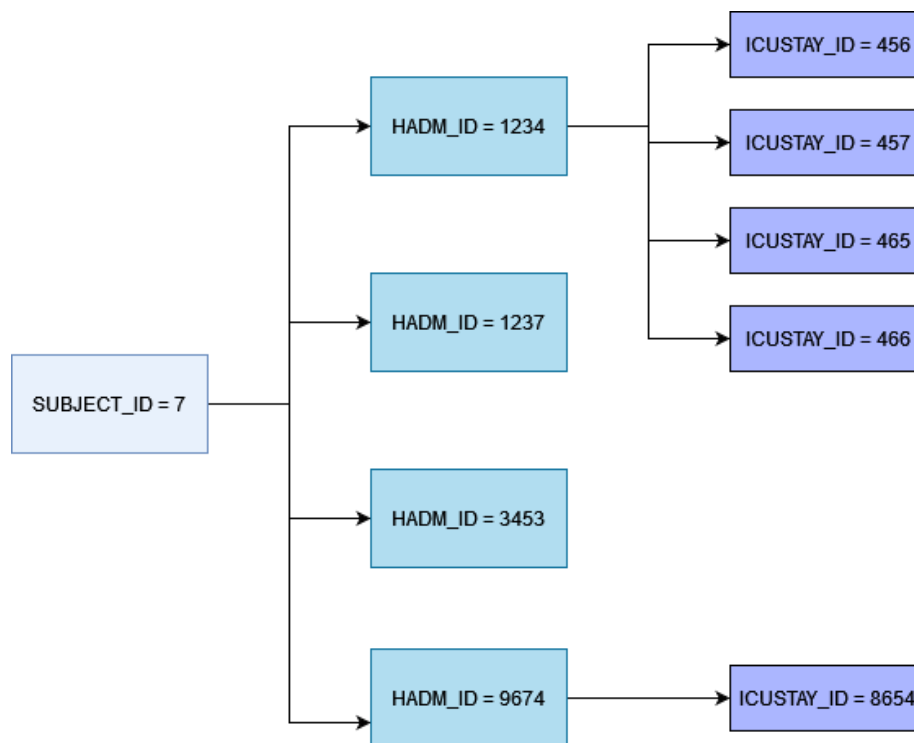


Figura 5 - Exemplo do funcionamento dos parâmetros únicos de pacientes, hospitalizações e internações. Os IDS são meramente exemplificativos.

2.5 Linguagem R

R é uma linguagem de programação, cuja última versão é 4.2.1 distribuída em 23 de junho de 2022. R é uma linguagem descendente de S, podendo mesmo ser considerada uma diferente implementação da mesma pois muito do código escrito para S corre inalterado em R. Trata-se de uma linguagem bastante capaz em manipulação de dados, cálculo e na criação de gráficos (MIT, 2017).

Das suas várias características destacam-se as seguintes que a levaram a ser escolhida para a realização deste trabalho:

- Manipulação de dados eficiente;
- Uma sintaxe fácil de compreender e com uma curva de aprendizagem não muito acentuada;
- Uma larga coleção de soluções para análise de dados;
- Recursos gráficos para análise e apresentação de dados.

R é assim uma linguagem altamente recomendável para análise de dados interativa, incluindo ainda *packages* que ampliam ainda mais as suas capacidades. Embora a sua semântica tenha algumas similaridades com C, R na sua base é uma linguagem funcional, a sua resolução de problemas é centrada em funções. No entanto, também contém uma componente de programação orientada a objetos (MIT, 2017).

2.5.1 R, Linguagem Funcional

Todas as linguagens de programação têm funções, mas isso não as fazem ser consideradas linguagens funcionais. Linguagens deste tipo incluem funções chamadas de *first-class functions*, que são funções que se comportam como qualquer outro tipo de estrutura de dados. Podemos fazer com funções aquilo que podemos fazer com outro tipo de estruturas de dados, como por exemplo vetores: armazená-las em listas, passá-las como argumento a outras funções, criá-las dentro de outras funções, retorná-las como resultado de outras funções e ainda atribuí-las a variáveis.

Outro aspeto que faz uma linguagem ser considerada funcional consiste no facto de requerer funções puras, ou seja:

1. Funções cujo *output* apenas dependa dos *inputs*. Para o mesmo *input* o *output* é o mesmo, ou seja, se chamarmos duas vezes uma função pura com os mesmos argumentos esta vai retornar o mesmo valor;
2. Não têm efeitos colaterais, tais como mudar o valor de uma variável global ou escrever para o disco. Existem claro funções que são exceção tais como *write.csv()* que escreve para o disco um ficheiro em formato CSV ou *<-* que permite mudar o valor de uma variável global.

Esta abordagem tem vantagens pois permite que os programas sejam facilmente testados, sendo possível testar cada componente isoladamente e, depois de perceber o seu funcionamento, tratá-los como uma espécie de caixa negra, com a garantia que não vai alterar outras partes do código.

De uma forma geral programar com uma linguagem funcional convida a dividir um problema em problemas mais pequenos, e resolvê-los com uma função ou combinação de várias funções.

Vejamos um exemplo: imaginemos que contemos o ficheiro CSV *alunos_csv* representado na Tabela 4 com 2 colunas e 5 linhas.

aluno_ID	Idade
1516	18
5165	16
2667	12
2879	17

Tabela 4 - Exemplo de um ficheiro de dados em formato CSV

Se quisermos obter o número de alunos com idade superior a 13 podemos fazer:

```
length(filter(alunos_csv, idade > 13))
```

2.5.2 R, Linguagem Orientada a Objetos

Em R, programação funcional é mais utilizada que programação orientada a objetos em resolução de problemas. A lógica é (como referido anteriormente) ser mais simples dividir um problema em outros mais simples, decompondo-os em simples funções, não em objetos. No entanto há espaço para este tipo de programação em R e tem a sua importância.

Primeiramente é necessário entender o que são classes em R. Em R tudo é um objeto. Um objeto é uma estrutura de dados que contém métodos que atuam sobre os seus atributos. Classes em R são utilizadas como um esboço ou *design* para o objeto e as funções são encapsuladas com os dados.

Existem vários sistemas de OOP, mas os mais relevantes são: S3 e S4. Ambos fazem parte da base do R, são duas gerações distintas de OOP. (Data Flair, ANO)

S3 é a original, a mais simples, menos formal, mais casual, não tem uma definição formal de classes e é menos aberta a validação. É usada para sobrecarregar qualquer função, permitindo diferentes nomes à função, dependendo do tipo de *input* ou do número de parâmetros. S3 é diferente das linguagens de programação convencionais tais como Java, C++ que implementam a passagem de mensagens OO, isto torna-a mais fácil de implementar. Neste tipo de classe, a função genérica chama o método. Resumindo, S3 permite que as funções retornem resultados *user-frindly* e seja mais *programmer-friendly* para o programador. (Data Flair, ANO)

Segue o exemplo de criação de uma classe deste tipo:

```
c <- list(name= "Aluno1232", idade=14,nota= 16)
class(c) <- "aluno"
```

S4 oferece as mesmas funcionalidades, mas requer mais programação, é mais formal e é mais difícil de fazer *debug*. Ao contrário de S3, existe uma definição formal que fornece descrição e representação de classes. Contém ainda funções especiais de auxílio para definição de métodos e as funções genéricas são capazes de captar métodos baseado na classe composta de múltiplos argumentos. S4 é assim um sistema rigoroso que obriga a pensar com cuidado no desenho do programa e está especialmente indicado na construção de sistemas maiores, que evoluem ao longo de mais tempo e recebem contribuições de outros programadores. Exemplo de classe S4:

```
Veiculo <- setClass(
  "Agent",
  # Definição das características
  specs= c(
    potencia = "numeric",
    eletrico  = "logical"
  ),
  # Atribuir valores por defeito para as specs
  (opcional)
  prototipo=list(
    potencia = c(320),
    eletrico  = FALSE
  ),
)
```

3. Estado da arte

Nas últimas duas décadas, a adoção de registos médicos eletrónicos nos hospitais tem vindo a aumentar, e é hoje padrão. Dados de 2015 indicam que nos Estados Unidos, quase 96% dos hospitais possuem um sistema digital de registo de saúde dos pacientes, (EHR). Além dos dados demográficos dos pacientes serem armazenados em formato digital (EMR, Electronic Medical Record), quase todos os hospitais do mundo possuem dispositivos eletrónicos com o intuito de monitorizar e registar dados fisiológicos dos seus pacientes, o que facilita a criação de base de dados como a MIMIC-III. Numa época onde existe um crescente interesse na *Big Data* e consequente análise, avanços em estudos sobre tomadas de decisão clínica fizeram com que o foco de investigação médica abandonasse um pouco a vertente de aquisição e armazenamento de dados e se focasse na sua análise. Um dos primeiros passos para este desenvolvimento é o acesso a largas quantidades de dados EMR, que é exatamente o que o projeto da MIMIC-III oferece. No entanto a sua utilização não é trivial. Existe uma curva de aprendizagem intrinsecamente ligada à mesma, que requer a qualquer investigador tempo para a ultrapassar e compreender. Esta pode ser estreitada com a criação de ferramentas de visualização.

Este capítulo pode ser dividido em duas partes: a primeira aborda a própria base de dados MIMIC-III, artigos e trabalhos que explicam a organização da mesma. A segunda foca-se na existência de ferramentas que permitam visualização de dados médicos, mais concretamente em ferramentas de visualização de dados da base de dados MIMIC.

Como referido anteriormente, a primeira parte da pesquisa feita para este capítulo centrou-se na pesquisa de artigos e trabalhos sobre a MIMIC-III. Foi utilizada a seguinte metodologia: pesquisa no *Google Scholar* compreendida de os seguintes termos ["MIMIC-III" "database"]. Devido à grande popularidade da MIMIC-III são retornados muito artigos, muitos deles de estudos clínicos que usam a MIMIC-III como suporte, o que não é o pretendido. No entanto existem 2 artigos que interessam ser detalhados: (Johnson et al., 2016) e (Miles, 2017).

O primeiro artigo aborda primeiramente o porquê de a base de dados ter sido criada, onde esperam que a mesma seja usada e que melhorias ofereça em relação à versão anterior.

De seguida é detalhada a origem dos dados (Figura 6) constituintes da MIMIC-III e são dadas algumas estatísticas gerais sobre os pacientes que dela compõem.

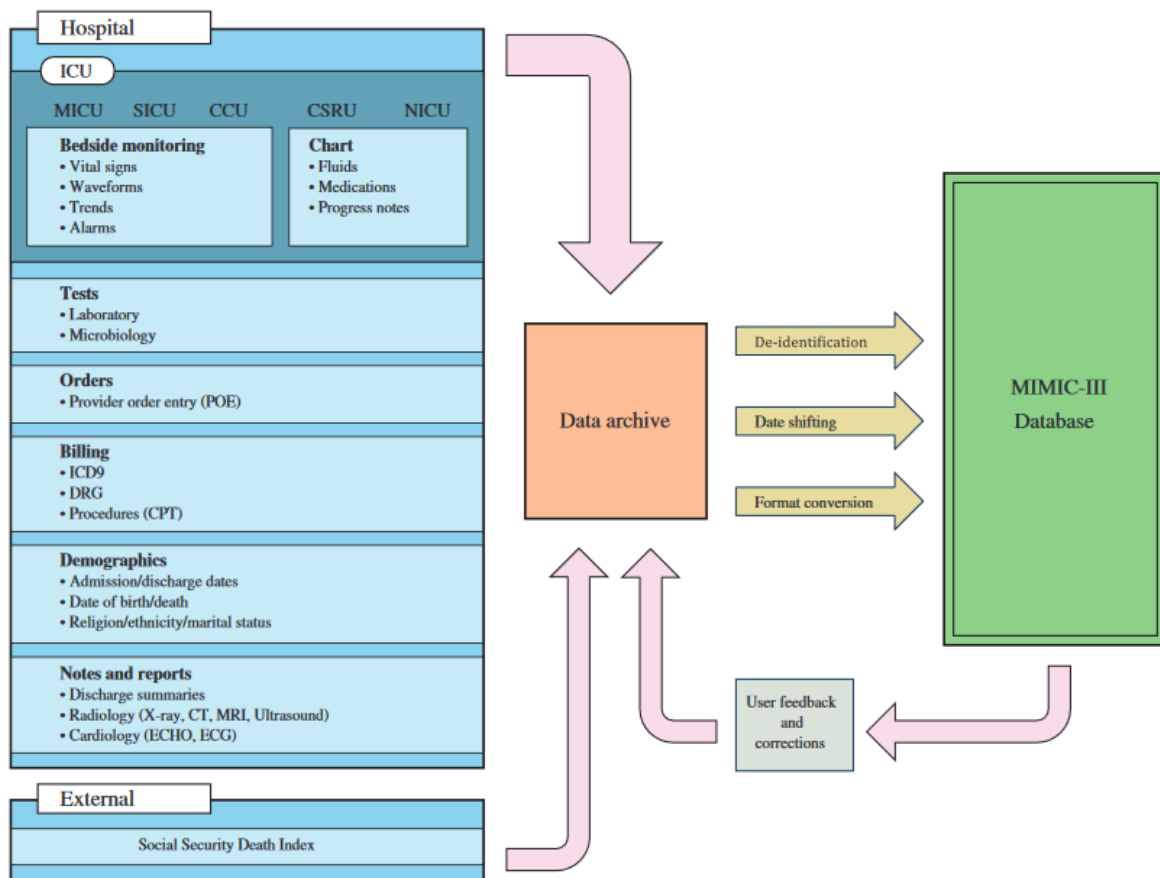


Figura 6- Origem dos dados da MIMIC-III retirado de (Johnson et al., 2016).

É apresentada uma descrição geral estatística sobre os pacientes constituintes da MIMIC-III, mas apenas daqueles cuja idade é superior a 16 anos, pacientes com mais de 16 anos são considerados adultos em (Johnson et al., 2016). Todas as estatísticas da ferramenta são geradas em tempo real com os dados da versão mais recente da MIMIC-III, no caso, versão 1.4, enquanto este artigo foi escrito abordando a versão 1.3 da mesma. Assim, estas estatísticas não serão de particular utilidade, visto não poderem ser meramente copiadas e serem de versões diferentes da base de dados. Servem, no entanto, para ajudar a ter uma ideia da veracidade dos cálculos feitos na ferramenta, pois grandes discrepâncias de valores serão um indicador de possíveis erros.

O número total de pacientes distintos na versão do artigo é 46 467, dos quais 7870 são neonatos e 38 597 são pacientes adultos (Johnson et al., 2016), a versão 1.4 contém números ligeiramente diferentes: 46 520 pacientes, dos quais 7874 são neonatos e 38 637 são pacientes com mais de 16 anos. Os restantes têm idade compreendida entre >1 anos e <16 anos.

Este artigo apresenta ainda estatísticas do número de pacientes por cada unidade de ICU diferente e grupo de doença ICD-9, algo que será implementado na nossa ferramenta de visualização.

O segundo artigo (Miles, 2017) aborda a desconstrução da base de dados MIMIC-III, de forma a torná-la mais leve e fácil de interagir. O objetivo é que seja capaz de receber e responder a *queries* e esteja disponível a todos os alunos da SUNY Oswego que pretendam adquirir e analisar dados para criação de hipóteses de estudo; validar projetos, comparando os seus resultados de estudo com os da MIMIC-III e ainda utilizar para fins de *data-mining*. Os utilizadores serão capazes de interagir com esta base de dados através de uma interface simples criada pelo próprio autor, depois de se autenticarem, utilizando comandos SQL.

O autor do artigo começou por criar um diagrama de *Entity Relationships* (ER), que incluía cada uma das 26 tabelas da base de dados, presente na Figura 7. Este diagrama permite-nos observar de que forma as tabelas da MIMIC-III estão relacionadas umas com as outras e a ainda entender, em cada uma delas, quais são as suas chaves primárias e externas de cada uma delas. Na construção do diagrama da Figura 7, o autor notou que poderiam ser implementadas certas alterações para aumentar a performance da base de dados. Por exemplo, embora

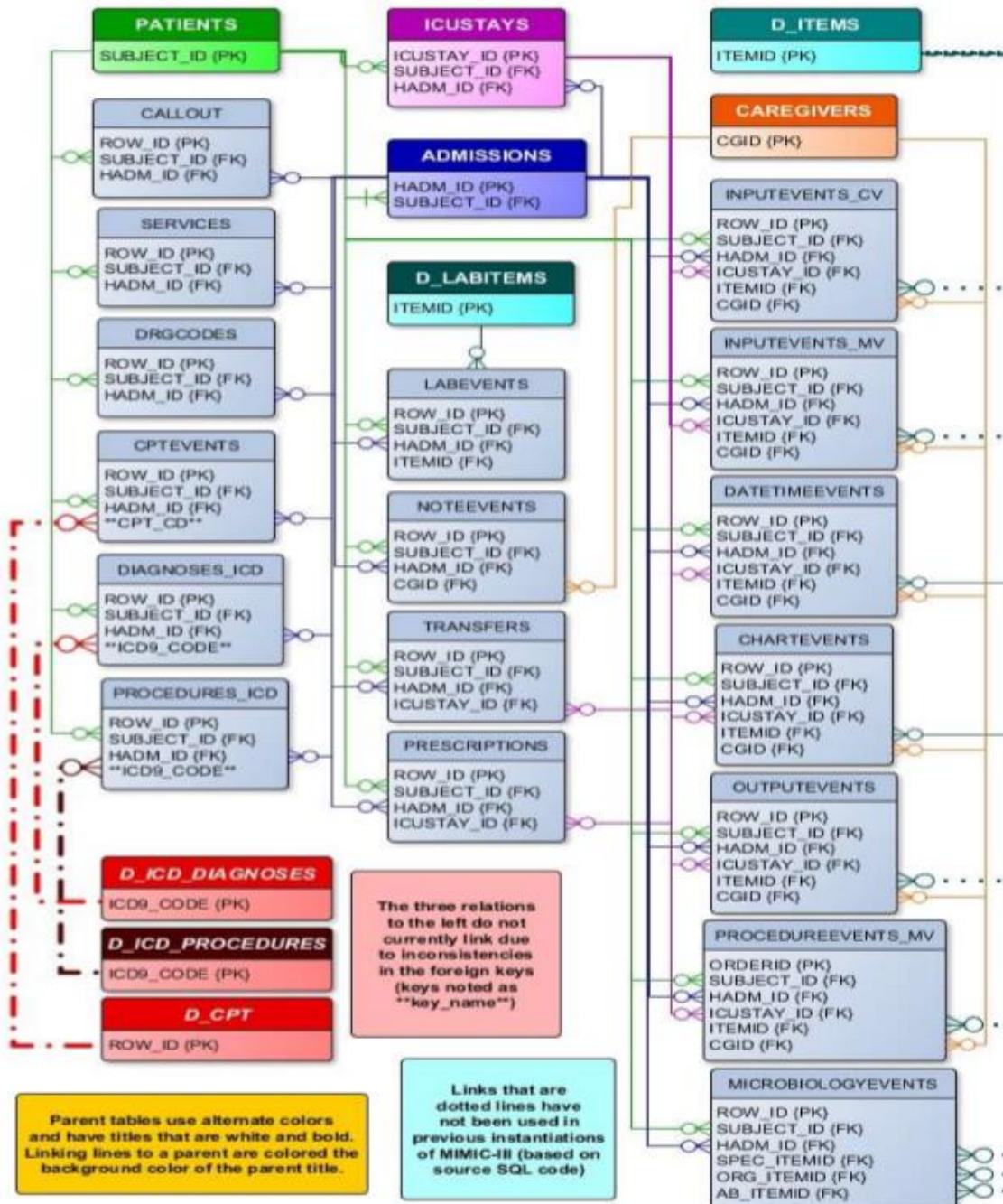


Figura 7- MIMIC-III Diagrama de entidade de relacionamento da MIMIC-III, extraído de (Miles, 2017).

todas as tabelas possuam um atributo "ROW_ID", este é descartado em todas as tabelas onde este não é uma chave primária. Como referido no capítulo 2, ROW_ID é um identificador único da tabela e não é possível ligá-lo a outras tabelas. Por exemplo: ROW_ID, da tabela PATIENTS não se refere ao mesmo ROW_ID da tabela ADMISSIONS.

Por esse mesmo motivo, no próprio diagrama da Figura 7, é excluído o atributo ROW_ID das tabelas onde isso acontece.

Ademais, foram incluídas outras alterações, chaves externas como ICD9_CODE de DIAGNOSES, por exemplo, contém *strings* que não correspondem a nenhum valor na sua tabela “pai”, D_ICD_DIAGNOSES. O mesmo acontece com o atributo CPT_D da tabela CPTEVENTS e ICD9_CD de PROCEDURES. Ambas também contém *strings* não correspondidas nas suas tabelas pai, D_CPT e D_ICD_PROCEDURES, respetivamente (Miles, 2017). Estes são alguns exemplos de alterações realizadas à base de dados MIMIC-III por parte do autor.

De forma a facilitar a interação dos utilizadores da base de dados implementada, foi criada uma interface baseada em PHP. A interface encontra-se disponível (ver Figura 8), e é possível utilizá-la, no entanto não é possível utilizá-la sem credenciais.



Figura 8- Interface gráfica de acesso à base de dados. Devido a falta de credenciais é impossível utilizá-la. Extraído de (Miles, 2017)

A interface suporta múltiplas *queries* ao mesmo tempo, dando o seu *output* sem ecrãs separados [Figura 9 e Figura 10, retirada de (Miles, 2017), devido a ser impossível testar a ferramenta sem as devidas credenciais]. Os utilizadores podem copiar a informação para outros programas, mas não foi implementada uma função de download dos dados.



Figura 9- Interface de acesso à base de dados, com duas queries cujo output é apresentado simultaneamente. Retirado de (Miles, 2017).

ROW_ID	ITEMID	LABEL	ABBREVIATION	DBSOURCE	LINKSTO
15001	227666	Immobilizer Location	Immobilizer Location	metavision	chartevents
15002	227667	Reason for Immobilizer	Reason for Immobilizer	metavision	chartevents
15003	227668	Side Rails_V2	Side Rails_V2	metavision	chartevents
15004	227669	Side Rails (Restraint)	Side Rails (Restraint)	metavision	chartevents

Figura 10- Output de dados da ferramenta, suporta mouse-over highlighting. Retirado de (Miles, 2017).

Para a obtenção de resultados da segunda parte foi utilizada a seguinte metodologia: pesquisa no *Google Scholar* e *PubMed* composta pelos termos ["mimic-iii" "data visualization" "tool"]. Esta pesquisa retorna 3 resultados no *PubMed*: (Johnson et al., 2016), (Festag & Spreckelsen, 2019) e (Alcaide & Aerts, 2021). Destes 3 artigos apenas selecionamos os primeiros dois devido ao facto de apenas estes implementarem interfaces gráficas, que é o objetivo do trabalho.

Existem diversas ferramentas implementadas ao longo dos anos com o objetivo de tornar os dados disponíveis na *PhysioNet* mais acessíveis a investigadores. No entanto, a maior parte delas foca-se nas bases de dados em formato Waveform da MIMIC, e não na base de dados que estamos a tratar. O seu objetivo é primordialmente o *output* dos sinais vitais dos pacientes.

PhysioBank Automated Teller Machine (ATM) é uma ferramenta oferecida pela *PhysioNet* especializada em visualização e extração de sinais temporais de mais de 50 base de dados disponíveis, onde se inserem também as bases de dados MIMIC ("PhysioBank ATM", 2022).

Os utilizadores são recebidos com uma caixa de input onde selecionam uma das várias bases de dados disponíveis, como a *MIMIC Database*. De seguida selecionam um registo de um único paciente, os sinais a serem visualizados, as anotações (se disponível), o formato de *output* (tempo de intervalo da visualização, formato da hora e da data) (Figura 11). Podem ainda selecionar o tipo de apresentação desejada, numa opção "Toolbox". Assim que estas opções são selecionadas a ferramenta permite o *download* de um ficheiro contendo a representação visual dos dados selecionados (Figura 12). Um dos problemas desta ferramenta, para além da seleção algo confusa dos registos a escolher, é a impossibilidade de extrair dados de mais de um paciente de cada vez (Festag & Spreckelsen, 2019).

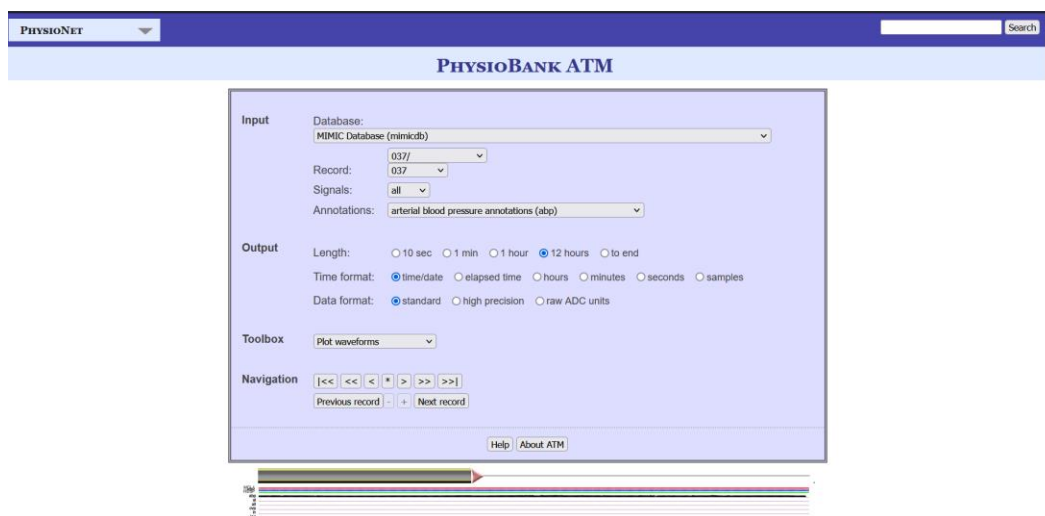


Figura 11- PhysioBank Automated Teller Machine ,retirado de ("PhysioBank ATM", 2022).

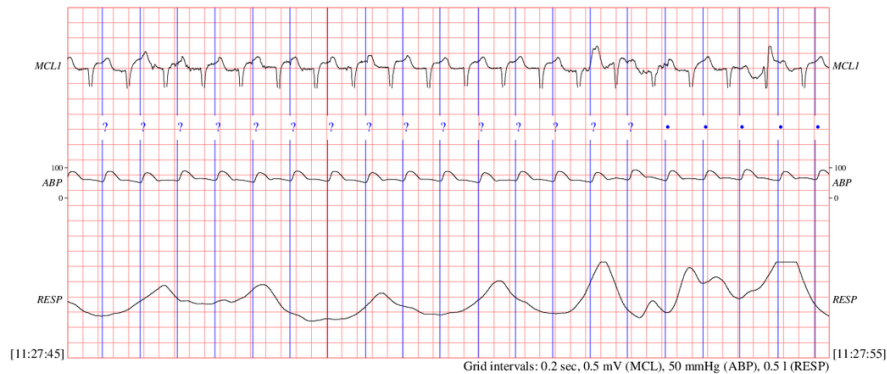
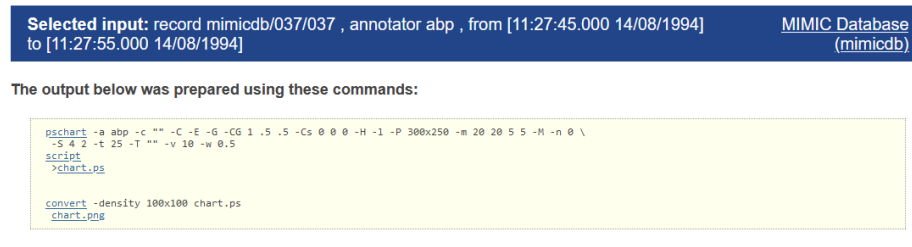


Figura 12- PhysioBank Automated Teller Machine, signal output. Retirado de ("PhysioBank ATM", 2022).

LightWAVE: *Waveform and Annotation Viewing and Editing* ("LightWAVE 0.71", 2022) in a *Web Browser* é outra ferramenta a mencionar e cujas funcionalidade são muito semelhantes à anterior. É um software *open-source*, capaz de correr em qualquer *web browser*, com a função de ler ECGs (eletrocardiogramas) e outros dados fisiológicos em formato *Waveform*. É semelhante a ATM, mas com uma vantagem é ainda capaz de criar e editar anotações armazenadas em conjunto com os dados em formato *Waveform* (Figura 13). Tal como no ATM, só é possível extrair um registo de cada vez. É possível utilizá-la no próprio website (Figura 14) da PhysioNet ("LightWAVE 0.71", 2022).

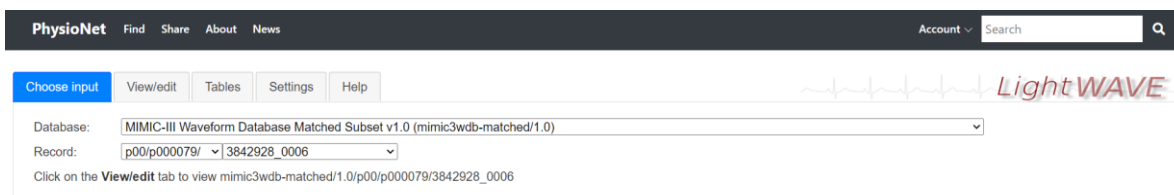


Figura 13- LightWAVE, seleção de input. Retirado de ("LightWAVE 0.71", 2022).



Figura 14- Output de sinal vital pelo LightWAVE. Retirado de ("LightWAVE 0.71", 2022).

Outra ferramenta a mencionar com o mesmo propósito e melhoramentos é referida em (Festag & Spreckelsen, 2019). A base desta ferramenta é implementada em KNIME e oferece duas vantagens em relação às anteriores: simplifica a comparação de dados, permitindo visualizar dados de vários pacientes de uma só vez (Figura 15 e Figura 16) e a visualização dinâmica dos gráficos é mais fácil de utilizar que a oferecida pela ATM. No entanto, existem desvantagens: requer a instalação de KNIME na máquina do utilizador, não bastando um *web-browser* para a utilizar.

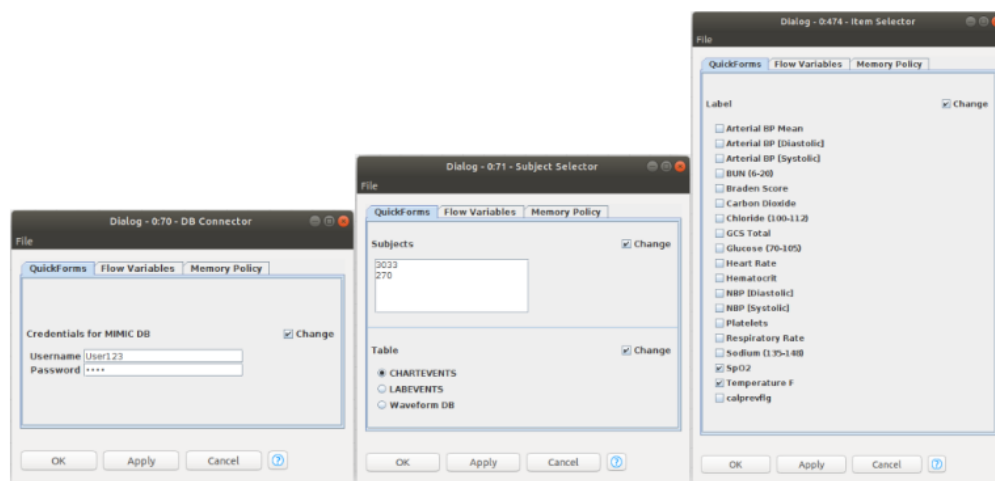


Figura 15- Ferramenta permite a seleção de múltiplos subjects de uma só vez. Retirado de (Festag & Spreckelsen, 2019).

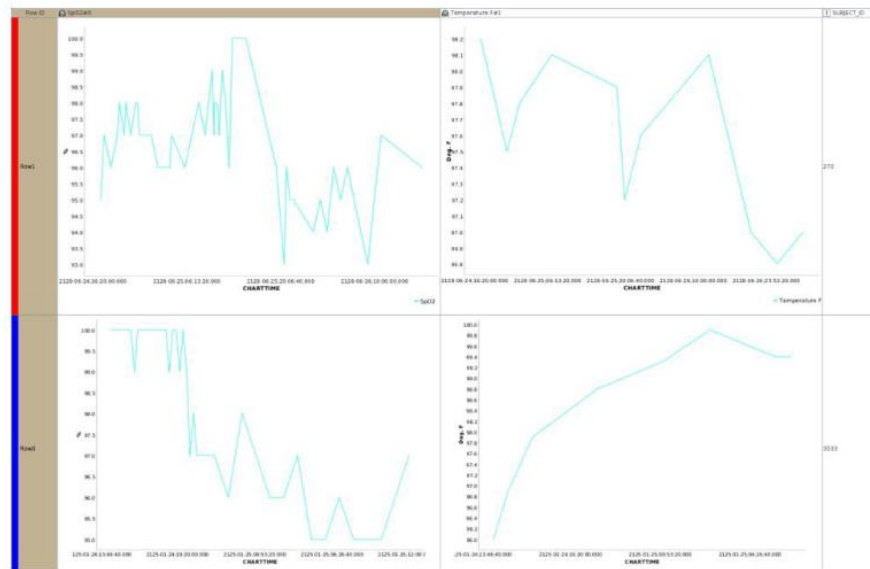


Figura 16- Output da ferramenta, retirado de (Festag & Spreckelsen, 2019).

O artigo mais relevante e alinhado com o objetivo do nosso trabalho é sem dúvida (Lee et al, 2015). Embora o objetivo do mesmo seja a criação de uma ferramenta de visualização de dados para a MIMIC-II, a versão anterior à MIMIC-III, este é provavelmente o trabalho mais similar ao nosso e como tal será detalhado mais pormenorizadamente. O seu objetivo consiste em facilitar a iniciação de investigação científica que se apoia e baseia na MIMIC-II, não só facilitando o acesso a dados da base de dados, como também tornar desnecessária a necessidade de os investigadores serem fluentes em linguagens de programação (tal como SQL). Quer os investigadores sejam fluentes ou não em programação, e conheçam ou não a base de dados, a ferramenta pretende permitir conduzir uma investigação preliminar da MIMIC-II e dos seus pacientes. Além disso, procura acelerar o progresso de investigadores (mesmo os fluentes em SQL/outras linguagens de programação), uma vez que elimina a necessidade de *queries* constantes para obter dados que poderiam ser obtidos de forma mais facilitada. Assim, o objetivo da proposta é permitir ao utilizador a capacidade de agregar estatísticas-chave (tamanho da amostra, distribuição de variáveis clínicas) de um grupo de pacientes da MIMIC-II utilizando apenas um *browser* da internet, tornando desnecessário o download de dados *raw* dos pacientes.

Esta ferramenta está dividida em duas funcionalidades: “explore” e “compare”.

“Explore” permite ao utilizador seleccionar um grupo de pacientes admitido em ICU, baseado no tipo de serviço recebido em ICU, género, idade e código ICD-9. Esta ferramenta não inclui a totalidade dos 5 tipos de ICU presentes na base de dados, os autores decidiram que pacientes pertencentes à NICU (neonatais) são excluídos devido às diferenças fisiológicas e clínicas quando comparado com adultos. Assim, nesta componente o utilizador pode seleccionar um conjunto de variáveis a serem visualizadas: informações demográficas (tais como idade, género ou religião); informações administrativas (tais como tipo de admissão em ICU ou tipo de seguro) ; resultados do paciente, (tais como mortalidade após 28 dias de alta, LOS, mortalidade hospitalar); sinais vitais, tais como batimento cardíaco e saturação de oxigénio; resultados de testes laboratoriais, tais como contagem e glóbulos brancos ou níveis de glicose; intervenções, tais como hemodiálises ou administrações de vasopressores e outros, tais como índice de massa corporal, altura ou ICD-9 primário. Todas estes dados são impressos em formato de histogramas ou gráficos circulares, no caso de serem valores categóricos. Alguns outros dados também são impressos em formato de texto.

A funcionalidade “compare”, tal como o nome indica, possibilita a comparação entre dois grupos seleccionados de pacientes, sendo os critérios de seleção dos pacientes e os dados apresentados iguais aos encontrados em “explore”.

Quanto à sua implementação, foi dividida em *front-end* e *back-end*. A *front-end* foi programada usando HTML, CSS e JavaScript e é aqui que é feito o processamento dos dados a visualizar. A *back-end*, por sua vez, foi programada utilizando PHP e PostgreSQL. A *front-end* comunica com um servidor web Apache hospedado no laboratório Health Data Science Lab(HDSL) na Universidade de Waterloo, que por sua vez submete um pedido a um servidor de base de dados, nesse mesmo laboratório, de uma cópia da base de dados MIMIC-II. Consequentemente, os dados são formatados do lado do servidor e depois visualizados do lado do cliente, no próprio *web browser*, sendo a velocidade de apresentação dos dados dependente da máquina que o cliente está a utilizar.

Infelizmente, ao longo da escrita desta tese, foram efetuadas inúmeras tentativas de consultar a ferramenta seguindo o link disponível no artigo e pesquisando noutros motores de busca, mas sem sucesso. O facto de a ferramenta não estar disponível para ser utilizada, impede que seja testada, assim iremos abordar o seu *design* baseando-nos nos em

imagens da mesma disponíveis no artigo. A Figura 17 retirada do artigo, mostra como é estruturada a página inicial da ferramenta.

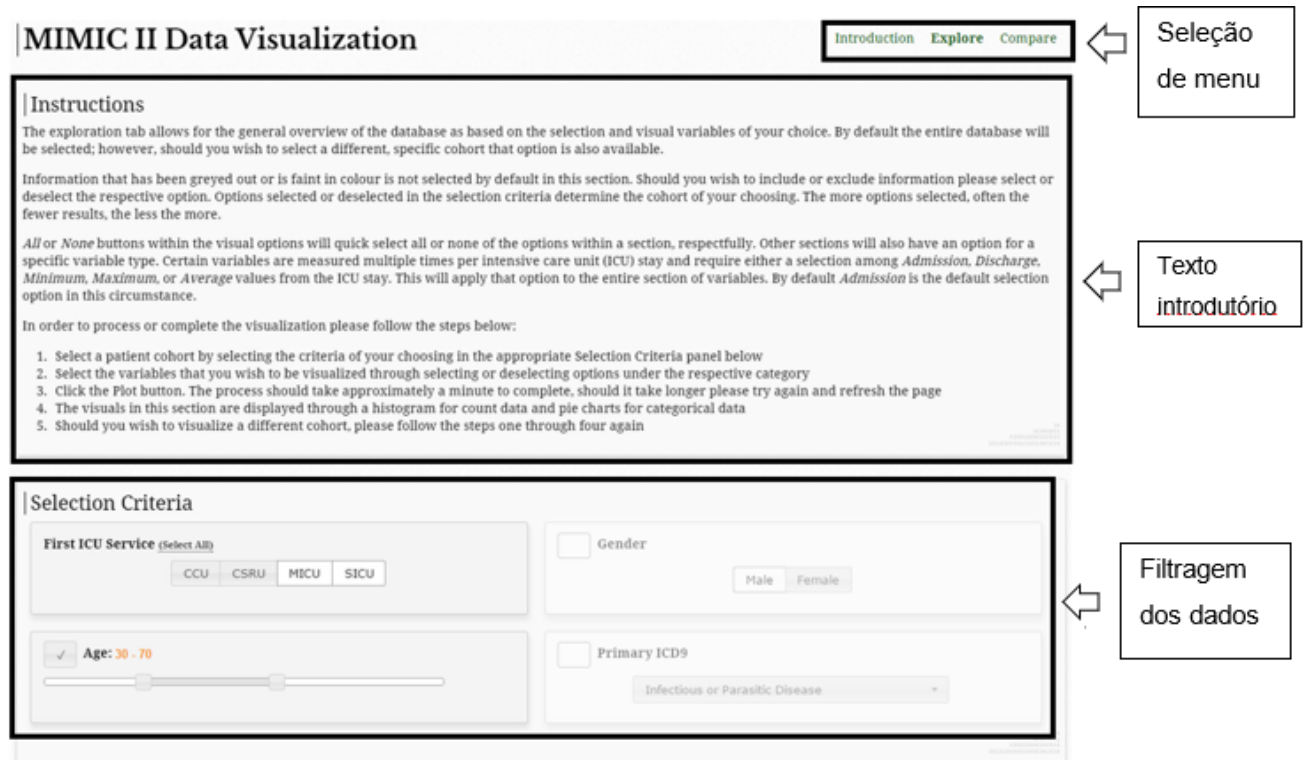


Figura 17- MIMIC-II Data Visualization Tool, Introdução retirado de (Lee et al,2015)



Figura 18-MIMIC-II Data Visualization Tool, Output do menu explore, retirado de (Lee et al,2015)

A Figura 18 mostra como são apresentados os dados dos pacientes resultantes dos filtros aplicados na Figura 17.

O resultado de “Compare” pode ser observado na Figura 19, que compara todos os pacientes em MICU (*Medical Intensive Care Unit*) com todos os de SICU (*Surgical Intensive Care Unit*), respetivamente, tendo sido selecionadas as seguintes variáveis: 3 tipos de mortalidade (28 dias, no hospital e na ICU), código ICD-9 e score SOFA. Tal como a Figura 19, demonstra, as duas últimas variáveis são categóricas e contínuas e assim são representadas através de gráficos (gráfico de barras e gráfico circular, respetivamente).

Este artigo aborda ainda limitações da ferramenta. A ferramenta apenas tem capacidade de oferecer dados para uma pesquisa preliminar da MIMIC-II, enquanto na base de vários estudos assentes sobre a MIMIC-II são requeridos outro tipo de estudos mais avançados que a ferramenta não é capaz de executar, tais como a geração de modelos para os dados, por exemplo, regressão linear. Outra limitação referida é o tempo de execução da mesma: certas ações podem demorar cerca de 60 segundos a serem executadas e consequentemente apresentadas ao utilizador, estando dependente da capacidade de processamento da máquina do mesmo. Por fim, outra limitação apontada pelos criadores da ferramenta encontra-se na própria base de dados e não tanto na ferramenta, mas que por consequência limita o tipo de observações possíveis na mesma. Todas as datas de admissão em ICU, tal como na MIMIC-III, são alteradas com um *offset random* de forma a obedecer às normas da HIPAA (Lee et al, 2015).

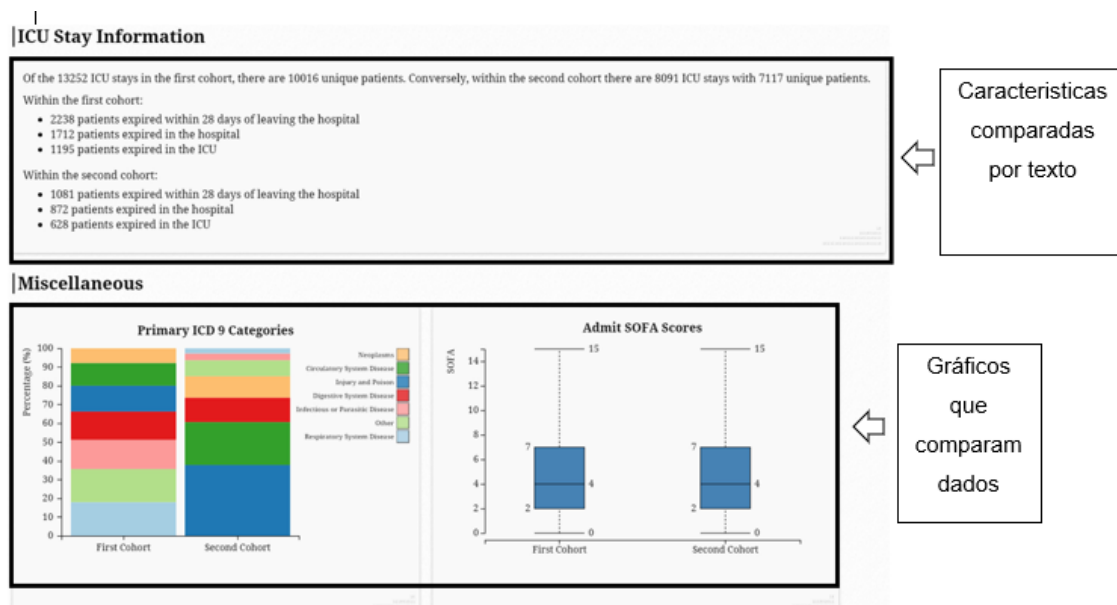


Figura 19- MIMIC-II Data Visualization Tool, menu Compare, retirado de (Lee et al,2015)

Como “*future work*” o artigo refere a criação de uma funcionalidade “*Analyze*” que permitiria a condução de variados testes estatísticos, tais como *t-test* ou *qui-quadrado*, bem como regressão multivariável utilizando covariáveis selecionadas pelo utilizador. É ainda referida a possibilidade de o utilizador desenvolver algoritmos de *machine learning*, utilizando Python ou R, bem como a criação de uma funcionalidade de exportação dos dados criados na ferramenta, o que permitiria ao utilizador fazer download da *raw data* que dá origem às figuras apresentadas pela ferramenta.

É de referir, no entanto, que existe um grande entrave à exploração e utilização desta ferramenta: não está disponível. No momento de escrita desta tese, nunca foi possível testar as funções reais da mesma, o seu correto funcionamento e apenas observar a sua interface baseando-nos em imagens da mesma inseridas no artigo.

No entanto, podemos retirar algumas conclusões para além das previamente retiradas: a sua interface não aparenta ser muito moderna, ou de fácil utilização; não existe indicação de como os restantes filtros são selecionados na página inicial; os gráficos gerados aparentam não ser iterativos mas sim estáticos e, aparentemente, não existe possibilidade de ver informação em pormenor sobre um determinado paciente, mas sim sob um grupo de pacientes e, finalmente, não é possível fazer uma pesquisa singular, não existe a possibilidade de pesquisar por uma determinada admissão, doença ou paciente utilizando o seu ID para visualizar mais informações.

4. Uma ferramenta de visualização dos dados do MIMIC-III

Neste capítulo abordamos a conceção de uma solução que vai em direção ao nosso objetivo: a criação de uma ferramenta com uma interface fácil de utilizar, que permita a visualização de dados da MIMIC-III relevantes, mas que não requeira ao utilizador saber programar.

Com exceção dos trabalhos descritos no capítulo anterior, esta base de dados não inclui em si mesma nenhum tipo de interface para quem a utiliza, é constituída unicamente por ficheiros em formato CSV, alguns com dezenas de *Gigabytes* de tamanho. Como vimos no capítulo anterior, não existem atualmente ferramentas públicas disponíveis que permitam uma exploração desta base de dados, e que contenham uma interface fácil de utilizar. As ferramentas de visualização atuais focam-se especialmente visualização de dados da MIMIC-III em formato Waveform. Das duas ferramentas que abordam a própria MIMIC-III: uma delas não está disponível, a sua interface é pouco iterativa e *user-friendly*, não contendo funções de *download* dos dados observados. A restante, exige que o utilizador tenha conhecimentos de programação para a utilizar e contém uma interface demasiado simplista, não permitindo também o *download* de informação. Pouco mais é que um mero terminal com que o utilizador interage.

Atualmente não existe uma interface *user-friendly* que permita tirar conclusões estatísticas imediatas dos dados presentes ou que, por exemplo, faça *merge* dos dados de algumas tabelas com o objetivo de conseguir ligar informações espalhadas em várias tabelas num só local e as apresente ao utilizador. Com o tipo de informação presente nesta base de dados, vários tipos de estudos podem ser feitos, mais ou menos específicos, de áreas ligadas à medicina ou não. As possibilidades são imensas devido à diversidade e riqueza de informação que esta base de dados contém.

Um dos principais objetivos na criação deste trabalho foi arquitetar uma interface que permitisse ao utilizador obter a informação pretendida sem este necessitar de qualquer tipo de conhecimento de programação. Todas as tarefas de manipulação e pesquisa de tabelas são ordenadas pelo utilizador, mas este não necessita de compreender exatamente o tipo de operações e etapas que estão a acontecer em *background*. Ademais, consideramos de

vital importância que a interface da ferramenta fosse o mais *user-friendly* possível, sem negligenciar funcionalidades. O utilizador da ferramenta tem de ser capaz de utilizar de a utilizar sem a necessidade de consultar a sua documentação.

4.1 R e motivação para a sua utilização

O principal motivo que levou R (R Core Team, 2020) a ser escolhido, foi a facilidade de criação de gráficos para analisar dados estatísticos.

Outro motivo foi a facilidade de acrescentar *packages* que adicionam funcionalidades extra ao próprio R, tais como *Tidyverse* (Wickham et al., 2022) (ferramenta de data science) e *Plotly* (Plotly Technologies Inc, 2015) (gráficos interativos), entre muitos outros.

R possui também bibliotecas para a criação de *dashboards*, em particular Shiny Dashboard, uma ferramenta que utiliza Shiny para criar *dashboards* mais apelativos. Esta foi mais uma das vantagens de utilizar R. Para poder começar a desenvolver a ferramenta neste dashboard foi apenas necessário instalar dois *packages* shiny e shinydashboard. Utilizando R no RStudio (RStudio Team, 2020). (IDE para R), pudemos ir incrementando o trabalho e testando a cada mudança feita.

A facilidade com que se consegue acrescentar funcionalidades através de *packages*, bem como a facilidade da sua instalação, é também uma das grandes vantagens da linguagem R e que mais tarde se veio a mostrar uma mais-valia. Com apenas um comando: `install.packages("packagename")` tinha ao dispor funcionalidades extra ao R.

Para além dos motivos referidos anteriormente, o facto de permitir utilizar comandos equivalentes a SQL em R, especialmente de manipulação de tabelas, facilitando por exemplo a junção de duas tabelas com *inner left join*. A versão do R utilizada na realização deste trabalho foi a versão 4.1.2. O IDE escolhido foi RStudio, sendo a última versão utilizada a 2022.02.3.

4.2 Dados

Esta secção irá abordar e detalhar as tabelas utilizadas pela ferramenta e a forma como foram obtidas tabelas derivadas das mesmas, já que estas não incluídas de raiz pela própria MIMIC-III.

4.2.1 Tabelas do MIMIC-III utilizadas

No desenvolvimento deste trabalho foram utilizadas as seguintes tabelas da MIMIC-III: ADMISSIONS, D_ICD_DIAGNOSES, DIAGNOSES, ICUSTAYS e PATIENTS.

Descreveremos com mais pormenor o conteúdo, organização e peculiaridades das tabelas utilizadas nos capítulos seguintes.

4.2.1.1 ADMISSIONS

Esta tabela tem como função definir a admissão hospitalar de cada paciente (entre 1 de junho de 2001 e 10 de outubro de 2012) e por consequência HADM_ID, que se trata de um número único que é atribuído a um paciente aquando da sua admissão. Todos os dados são obtidos de admissões, altas e transferências constantes da base de dados do hospital. Doadores de órgãos de pacientes que faleceram no hospital aparecem nesta tabela, contêm tempos de estadia negativos e o seu DEATHTIME é cópia do DEATHTIME do paciente anterior.

Nesta tabela encontram-se as seguintes colunas:

- SUBJECT_ID e HADM_ID: A primeira representa um único paciente e é uma chave secundária da tabela (Miles, 2017) e a segunda uma única admissão no hospital tratando-se uma chave primária (Miles, 2017). Esta tabela pode ter linhas SUBJECT_ID duplicadas, pois um paciente pode ter mais que uma admissão hospitalar;
- ADMITTIME, DISCHIME e DEATHTIME: Os dois primeiros correspondem a data e hora em que o paciente foi admitido ao hospital e recebeu alta do mesmo, respetivamente. DEATHTIME está presente apenas se o paciente morreu no hospital;
- ADMISSION_TYPE: descreve o tipo de admissão. Existem 4 tipos de admissão, nomeadamente: "ELECTIVE", "URGENT", "NEWBORN" ou "EMERGENCY". EMERGENCY e URGENTE indicam ambas idas não planeadas ao hospital,

ELECTIVE indica uma admissão planeada e por sua vez NEWBORN indica o HADM_ID do recém-nascido.

- ADMISSION_LOCATION: fornece informação sobre a localização do paciente antes de chegar ao hospital. Existem 9 possíveis:
 - EMERGENCY ROOM ADMIT;
 - TRANSFER FROM HOSP/EXTRAM
 - TRANSFER FROM OTHER HEALT
 - CLINIC REFERRAL/PREMATURE:
 - ** INFO NOT AVAILABLE **;
 - TRANSFER FROM SKILLED NUR:
 - TRSF WITHIN THIS FACILITY
 - HMO REFERRAL/SICK:
 - PHYS REFERRAL/NORMAL DELI
- INSURANCE, LANGUAGE, RELIGION, MARITAL_STATUS, ETHNICITY: Estas colunas abordam a demografia dos pacientes e têm origem em dados de admissão, alta e transferência de base de dados do hospital. Alguns valores como estado matrimonial e religião podem mudar entre admissões para o mesmo paciente;
- EDREGTIME, EDOUTTIME: Data em que paciente é registado e efetivamente recebe alta do hospital;
- DIAGNOSIS: Esta coluna contém um diagnóstico preliminar, em formato de texto, realizado quando o paciente dá entrada no hospital. Este diagnóstico preliminar é realizado pelo pessoal do corpo clínico e não utiliza uma ontologia sistemática, no entanto pode ser muito específico ou muito vago. Não substitui DIAGNOSES_ICD.
- HOSPITAL_EXPIRE_FLAG: é igual a 0, exceto se o paciente tiver morrido nessa hospitalização, onde é igual a 1.

4.2.1.2 D_ICD_DIAGNOSES

Esta tabela tem como função definir cada um dos diagnósticos ICD-9. O código ICD é convertido tanto numa versão curta como numa versão longa.

Contêm as seguintes colunas:

- ICD-9 CODE: código ICD-9, funcionando como chave primária da tabela (Miles, 2017);
- SHORT_TITLE, LONG_TITLE: descrição curta e longa do código.

É importante lembrar que todos os códigos ICD-9 presentes na MIMIC-III não contêm ponto decimal.

4.2.1.3 DIAGNOSES

DIAGNOSES contém diagnósticos dos pacientes, em formato ICD-9.

Contêm os seguintes atributos:

- ROW_ID: chave primária da tabela (Miles, 2017);
- SUBJECT_ID, HADM_ID: ambas chaves externas (Miles, 2017);
- SEQ_NUM: este número indica a prioridade do diagnóstico para o paciente. Quanto menor o número maior é a prioridade;
- ICD9_CODE: chave externa (Miles, 2017).

4.2.1.4 ICUSTAYS

ICUSTAYS é derivada da tabela TRANSFERS. Contém as linhas onde existe um ICUSTAY_ID e descarta as que não o contém. Esta tabela tem como função definir uma única admissão em ICU, ICUSTAY_ID. Agrupa todas admissões de ICU em 24horas. ICUSTAY_ID é um identificador não baseado em qualquer outro, pois as bases de dados do hospital e da ICU não estão inerentemente ligadas.

Contém as seguintes colunas:

- SUBJECT_ID, e HADM_ID: ambas chaves externas da tabela (Miles, 2017);
- ICUSTAY_ID: chave primária da tabela (Miles, 2017);

- DBSOURCE: indica a base de dados de ICU de onde a informação foi retirada. Como referido anteriormente, pacientes entre 2001 e 2008 tiveram os seus dados oriundos do sistema de informação CareVue, são identificados com “carevue”. Entre 2008-2012 foi retirado do sistema MetaVision;
- FIRST_CAREUNIT, LAST_CAREUNIT: contêm, respetivamente, o primeiro e último tipo de ICU onde o paciente foi tratado. É possível que diferentes transferências entre diferente tipo de ICU tenham o mesmo ICUSTAY_ID;
- FIRST_WARDID, LAST_WARDID: indica, respetivamente, a primeira e última UCI onde o paciente permaneceu. Na base de dados do próprio hospital, as diferentes localizações do mesmo são denominadas de enfermarias (“ward”), embora não o sejam na prática. Isto explica o porquê de cada ICU distinta ter um WARID;
- INTIME, OUTTIME: data e hora em que o paciente foi transferido para a ICU e desde a ICU, respetivamente;
- LOS: um parâmetro importante que indica o período temporal em ICU, em dias fracionais. Indica assim o tempo em ICU de determinado ICUSTAY_ID.

4.2.1.5 PATIENTS

Esta tabela tem como objetivo definir SUBJECT_ID, ou seja, um único paciente. O número total de linhas desta tabela indica aí o número de diferentes pacientes admitidos no hospital (46 520). É a responsável por fornecer detalhes sobre cada paciente.

Esta tabela contém as seguintes colunas:

- SUBJECT_ID: é a chave primária desta tabela (Miles, 2017), cada linha contém apenas um único SUBJECT_ID;
- GENDER: género do paciente;
- DOB: data de nascimento do paciente. Este parâmetro, de forma a respeitar as normas de privacidade da HIPAA, teve de ser modificado. Pacientes com idade superior a 89 anos têm a sua data de nascimento mudada de forma a esconder a sua real idade. A data de nascimento destes pacientes foi mudada para exatamente 300 anos antes da data da sua primeira admissão. Este procedimento torna assim mais difícil a identificação destes doentes, visto serem poucos e mais facilmente identificáveis (Johnson et al., 2016);

- DOD, DOD_HOSP, DOD_SSN: DOD representa a data de falecimento de um dado paciente e resulta da união de DOD_HOSP e DOD_SSN, sendo a primeira priorizada. DOD_HOSP é a data de falecimento que está guardada nas bases de dados do hospital. DOD_SSN é a data de morte que consta na base de dados da segurança social;
- EXPIRE_FLAG: valor binário que indica se o paciente morreu ou não. O valor é igual a 1 se tiver morrido, 0 se o seu contrário. São ainda tidas em consideração mortes dentro do hospital, DOD_HOSP, e mortes identificadas fazendo o cruzamento do paciente com o *social security master death index*, DOD_SSD (constante no ficheiro DMF).

4.2.2 Tabelas derivadas

Como referido no capítulo 4.2.1, nem todas as tabelas do MIMIC-III foram utilizadas. No âmbito deste trabalho foram utilizadas as seguintes: ADMISSIONS, D_ICD_DIAGNOSES, DIAGNOSES, ICUSTAYS e PATIENTS. Todas as tabelas estão no formato CSV.

Foi criado um diretório com estes ficheiros e todos eles foram importados para o IDE RStudio, onde aí foram importados utilizando a função `read.csv()`. Esta função lê um ficheiro em formato CSV presente no disco rígido do computador e converte o para o formato *dataframe*. Um *data frame* é uma estrutura de dados fundamental em R, fazendo parte da sua base. Trata-se de uma tabela ou *array* bidimensional, onde cada coluna contém valores de uma variável e cada linha contém conjuntos de valores para cada coluna. Os valores guardados num *dataframe* podem ser do tipo numérico, factor ou character. Embora cada coluna possa ter qualquer tipo de dados, cada coluna deve ter apenas um tipo de dados.

Durante a implementação desta ferramenta os *dataframes* originais da MIMIC-III nunca são alterados. Quando necessário obter mais informação e cruzar dados de outros *dataframes*, são criados com essa mesma informação:

- ADMISSIONS_UNIQUE: um *dataframe* com os mesmos atributos que ADMISSIONS, mas que contém apenas uma única admissão por paciente. No caso de pacientes com múltiplas admissões apenas a última é utilizada, de forma a não causar problemas com os valores HOSPITAL_EXPIRE_FLAG, ou seja, se fosse mantida outra admissão que não a última, poderíamos estar a apresentar valores errados de mortalidade hospitalar.

```
ADMISSIONS_UNIQUE<- ADMISSIONS[!rev(duplicated(rev(ADMISSIONS$SUBJECT_ID))),]
```

- *df*: *dataframe* resultante de um `left_join` de `ADMISSIONS_UNIQUE` com o *dataframe* `PATIENTS`. Este *dataframe* permite assim obter um *dataframe* onde cada linha pertence a cada um dos 46520 pacientes diferentes, mas que contém mais informação que a disponibilizada na tabela `PATIENTS`. Contém os seguintes atributos:

- `SUBJECT_ID`;
- `HADM_ID`;
- `ADMITTIME`;
- `DISCHTIME`;
- `DEATHTIME`;
- `ADMISSION_TYPE`;
- `ADMISSION_LOCATION`;
- `DISCHARGE_LOCATION`;
- `INSURANCE`;
- `LANGUAGE`;
- `RELIGION`;
- `MARITAL_STATUS`;
- `ETHNICITY`;
- `EDREGTIME`;
- `EDOUTTIME`;
- `DIAGNOSIS`;
- `HOSPITAL_EXPIRE_FLAG`;
- `HAS_CHARTED_EVENTS`;
- `GENDER`;
- `DOB`;
- `DOD`;
- `DOD_HOSP`;
- `DOD_SSN`;
- `EXPIRE_FLAG`;

- AGE.

Uma das informações mais importantes para conhecer o paciente não estava presente em nenhuma das tabelas, quer `ADMISSION_UNIQUE` ou `PATIENTS`, a idade do paciente. A data de nascimento foi alterada, como explicado anteriormente, por motivos de privacidade, assim, a idade dos pacientes não é incluída na MIMIC-III, mas existe um procedimento para a calcular. O procedimento é o seguinte: à data da sua primeira admissão em ICU subtrair a data de nascimento (Johnson et al., 2022). Para isto a utilização da biblioteca *eeptools* foi de grande utilidade, mais concretamente a sua função *age_calc()* (Knowles, 2020). Esta função calcula idades, tanto em anos, meses ou dias, recebendo como *input* uma data de nascimento e outra data arbitrária, neste caso sendo a data de primeira admissão em ICU.

- `diagnoses_with_LOS`: este *dataframe* junta informações de duas tabelas distintas, `DIAGNOSES` e `ICUSTAYS`.
- `ICUcomplete`: este *dataframe* é resultante de um *left join* através do `SUBJECT_ID` de `ICUSTAYS` com *dataframe* `df`, mas apenas com os atributos "age", "SUBJECT_ID", "GENDER", "ETHNICITY", "EXPIRE_FLAG" e "HOSPITAL_EXPIRE_FLAG" de `df`.

4.3 Desenho – R Shiny dashboard Framework

A decisão de utilizar Shiny dashboard (ShinyInc, 2020) influenciou fortemente o desenho da ferramenta, pois tivemos de seguir a sua estrutura própria que se divide em dois componentes: UI (user interface) e Server. De uma forma geral a primeira componente é responsável pelo aspeto estético, a segunda pelo seu funcionamento. Ambas são depois chamadas como argumentos à função `shinyApp()`, resultando na criação de um *dashboard*.

Na UI são construídos o layout e aparência da ferramenta. Descreve que dados serão exibidos, que controlos estarão disponíveis e outras informações relevantes, como títulos de páginas, tipos e títulos de gráficos, caixas de texto com descrições em texto formatado

ou interface de seleção de *inputs*, tais como *sliders* ou caixas de inserção de texto para pesquisa.

A UI de uma shinyApp é composta por três componentes: *header*, *sidebar* e *body*.

Header contém o título da ferramenta/app no canto superior esquerdo e menus do tipo *dropdown* no canto superior direito. Estes menus podem conter três tipos de informação: mensagens, notificações ou tarefas. Foi escolhido utilizar apenas um menu deste tipo para oferecer ajuda em como utilizar a interface.

Sidebar consiste numa barra lateral, que pode ser escondida ou expandida clicando no símbolo de hambúrguer menu. É nesta *sidebar* que as principais categorias da ferramenta estão presentes e são divididas em 4 grandes grupos(*tabItem*): Home, Patients, Admissions e Diagnoses. Cada uma delas contém depois submenus(*menuSubItem*) que dividem cada categoria em subcategorias, com mais funcionalidades. Cada vez que o utilizador seleciona um menu lateral o conteúdo do *main body* muda para apresentar informações do *tabItem* escolhido. ("Shiny - Create a page with fluid layout — fluidPage", 2020)

Por último, o *body* é o principal bloco de construção da UI de um Shiny dashboard. Aqui estão todos os gráficos, tabelas e blocos de texto da ferramenta, podendo conter qualquer tipo de conteúdo disponível no shiny. Na implementação desta ferramenta, o *body* contém os conteúdos de cada *tabItem* e *menuSubItem* da *Sidebar*. ("Shiny - Create a page with fluid layout — fluidPage", 2020)

É no *body* que cada *tabItem* da barra lateral é desenvolvido, ou seja, que cada página da ferramenta é gerada. Podem-se usar diversas formas de apresentar os dados e organizar a informação, sendo duas utilizadas neste dashboard:

- *sidebarLayout*: cria uma página com um painel lateral (*sidebarPanel()*) e uma área principal (*mainPanel()*), contendo a primeira espaço para *inputs* do utilizador e sendo de uma cor diferente. A zona principal, do *mainPanel*, ocupa mais de metade da página inteira, cerca de dois terços.
- *fluidRow*: permite dispor elementos do R de forma fluída e mais organizada, tais como *boxes*, permitindo que estejam todos exatamente à mesma altura. Permite ainda que mesmo diminuindo ou aumentando o tamanho da janela do *dashboard* os

seus elementos se organizem em tempo real para caber no tamanho da janela; ("Shiny - Create a page with fluid layout — fluidPage", 2020)

Como referido anteriormente, um Shiny Dashboard é também constituído por uma componente chamada *server*. Este componente trata-se de uma função que contém as instruções de como construir a aplicação, contendo o código que controla a forma como o *dashboard* atua. É responsável por interagir com os *inputs* recebidos e os *outputs* da aplicação, enviando-os para a UI. Esta componente utiliza programação reativa para automaticamente atualizar os *outputs* quando os *inputs* mudam ao longo da execução do dashboard.

Por exemplo, é a UI que contém a chamada de função de criação de um gráfico, mas é no *server* que a mesma é executada e os cálculos são efetuados. A execução do gráfico é assim realizada do lado do servidor e depois o seu *output* é apresentado na UI. Outro exemplo: imaginemos que é realizada a seleção de um valor de input na UI através de um dropdown menu. Esse valor de *input* é guardado e recebido no *Server*, que o processa, realiza uma operação sobre o mesmo e fornece um *output* que é depois impresso na UI.

Por último, a função *shinyApp()* recebe como argumento tanto a UI como o *Server*, resultando na criação de uma aplicação Shiny, neste caso um *dashboard* e a sua execução.

4.3.1 Requisitos de sistema

Durante o desenvolvimento inicial foram utilizados dois computadores com dois sistemas operativos distintos. A primeira máquina correndo Windows 10 64-bit, equipada com um CPU i7, 8GB de RAM e SSD de 500GB. A segunda máquina corria Linux Mint 20 64-bit, i7, 4GB de RAM e HDD de 500GB. Depois de combater pequenos *bugs* que se verificavam em Linux mas não em Windows (por exemplo no *package* Plottly, certas legendas apareciam desformadas), foi tomada a decisão de migrar a segunda máquina para Windows 10 64-bit também e de substituir o seu HDD por um SSD de 250GB.

Com referido na secção 4.1, o programa foi desenvolvido e é executado no próprio IDE RStudio. Todos os *packages* extra utilizados e necessários para a execução do programa

podem ser instalados automaticamente, bastando aceitar a sua instalação numa mensagem prompt enviada pelo IDE. É necessário ter o R versão 4.1.2 instalado.

Não existem requisitos mínimos rigorosos, no entanto o fator mais crítico para a execução deste programa é a RAM disponível. Foi desenvolvido e testado numa máquina com 4GB de RAM, portanto uma máquina que cumpra este requisito será capaz de o correr. É, por fim, fortemente recomendada a utilização de um SSD, são frequentes as operações de leitura e escrita pelo que tornará a execução do programa mais ágil.

4.3.2 Navegação e menu lateral

Neste capítulo será abordada a forma como é feita a navegação entre os diferentes menus e que escolhas no design do R Shiny dashboard foram tomadas para facilitar essa mesma navegação. Nos capítulos seguintes cada menu e página da ferramenta serão explicados em detalhe. Podemos dividir esta ferramenta em 5 grandes grupos: Dashboard, Patients, Admissions, Diagnoses e ICU por esta ordem.

Cada um destes grupos é acessível através de um menu lateral, contendo cada um deles, para além do seu nome, um símbolo para tornar a identificação mais rápida. Este menu é acedido através de um símbolo “hambúrguer menu” localizado no topo esquerdo e pode ser escondido clicando nesse mesmo botão, recolhendo-se com um movimento de deslize para a esquerda, ficando oculto. Isto permite uma maior área de visualização visto o conteúdo adaptar-se ao espaço de ecrã maior e, por exemplo, apresentar os gráficos com maior dimensão visto ter mais espaço horizontal. Estando o menu oculto o símbolo do “hambúrguer menu” fica visível na barra superior azul do *dashboard*, no canto superior esquerdo, sendo apenas necessário clicar no mesmo para expandir novamente o menu lateral com um movimento de deslize para a direita. O menu pode ser mostrado ou ocultado múltiplas vezes, sem que a página atual seja alterada para outra.

Quando o programa é executado a primeira vez, a página inicial, chamada de Dashboard é apresentada. Aqui podem ser encontradas estatísticas rápidas e gerais sobre a MIMIC-III, bem como um texto introdutório para o utilizador ficar a par do funcionamento da ferramenta. Se este desejar, poderá consultar outros menus da ferramenta clicando na opção desejada nesse mesmo menu lateral, sendo apresentada a azul mais escuro a opção do menu selecionada, para que o utilizador saiba sempre em que menu está.

No caso de menus com submenus, a existência de submenus é anunciada por uma seta na horizontal, e clicando sobre os mesmos é expandida na vertical a lista de submenus disponíveis. Quando um submenu é selecionado, o texto do mesmo passa de cinzento-claro para branco, indicando que está aberto e a ser exibido nesse momento, o que permite ao utilizador saber sempre, em qualquer momento da execução do *dashboard*, em que página está e a sua localização geral na ferramenta. Para fechar o submenu na barra lateral basta selecionar novamente o menu principal, sendo possível abrir submenus de outros menus sem mudar a página atual. Um fator importante na fluidez de utilização deste *dashboard* é a possibilidade de abrir múltiplas páginas e o seu progresso ser guardado se o utilizador a fechar para abrir outra. Por exemplo: imaginemos que o utilizador deseja conhecer o histórico de hospitalizações de um doente, no menu Diagnoses, se em seguida desejar consultar informação demográfica e pessoal sobre o mesmo, poderá abrir o menu Patients e consultá-la, sem que o progresso feito no separador Diagnoses seja afetado. É assim possível alternar entre menus e submenus sem perder informação e progresso em cada um deles.

Assim, o utilizador é capaz de navegar através da ferramenta e aceder qualquer página/menu do *dashboard* estando em que menu estiver utilizando o menu lateral da mesma.

Este menu lateral faz parte do R Shiny Dashboard, sendo desenvolvido na UI do dashboard, utilizando a função `dashboardSidebar`, que recebe como *input* outra função, `sidebarMenu`, responsável pela esquematização do menu lateral. É na função `sidebarMenu` que cada menu é criado, através da utilização da função `menuItem`, onde é descrito e criado cada um dos menus. É nesta função que definimos o nome do menu, o nome da variável `tabName` que o identifica no R Shiny dashboard e outros parâmetros, tais como o ícone que apresenta. Se quisermos criar um submenu dentro de um menu chamamos a função `menuItem` no interior da `menuItem` correspondente, recebendo os mesmos parâmetros. É possível definir se o submenu é automaticamente expandido ou não, atribuindo o valor booleano `TRUE` à variável `StartExpanded` na função `menuItem`.

A função `dashboardSidebar` é depois enviada para a componente UI, como explicado na Secção 4.3.

```

sidebar <- dashboardSidebar(
  sidebarMenu(
    menuItem("Menu name", tabName, icon),
    menuItem("Menu name", tabName, icon, startExpanded = boolean
      menuSubItem("SubMenu name",
        tabName)
    )
  )
ui <- dashboardPage(header, sidebar, body)

```

4.3.3 Ecrãs do dashboard e porquê

Neste capítulo serão abordadas em detalhe todas as páginas e menus da ferramenta, Figura 20. Os detalhes da sua implementação, razão de terem sido criados bem como a forma de as utilizar. Para uma melhor organização cada uma das páginas/menus será um capítulo distinto.

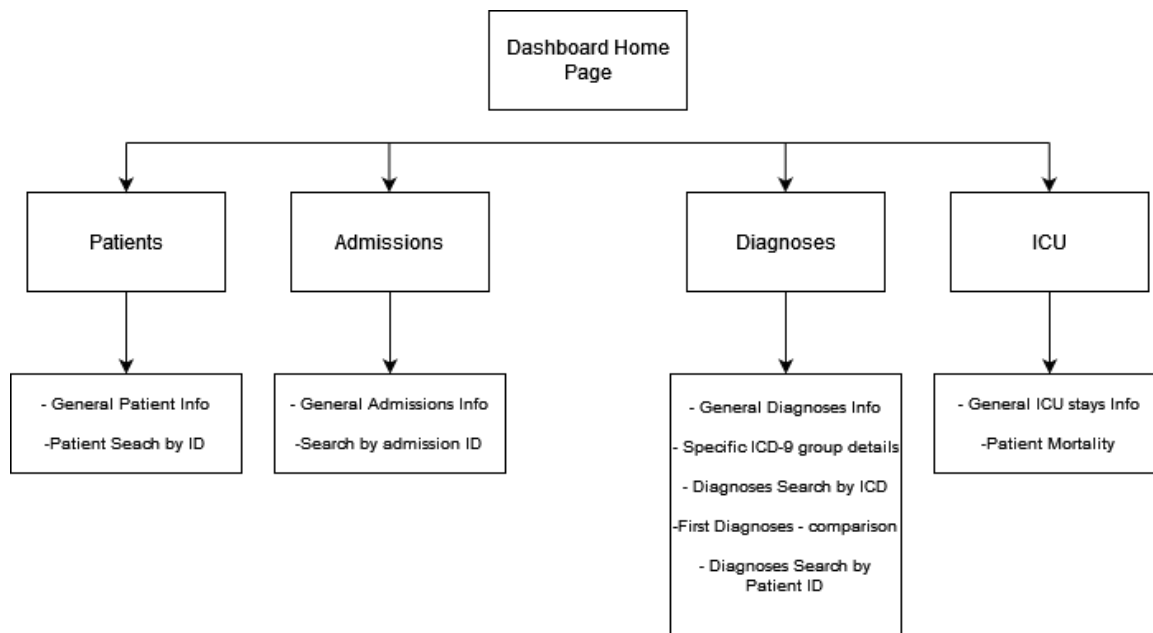


Figura 20- Diagrama do Dashboard

4.3.4.1 Dashboard

Dashboard é a página inicial da ferramenta, funcionando como *homepage* do *dashboard*. Para esta página foi utilizada a estrutura *fluidRow*, sendo as informações contidas no interior de boxes. Dashboard está dividido em 8 componentes (Figura 21), sendo cada uma delas contida numa *box*: estatísticas e dados gerais sobre a base de dados; texto introdutório à ferramenta; histograma de LOS de todo o hospital, idades e mortalidade por idade e ainda gráficos circulares da distribuição de etnia, género e mortalidade por etnia.

Na *box* contendo as estatísticas são apresentas dados e estatísticas gerais em formato de texto sobre a base de dados MIMIC-III, tais como:

- Número total de doentes;
- Número de admissões hospitalares;
- Número de admissões em ICU;
- Percentagem de pacientes masculinos;
- Idade média dos pacientes em anos;
- LOS médio de toda a MIMIC-III em dias;
- Mortalidade hospitalar em percentagem;

Para apresentar os histogramas e gráficos circulares, foi utilizado uma biblioteca disponível para R, *plotly*. *Plotly* permite a criação de gráficos interativos (e não apenas estáticos como seriam os criados originalmente pela função *hist()* do R), sendo possível fazer *zoom* nos gráficos, seleccionar apenas uma pequena parte dos mesmos para visualização ou, por exemplo, existir interação com o gráfico e informações extra durante *mouse-over*.

Para além das vantagens que oferece na visualização é mais personalizável e permite facilmente a utilização de várias cores se pretendido. Todos os gráficos utilizados nesta ferramenta são produzidos pelo *Plotly*. Nesta página inicial são apresentados 6 gráficos, sendo 3 deles histogramas e os restantes gráficos circulares. Cada um destes gráficos é oriundo do dataframe *df*, exceto o histograma relativo à LOS hospitalar, que é oriundo de *ICUSTAYS*. Destes gráficos o menos triviais são os gráficos relativo à mortalidade, quer hospitalar, quer hospitalar por etnia. Em ambos foi necessário filtrar o dataframe *df* por todos os casos onde a *HOSPITAL_EXPIRE_FLAG* é igual a 1. É de referir ainda que no histograma de idades os pacientes com mais de 89 anos foram retirados, de forma que o

gráfico não apresente idades iguais ou superiores a 300 anos. Estes pacientes não foram retirados da base de dados, apenas omitidos neste gráfico em particular.

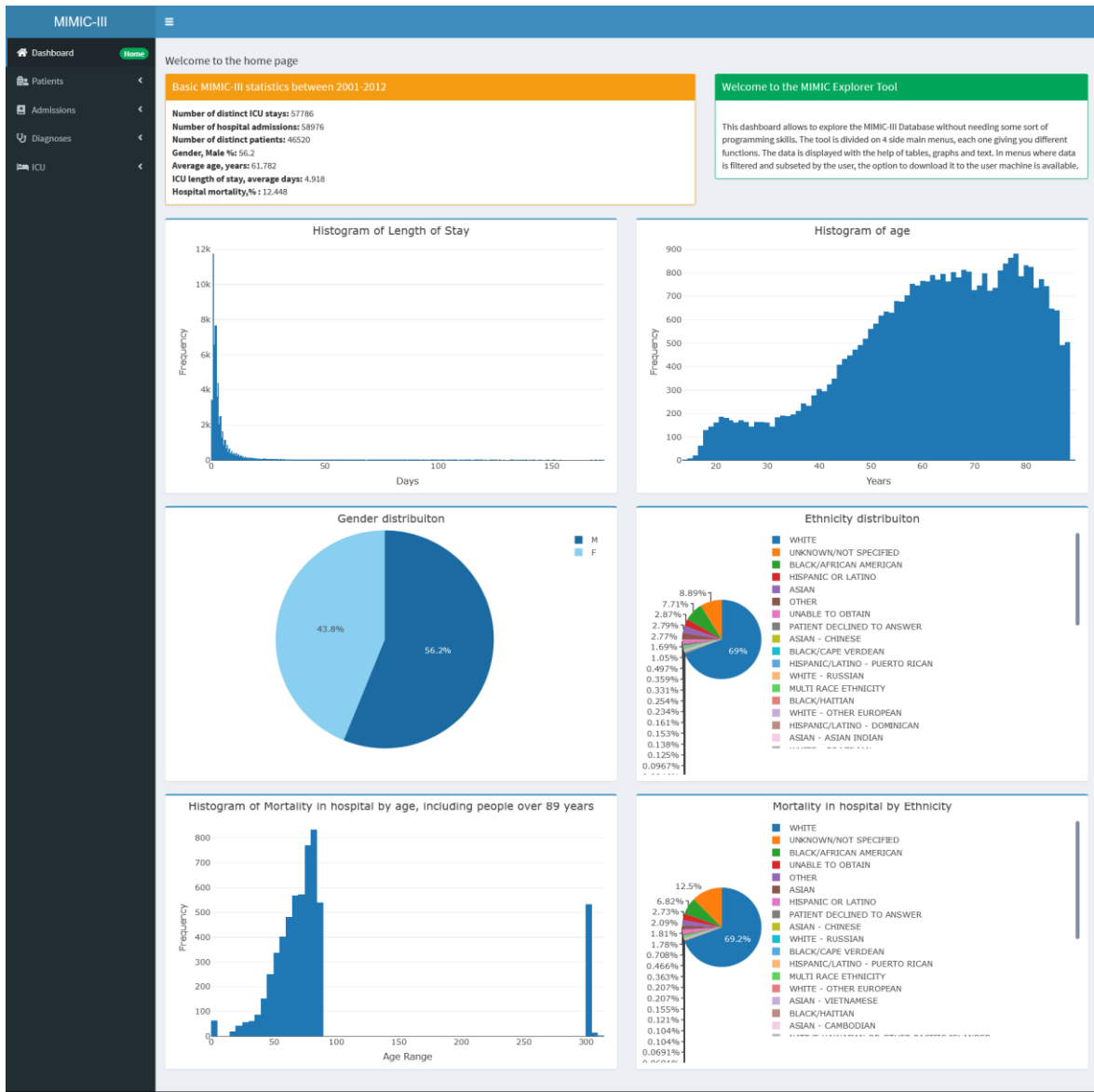


Figura 21- Dashboard, página inicial

4.3.4.2 PATIENTS

Esta página do *dashboard* foca-se na apresentação de informação relativa a todos os pacientes da base dados, sendo a maioria desses mesmo dados baseados na tabela

PATIENTS da MIMIC-III, mas complementada com informação de outras tabelas quando necessário, de forma que um maior conhecimento do paciente seja oferecido. Aqui, o utilizador pode procurar dados de forma precisa, quer procure dados de um paciente em específico ou procure de um grupo dos mesmos.

Este menu está dividido em dois sub-menus: General Patient Info e Patient Search by ID. Cada um deles com uma função distinta e, consequentemente, implementações diferentes. O primeiro foca-se na apresentação de dados sobre um paciente específico, o segundo na apresentação de sumários de informação de um grupo de pacientes.

4.3.4.2.1 PATIENT SEARCH BY ID

Patient Search by ID, tal como o nome indica, permite ao utilizador inserir ou seleccionar o ID (SUBJECT_ID) de um paciente numa caixa de seleção/introdução de *input*. Através do ID do paciente podem ser consultadas informações detalhadas do paciente e ainda todas as admissões hospitalares do mesmo. Durante a inserção do SUBJECT_ID, não só é possível seleccionar tanto o ID do paciente da lista de opções disponíveis, como também introduzir o mesmo, sendo que à medida que cada algarismo do ID é introduzido, são apresentadas sugestões de ID's existentes na tabela PATIENTS para tornar este procedimento mais rápido.

Assim que um ID válido é introduzido é apresentada a informação detalhada desse mesmo paciente à direta numa tabela, oriunda do dataframe df:

- número único de identificação,
- género,
- data de nascimento,
- data de falecimento, data de falecimento registada pelo hospital e pelo SSN;
- expire flag (indicador de que o paciente faleceu no hospital se for igual a 1.
- idade,
- data de primeiro internamento,
- data e hora de alta hospitalar,
- hora do óbito,

- tipo e localização da primeira admissão,
- local para onde o paciente foi transferido após receber alta,
- seguro do paciente,
- linguagem (se presente),
- religião,
- estado matrimonial,
- etnia,
- data de alta de ICU,
- diagnóstico rápido realizado na admissão ao hospital,

Numa outra tabela, são apresentadas todas as admissões hospitalares do paciente. A informação aí apresentada é oriunda de uma filtragem do dataframe ADMISSIONS. Assim que o utilizador insere um ID (na caixa de *input*), este é guardado e utilizado para filtrar todas as linhas de ADMISSIONS cujo SUBJECT_ID é igual a esse valor.

Este submenu utiliza uma estrutura de sidebarLayout, sendo o sidebarPanel constituído por uma caixa de seleção de input e o mainPanel pelas tabelas onde são apresentadas as informações do paciente (Figura 22- Estrutura do submenu patient search by ID).

Search by patient ID

Choose or type patient ID: 2

Patient info

Show 10 entries

SUBJECT_ID	age	GENDER	ETHNICITY	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE	LANGUAGE	RELIGION
1	2	0	M	ASIAN	163353	2138-07-17 19:04:00	2138-07-21 15:45:00	NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME	Private	NOT SPECIFIED	

Showing 1 to 1 of 1 entries

Previous 1 Next

All admissions of patient

Show 10 entries

SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE	LANGUAGE	RELIGION	MARITAL_STATUS	ETHNICITY
1	2	163353	2138-07-17 19:04:00	2138-07-21 15:45:00	NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME	Private	NOT SPECIFIED			ASIAN

Showing 1 to 1 of 1 entries

Previous 1 Next

Figura 22- Estrutura do submenu patient search by ID

No entanto existem algumas particularidades na implementação desta página que devem ser referidas e que serviram depois de referência para a criação dos restantes menus.

A primeira é o facto de a seleção de input, do ID do paciente, ser feita através da função `selectizeInput` que contém uma componente de implementação no servidor, ao contrário da função `selectInput`, que logicamente seria a preferida para este tipo operação, que apenas é executada no cliente. Numa primeira abordagem foi utilizado `selectInput` como caixa de introdução de *input*, mas foi imediatamente notório que o facto de a lista de opções de IDS de pacientes ser muito grande (mais de 46 000 IDS) traria problemas graves de performance. No momento em que o utilizador começa a inserir o ID, `selectInput` começa de imediato a procurar opções que completam o *input* do utilizador, e devido ao tamanho enorme da lista a percorrer a performance é afetada (Xie, 2017). Assim, foi necessário encontrar uma solução para este problema: `selectizeInput` é preferido para este tipo de operação onde a lista de opções de *input* é extensa. `selectInput` faz uso exclusivamente de JavaScript para processar a pesquisa durante a digitação, enquanto `selectizeInput` implementado no servidor usa R para processar a pesquisa e é também R que retornará os dados filtrados para serem selecionados.

Para utilizar `selectizeInput` foi necessário criar uma instância da mesma no cliente:

```
selectizeInput("patid", "Choose or type patient ID:", choices = NULL)
```

E atualizá-la no servidor:

```
updateSelectizeInput(session, "patid", choices = PATIENTS$SUBJECT_ID,  
server = T, selected = "2")
```

É de notar que a opção `choices = NULL` no cliente foi utilizada de modo a acelerar mais ainda a inicialização inicial da página, sendo necessário utilizar a opção `server = TRUE` no servidor de forma a passar a data das opções para o R.

Importante também referir que todos as opções de ID foram obtidas por: `PATIENT$SUBJECT_ID`, cujo output são exatamente 46 520 IDs diferentes.

A segunda particularidade é a utilização da biblioteca DT. Após a seleção do ID do paciente é consequentemente apresentada numa tabela com que o utilizador pode interagir, utilizando a função `renderDataTable`, pertencente ao pacote DT. Esta tabela foi escolhida

para apresentar a formação pois permite definir o número de linhas a apresentar de uma só vez, ordenar por atributo (por exemplo, SEQ_NUM) e pesquisar numa barra de pesquisa no topo direito por alguma informação contida na mesma.

De agora em diante, todas as tabelas existentes são originárias de `DT::renderDataTable`.

4.3.4.2.2 General Patient Info

General Patient Info, é o segundo submenu de PATIENTS. Ao contrário do primeiro, o foco aqui é apresentar informação o mais detalhada possível de um grupo de pacientes, observá-la e se pretendido, fazer *download* da mesma. Para isto o utilizador pode filtrar esses mesmos pacientes por vários atributos (atributos do dataframe `df`), sendo apresentada informação sumariada consoante os filtros aplicados.

Além, disso sob essa mesma informação sumariada, é apresentada uma tabela constituída por todos os pacientes afetados na seleção anterior e ainda 6 gráficos cujo objetivo é ajudar o utilizador a rapidamente entender a distribuição e características dos dados que selecionou.

Os dados da tabela podem ser descarregados para o computador do utilizador, mediante o clique no botão “Download the data”. Ao iniciar o download dos dados selecionados, é aberta uma janela do sistema a perguntar onde o utilizador deseja guardar o ficheiro. Este ficheiro tem o formato CSV. Os 6 gráficos referidos anteriormente consistem em 5 gráficos circulares e um histograma. O histograma contém as idades dos pacientes, incluindo os pacientes com mais de 89 anos. Os restantes gráficos circulares representam: distribuição étnica, tipo de admissão, tipo de seguro, local de admissão e local após alta. Este menu tal como o anterior utiliza uma estrutura de `sidebarLayout`, sendo a componente `sidebarPanel` constituído por uma caixa de inputs e o `mainPanel` pelo output dos filtros selecionados na `sidebarPanel` (Figura 23).

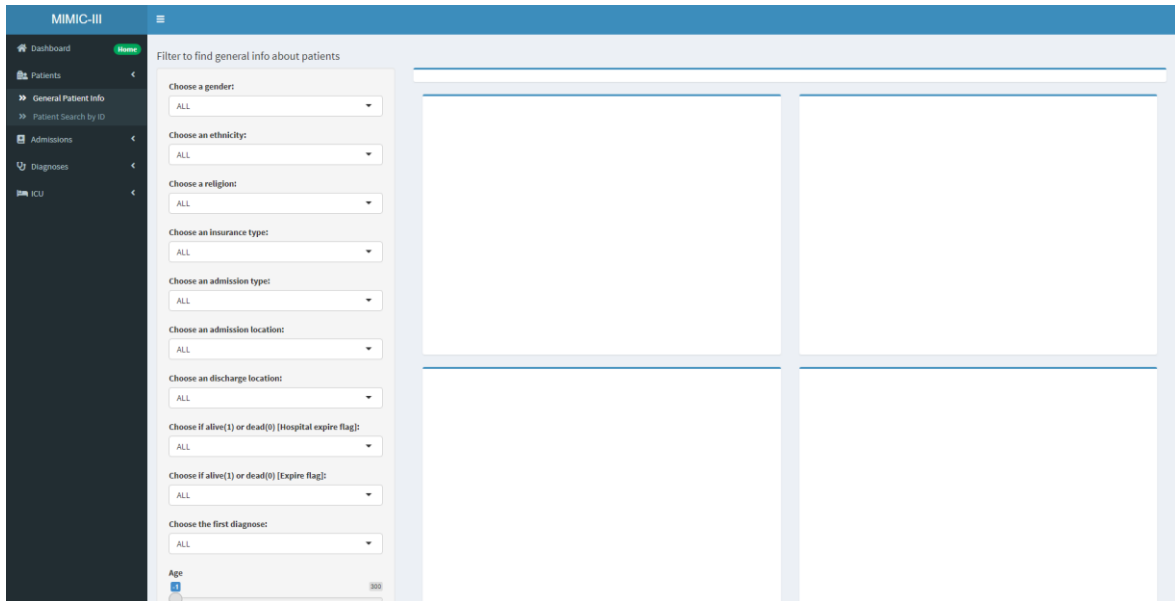


Figura 23- General Patient Info, antes de seleccionar opções no filtro as boxes estão vazias

Utilizando o *dataframe* do menu anterior, resultante da junção de dados múltiplas tabelas da MIMIC-III, o utilizador pode filtrar os dados dos pacientes pelos seguintes atributos, respetivamente:

- género;
- etnia;
- religião;
- tipo de seguro;
- tipo e local de admissão;
- local para onde o paciente foi deslocado após receber alta;
- verificação de falecimento no hospital;
- verificação de falecimento;
- idade.

À direita deste caixa de *input* é dado o *output* resultante dos inputs selecionados. Estes são impressos no interior de *boxes* do R Shiny. O sumário dos dados selecionados, através da função *summary()* permite ao utilizador ficar a conhecer melhor os pacientes que correspondem ao filtro anteriormente aplicado (Figura 24).

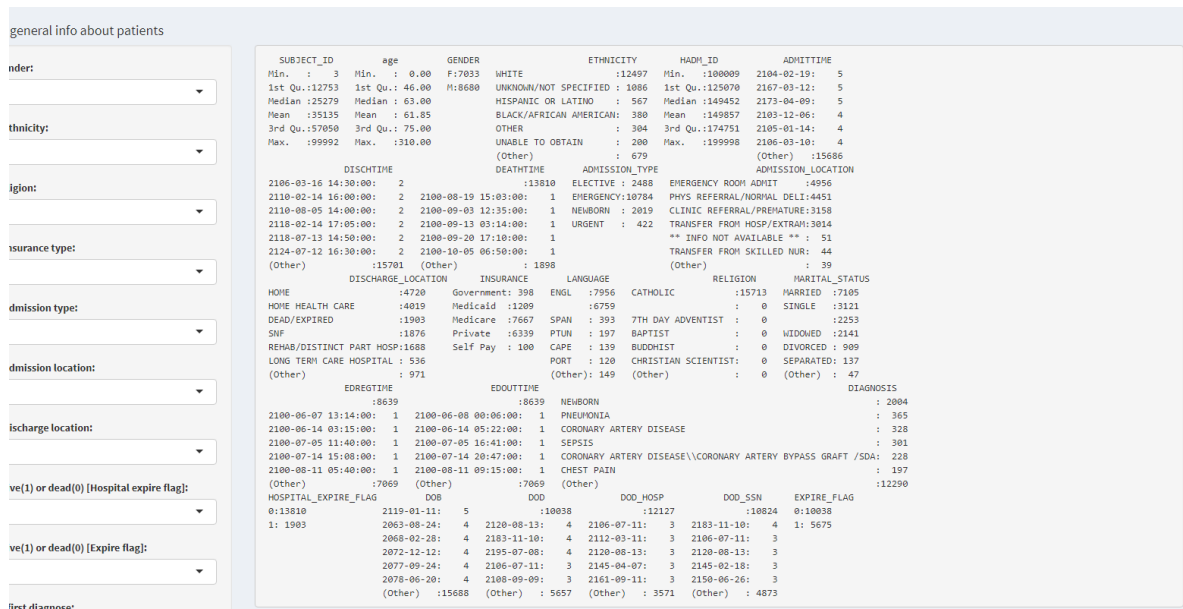


Figura 24- Output de summary() dos dados filtrados do submenu General Patients Info

Sob esta box, existe uma outra, contendo uma tabela (Figura 25) com o registo de todos os pacientes filtrados. Utilizando a biblioteca DT e a sua função dataTable (DT::dataTable), é apresentada uma tabela com todos os atributos presentes no dataframe df.

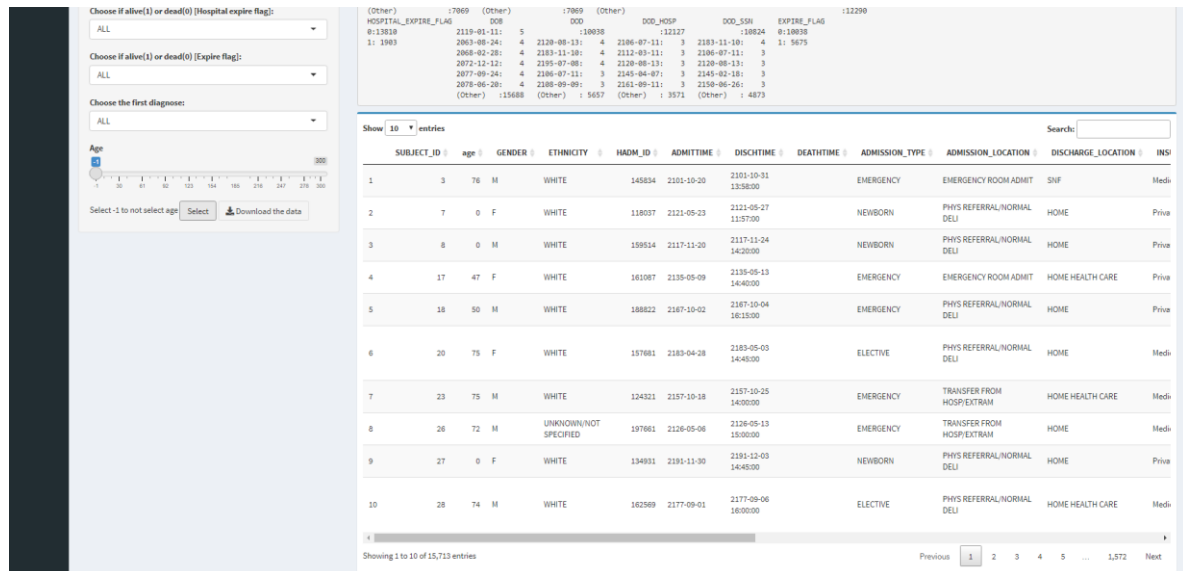


Figura 25- Output dos dados filtrados apresentado numa tabela do submenu General Patients Info

Sob esta encontram-se os gráficos que representam estes dados, (Figura 26).

Vamos agora mergulhar em alguns detalhes importantes de como esta página da ferramenta foi implementada.

Primeiramente foi construída um painel lateral, contendo uma série de caixas de input, uma para cada atributo citado acima. Neste caso, ao contrário do submenu Search Patient by ID, é utilizado como método de *input* o elemento SelectInput visto que as opções de escolha não são tantas como anteriormente. A exceção é a seleção da idade onde é utilizado o elemento SliderInput. Cada uma das opções para a caixa de menu foi obtida utilizando: `unique( as.character(dataframe$atributo))`. Isto assegura que não existem elementos iguais ou repetidos na seleção de *input* e que o tipo é o correto, neste caso em particular, do tipo `caracter`.

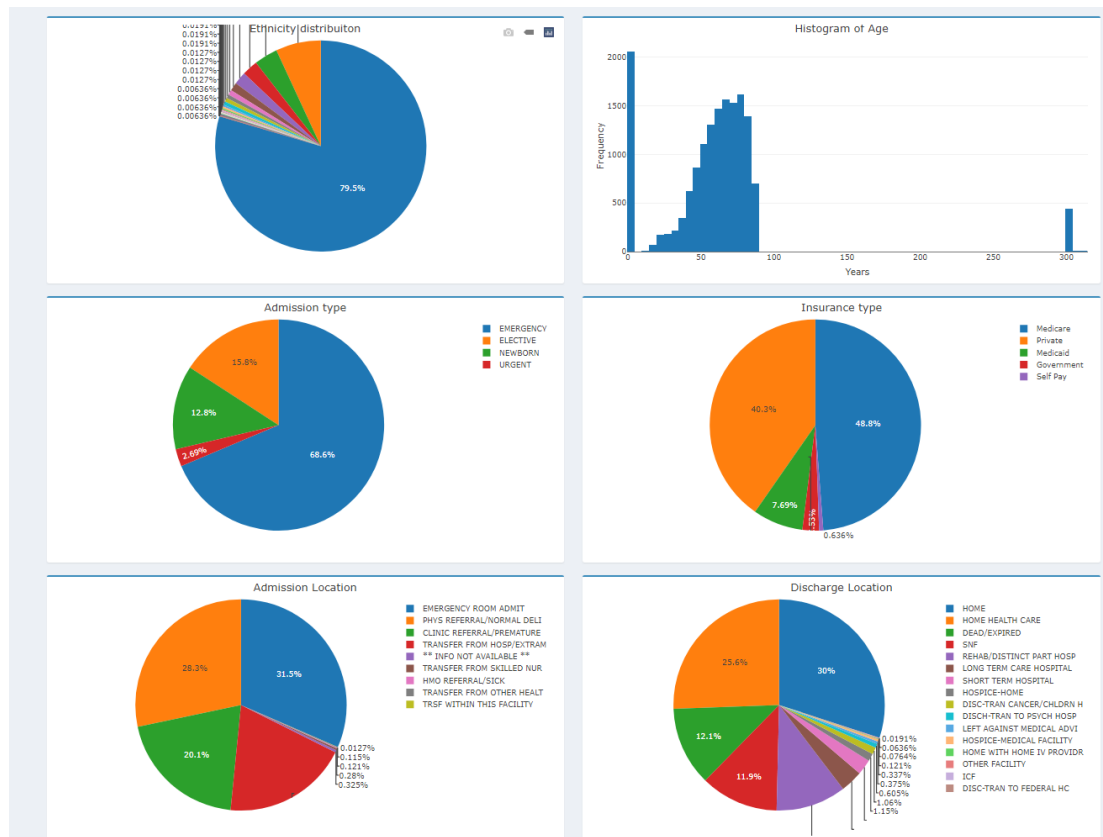


Figura 26- Output dos dados filtrados apresentados em gráficos, submenu General Patients Info

Como referido anteriormente, para ser possível filtrar por idade foi utilizado o elemento SliderInput, cujo *range* de idades é -1 a 300 anos (de forma a abranger os pacientes com

mais de 89 anos). Este método de *input* consiste de uma barra horizontal com 2 *sliders* sobrepostos: interagindo a primeira vez, selecionando e movendo o cursor para a direita o limite máximo de idade é definido. Movendo o cursor mais à esquerda é definido o limite mínimo de idade, sendo o range da seleção dado por uma faixa azul sobre a barra horizontal das idades. No caso de querer anular a seleção de idade e mostrar os dados de qualquer idade, basta colocar ambos os cursores na posição -1. Esta é posição inicial do slider, dado pelo argumento `value = c(-1,-1)` no interior do elemento `sliderInput()`.

Após todas as caixas de seleção de *input* e de interação com o utilizador, foi colocado um botão no canto inferior da caixa de *inputs* com o nome “Select”. O *output* dos filtros selecionados só é impresso após este botão ser selecionado. Esta opção foi tomada para solucionar um problema encontrado no desenvolvimento desta página: assim que um filtro era selecionado, era imediatamente dado o output no `mainPanel` e os restantes filtros apenas operavam sobre o dataframe resultante do filtro anterior, sendo impossível desselecionar um filtro previamente utilizado. O dataframe sobre onde estas operações eram realizadas não era capaz de “anular” uma operação de filtragem. Por exemplo: suponhamos que queríamos selecionar todos os pacientes de sexo feminino e que possuíam o seguro de saúde Medicare. Selecionando sexo feminino o dataframe passava a conter apenas pacientes do sexo feminino e após selecionar seguro de saúde Medicare, esse mesmo dataframe filtrado apresentava informações apenas sobre pacientes com seguro de saúde Medicare do sexo feminino. No entanto, se o utilizador quisesse desselecionar o primeiro filtro e visualizar os dados para apresentar todos os pacientes, independentemente do seu género, resultaria num erro, pois era impossível anular as alterações feitas no dataframe original e assim criar uma opção “ALL” que fizesse o reset do filtro.

Com o objetivo de solucionar este problema foi utilizada programação reativa do R Shiny. O utilizador pode alterar os valores dos variados filtros, mas apenas é executada uma ação após a seleção do botão “Select” referido anteriormente. Através da função `eventReactive()`, uma função reativa despoletada por um evento que neste caso é o clicar do botão “Select”.

```
fully_filtered -> eventReactive( click of Select button, subset the data based on
                                user inputs)
```

Assim que o botão Select é selecionado o segundo argumento de *eventReactive* ocorre:

```
fully_filtered <- eventReactive(input$select, {  
  
  filter_attribute()  
  
})
```

Cada atributo contém uma função reativa cujo funcionamento é o seguinte:

```
filter_previous_atribute <- reactive({  
  if(user_input == "ALL"){  
    original_dataframe  
  }  
  else {  
    original_dataframe %>%  
      filter(ATTRIBUTE_1 == originaldataframe$user_input)  
  }  
})  
  
filter_attribute <- reactive({  
  if(user_input == "ALL"){  
    filter_previous_atribute()  
  }  
  else {  
    filter_previous_atribute() %>%  
      filter(ATTRIBUTE_2 == originaldataframe$user_input)  
  }  
})
```

Esta chamada em cadeia de funções permite a não utilização de extensos *if else statements*, o que neste caso poderia ser problemático devido ao número considerável de atributos a serem filtrados. A mesma solução reativa é utilizada para o botão de *Download*, isto assegura que quando o utilizador clicar em “Download this data”, lhe será dado um ficheiro CSV com os dados dos pacientes filtrados, e não o *dataframe* df apenas.

4.3.4.3 ADMISSIONS

Este menu do dashboard foca-se na apresentação de informações estatísticas relativas às 58 976 admissões no hospital, sendo a sua fonte de dados exclusivamente a tabela ADMISSIONS. É composto por 2 submenus: General Admissions Info e Search by Admission ID.

4.3.4.3.1 General Admissions Info

Esta página é composta por 5 separadores no topo, cada uma delas focada em diferentes atributos:

- Admission type;
- Admission location;
- Discharge location;
- Insurance;
- Diagnosis.

A página segue o design de uma fluidRow, no entanto contém apenas uma *box*. No interior desta *box* está todo o conteúdo apresentado nesta página.

De forma a dividir este painel em separadores foi utilizado um `tabsetPanel()`, o que permite separar a visualização dos 5 atributos de ADMISSIONS referidos anteriormente. Cada um destes separadores é um `tabPanel()` constituído por uma tabela (DT datatable) e um histograma (Plotly histogram) com dados referentes a cada um dos atributos (Figura 27).

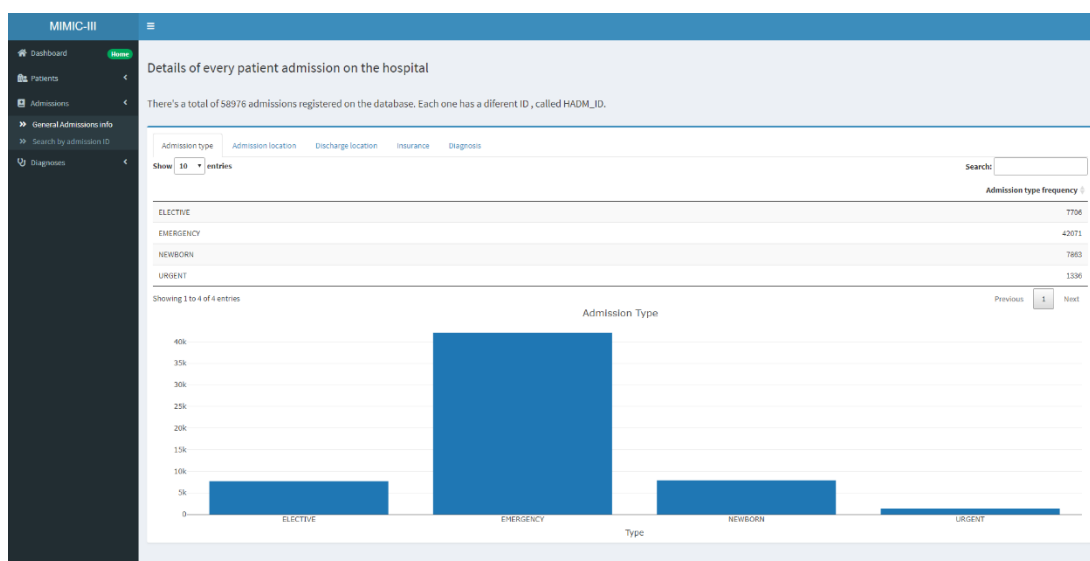


Figura 27 -General Admissions Info

De forma a obter os dados para a tabela os seguintes passos foram feitos para cada um dos atributos:

- Conversão do tipo de dados para factor, utilizando a função `as.factor` com `as.factor(ADMISSIONS$ATRIBUTO)`;
- Sumarização do passo anterior, utilizando a função `summary()`: `summary(as.factor(ADMISSIONS$ATRIBUTO))`;
- Conversão para dataframe utilizando `data.frame(summary(as.factor(ADMISSIONS$ATRIBUTO)))`;

O resultado é um dataframe (Figura 28) deste tipo:

	Admission location frequency
** INFO NOT AVAILABLE **	204
CLINIC REFERRAL/PREMATURE	12032
EMERGENCY ROOM ADMIT	22754
HMO REFERRAL/SICK	102
PHYS REFERRAL/NORMAL DELI	15079
TRANSFER FROM HOSP/EXTRAM	8456
TRANSFER FROM OTHER HEALT	71
TRANSFER FROM SKILLED NUR	273
TRSF WITHIN THIS FACILITY	5

Figura 28- Dataframe para realização de histograma

Já os histogramas desenhados pelo Plotly representam os dados desse mesmo dataframe (Figura 29).

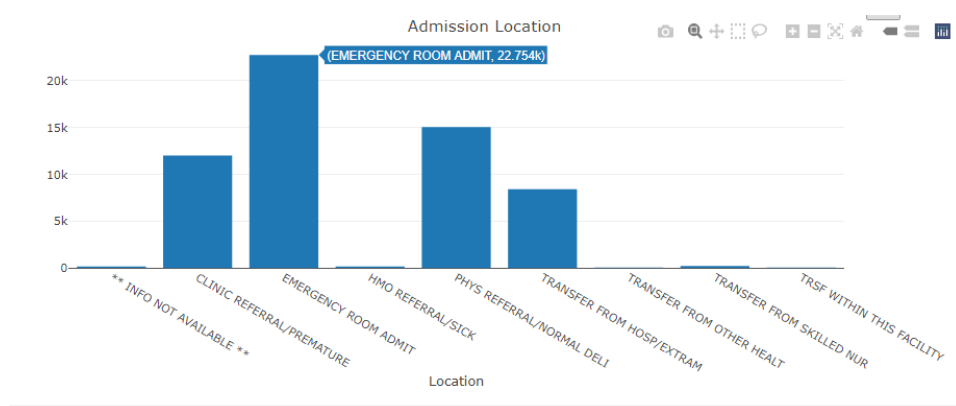


Figura 29- Histogramas com mouse-over do submenu Admissions Info

4.3.4.3.2 Search by admission ID

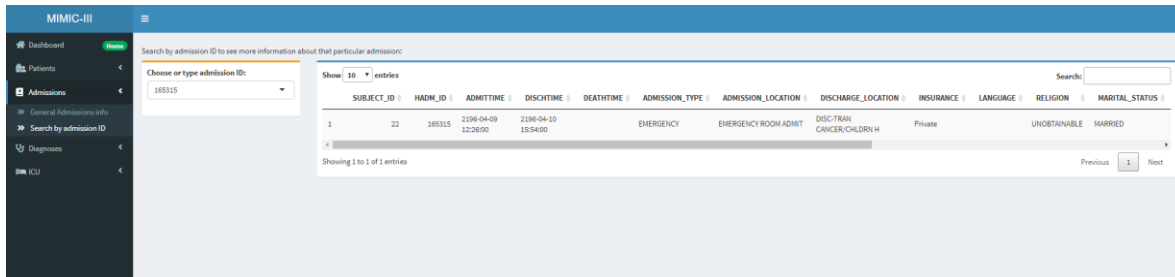
Search by admission ID é um submenu de Admissions que permite ao utilizador, mediante a introdução de um HADM_ID válido, obter detalhes sobre uma determinada admissão hospitalar. Os detalhes oferecidos por esta pesquisa são os seguintes, por esta ordem:

- SUBJECT_ID do paciente;
- HADM_ID do paciente;
- Data e hora de admissão;
- Data e hora de alta;
- Data e hora de óbito;
- Tipo de admissão;
- Local de admissão;
- Local após alta;
- Seguro;
- Linguagem;
- Religião;
- Estado Matrimonial;
- Etnia;
- Edregtime;
- Edoutime;
- Diagnóstico preliminar;
- Morte hospitalar ou não;
- Existência ou não de dados em chart events.

Este submenu segue uma estrutura de fluidRow, sendo constituído por 2 elementos do tipo *box*. Um contendo uma caixa de *input*, colocado no canto esquerdo da página e outro à direita do primeiro contendo uma tabela com o resultado do *input* recebido.

Tal como no menu anterior, também aqui foi utilizado *selectizeInput* para acelerar o processo de seleção e inserção de *input*, visto serem quase 59 000 as opções disponíveis. As opções consistem em todos os Ids de admissão, obtidos por `ADMISSIONS$HADM_ID`.

Na construção da tabela que contém a informação de cada admissão hospitalar, foi utilizado `DT::dataTable`, sendo depois impresso a linha de `ADMISSIONS` com `HADM_ID` igual ao input selecionado na caixa de input (Figura 30).



SUBJECT_ID	HADM_ID	ADMTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE	LANGUAGE	RELIGION	MARITAL_STATUS
1	22	185315	2198-04-09 12:26:00	2198-04-10 15:54:00	EMERGENCY	EMERGENCY ROOM ADMIT	DISC/TRAN CANCER/CHLDREN H	Private		UNOBTAINABLE	MARRIED

Figura 30- Submenu Search by Admission ID

4.3.4.4 DIAGNOSES

O menu diagnoses é o menu com o maior número de submenus do dashboard. Reúne informações detalhadas sobre todos os diagnósticos existentes na MIMIC-III, sendo capaz de não só oferecer informação detalhada acerca de cada paciente em particular, mas também dados estatísticos e de comparação.

As funcionalidades deste menu foram separadas em submenus:

- Informação geral de diagnóstico e estatística;
- Informação estatística por grupo de código ICD;
- Pesquisa por código ICD específico;
- Informação e comparação sobre primeiros diagnósticos;
- Diagnósticos de cada paciente através de pesquisa por ID do mesmo.

4.3.4.4.1 GENERAL DIAGNOSES INFO

Este submenu foca-se especialmente nos diferentes grupos de códigos ICD-9, apresentando uma visão geral sobre a frequência de cada um, a sua LOS e ainda informações mais detalhadas, sendo feito um sumário de cada grupo de código ICD-9.

Como explicado nos capítulos anteriores, existem 19 grupos de códigos ICD. Neste menu é apresentada a frequência absoluta de cada um deles através de um histograma e um *boxplot* do LOS para cada um desses grupos. Ambos são colocados lado a lado na mesma

linha, sendo ainda colocado sob os mesmos um menu onde o utilizador pode inserir um grupo de código ICD e visualizar informação sumarizada do mesmo.

Além disso é apresentado um histograma contendo as 20 de doenças mais diagnosticadas, independentemente do seu grupo (Figura 30).



Figura 30- General Diagnoses Info

Esta página utiliza uma estrutura de fluidRow e está dividida em duas grandes partes: uma linha superior onde são dispostos os dois gráficos (contidos num elemento column()), e outra abaixo desta, onde é colocado o gráfico dos 20 ICD's mais frequentes e é feita a escolha da informação a ser sumarizada (contidos cada um deles num elemento box()).

Para a realização desta página, com a ajuda da função *filter()* a tabela ADMISSIONS foi dividida em 19 dataframes mais pequenos, cada um deles contendo um grupo de código ICD-9. Cada dataframe contém apenas um *subset* de pacientes diagnosticados com um determinado *range* de códigos iCD-9, mantendo todos os mesmos atributos do dataframe original (ADMISSIONS). Estes dataframes foram assim utilizados para a realização do gráfico que compara os LOS de cada grupo de ICD usando *boxplots* (Figura 30) e ainda para o sumário de informação relevante para cada ICD-9:

```
plot_ly(
  data = ADMISSIONS_subset_1,
  y = ~LOS,
  type = "box",
  name = "name"
)
```

No entanto, a realização do gráfico histograma usando plotly (que apresenta a frequência de cada grupo ICD) mostrou-se não trivial. Era necessário enviar um dataframe com o nome do grupo ICD e a sua frequência. Assim foi necessário solucionar este problema.

Primeiramente foi criada uma lista Frequency com o número de linhas de cada dataframe subset de ADMISSIONS:

```
Frequency <- c(nrow(infections139), nrow(neoplasms239), nrow(endocrine279),
nrow(blood289), nrow(mental319), nrow(nervous389), nrow(circulatory459),
nrow(respiratory519), nrow(digestive579), nrow(genitourinary629), nrow(pregnancy679),
nrow(skin709), nrow(muscle739), nrow(congenital759), nrow(perinatal779),
nrow(symptoms799), nrow(injury999), nrow(v), nrow(e))
```

Foi feita uma outra lista ICD9CODE contendo na mesma ordem o *range* de ICD correspondente a cada um dos elementos da lista anterior:

```
ICD9CODE <- c("0-139", "140-239", "240-279", "280-289", "290-319", "320-389",
"390-459", "460-519", "520-579", "580-629", "630-679", "680-709", "710-739",
"740-759", "760-779", "780-799", "800-999", "V01-V091", "E000-E999")
```

Assim, foi possível criar um dataframe (Figura 33) contendo na primeira coluna o range de ICD e na segunda a frequência desse mesmo range, o que facilitou a criação do histograma:

	ICD9CODE	Frequency
1	0-139	20311
2	140-239	14235
3	240-279	69592
4	280-289	25262
5	290-319	24876
6	320-389	23334
7	390-459	140256
8	460-519	44825
9	520-579	38527
10	580-629	32867
11	630-679	654
12	680-709	8789
13	710-739	13392
14	740-759	4078
15	760-779	20160
16	780-799	29691
17	800-999	42948
18	V01-V091	74705
19	E000-E999	22544

```
diagnosesPlot <- data.frame(ICD9CODE,
Frequency)
```

```
plot_ly(
  data = diagnosesPlot,
  x = ~ICD9CODE,
  y = ~Frequency,
  type = "bar",
  text = ~Frequency,
  textposition = "auto",
  hoverinfo = "text",
  hovertext = paste("IC9-Code:",
diagnosesPlot$ICD9CODE)
)
```

Figura 31- Dataframe DiagnosesPlot

Cada barra do histograma apresenta a frequência absoluta e fazendo *mouse-over* sobre qualquer uma das barras é apresentada descrição curta por texto do ICD9 condizente.

O mesmo sucede com o outro histograma desta página, com o objetivo de rapidamente identificar a doença (Figura 32).

Para a seleção do grupo ICD-9 e consequente impressão do sumário de cada um, foi utilizado `selectInput()`, sendo as 19 opções manualmente inseridas. Após uma opção selecionada é imediatamente impresso numa *box* ao lado direito o *output* de `summary(ADMISSIONS_subset)`.

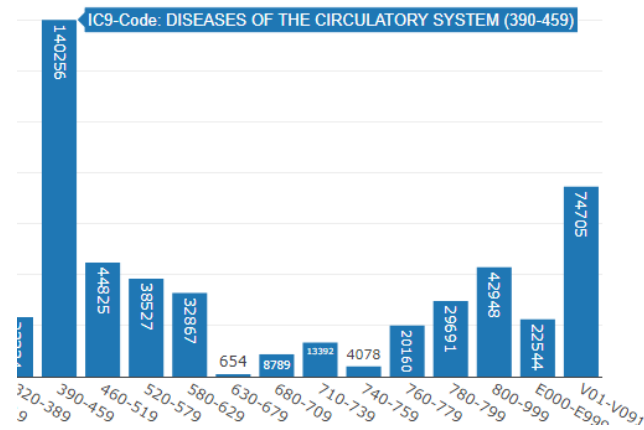


Figura 32- Mouse-over histograma

4.3.4.4.2 Specific ICD9 Group Details

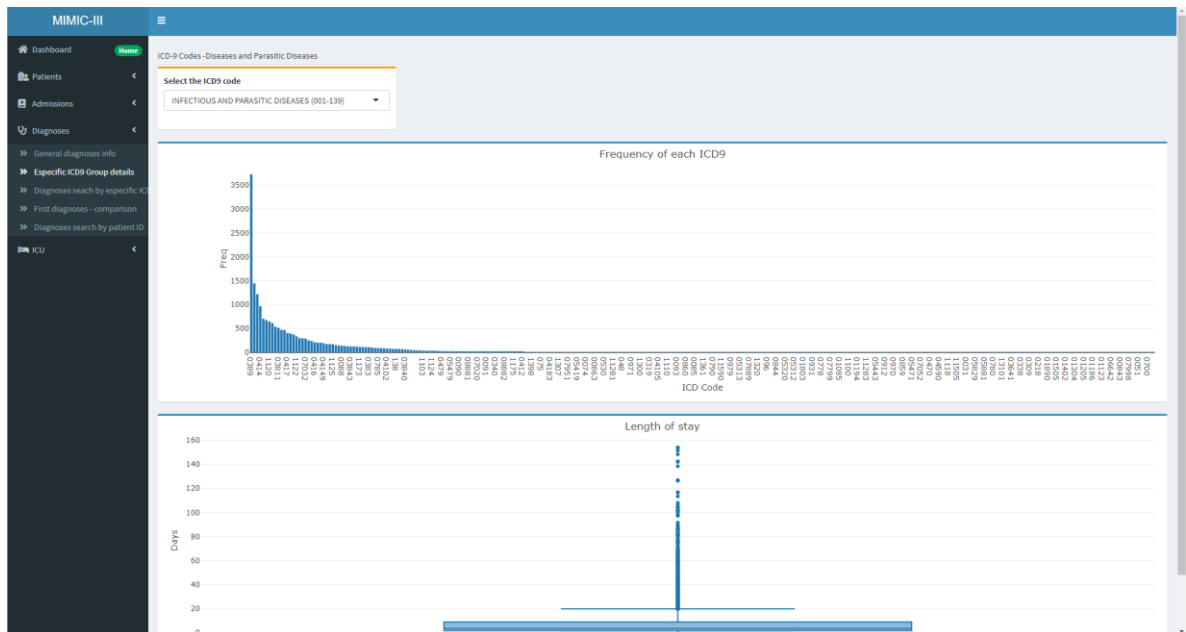


Figura 33- Submenu Search Specific ICD9 group details

Este submenu apresenta gráficos de frequência e de LOS, no entanto, ao contrário do submenu anterior, foca-se na apresentação de cada um dos códigos ICD-9 em específico e não apenas num grupo de códigos (Figura 33).

Através de um menu lateral o utilizador pode seleccionar um grupo que deseja observar, por exemplo “Diseases of the respiratory system(460-519)”.

Este submenu utiliza a estrutura *fluidRow*, sendo cada um dos três componentes da mesma colocados numa *box*. A primeira é uma caixa de *input*, utilizando *selectInput*, cujas *choices* são os 19 grupos de ICD, e as restantes 2 *boxes* são constituídas por um gráfico de barras e um *boxplot*.

Assim que essa opção é seleccionada é gerado um gráfico de barras com o número de pacientes no eixo dos Y e o código ICD-9 no eixo dos X. Cada barra apresenta o número de pacientes desse código em específico, sendo colocado por ordem decrescente de casos. Ou seja, do lado esquerdo do gráfico estão os códigos mais frequentes dentro do grupo ICD escolhido e do direito os menos. Ao mesmo tempo, sob este gráfico de barras é também gerado um *boxplot* correspondente ao LOS desse mesmo grupo. No menu anterior era apresentado o mesmo tipo de gráfico, mas com o objetivo de comparar com outros ICDs. Ao contrário, este agora permite uma visão mais detalhada do mesmo, permitindo ver claramente os valores inscritos no gráfico fazendo *mouse-over* sobre o mesmo. Ao fazer *mouse-over* é possível observar:

- Valor mínimo;
- Valor máximo;
- 1º quartil;
- 3ª quartil;
- Mediana;
- Limite Superior;
- Outliers/Valores atípicos;

Além disso a possibilidade oferecida pelo Plotly de fazer zoom no gráfico e seleccionar apenas uma zona a observar permite observar com detalhe os dados fornecidos pelo *boxplot* (Figura 34 e Figura 35).

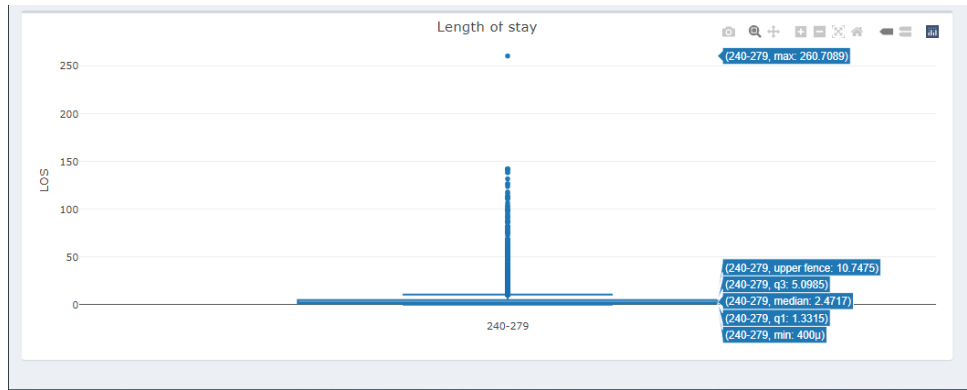


Figura 34- BoxPlot de LOS

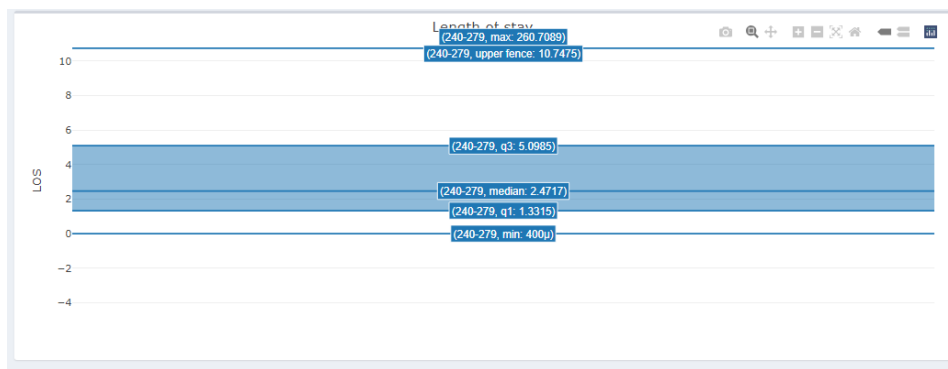


Figura 35 -BoxPlot de LOS, informação em mouse-over

Neste capítulo importa mencionar de que forma os dados foram preparados para a criação do gráfico de barras e o porquê de utilizar um e não um histograma. O gráfico de barras é mais apropriado para apresentar dados categóricos, sendo este o caso. O objetivo é apresentar a frequência de cada um dos códigos ICD-9. Para criar um gráfico deste tipo utilizando Plotly é necessário fornecer os seguintes campos:

```
plot_ly(
  data = counts,
  x = ~Var1,
  y = ~Freq,
  type = "bar"
)
```

Counts é um *dataframe* criado em tempo real sempre que uma opção diferente é selecionada. Este dataframe temporário é construído da seguinte forma:

```
counts <- as.data.frame(table(ADMISSIONS_subset_1$ICD9CODE))
```

A função `table`, da base de R, permite criar uma representação categórica (o pretendido para utilizar no gráfico de barras do Plotly) dos dados, criando o nome de uma variável (Var1) e a frequência (Freq). Depois basta converter esta tabela para dataframe, visto Plotly apenas aceitar dados deste tipo, utilizando `as.data.frame()`, também pertencente à base do R.

4.3.4.4.3 Diagnoses search by Specific ICD

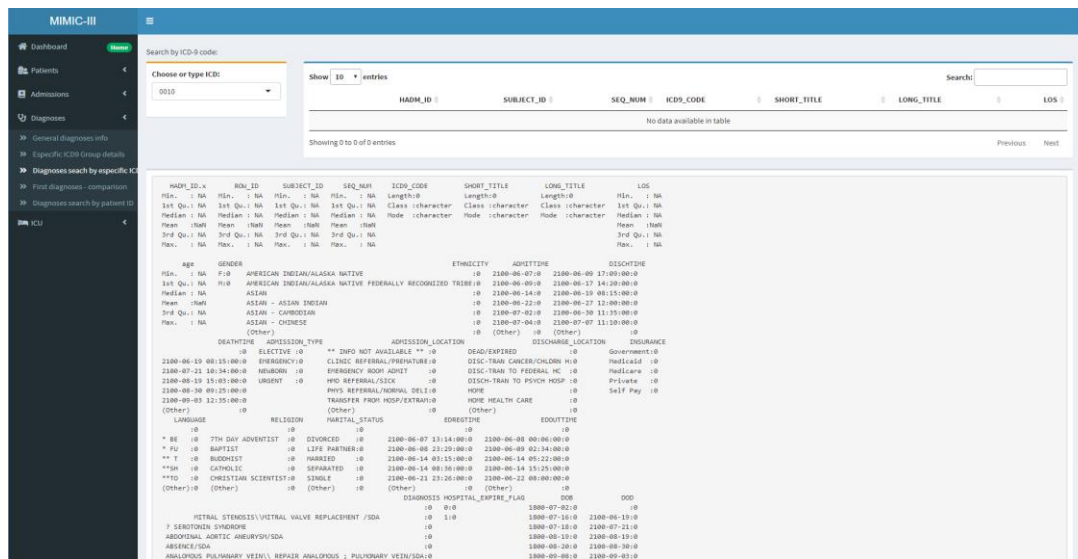


Figura 36- Diagnoses Search by Specific ICD

Um dos principais requisitos do utilizador desta ferramenta é a capacidade obter dados sobre todos os pacientes que foram diagnosticados com uma determinada doença. Para além de estatísticas gerais sobre cada grupo geral de ICD, nesta ferramenta é possível pesquisar exatamente por um código ICD específico e assim obter informação sobre que pacientes foram diagnosticados com essa mesma doença, o número de vezes que cada doente foi diagnosticado com a mesma nas diversas hospitalizações e ainda dados

sumarizados sobre esses mesmos doentes, por exemplo: idade, género, a severidade/prioridade desse ICD para os pacientes baseado no valor SEQ_NUM, taxa de mortalidade, entre outras (Figura 36).

Com esse objetivo, foram reunidos dados de quatro tabelas: PATIENTS, DIAGNOSES, D_ICD_DIAGNOSES e ICUSTAYS.

Esta página utiliza novamente uma estrutura de fluidRow, sendo composta por três boxes: seleção de input, tabela com dados correspondentes ao input e sumário dos dados apresentados na tabela.

Na seleção de input, tal como no menu Patients é utilizada novamente a função selectizeInput(), sendo as opções todas carregadas no lado do servidor. Esta decisão foi tomada de forma a evitar perdas de performance visto as opções de código ICD-9 serem mais de 14 500. Por mais, foi decidido que apareceriam como opção para selecionar ou auto-completar (assim que o utilizador começa a escrever na caixa de input) todos os códigos ICD-9 presentes em D_ICD_DIAGNOSES. Isto permite que o utilizador rapidamente possa consultar também se existe algum paciente diagnosticado com determinado código ICD-9. Em caso negativo a tabela aparecerá vazia, tal como o sumário.

O código ICD introduzido pelo utilizador serve de input para a função filter() (e guardado em input\$diagnose_ICD) sobre a tabela DIAGNOSES, que por sua vez foi estendida com a definição por texto do código ICD, oriundo da tabela D_ICD_DIAGNOSES (resultado de *left join* de DIAGNOSES com D_ICD_DIAGNOSES), Figura 37:

ROW_ID	SUBJECT_ID	HADM_ID	SEQ_NUM	ICD9_CODE	SHORT_TITLE	LONG_TITLE
1297	109	172335	1	40301	Mal hyp kid w cr kid V	Hypertensive chronic kidney disease, malignant, with chroni...
1298	109	172335	2	486	Pneumonia, organism NOS	Pneumonia, organism unspecified
1299	109	172335	3	58281	Chr nephritis in oth dis	Chronic glomerulonephritis in diseases classified elsewhere

Figura 37- Dataframe resultante de *left join* de DIAGNOSES com D_ICD_DIAGNOSES

Além disso é acrescentado o campo LOS, oriundo da tabela ICUSTAYS. Para isto foi necessário utilizar um *left join* da tabela DIAGNOSES com a tabela ICUSTAYS, através do atributo HADM_ID, como explicado no capítulo anterior.

Mais uma vez, foi escolhido utilizar a tabela da biblioteca DT pois permite definir o número de linhas a apresentar de uma só vez, ordenar por atributo (por exemplo, SEQ_NUM) e

pesquisar numa barra de pesquisa no topo direito por alguma informação contida na mesma (por exemplo, de entre todos 500 pacientes diagnosticados com ICD-9 1645 , o paciente com SUBJECT_ID igual a 225).

Localizada no separador “Diagnoses”, no submenu “Search by ICD”, é assim possível seleccionar dos vários disponíveis ou inserir manualmente o código ICD num menu lateral. De notar que foi decidido apresentar todos os códigos ICD neste menu de selecção, e não apenas aqueles com que os pacientes foram diagnosticados, sendo assim possível seleccionar um código ICD não presente na tabela ADMISSIONS, e o *output* seja uma tabela vazia. Esta escolha permite assim auferir rapidamente se existe na base de dados MIMIC-III algum paciente que tenha sido diagnosticado com tal código ICD.

À direita do menu onde é realizado o input, é apresentada uma tabela, dentro de uma box do R Shiny, com todos os pacientes diagnosticados com a mesma doença, sendo apresentadas as seguintes informações para cada paciente:

- SUBJECT_ID;
- HADM_ID;
- SEQ_NUM;
- ICD-9 CODE;
- SHORT TITLE;
- LONG TITLE;
- LOS;

É, no entanto, ter em atenção que o valor de LOS apresentado diz respeito à admissão inteira. Um paciente que tenha sido hospitalizado múltiplas vezes e diagnosticado com o mesmo ICD aparecerá nesta lista múltiplas vezes também (sendo as distintas hospitalizações distinguidas pelo HADM_ID), e o correspondente LOS variará consoante a hospitalização.

Para além desta informação, é ainda apresentada mais informação através de um sumário (utilizando a função `summary()`) de todos os dados seleccionados anteriormente. Este sumário encontra-se numa nova box do R Shiny. Para conter mais informação relevante foi feito um *left join* da tabela DIAGNOSES_ICD com o *dataframe* df através do SUBJECT_ID.

```
summary(filter(left_join(DIAGNOSES_ICD, df, by = "SUBJECT_ID"),
  ICD9_CODE == input$diagnose_icd))
```

Neste sumário temos acesso a várias informações, onde destacamos:

- SEQ_NUM: é apresentado valor mínimo, máximo, valor do 1º e 3º quartil e ainda média e mediana, o que permite inferir sobre o grau de prioridade desta doença em específico;
- Idade: idade mínima, máxima, média, mediana e valores de 1º e 3º quartil, o que permite conhecer melhor o tipo de pacientes que normalmente são diagnosticados com esse código ICD;
- Género: número de pacientes do sexo feminino ou masculino;
- Admission type e localization: permite auferir o tipo de situação que leva um paciente a deslocar-se ao hospital com este código ICD;
- Discharge location: dá informações sobre para onde os pacientes são levados após receberam alta, tendo sido diagnosticados com este ICD em particular;
- Hospital expire flag e expire flag: permite saber qual a mortalidade no hospital e também que pacientes faleceram tendo sido diagnosticados com esta doença podendo esta ser, ou não, a causa direta.

Para além destas informações existem outras que embora menos relevantes também fazem sentido analisar tais como: etnia pacientes, linguagem, tipo de seguro, estado matrimonial, diagnóstico rápido atribuído dado na entrada do hospital, religião, entre outros.

4.3.4.4.3 First Diagnoses – comparison

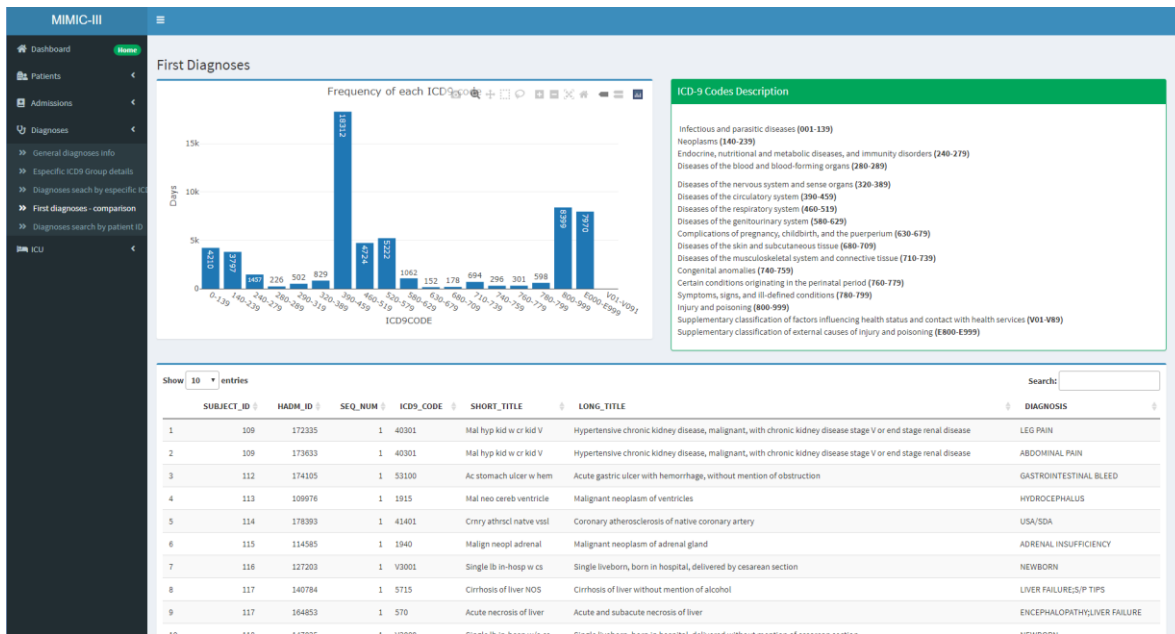


Figura 38- Submenu First Diagnoses

Este submenu foca-se em analisar quais são os primeiros diagnósticos de cada admissão hospitalar. Como explicado em capítulos anteriores, cada um dos diagnósticos possui um SEQ_NUM, responsável por indicar o nível de prioridade da doença para o doente, assim este submenu possui como objetivo ajudar a compreender quais são os diagnósticos mais relevantes, selecionando todos os diagnósticos presentes na tabela DIAGNOSES_ICD cujo SEQ_NUM é igual a 1. Para além disso, neste submenu foi criada uma tabela que compara o diagnóstico inicial feito na entrada do hospital para cada admissão hospitalar (presente na tabela ADMISSIONS) com a real doença diagnosticada (presente com SEQ_NUM igual a 1 na tabela DIAGNOSES_ICD). (Figura 38)

Baseando-se numa estrutura fluidRow este menu é constituído por três boxes: a primeira é responsável por apresentar um gráfico de barras utilizando Plotly, mostrando o número de diagnósticos com SEQ_NUM igual a 1 para cada grupo ICD; a segunda contém texto em formato HTML explicando o funcionamento e propósito desta página e, finalmente, a terceira contendo a tabela que compara os diagnósticos feitos na entrada no hospital com o verdadeiro diagnóstico.

Para a criação do gráfico de barras foi novamente necessário criar um dataframe novo (tal como no submenu General Diagnoses Info), contendo apenas 19 linhas (cada grupo ICD) (Figura 41).

	a	Frequency2
1	0-139	4210
2	140-239	3797
3	240-279	1457
4	280-289	226
5	290-319	502
6	320-389	829
7	390-459	18312
8	460-519	4724
9	520-579	5222
10	580-629	1062
11	630-679	152
12	680-709	178
13	710-739	694
14	740-759	296
15	760-779	301
16	780-799	598
17	800-999	8399
18	V01-V091	7970
19	E000-E999	0

Figura 41- Dataframe DiagnosesPlot2

```
plot_ly(
  data = diagnosesPlot2,
  x = ~ICD9CODE,
  y = ~Frequency2,
  type = "bar",
  text = ~Frequency2,
  textposition = "auto",
  hoverinfo = "text",
  hovertext = paste("IC9-Code:",
diagnosesPlot2$a)
)
```

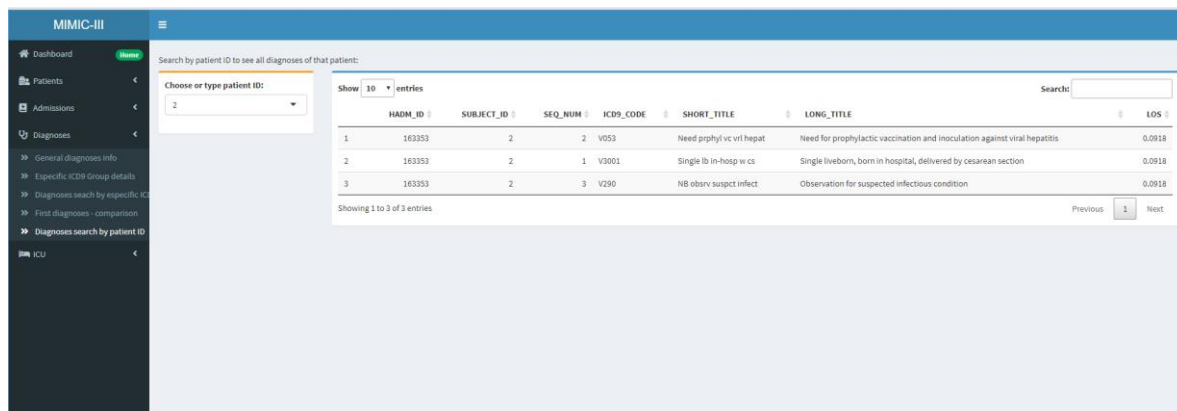
Finalmente, a *box* contendo uma tabela contendo todas as admissões hospitalares e consequente código ICD com SEQ_NUM resulta de um *left join* da tabela ADMISSIONS com a tabela previamente usada neste submenu contendo apenas diagnósticos cujo SEQ_NUM é igual a 1, através do HADM_ID. Assim esta tabela contém os seguintes argumentos:

- SUBJECT_ID;
- HADM_ID;
- SEQ_NUM;
- ICD9_CODE;
- SHORT TITLE;
- LONG_TITLE;
- DIAGNOSIS;

Este último atributo, diagnosis, é importado da tabela ADMISSIONS e permite fazer a comparação com o real diagnóstico presente na tabela DIAGNOSES.

Esta tabela contém mais de 58 000 linhas contendo múltiplas admissões hospitalares, no entanto, o facto de ter sido utilizado novamente dataTable da biblioteca DT permite uma conveniente utilização desta tabela, sendo possível pesquisar qualquer tipo de dado contido na mesma, escolher quantas linhas ver de uma só vez sem mudar de página da tabela, ordenar por qualquer tipo de atributo e ainda a existência de uma barra horizontal para uma maior facilidade de utilização.

4.3.4.4 Diagnose Search by Patient ID



	HADM_ID	SUBJECT_ID	SEQ_NUM	ICD9_CODE	SHORT_TITLE	LONG_TITLE	LOS
1	163353	2	2	V053	Need prophyl vc vrt hepat	Need for prophylactic vaccination and inoculation against viral hepatitis	0.0918
2	163353	2	1	V3001	Single lb in-hosp w ca	Single liveborn, born in hospital, delivered by cesarean section	0.0918
3	163353	2	3	V290	HB obsrv suspect infect	Observation for suspected infectious condition	0.0918

Figura 41- Submenu Diagnoses Search by patient ID

Este submenu tem como função, mediante a introdução do ID do paciente, fornecer todos os diagnósticos do paciente, oriundo da tabela DIAGNOSES.

Uma das necessidades do utilizador é a visualização de todo o histórico médico de um paciente através do seu ID num só local. Para isto apenas necessita de inserir o seu ID numa caixa de pesquisa e toda a informação é apresentada numa tabela que permite ordenar os dados e ainda filtrá-los. Como explicado anteriormente, quando um paciente é admitido/internado é associado a esse internamento um código ICD-9 de doença, que identifica de forma precisa qual o problema do doente, no entanto, muitas das vezes, um paciente não tem apenas uma só doença associada, mas várias. Ou seja, um internamento está associado a múltiplos códigos ICD-9. O seu grau de importância e gravidade no caso

de cada paciente é definido pelo campo SEQ_NUM, sendo esse número maior quanto menor for a sua relevância.

Se o utilizador quiser pesquisar todo o histórico de um paciente na MIMIC-III teria de consultar a tabela DIAGNOSES e filtrar pelo seu SUBJECT_ID. No entanto, desta forma apenas são apresentados os códigos ICD-9, que não sendo traduzidos tornam a identificação rápida da doença mais difícil. Assim, nesta secção, todos os internamentos têm os seus códigos traduzidos, tendo sido feito merge da tabela DIAGNOSES já filtrada pelo SUBJECT_ID do paciente selecionado com a tabela D_ICD_DIAGNOSES, que contém a definição curta e longa de cada um dos códigos ICD-9.

Para além desta informação, é incluída ainda o LOS de cada admissão hospitalar. LOS é um campo importante, que permite tirar vários tipos de conclusões, tais como a dificuldade que o paciente teve em recuperar de uma doença (por exemplo, quando comparado com um paciente com características muito semelhantes) ou ter uma ideia do tempo de internamento para uma dada enfermidade. Como explicado no capítulo 2, LOS é um campo apenas presente na tabela ICUSTAYS, para o apresentar foi necessário realizar um merge dos dados presentes na tabela DIAGNOSES com ICUSTAYS, no entanto, surgiu um problema. Existem pacientes que numa só admissão (HADM_ID) tiveram duas ou mais admissões em ICU distintas (ICUSTAY_ID) (Figura 5) .

Como explicado no capítulo 2, este último identificador diz apenas respeito à tabela ICUSTAYS e não está presente em mais nenhuma tabela da MIMIC-III. Isto fez com que na primeira tentativa de merge das duas tabelas, DIAGNOSES e ICUSTAYS, existissem problemas e entradas duplicadas. Para um mesmo HADM_ID podem existir dois ou mais valores diferentes de LOS (derivados de ICUSTAY_ID's diferentes), isto faz com que seja impossível saber exatamente que código ICD-9 esteve ligado a esse diagnóstico.

Por exemplo, na Figura 42, podemos observar que o paciente com o SUBJECT_ID 305 teve 4 admissões distintas no hospital. Na sua terceira, ou seja, na admissão com HADM_ID igual a 122211, existiram 2 internamentos em ICU, cujos códigos ICUSTAY_ID foram 224571 e 232248. A primeira teve um LOS de 0.7925 dias, enquanto a última 5.8118 dias.

Se verificarmos na tabela ADMISSIONS, e procurarmos pelo HADM_ID cujo número é igual a 12211 verificamos que é impossível distinguir em qual das duas permanências em ICU o ICD-9 correspondente foi responsável pelo LOS (Figura 43)

ROW_ID	SUBJECT_ID	HADM_ID	ICUSTAY_ID	DBSOURCE	FIRST_CAREUNIT	LAST_CAREUNIT	FIRST_WARDID	LAST_WARDID	INTIME	OUTTIME	LOS
406	305	133059	232719	carevue	CCU	CCU	7	7	2125-04-27 18:47:43	2125-05-01 16:50:19	3.9185
407	305	108015	297009	carevue	CSRU	CSRU	15	15	2126-01-02 11:28:23	2126-01-07 10:07:29	4.9438
408	305	122211	224571	carevue	MICU	MICU	50	50	2127-06-19 23:50:45	2127-06-20 18:51:59	0.7925
409	305	122211	232248	carevue	MICU	MICU	50	50	2127-07-03 20:02:34	2127-07-09 15:31:36	5.8118
410	305	194340	217232	metavision	SICU	SICU	33	33	2129-09-03 12:31:31	2129-09-05 23:00:50	2.4370

Figura 42- Dataframe ADMISSIONS, paciente 305 com 2 admissões

ROW_ID	SUBJECT_ID	HADM_ID	SEQ_NUM	ICD9_CODE
3593	305	108015	7	4019
3594	305	108015	8	2720
3595	305	108015	9	25000
3596	305	122211	1	42820
3597	305	122211	2	99859
3598	305	122211	3	5070
3599	305	122211	4	2639
3600	305	122211	5	4271
3601	305	122211	6	4280
3602	305	122211	7	2859
3603	305	122211	8	4148
3604	305	122211	9	25060
3605	305	122211	10	3572
3606	305	122211	11	4019
3607	305	122211	12	53081
3608	305	122211	13	41400
3609	305	122211	14	V4581
3610	305	133059	1	41071
3611	305	133059	2	99672
3612	305	133059	3	42820
3613	305	133059	4	4280
3614	305	133059	5	4240

Figura 43- Dataframe ADMISSIONS, contendo todas as admissões do paciente 305

A solução encontrada para este problema foi a seguinte: iterar sobre a tabela ICUSTAYS e somar os LOS de HADM_ID iguais. Em R, a função `aggregate` permite fazê-lo de forma sucinta:

```
aggregate(LOS~HADM_ID, data=data_frame, FUN=sum)
```

Tal como os submenus anteriores está organizado numa `fluidRow`, sendo constituído por duas boxes: uma com a função de receber input do utilizador e outra de apresentar o output, apresentando uma tabela.

O funcionamento deste submenu é equivalente ao menu “Search by ID”, sendo aqui o mesmo dataframe filtrado por `SUBJECT_ID`. Assim, o output na tabela apresenta todas as admissões hospitalares do doente e respetivos diagnósticos com LOS.

4.3.4.5 ICU

Este menu contém informação relativa a todas as visitas à ICU do hospital e baseia-se especialmente na tabela ICUSTAYS da MIMIC-III. Está dividido em 2 submenus, cada um com uma função diferente: General ICU Stays Info e Patient Mortality. Iremos descrever a função e implementação de cada um deles de seguida.

4.3.4.5.1 General ICU Stays Info

Este submenu tem como função possibilitar ao utilizador filtrar por determinados atributos todas as hospitalizações em ICU, presentes na tabela ICUSTAYS. Os atributos são os seguintes:

- Género;
- Etnia;
- Morte em hospital;
- Morte;
- Primeira e última unidade ICU;
- Primeira e última enfermaria;
- LOS;
- Idade.

Utilizando a mesma técnica do submenu Patients General Info, é capaz de permitir a utilização de filtros sobre o dataframe ICUcomplete (detalhado anteriormente). Através de programação reativa, só quando o utilizador clicar em Select os filtros são aplicados, e automaticamente tanto o sumário, a tabela e os gráficos são atualizados (Figura 44).

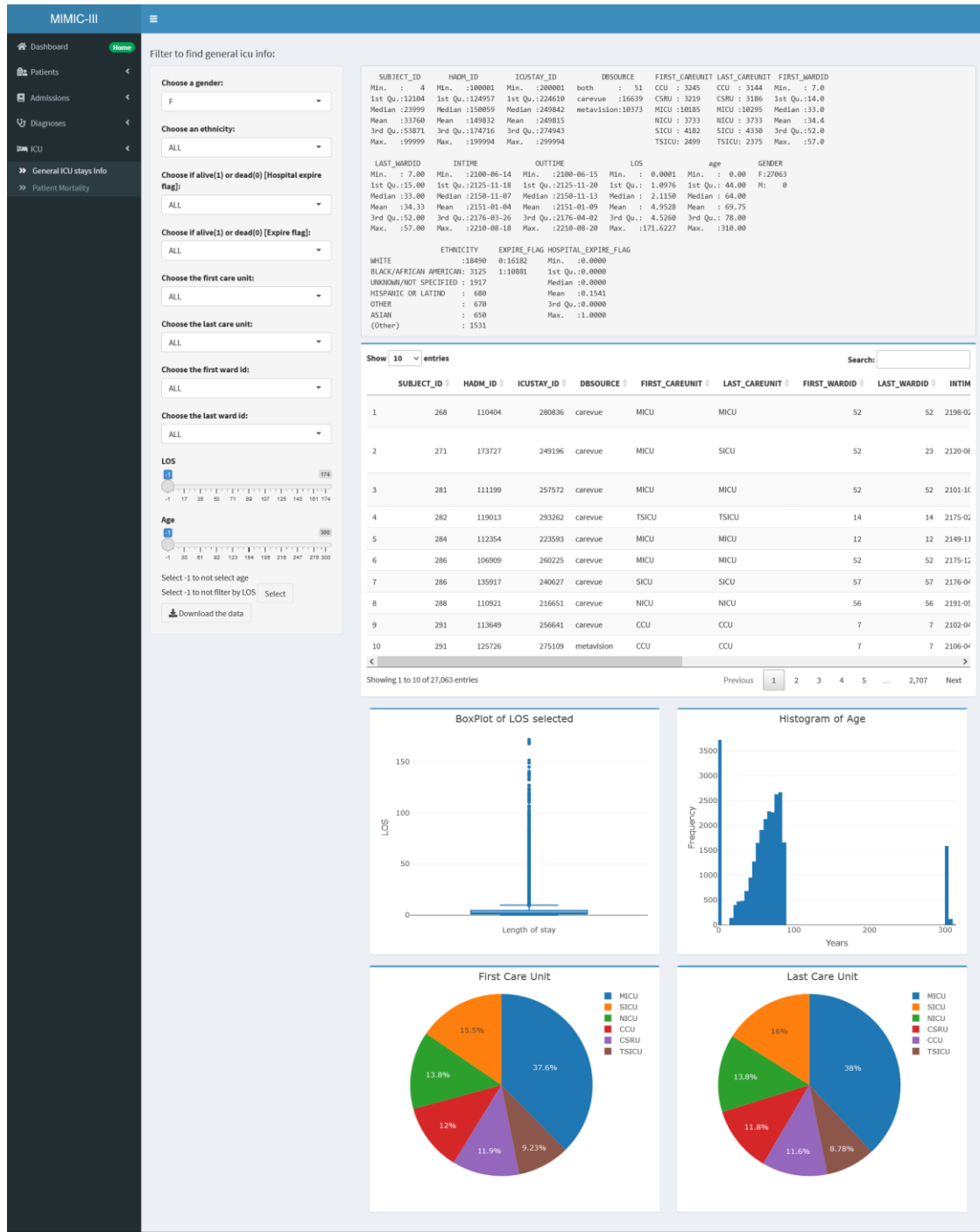


Figura 44- Submenu general ICU stays info

```

filtered_gender2 <- reactive({
  if(input$in_gender2 == "ALL"){
    ICUcomplete
  } else {
    ICUcomplete %>%
      filter(GENDER == input$in_gender2)
  }
})

(...)

filtered_lward <- reactive({
  if(input$in_lward == "ALL"){
    filtered_fward()
  } else {
    filtered_fward() %>%
      filter(LAST_WARDID == input$in_lward)
  }
})

fully_filtered2<- eventReactive(input$select2, {
  filtered_lward()
})
output$download2 <- downloadHandler(
  filename = function(){"data.csv"},
  content = function(fname){
    write.csv(fully_filtered2(), fname)
  }
)

```

Além disso, o botão Download permite descarregar para a máquina do utilizador os dados a serem exibidos nesse momento, em formato CSV. Este submenu utiliza a estrutura sidebarLayout, sendo a caixa de seleção dos filtros colocado em sidebarPanel e todo o output resultante em mainPanel. Na caixa de seleção estão disponíveis os filtros referidos no parágrafo anterior. Após o utilizador escolher os filtros a aplicar e clicar em Select (no caso de não selecionar nenhum, um *click* em Select apresenta ICUcomplete completa) o output dos mesmos é apresentado em mainPanel. Em mainPanel existem três divisões: sumário dos dados, tabela e gráficos. Uma *box* apresenta o output de summary(), tendo ICUcomplete filtrada como input. Abaixo desta, outra *box* contém um DT dataTable contendo o dataframe ICUcomplete filtrado. Sob estas 2 boxes anteriores são apresentados 4 gráficos: *boxplot* de LOS, histograma de idades, gráficos circulares de primeira unidade

e de última unidade de ICU. Cada um dos gráficos recebe a informação filtrada tal como no menu Patients Info.

```
output$icutable <- DT::renderDataTable({
  DT::datatable(fully_filtered2(),options = list(scrollX = TRUE) )
})

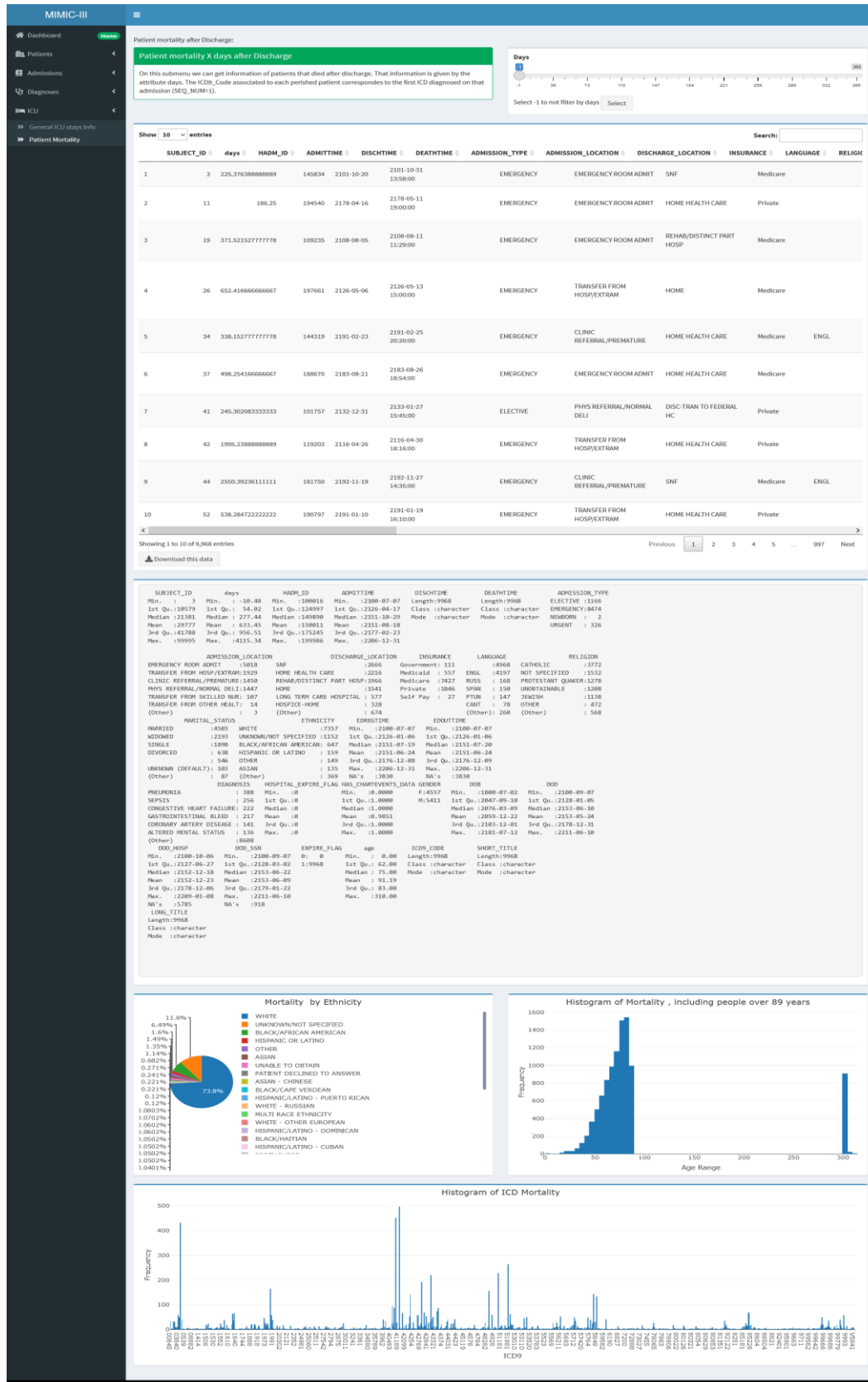
output$summary3 <- renderPrint({
  summary(fully_filtered2())
})

output$icugrafico <- renderPlotly({
  plot_ly(
    data = fully_filtered2(),
    y = ~LOS,
    type = "box",
    name = "Length of stay"
  ) %>%
  layout(title= "BoxPlot of LOS selected"
  )
})
```

4.3.4.5.2 Patient Mortality

Este submenu fornece informação sobre pacientes que não morreram no hospital, mas que morreram após receber alta. Para este submenu foi criado um *dataframe*, *days_before_death*, que regista todos os pacientes que morreram após receber alta.

Este *dataframe* resulta de um *filter()*, onde são seleccionados todos os pacientes que não morreram no hospital, mas que morreram após sair do hospital. A este *dataframe* depois é adicionado uma nova coluna “days”, onde é calculada a diferença entre a data de morte e a data de alta hospitalar. De seguida é feito um *left join* desse mesmo *dataframe* com um *dataframe* contendo todas as admissões e respetivos códigos ICD cujo *SEQ_NUM* é igual a 1. Isto permite obter o código ICD da doença principal que levou cada um dos pacientes ao hospital, e mais tarde a falecer (Figura 45).



```
days_before_death <- filter(df, EXPIRE_FLAG == "1" &
HOSPITAL_EXPIRE_FLAG == "0")

days_before_death$days<- difftime( days_before_death$DOD,
days_before_death$DISCHTIME , units = c("days"))

days_before_death <- left_join(days_before_death,
firstseq_num[,c("ICD9_CODE","SHORT_TITLE","LONG_TITLE","HADM_ID")
], by = "HADM_ID")
```

Este submenu está organizado numa estrutura fluidRow(), onde o utilizador pode seleccionar um range do número de dias entre a data de alta e a data de morte numa barra de seleção de input horizontal. Após clicar no botão Select, o dataframe days_before_death é filtrado e é dado o output em 3 formatos: tabela contendo todos os pacientes abrangidos pelo filtro, summary desses mesmos pacientes e gráficos que ajudam o utilizador a conhecer melhor esses doentes (Figura 45).

Além do mais o paciente pode fazer o download do dataframe filtrado para a sua máquina clicando em Download, sendo feito o download de um ficheiro em formato CSV.

Os gráficos (Figura 46) contidos neste submenu são os seguintes: mortalidade por etnia num gráfico circular, histograma de mortalidade por idade e histograma de mortalidade por ICD.

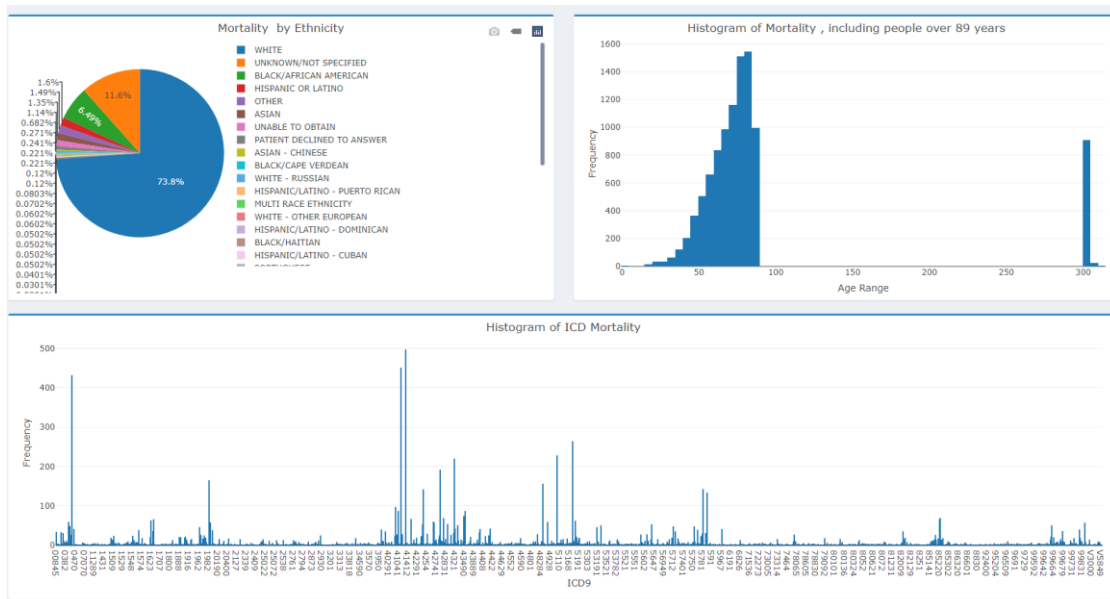


Figura 46 - Gráficos do submenu Patient Mortality

5. Resultados

Neste capítulo serão apresentados e detalhados os tipos de resultados possíveis com esta ferramenta, de que forma pode ser usada para facilitar e servir de base para investigação sobre a MIMIC-III e ainda o porquê de ser uma ferramenta que auxilia um utilizador da MIMIC-III a mais rapidamente ganhar conhecimento da base de dados.

A criação desta ferramenta como um R dashboard tem dois objetivos primordiais: o primeiro foi criar uma ferramenta para o utilizador da MIMIC-III ficar a conhecer melhor a base de dados sem necessitar de conhecimentos de programação e o segundo foi criar uma ferramenta que pudesse servir de base para estudos clínicos na MIMIC-III. Com isto em mente a aplicação/dashboard teve de ser desenhada de forma a manter uma interface simples de utilizar, mas ao mesmo tempo com funcionalidades que fossem além da mera apresentação de dados estatísticos.

5.1 Sistema de visualização simples, direto, objetivo

No momento de desenhar esta ferramenta, para além das funções e funcionalidades da mesma, foi necessário ponderar que tipo de interface faria mais sentido para uma aplicação deste género, e como poderia ser implementada uma ferramenta que fosse visualmente mais apelativa que as já existentes. Um dos pormenores pensados foi a paleta de cores escolhida para a ferramenta, a utilização de tons de azul-claro, azul-escuro e verde foi propositada, visto serem cores comumente presentes em ambientes hospitalares por serem associadas a tranquilidade e bem-estar (Coad, 2008). Além do mais, cada elemento inserido numa *box* apresenta cores diferentes consoante o seu tipo. Este padrão é mantido ao longo de toda a ferramenta, mas pode, por exemplo, ser observado na Figura 21 a amarelo são dados estatísticos, a verde texto informativo e a azul boxes que contém gráficos ou tabelas. Isto foi pensando de forma a aumentar a agradabilidade do uso da ferramenta.

Com o objetivo de manter a interface o menos poluída possível, esta é dividida em submenus escondidos numa barra lateral, colocados no interior de 5 menus principais (Figura 47), evitando assim apresentar uma interface gráfica desnecessariamente complexa. Esta escolha de design permite também aceder a qualquer menu do dashboard

sem necessitar de passos intermédios, como por exemplo, navegar até à homepage e seleccionar outro menu, agilizando a navegação e diminuindo o número de *clicks* necessários para aceder uma determinada informação. Para isto contribui ainda o facto de o progresso em cada página do menu ser guardado, ou seja, é possível “saltar” entre vários menus livremente sem nunca perder os dados de uma pesquisa ou ser necessário reinserir um input.

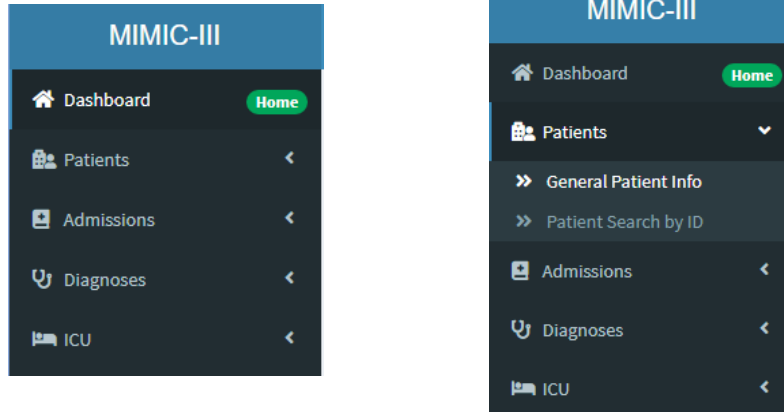


Figura 47- Menu Lateral, organizado em menus e submenus

Para mais, a possibilidade de esconder o menu foi pensada e aproveitada para que em certos casos seja possível visualizar melhor a informação apresentada, aumentando o espaço horizontal da página, o que pode ser útil consoante o ecrã onde a ferramenta esteja a ser visualizada. Um exemplo de uma página que beneficia em larga escala desta funcionalidade é a presente submenu “General Diagnoses Info”, do menu Diagnoses (Figura 48 e Figura 49).

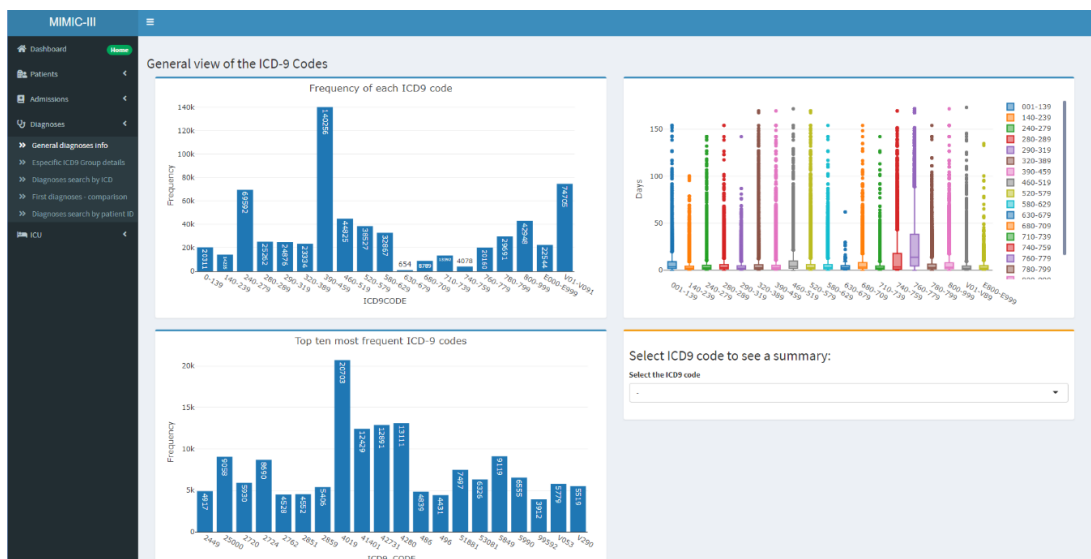


Figura 48- Menu lateral exposto



Figura 49- Menu lateral oculto

Atente-se especialmente ao gráfico onde são comparados *boxplots* referentes a LOS, com a barra lateral oculta é consideravelmente mais prático visualizar o gráfico. Este trabalho foi desenvolvido em computadores com um ecrã 15.6", no entanto, em computadores mais pequenos esta função é ainda mais útil.

Ademais, todas as páginas da ferramenta seguem uma linha de *design* em comum. A grande maioria das componentes da página são contidas em boxes, sendo os limites das mesmas visíveis de forma a tornar mais fácil entender de que forma está organizada a página e sempre que é requerida a inserção de input por parte do utilizador, esta é colocada no canto esquerdo da página de forma que não haja diferenças grandes de *design* de página para página e assim seja mantida uma coerência ao longo de todo o *dashboard*. (Figura 50, Figura 51, Figura 52)

Filter to find general info about patients

Choose a gender: ALL

Choose an ethnicity: ALL

Choose a religion: ALL

Choose an insurance type: ALL

Choose an admission type: ALL

Choose an admission location: ALL

Choose a discharge location: ALL

Choose if alive(1) or dead(0) [Hospital expire flag]: ALL

Choose if alive(1) or dead(0) [Expire flag]: ALL

Choose the first diagnosis: ALL

Age: 300

SUBJECT_ID	age	GENDER	ETHNICITY	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE	LANGUAGE	RELIGION	MARITAL_STATUS
1210-05-10 16:53:00	2	M	ASIAN	163353	2138-07-17 15:48:00	2138-07-21 15:48:00		NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME	Private	NOT SPECIFIED	NOT SPECIFIED	

Figura 50- Submenu General Patients Info, localização do menu lateral de input na esquerda

Search by patient ID

Choose or type patient ID: 2

Patient info

Show 10 entries

SUBJECT_ID	age	GENDER	ETHNICITY	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE	LANGUAGE	RELIGION
1	2	M	ASIAN	163353	2138-07-17 15:48:00	2138-07-21 15:48:00		NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME	Private	NOT SPECIFIED	NOT SPECIFIED

Showing 1 to 1 of 1 entries

All admissions of patient

Show 10 entries

SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE	LANGUAGE	RELIGION	MARITAL_STATUS	ETHNICITY
1	2	163353	2138-07-17 15:48:00	2138-07-21 15:48:00	NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME	Private	NOT SPECIFIED	NOT SPECIFIED		ASIAN

Showing 1 to 1 of 1 entries

Figura 51- Patient Search by ID, menu de input na esquerda

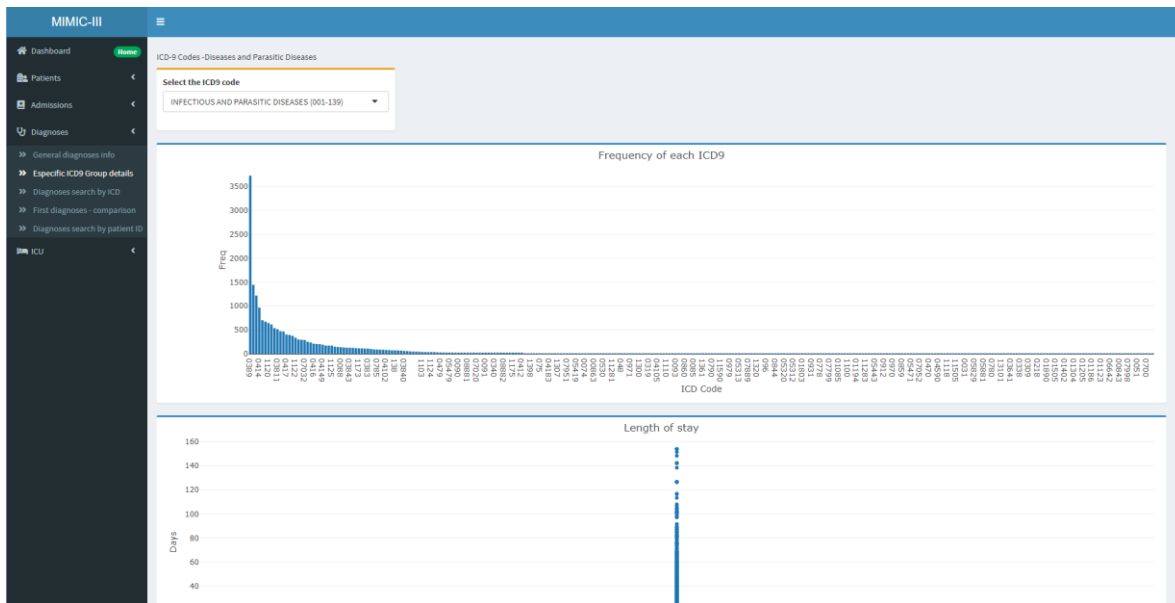


Figura 52- Submenu Specific ICD9 group details, menu de input na esquerda

Além disso, manter a coerência no posicionamento das caixas de input ao longo da aplicação permite a utilização mais rápida da ferramenta, pois elimina a necessidade do utilizador, em cada menu diferente, precisar de as procurar.

5.2 Tipos de visualizações possíveis

No capítulo 4 foram detalhados os menus existentes: a forma como foram implementados e que dados são apresentados. Neste capítulo pretendemos aprofundar em mais detalhe o tipo de informação que é possível extrair dos dados que são apresentados e que conclusões e possíveis ilações se podem retirar. Além disso, serão dados exemplos de utilização de alguns menus desta ferramenta.

Por questões de organização vamos dividir este capítulo em subcapítulos.

5.2.1 Dashboard

Esta página inicial é o primeiro contacto que o utilizador tem com a ferramenta, assim é feito não só um texto instrutório à mesma, mas também é dado a conhecer informações sobre a MIMIC-III que permitem melhor entender o tipo de pacientes representados na base de dados.

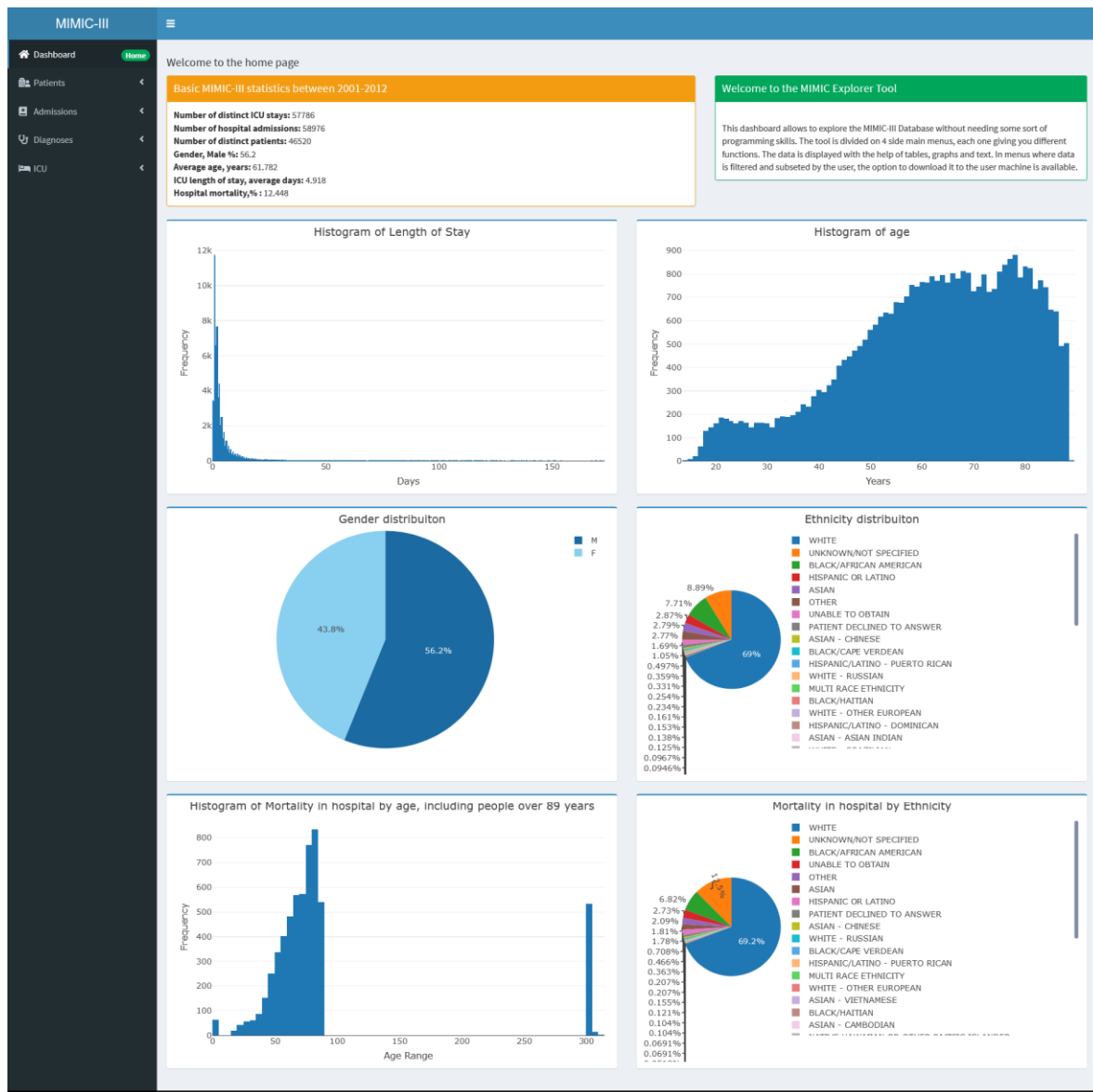


Figura 53 -Dashboard, página inicial

Na box a amarelo (Figura 53) é dada informação resumida da tabela, de forma que o utilizador rapidamente possa conferir números relevantes da MIMIC-III e entender melhor o

tamanho da base de dados que está a ser trabalhada. São exemplos: o número total de pacientes, a idade média dos mesmos, o número de hospitalizações ou até mesmo a taxa de mortalidade. Aqui o objetivo não é entrar em detalhe, mas sim dar uma visão geral da base de dados.

Para tornar a página inicial mais apelativa e ao mesmo tempo fornecer informação relevante, são apresentados dois gráficos, um dando informação relativa ao LOS e outro à idade. Estes gráficos são interativos, passando o ponteiro do rato sobre eles é mostrada mais informação, por exemplo na Figura 54, passando o rato sobre cada barra do histograma, é dada a idade e frequência de pacientes com essa mesma idade.

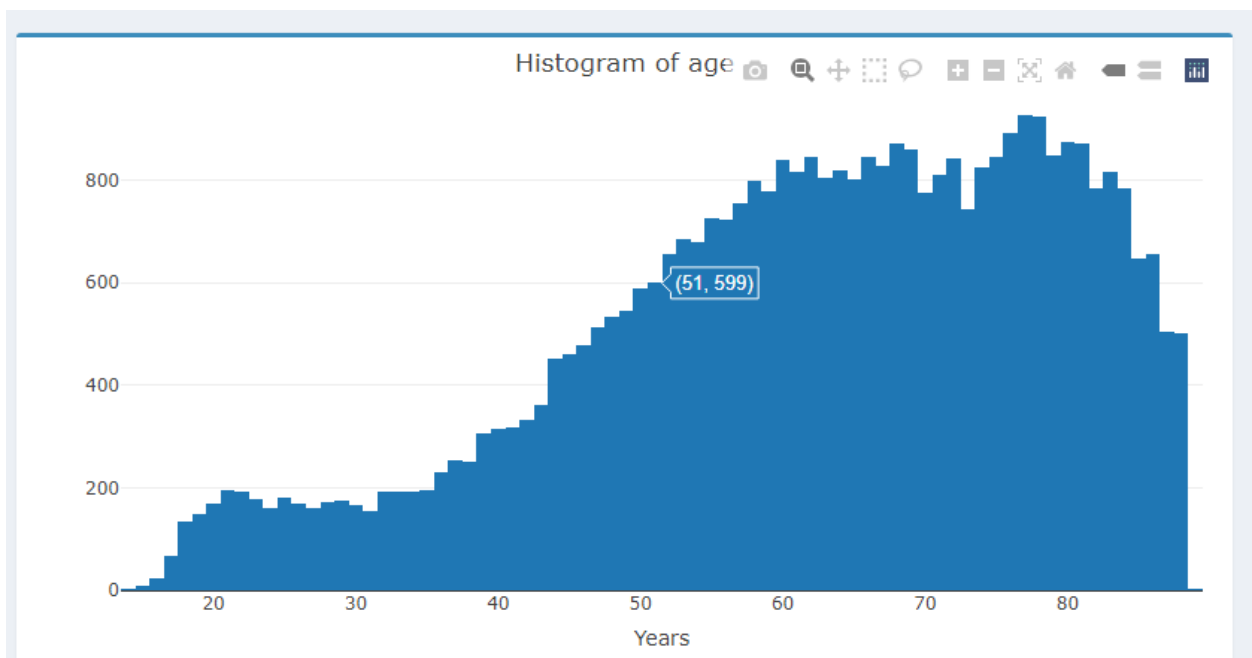


Figura 54- Histograma de idades, mouse-over na idade 51 anos

O mesmo se passa no histograma relativo a LOS (Figura 55), é possível fazer zoom, por exemplo, focar nos LOS de menores dias (Figura 56)

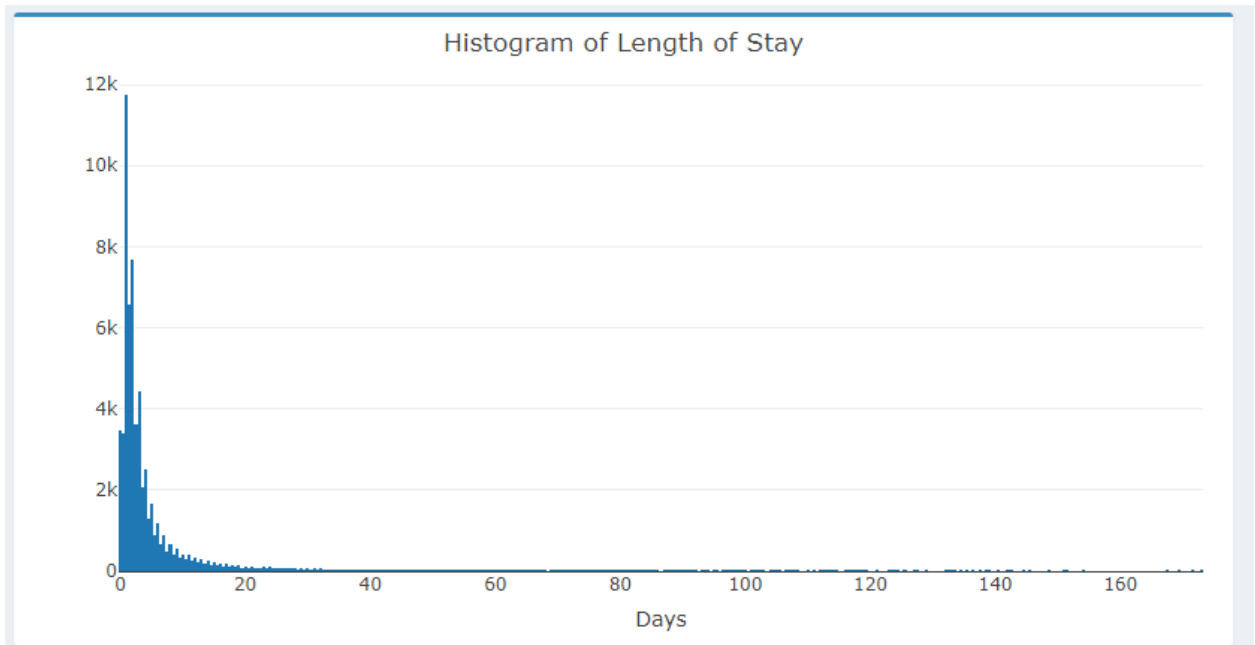


Figura 55- Histograma de LOS

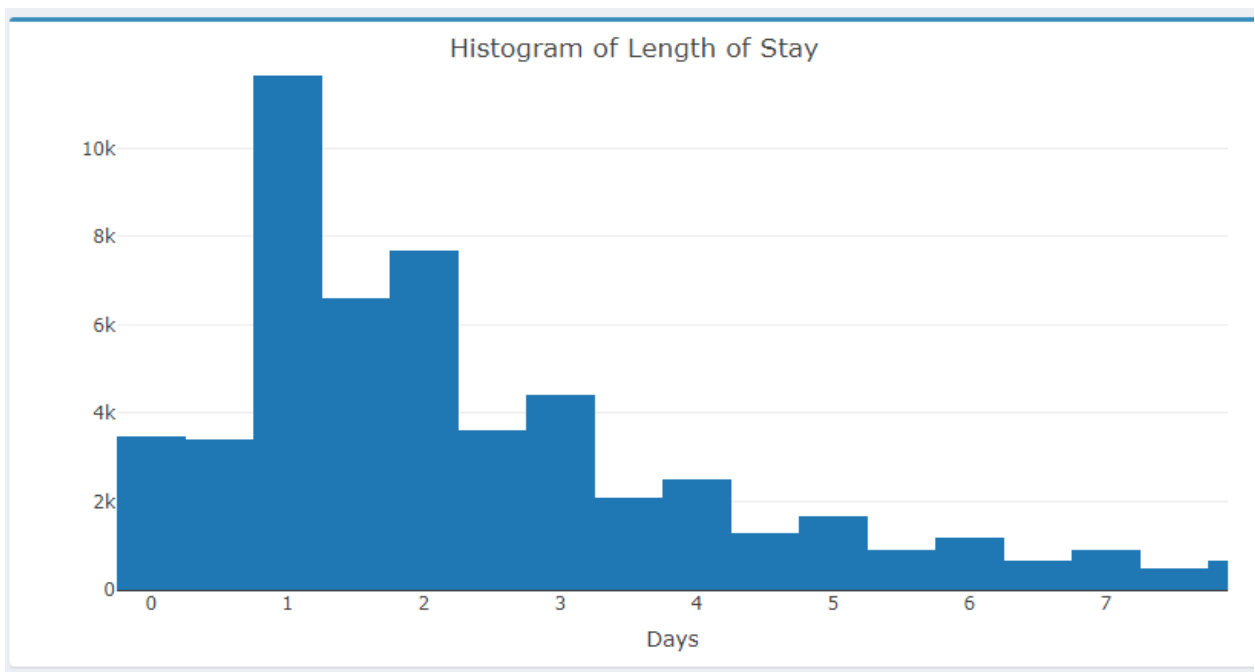


Figura 56-Zoom in no histograma de LOS

E passando o indicador do rato sobre a barra são apresentadas informações que ajudam a visualização do gráfico (Figura 57).

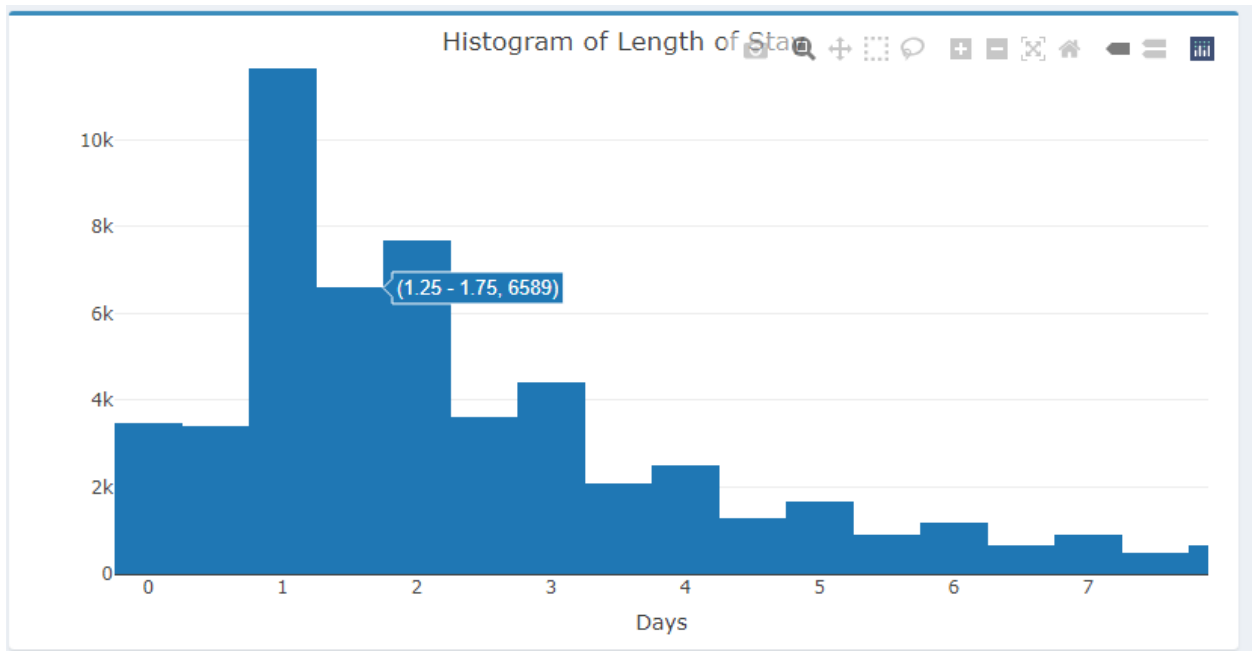


Figura 57- Histograma de LOS, mouse-over

No intervalo de 1.25 a 1.75 dias existem 6589 pacientes (Figura 57).

5.2.2 Patients

Este menu contém submenus que proveem informação importante e relevante, no entanto, o submenu que mais se destaca é General Patients Info. Neste menu é possível filtrar os pacientes por diversos filtros. Manipulando-os é possível conhecer com exatidão uma série de informações. Por exemplo:

Suponhamos que queremos saber qual é a média de idades de todos os pacientes brancos, do sexo masculino e sem seguro (“Self-pay”) (Figura 58).

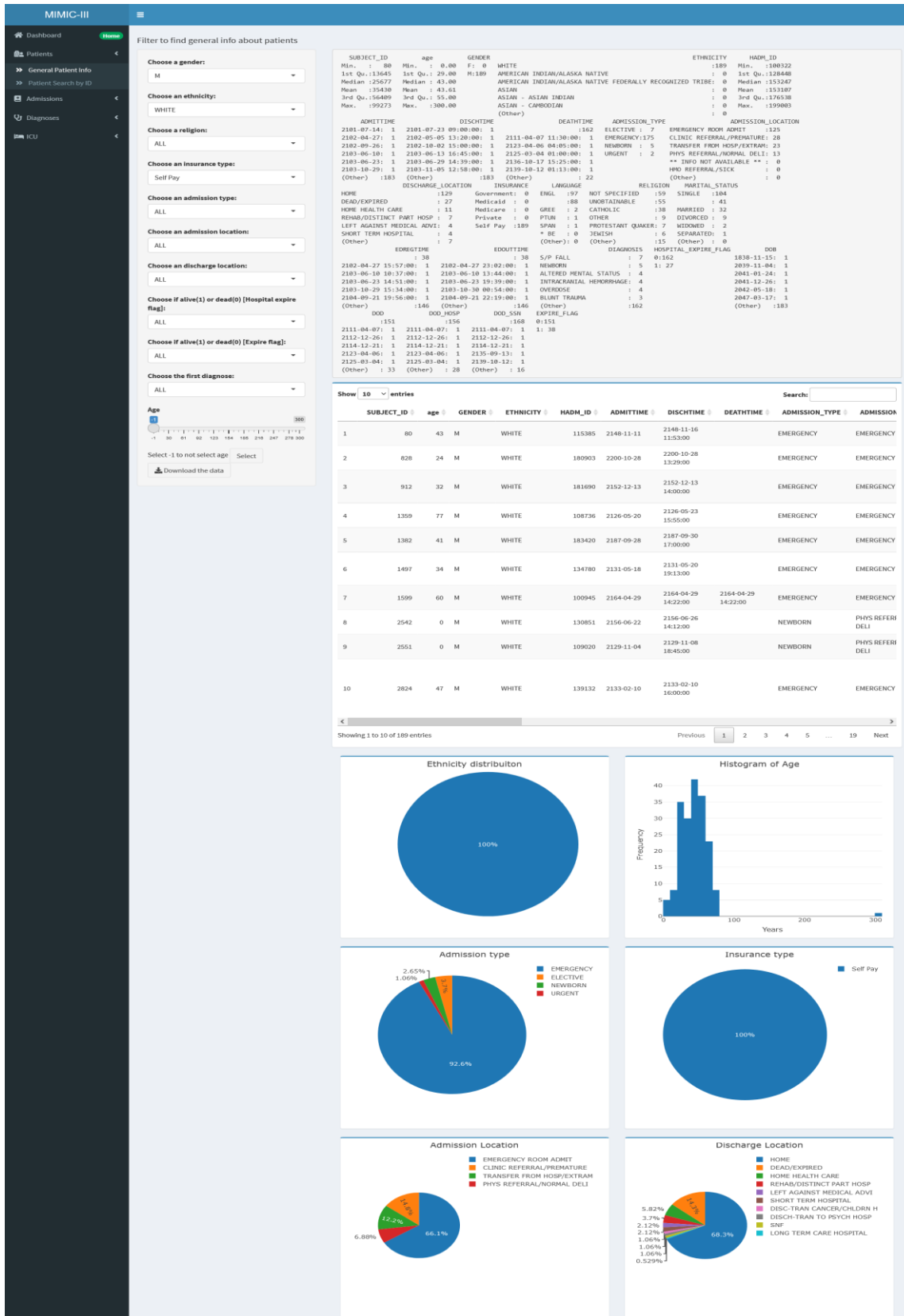


Figura 58- Submenu general patient info, seleção de género Masculino, etnia branca e seguro "selfpay"

Como podemos ver na Figura 58, podemos obter vários tipos de informações relevantes sobre este grupo de pacientes:

- Existem 200 pacientes com estas características;
- A idade média é de 44.62 anos, sendo a mediana das idades 43 anos. A idade mínima é 0 sendo a idade máxima dos pacientes selecionados 300 anos (indicando que existem pacientes com mais de 89 anos neste grupo);
- Dos 200 pacientes, 185 entraram no hospital como urgência, ou seja, 92.5%;
- Após receber alta, 66.5% dos pacientes foram para casa, enquanto 14% faleceram;
- 14% destes pacientes faleceram no hospital;
- 21,5% dos pacientes faleceram;

Para além destes dados é ainda possível saber, com exatidão, que pacientes estão neste grupo, pois todos os pacientes contidos no mesmo são apresentados numa tabela (Figura 59).

Show 10 entries

Search:

SUBJECT_ID	age	GENDER	ETHNICITY	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOC
1	80	43 M	WHITE	115385	2148-11-11	2148-11-16 11:53:00		EMERGENCY	EMERGENCY ROOM
2	828	24 M	WHITE	180903	2200-10-28	2200-10-28 13:29:00		EMERGENCY	EMERGENCY ROOM
3	912	32 M	WHITE	181890	2152-12-13	2152-12-13 14:00:00		EMERGENCY	EMERGENCY ROOM
4	1359	77 M	WHITE	108736	2126-05-20	2126-05-23 15:55:00		EMERGENCY	EMERGENCY ROOM
5	1382	41 M	WHITE	183420	2187-09-28	2187-09-30 17:00:00		EMERGENCY	EMERGENCY ROOM
6	1497	34 M	WHITE	134780	2131-05-18	2131-05-20 19:13:00		EMERGENCY	EMERGENCY ROOM
7	1599	60 M	WHITE	100945	2164-04-29	2164-04-29 14:22:00	2164-04-29 14:22:00	EMERGENCY	EMERGENCY ROOM
8	2542	0 M	WHITE	130851	2156-06-22	2156-06-26 14:12:00		NEWBORN	PHYS REFERRAL/N DELI
9	2551	0 M	WHITE	109020	2129-11-04	2129-11-08 18:45:00		NEWBORN	PHYS REFERRAL/N DELI
10	2824	47 M	WHITE	139132	2133-02-10	2133-02-10 16:00:00		EMERGENCY	EMERGENCY ROOM

Figura 59- Tabela dos pacientes incluídos no filtro

É possível percorrer a tabela utilizando a barra horizontal para mostrar outras informações e até mesmo utilizar a barra de pesquisa para estreitar a nossa busca. Por exemplo: dos

pacientes apresentados queremos apenas ver os que deram entrada no ano (com offset) 2187 (Figura 61). Sem alterar os campos de input anteriores a Figura 60 mostra como isto é possível, utilizando o campo *search* da própria tabela.

The screenshot shows a web interface for querying MIMIC-III data. On the left, there are filters for 'Choose if alive(1) or dead(0) [Hospital expire flag]: ALL', 'Choose if alive(1) or dead(0) [Expire flag]: ALL', 'Choose the first diagnosis: ALL', and an 'Age' slider set to 1. Below these is a 'Download the data' button. The main area displays a table with 10 columns: SUBJECT_ID, age, GENDER, ETHNICITY, HADM_ID, ADMITTIME, DISCHTIME, DEATHTIME, ADMISSION_TYPE, ADMISSION_LOCATION, and DISCHARGE. The table contains 10 rows of patient data. A search bar at the top right of the table is empty.

SUBJECT_ID	age	GENDER	ETHNICITY	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE
1	80	43	M	WHITE	115385	2148-11-11 11:53:00	2148-11-16 11:53:00	EMERGENCY	EMERGENCY ROOM ADMIT	HOME
2	828	24	M	WHITE	180903	2200-10-28 13:29:00	2200-10-28 13:29:00	EMERGENCY	EMERGENCY ROOM ADMIT	HOME
3	912	32	M	WHITE	181690	2152-12-13 14:00:00	2152-12-13 14:00:00	EMERGENCY	EMERGENCY ROOM ADMIT	HOME
4	1359	77	M	WHITE	108736	2126-05-20 15:55:00	2126-05-23 15:55:00	EMERGENCY	EMERGENCY ROOM ADMIT	HOME
5	1382	41	M	WHITE	183420	2187-09-28 17:00:00	2187-09-30 17:00:00	EMERGENCY	EMERGENCY ROOM ADMIT	HOME
6	1497	34	M	WHITE	134780	2131-05-18 19:13:00	2131-05-20 19:13:00	EMERGENCY	EMERGENCY ROOM ADMIT	HOME
7	1599	60	M	WHITE	100945	2164-04-29 14:22:00	2164-04-29 14:22:00	EMERGENCY	EMERGENCY ROOM ADMIT	DEAD/EXPI
8	2542	0	M	WHITE	130851	2156-06-22 14:12:00	2156-06-28 14:12:00	NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME
9	2551	0	M	WHITE	109020	2129-11-04 18:45:00	2129-11-08 18:45:00	NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME
10	2624	47	M	WHITE	139132	2133-02-10 16:00:00	2133-02-10 16:00:00	EMERGENCY	EMERGENCY ROOM ADMIT	HOME

Figura 60- Pesquisa na tabela pelo ano 2187. Download destes dados é possível.

The screenshot shows the same interface as Figure 60, but with the search bar set to '2187'. The table now displays 4 entries filtered from 200 total entries. The columns are: RACE, LANGUAGE, RELIGION, MARITAL_STATUS, EDRETIME, EDOUTTIME, DIAGNOSIS, HOSPITAL_EXPIRE_FLAG, DOB, and DOD. The table contains 4 rows of patient data.

RACE	LANGUAGE	RELIGION	MARITAL_STATUS	EDRETIME	EDOUTTIME	DIAGNOSIS	HOSPITAL_EXPIRE_FLAG	DOB	DOD
ay			SEPARATED	2187-09-28 16:53:00	2187-09-28 22:13:00	OVERDOSE	0	2145-10-28	
ay				2187-11-26 09:36:00	2187-11-26 10:20:00	ACUTE MYOCARDIAL INFARCTION;S/P VFIB ARREST	0	2148-09-30	
ay	PTUN	CATHOLIC	DIVORCED			ST ELEVATION	0	2133-11-01	
ay	ENGL	CATHOLIC	MARRIED			ABDOMINAL AORTIC ANEURYSM	0	2126-01-19	

Showing 1 to 4 of 4 entries (filtered from 200 total entries)

Figura 61- Pesquisa na tabela pelo ano 2187.

Podemos verificar que dos 200 pacientes selecionados, 4 deles deram entrada neste ano, tendo nenhum deles falecido. Podemos fazer download destes dados clicando em “Download the data” (Figura 60).

Este menu também suporta a pesquisa por SUBJECT_ID, no submenu Patient Search by ID. Se o utilizador assim desejar, basta inserir o seu ID para ficar a saber mais sobre o paciente. No exemplo da Figura 62 , é inserido o ID do paciente 1382.

The screenshot shows the MIMIC-III web interface. On the left is a dark sidebar with navigation links: Dashboard (Home), Patients, General Patient Info, Patient Search by ID, Admissions, Diagnoses, and ICU. The main area is titled 'Search by patient ID' and has a search box containing '1382'. Below this, there are two sections:

Patient info

	SUBJECT_ID	age	GENDER	ETHNICITY	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANC
1	1382	41	M	WHITE	183420	2187-09-28	2187-09-30 17:00:00		EMERGENCY	EMERGENCY ROOM ADMIT	HOME	Self Pay

Showing 1 to 1 of 1 entries

All admissions of patient

	SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE	LANGUAGE	RELIGION	I
1	1382	183420	2187-09-28 19:22:00	2187-09-30 17:00:00		EMERGENCY	EMERGENCY ROOM ADMIT	HOME	Self Pay			SI

Showing 1 to 1 of 1 entries

Figura 62- Pesquisa pelo SUBJECT_ID 1382 no submenu Patient Search y ID

A informação apresentada é a mesma da Figura 62, mas apenas para esse doente. Durante o desenvolvimento desta tarefa foi ponderada a integração desta função no submenu General Patient Info. Esta ideia foi descartada pois o objetivo desse submenu é permitir ao utilizador obter dados sobre um conjunto de pacientes, fazendo uma pesquisa granular sobre todo o conjunto de pacientes da MIMIC-III. Para pesquisas singulares foi escolhida a criação de um submenu só para essa tarefa, de forma a tornar a interface mais rápida de utilizar.

5.2.3 ADMISSION

Neste menu, como explicado no capítulo 4, existem 2 submenus: General Admissions Info e Search by Admission ID. O segundo tem um funcionamento simples e uma só função. Dado um HADM_ID é retribuído uma tabela com dados sobre essa mesma admissão. Este submenu foi criado numa fase posterior do desenvolvimento, quando foi notada a necessidade de existir um local onde se pudesse pesquisar pelo ID de uma admissão. Um exemplo da sua utilização está presente na Figura 63.

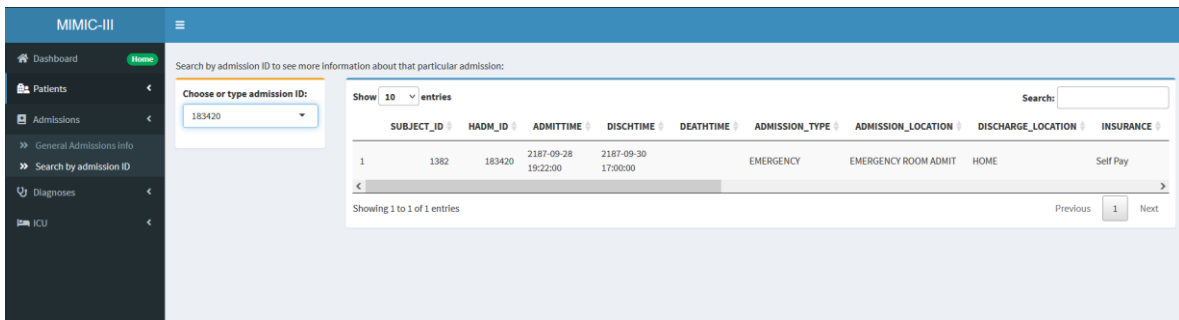


Figura 63- Pesquisa pela admissão HADM_ID 183420

O segundo submenu, no entanto, tem como objetivo oferecer uma visão estatística sobre as admissões neste hospital. O seu foco é inserido sobre 5 atributos das admissões: tipo de admissão, localização de admissão, local de alta após admissão, tipo de seguro do paciente na sua admissão e diagnóstico preliminar aquando da entrada no hospital. (Figura 64)



Figura 64- 5 separadores do submenu general admissions info

Inicialmente apenas existia uma tabela para cada separador, para apresentar os diferentes dados, no entanto, para melhor legibilidade e mais rápida perceção dos dados para o utilizador, foi criado um gráfico que os representa (Figura 65).



Figura 65- Submenu General Admissions Info, separador Admissions Type

Mais uma vez, estes gráficos são interativos, sendo possível fazer zoom e obter legendas para cada uma das barras do histograma (Figura 66)

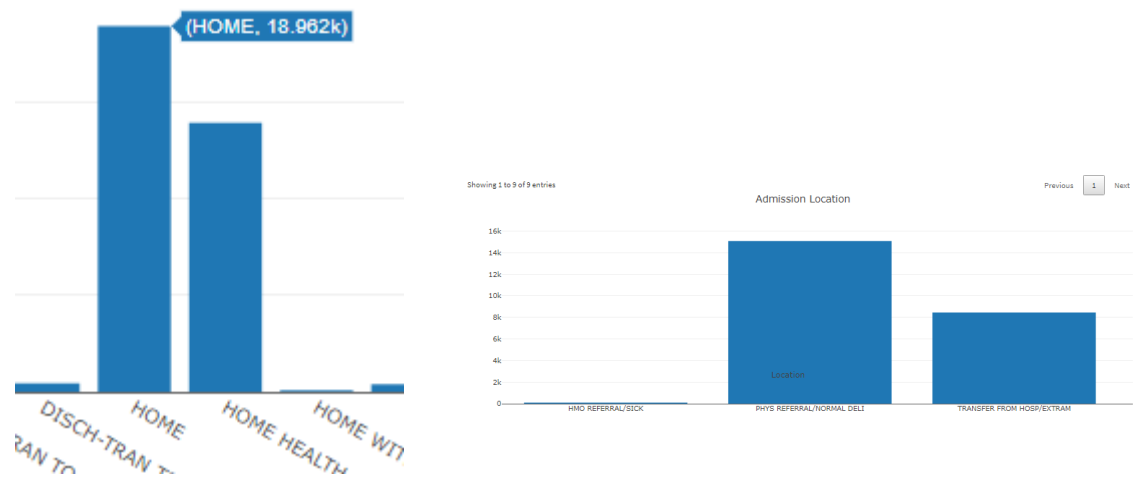


Figura 66- Mouse-over sob barras do histograma apresenta informação

5.2.4 DIAGNOSES

Um dos menus mais ricos da ferramenta é Diagnoses. Este contém submenus que permitem não só fazer pesquisas singulares (submenu Search by ICD ou Diagnoses Search by patient ID), conhecer dados estatísticos (submenu Specific ICD-9 code) e ainda comparar atributos como LOS de diferentes doenças (General Diagnoses Info e First Diagnoses).

O submenu principal de Diagnoses, General Diagnoses Info, permite rapidamente perceber qual o grupo de ICD mais frequente, através do histograma da Figura 67. Facilmente o utilizador pode verificar que as doenças com código ICD 390-459 (Diseases of the circulatory system) são as mais frequentes, seguidas de V01-V091 (Supplementary classification of factors influencing health status and contact with health services) e 240-279 (Endocrine, nutritional and metabolic diseases, and immunity disorders) (Figura 67)

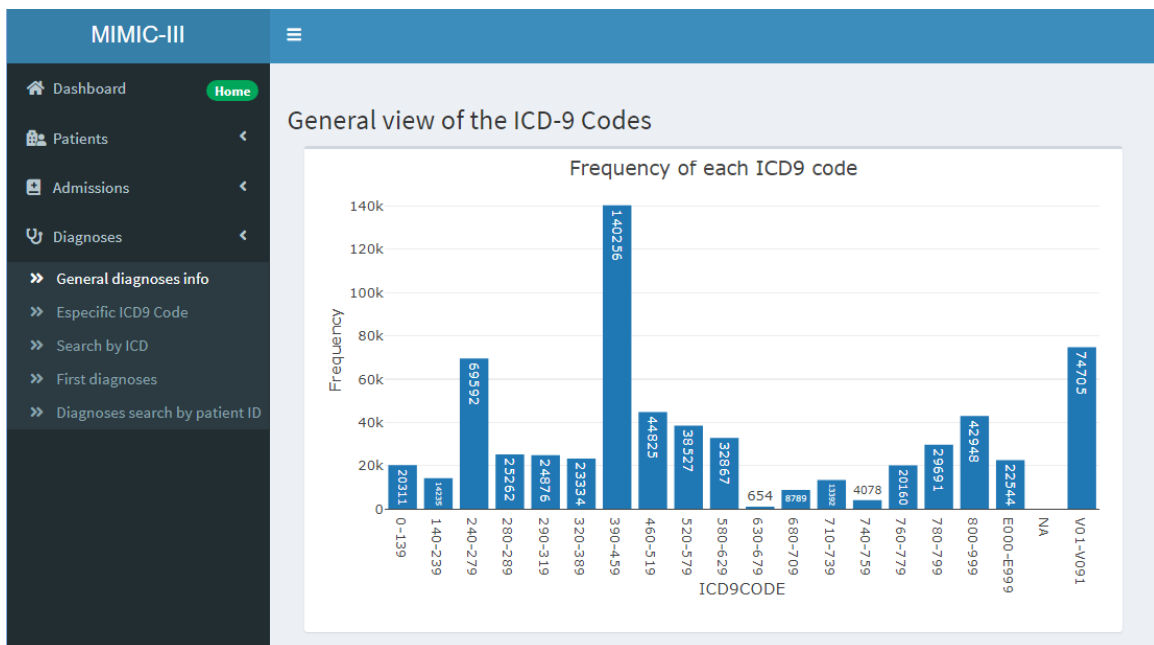


Figura 67- Histograma da frequência dos grupos de ICD

Ao fazer mouse-over sob as barras é possível obter mais informação (Figura 68).

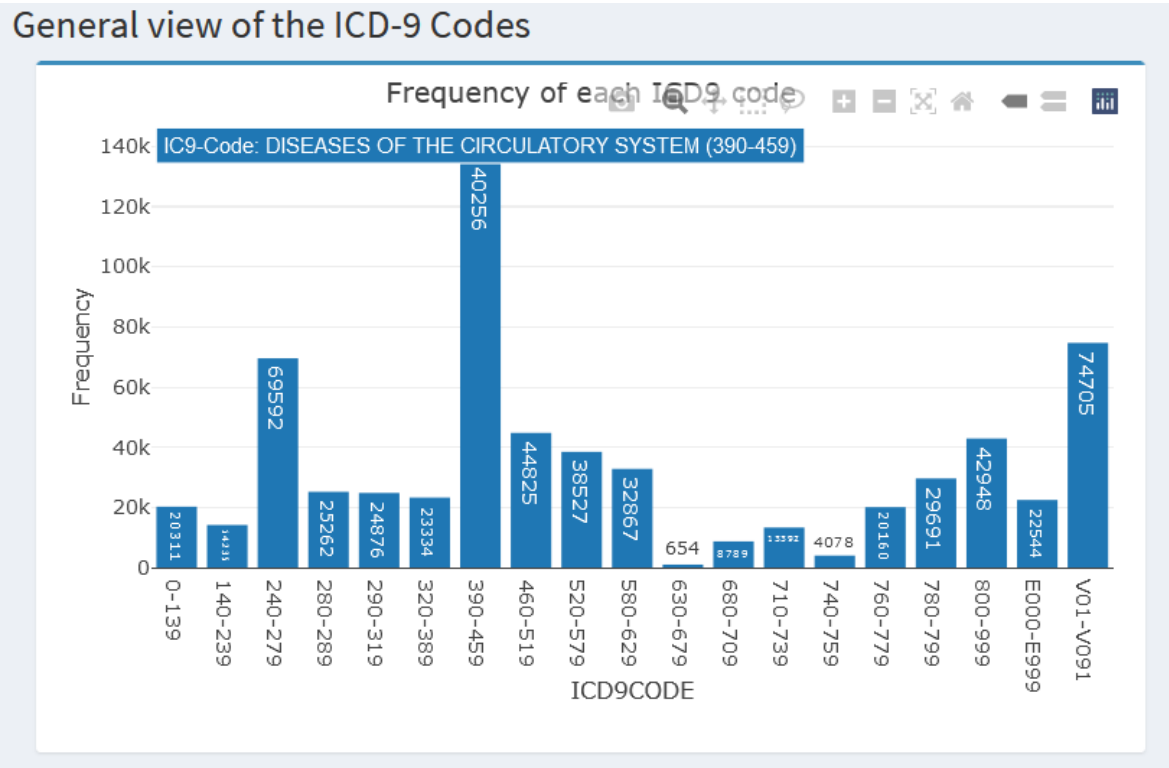


Figura 68- Mouse-over sobre a barra 390-459 do histograma

Ao lado deste histograma é possível rapidamente comparar os LOS, (Figura 69), de cada um dos diferentes grupos de códigos ICD, através de diferentes boxplots. Cada um tem uma cor diferente para facilitar a sua distinção. Passando o rato sobre cada um deles é possível ver em maior detalhe os valores de cada um. É ainda possível fazer zoom para assim mais facilmente comparar valores, Figura 69, 70 e 71.

Podemos observar pelas Figura 70 e Figura 71, que a mediana de LOS do grupo 001-139 (Infectious And Parasitic Diseases) é superior ao grupo 140-239 (Neoplasms). Em média, os pacientes passam mais tempo no hospital com o primeiro grupo de deonças.

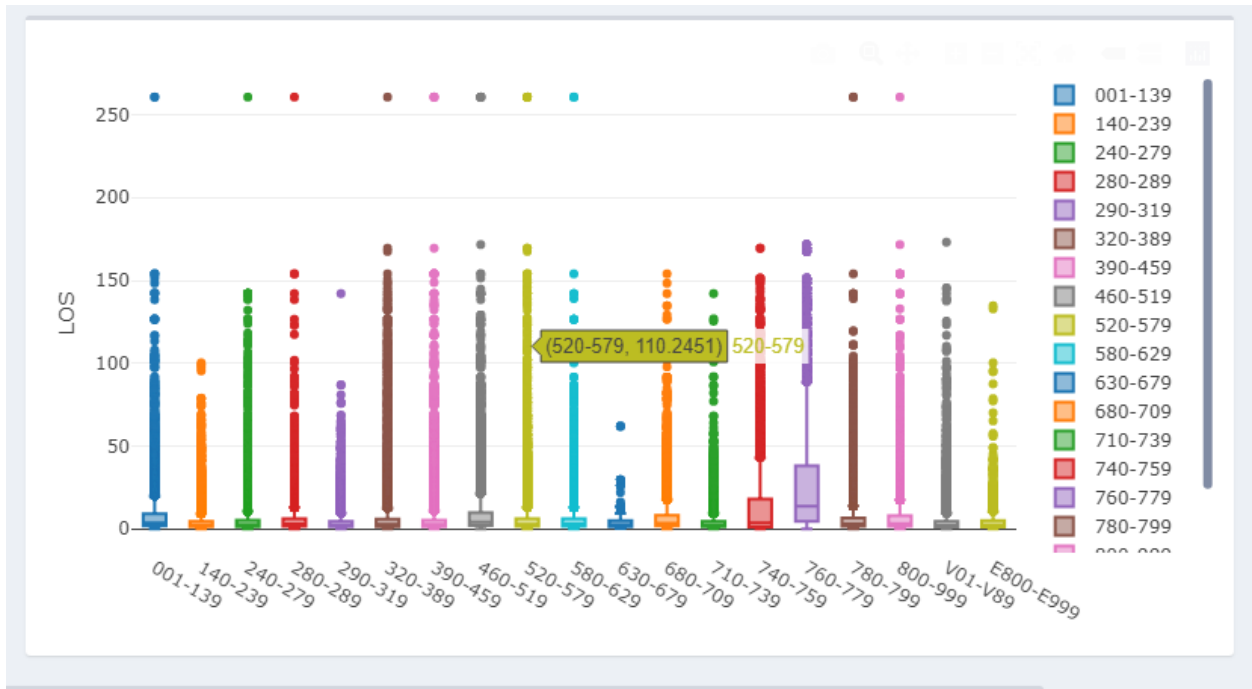


Figura 69- Comparação de boxplots do gráfico com mouse-over sob a barra 520-579

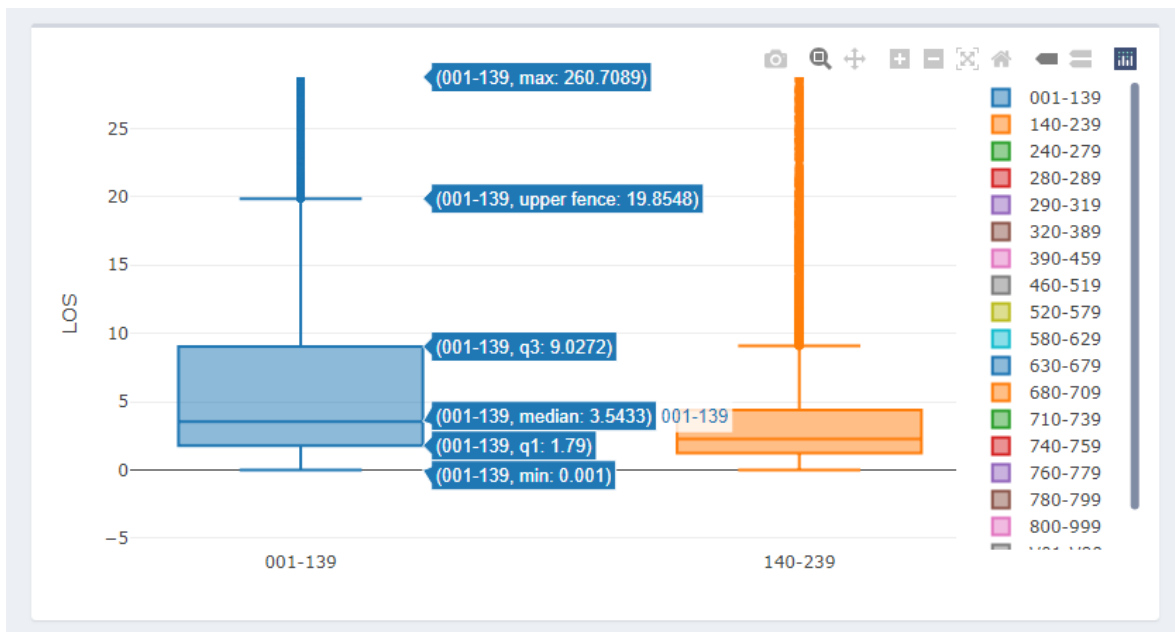


Figura 70- Zoom com mouse-over na barra 001-139

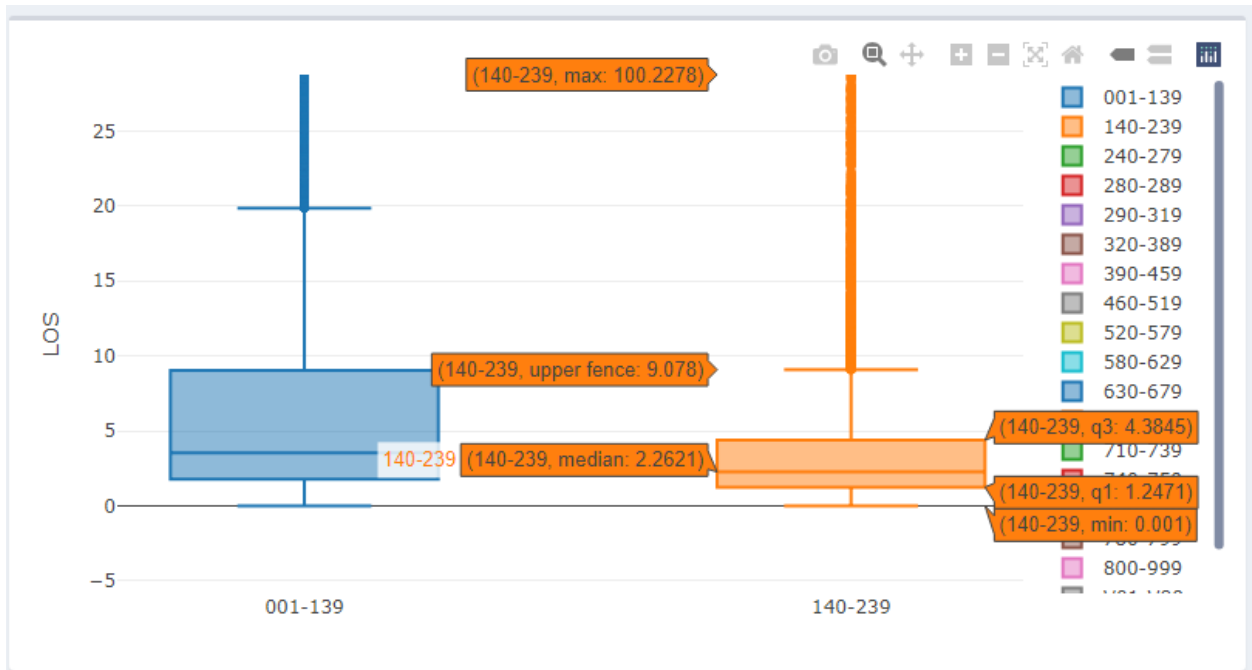


Figura 71- Zoom com mouse-over na barra 140-239

Se o utilizador assim desejar, pode ainda ver um sumário com informação sobre cada um dos grupos de ICD, selecionando na caixa de input o grupo sobre qual pretende obter mais informação (Figura 72)

Select ICD9 code to see a summary:

Select the ICD9 code

INFECTIOUS AND PARASITIC DISEASES (001-139)

NEOPLASMS (140-239)

ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS (240-279)

DISEASES OF THE BLOOD AND BLOOD-FORMING ORGANS (280-289)

MENTAL DISORDERS (290-319)

DISEASES OF THE NERVOUS SYSTEM AND SENSE ORGANS (320-389)

DISEASES OF THE CIRCULATORY SYSTEM (390-459)

Figura 72- Seleção de input no submenu general diagnoses info



- Uma média de SEQ_NUM superior em doenças circulatórias 6.894 versus 9.702, indica que doenças do foro mental contém menor prioridade e são menos vezes diagnosticada como causas principais de deslocação ao hospital;
- Em ambas o sexo masculino é mais afetado;

- A idade média dos pacientes diagnosticados com doenças do foro circulatório é superior às doenças do foro mental (83,48 anos vs 66.06 anos)
- Nas doenças do coração, a maleita mais comum é hipertensão, sendo depressão a mais comum nas mentais.

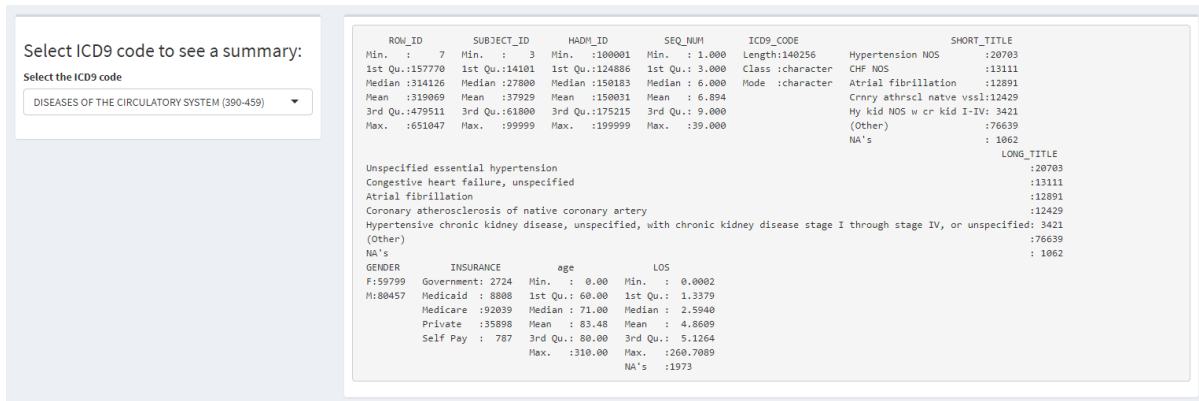


Figura 74- Seleção de doenças do sistema circulatório

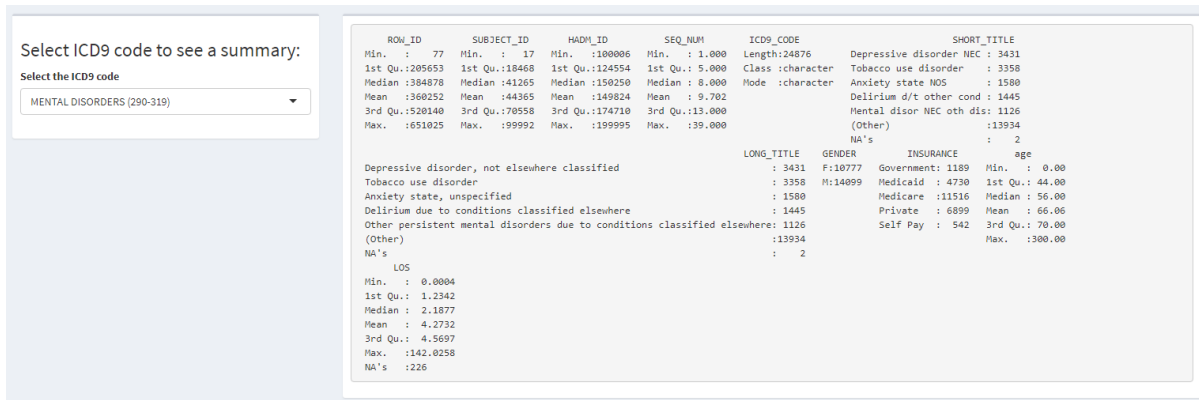


Figura 75- Seleção de doenças mentais

Se o utilizador desejar saber mais detalhes sobre um determinado ICD, e não sobre todo o conjunto deles, terá de utilizar o submenu: Search by ICD. Este menu além de apresentar uma tabela com todos os pacientes diagnosticados com essa doença, também faz um sumário das características dos doentes que a contraíram. Suponhamos que o utilizador pretende saber mais sobre hipertensão, a doença mais comum no campo das doenças

circulatórias, cujo código ICD é 4019. Como se pode ver Figura 76, a ferramenta sugere opções para completar a pesquisa baseado no que á foi inserido.

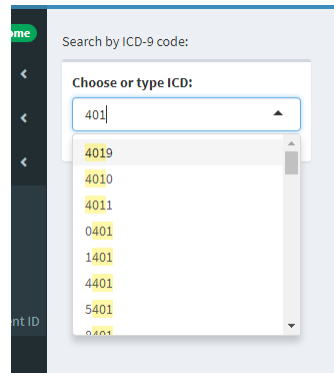


Figura 76- Sugestão de opções durante inserção do código ICD

MIMIC-III

Search by ICD-9 code:

Choose or type ICD: 4019

Show 10 entries

HADM_ID	SUBJECT_ID	SEQ_NUM	ICD9_CODE	SHORT_TITLE	LONG_TITLE	LOS
1	100003	54610	7 4019	Hypertension NOS	Unspecified essential hypertension	1.9425
2	100007	23018	5 4019	Hypertension NOS	Unspecified essential hypertension	4.0988
3	100009	533	12 4019	Hypertension NOS	Unspecified essential hypertension	2.4908
4	100012	60039	4 4019	Hypertension NOS	Unspecified essential hypertension	3.5763
5	100020	9973	15 4019	Hypertension NOS	Unspecified essential hypertension	1.1011
6	100021	29971	23 4019	Hypertension NOS	Unspecified essential hypertension	3.6213
7	100024	77221	7 4019	Hypertension NOS	Unspecified essential hypertension	1.0799
8	100031	6892	6 4019	Hypertension NOS	Unspecified essential hypertension	6.8557
9	100034	51495	4 4019	Hypertension NOS	Unspecified essential hypertension	1.1399
10	100037	58947	15 4019	Hypertension NOS	Unspecified essential hypertension	5.0026

Showing 1 to 10 of 20,703 entries

Previous 1 2 3 4 5 ... 2,071 Next

HADM_ID_x	ROW_ID	SUBJECT_ID	SEQ_NUM	ICD9_CODE	SHORT_TITLE	LONG_TITLE	LOS
Min. : 100003	Min. : 47	Min. : 9	Min. : 1,000	Length: 20703	Length: 20703	Length: 20703	
1st Qu.: 124882	1st Qu.: 157197	1st Qu.: 14046	1st Qu.: 5,000	Class : character	Class : character	Class : character	
Median : 158469	Median : 312419	Median : 27666	Median : 7,000	Mode : character	Mode : character	Mode : character	
Mean : 158053	Mean : 318259	Mean : 137852	Mean : 8.215				
3rd Qu.: 175299	3rd Qu.: 479919	3rd Qu.: 161915	3rd Qu.: 10,000				
Max. : 199998	Max. : 651845	Max. : 99999	Max. : 39,000				
LOS	age	GENDER	ETHNICITY	ADMITTIME	DISCHTIME		
Min. : 0.0003	Min. : 0.00	F: 9370	WHITE	154909	2108-04-16	22	2108-04-19 16:58:00
1st Qu.: 1.2899	1st Qu.: 39.00	M: 13333	UNKNOWN/NOT SPECIFIED	1681	2137-07-14	17	2137-07-20 15:57:00
Median : 2.1553	Median : 69.00		BLACK/AFRICAN AMERICAN	1616	2157-08-19	14	2157-08-21 18:50:00
Mean : 3.9765	Mean : 80.85		OTHER	505	2195-06-06	13	2202-10-11 16:30:00
3rd Qu.: 4.4640	3rd Qu.: 79.00		HISPANIC OR LATINO	441	2202-10-01	13	2206-08-21 16:30:00
Max. : 2608.7089	Max. : 310.00		UNABLE TO OBTAIN	383	2206-08-19	11	2195-06-18 18:30:00
NA's : 427			(Other)	1168	(Other)	20613	(Other)
DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	HOME_HEALTH_CARE	DISCHARGE_LOCATION			
117884	ELECTIVE	3640	EMERGENCY ROOM ADMIT	17878	HOME HEALTH CARE	15555	
2147-09-29 19:38:00	8	EMERGENCY:16578	CLINIC REFERRAL/PREATURE:4925	HOME		4479	
2156-11-06 20:00:00	7	HEMORRH	11	PHYS REFERRAL/INTERNAL DELT:14233	SNF	16117	
2116-10-05 02:30:00	6	URGENT	474	TRANSFER FROM HOSP/EXTERN:15088	DEAD/EXPIRED	2819	
2116-02-07 16:58:00	5			TRANSFER FROM SKILLED NURS	116	REHAB/DISTRICT PART HOSP:23997	
2125-01-08 01:20:00	5			TRANSFER FROM OTHER HEALT	36	LONG TERM CARE HOSPITAL: 967	
(Other)	1788			(Other)	7	(Other)	869
INSURANCE	LANGUAGE	RELIGION	MARITAL_STATUS	EDREGTIME			
Government: 395	ENG	12445	CATHOLIC	17874	MARRIED	18548	8827
Medicaid: 1387		6495	NOT SPECIFIED	13877	SINGLE	4086	2108-04-15 20:36:00
Medicare: 13809	SPAN	431	PROTESTANT QUAKER:2817	WIDOWED	3671	2137-07-14 03:08:00	17
Private: 15804	RUSS	363	UNORTHAXABLE	2127	DIVORCED	1264	2202-10-03 01:19:00
Self Pay: 108	PTUN	246	JEWISH	20814		885	2206-08-19 19:51:00
	CANT	156	OTHER	983	SEPARATED	197	2195-06-06 16:14:00
	(Other)	657	(Other)	1891	(Other)	132	(Other)
EDOUTTIME				DIAGNOSIS	HOSPITAL_EXPIRE_FLAG		
1	8827	PNEUMONIA			608	0:17884	
2108-04-16 02:40:00	22	CORONARY ARTERY DISEASE			525	1: 2819	
2137-07-14 00:30:00	17	SEPSIS			489		
2202-10-03 01:40:00	13	CHEST PAIN			406		
2206-08-20 00:27:00	11	CORONARY ARTERY DISEASE/CORONARY ARTERY BYPASS GRAFT /SDA			406		
2195-06-06 18:54:00	9	INTRACRANIAL HEMORRHAGE			381		
(Other)	11884	(Other)			17856		
DOB	DOB_HOSP	DOB_SSN	EXPIRE_FLAG				
2039-03-10:	23	12378	115473	13635	0:12370		
2078-12-05:	19	2108-06-07:	22	2108-06-07:	22	1: 8333	
2092-04-12:	16	2202-12-05:	13	2147-09-29:	8	2202-12-05:	13
2136-07-28:	13	2200-08-21:	9	2207-01-14:	8	2200-08-21:	9
2143-06-19:	11	2147-09-29:	8	2130-06-19:	7	2147-09-29:	8
2009-12-13:	9	2207-01-14:	8	2143-03-06:	7	2133-10-07:	7
(Other)	20612	(Other)	8273	(Other)	5178	(Other)	7089

Figura 77- Seleção da patologia com doença com ICD 4019

Na parte superior desta página, o utilizador tem acesso aos dados de todos os pacientes diagnosticados com esta doença, indicando no canto inferior esquerdo da tabela que existe um total de 20,703 diagnósticos. No canto inferior esquerdo, é impresso um sumário das características desses doentes (Figura 77).

Um outro submenu relevante é “First Diagnoses” este menu tem como objetivo informar de qual é a principal causa de entrada no hospital, selecionando apenas ICD cujo SEQ_NUM é igual a 1, no entanto, uma outra função foi acrescentada.

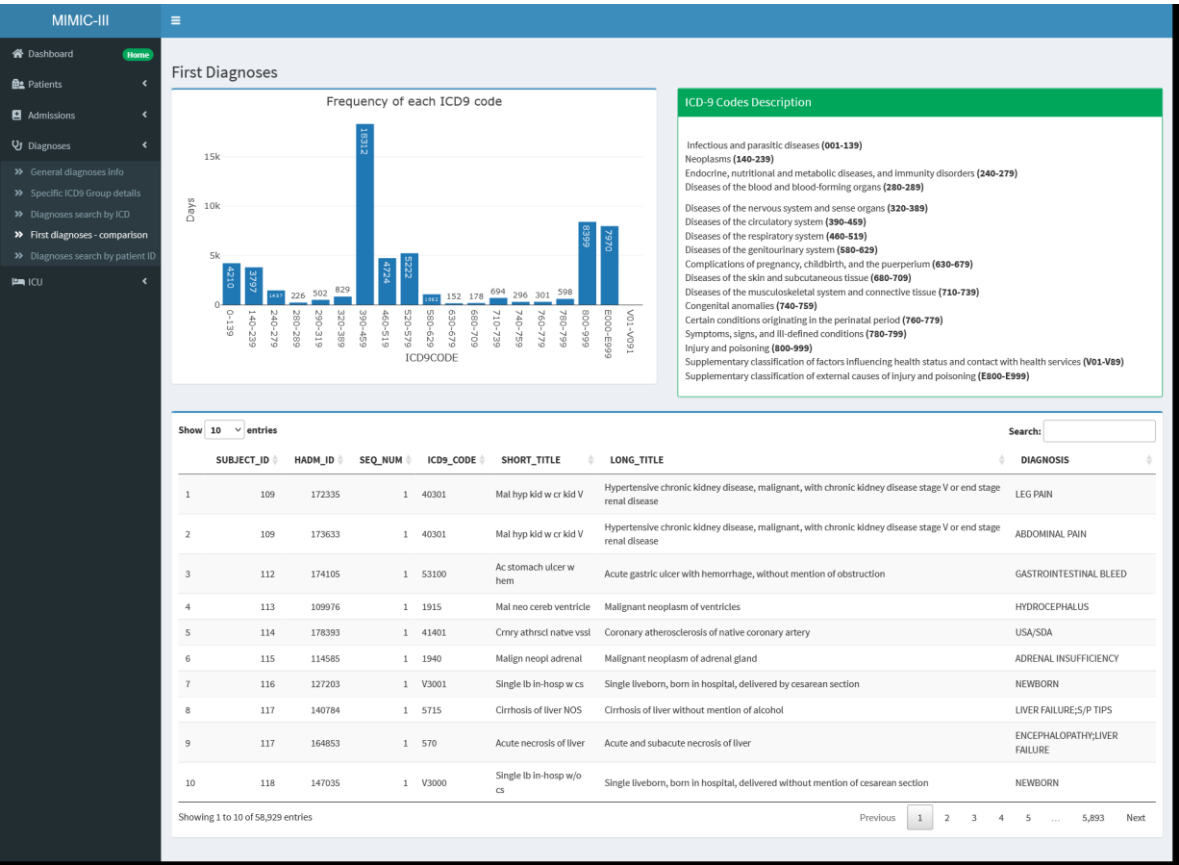


Figura 78- Submenu First Diagnoses

Sob o histograma da Figura 78, é apresentada uma tabela com todos os pacientes e respetivas 1ª causas de deslocação ao hospital. Nesta tabela através de cruzamento de dados com a tabela ADMISSIONS, é colocada uma outra coluna chamada “Diagnosis”, esta permite comparar o primeiro diagnóstico real com o feito na triagem (Figura 79)

Supplementary classification of external causes of injury and poisoning (E800-E999)

Show 10 entries

	SUBJECT_ID	HADM_ID	SEQ_NUM	ICD9_CODE	SHORT_TITLE	LONG_TITLE	DIAGNOSIS
1	109	172335	1	40301	Mal hyp kid w cr kid V	Hypertensive chronic kidney disease, malignant, with chronic kidney disease stage V or end stage renal disease	LEG PAIN
2	109	173633	1	40301	Mal hyp kid w cr kid V	Hypertensive chronic kidney disease, malignant, with chronic kidney disease stage V or end stage renal disease	ABDOMINAL PAIN
3	112	174105	1	53100	Ac stomach ulcer w hem	Acute gastric ulcer with hemorrhage, without mention of obstruction	GASTROINTESTINAL BLEED
4	113	109976	1	1915	Mal neo cereb ventricle	Malignant neoplasm of ventricles	HYDROCEPHALUS
5	114	178393	1	41401	Coron atheroscl native vssl	Coronary atherosclerosis of native coronary artery	USA/SDA
6	115	114585	1	1940	Malign neopl adrenal	Malignant neoplasm of adrenal gland	ADRENAL INSUFFICIENCY
7	116	127203	1	V3001	Single lb in-hosp w cs	Single liveborn, born in hospital, delivered by cesarean section	NEWBORN
8	117	140784	1	5715	Cirrhosis of liver NOS	Cirrhosis of liver without mention of alcohol	LIVER FAILURE;S/P TIPS
9	117	164853	1	570	Acute necrosis of liver	Acute and subacute necrosis of liver	ENCEPHALOPATHY/LIVER FAILURE
10	118	147035	1	V3000	Single lb in-hosp w/o cs	Single liveborn, born in hospital, delivered without mention of cesarean section	NEWBORN

Showing 1 to 10 of 59,929 entries

Previous 1 2 3 4 5 ... 5,993 Next

Figura 79- Tabela comparando código ICD9 com Diagnosis

Este facto serve introdução para a próxima secção, a possibilidade de usar esta ferramenta para evidenciar erros.

5.3 Erros que a ferramenta pode evidenciar

Como referido no capítulo anterior, esta ferramenta pode, por vezes, expor possíveis erros efetuados no hospital.

Um desses exemplos pode ser observado no submenu “First Diagnoses”, comparando o diagnóstico feito na triagem com a real razão de hospitalização do doente. Observemos o exemplo da Figura 80:

16	124	134369	1	44021	Ath ext nrv at w claudct	Atherosclerosis of native arteries of the extremities with intermittent claudication	CLAUDICATION
17	124	138376	1	56983	Perforation of intestine	Perforation of intestine	PNEUMONIA/CHRONIC OBST PULM DISEASE
18	124	172461	1	43331	Ocl mlt bi art w infct	Occlusion and stenosis of multiple and bilateral precerebral arteries with cerebral infarction	CVA

Figura 80- Paciente 124 com perfuração do intestino

O paciente cujo SUBJECT_ID é 124, na sua admissão hospitalar com HADM_ID 138376, é diagnosticado ao dar entrada no hospital com pneumonia/obstrução crónica do pulmão. No entanto, concluiu-se que a verdadeira causa de hospitalização do mesmo deveu-se a uma perfuração do intestino, iCD-9 56983.

Verificando todo o historial do paciente 123 no submenu Diagnoses Search by patient ID, após inserir o seu ID na caixa de input e filtrar os dados por “138376”, obtemos os dados da Figura 81.

Show 10 entries

Search: 138376

	HADM_ID	ROW_ID	SUBJECT_ID	SEQ_NUM	ICD9_CODE	SHORT_TITLE	LONG_TITLE	LOS
25	138376	1599	124	1	56983	Perforation of intestine	Perforation of intestine	22.9075
40	138376	1600	124	2	56089	Intestinal obstruct NEC	Other specified intestinal obstruction	22.9075
27	138376	1601	124	3	49121	Obs chr bronc w(ac) exac	Obstructive chronic bronchitis with (acute) exacerbation	22.9075
41	138376	1602	124	4	99859	Other postop infection	Other postoperative infection	22.9075
26	138376	1603	124	5	0389	Septicemia NOS	Unspecified septicemia	22.9075
29	138376	1604	124	6	99592	Severe sepsis	Severe sepsis	22.9075
31	138376	1605	124	7	99831	Disrup internal op wound	Disruption of internal operation (surgical) wound	22.9075
34	138376	1606	124	8	486	Pneumonia, organism NOS	Pneumonia, organism unspecified	22.9075
37	138376	1607	124	9	5849	Acute kidney failure NOS	Acute kidney failure, unspecified	22.9075
30	138376	1608	124	10	5854	Chr kidney dis stage IV	Chronic kidney disease, Stage IV (severe)	22.9075

Showing 1 to 10 of 19 entries (filtered from 52 total entries)

Previous 1 2 Next

Figura 81- Filtragem da tabela pelo HADM_ID 138376

Para facilitar a visualização ordenamos por SEQ_NUM. Aqui, podemos verificar que pneumonia é apenas a 8ª enfermidade do paciente: “Pneumonia, organism unspecified”.

Desta forma podemos levantar a questão: terá uma triagem imprecisa prejudicado a recuperação deste paciente? A gravidade e o local do hospital onde é admitido um paciente com pneumonia, certamente que não será o mesmo que um paciente com uma perfuração no intestino.

Inserindo o ID do paciente no Menu Patient Search by ID, verificamos que o paciente tinha 69 anos quando entrou no hospital com a data anonimizada de 2160 (Figura 82), a admissão foi em 2165 (Figura 83), logo teria cerca de 74 anos no momento dessa hospitalização.

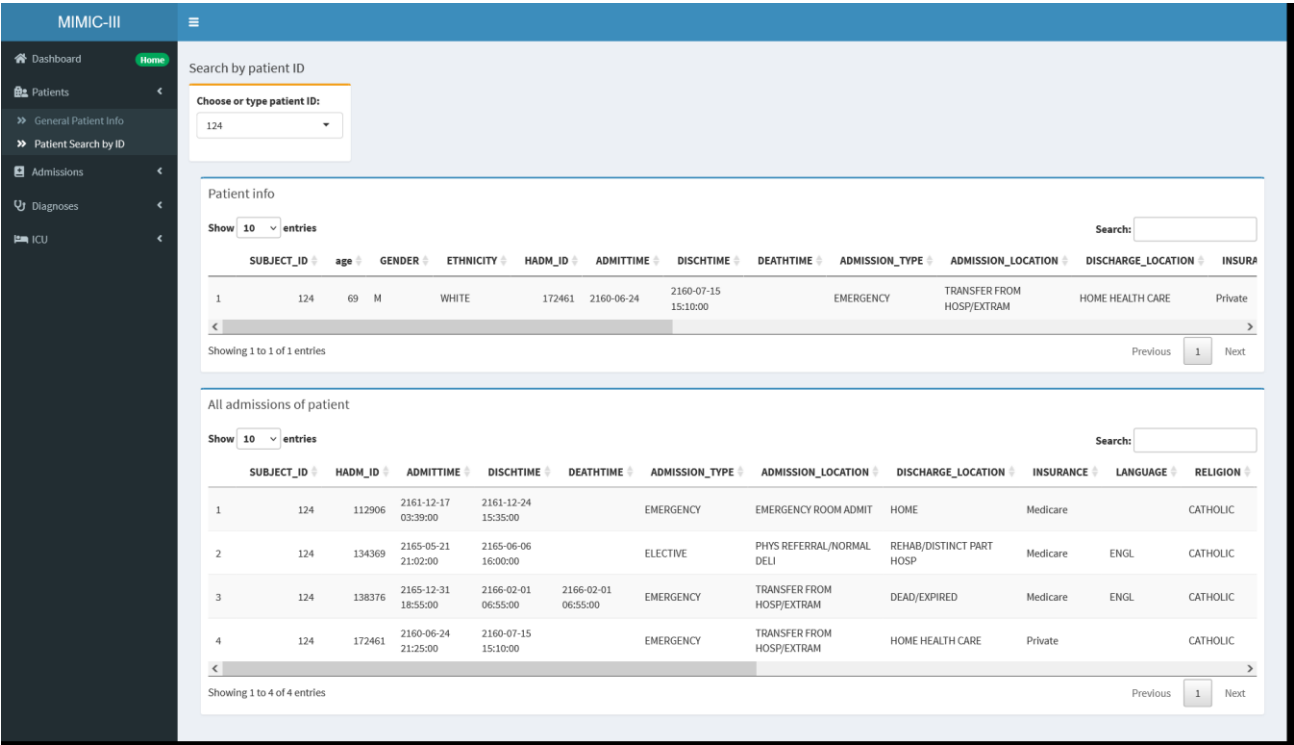


Figura 82 - Data de admissão do paciente 124 no submenu Patient Search by ICD

2	124	134369	2165-05-21 21:02:00	2165-06-06 16:00:00	
3	124	138376	2165-12-31 18:55:00	2166-02-01 06:55:00	2166-02-01 06:55:00
4	124	172461	2160-06-24 21:25:00	2160-07-15 15:10:00	

Figura 83- Admissão 138376 do paciente 124

O paciente acabou por falecer no hospital nesta hospitalização que durou cerca de 22 dias.

Estes tipos de problemas evidenciados por esta ferramenta podem ajudar a perceber se uma triagem mais atenta não poderia acelerar o início do tratamento deste doente e contribuído para um desfecho mais positivo.

5.4 Patient Mortality

O submenu Patient Mortality do menu ICU permite ao utilizador iniciar a investigação e criar estatísticas de mortalidade após ser dada alta do hospital. Este submenu foi inspirado no trabalho de (Lee et al, 2015), que oferece estudos sobre a mortalidade de pacientes 28 dias após receberem alta. Esta ferramenta permite não só estudar isso, como também é dado a escolher ao utilizador qualquer intervalo de dias que procura filtrar, através de uma barra de *input horizontal*.

A Figura 84 mostra o output de um estudo sobre todos os pacientes que faleceram até 28 dias após receberam alta. O utilizador pode visualizar estes dados numa tabela, ordená-los, observar os gráficos criados pelo grupo de pacientes escolhidos, e se desejar, efetuar o download desses mesmos dados para a sua máquina (em formato CSV), clicando no botão “Download this data” que se encontra sob a tabela.

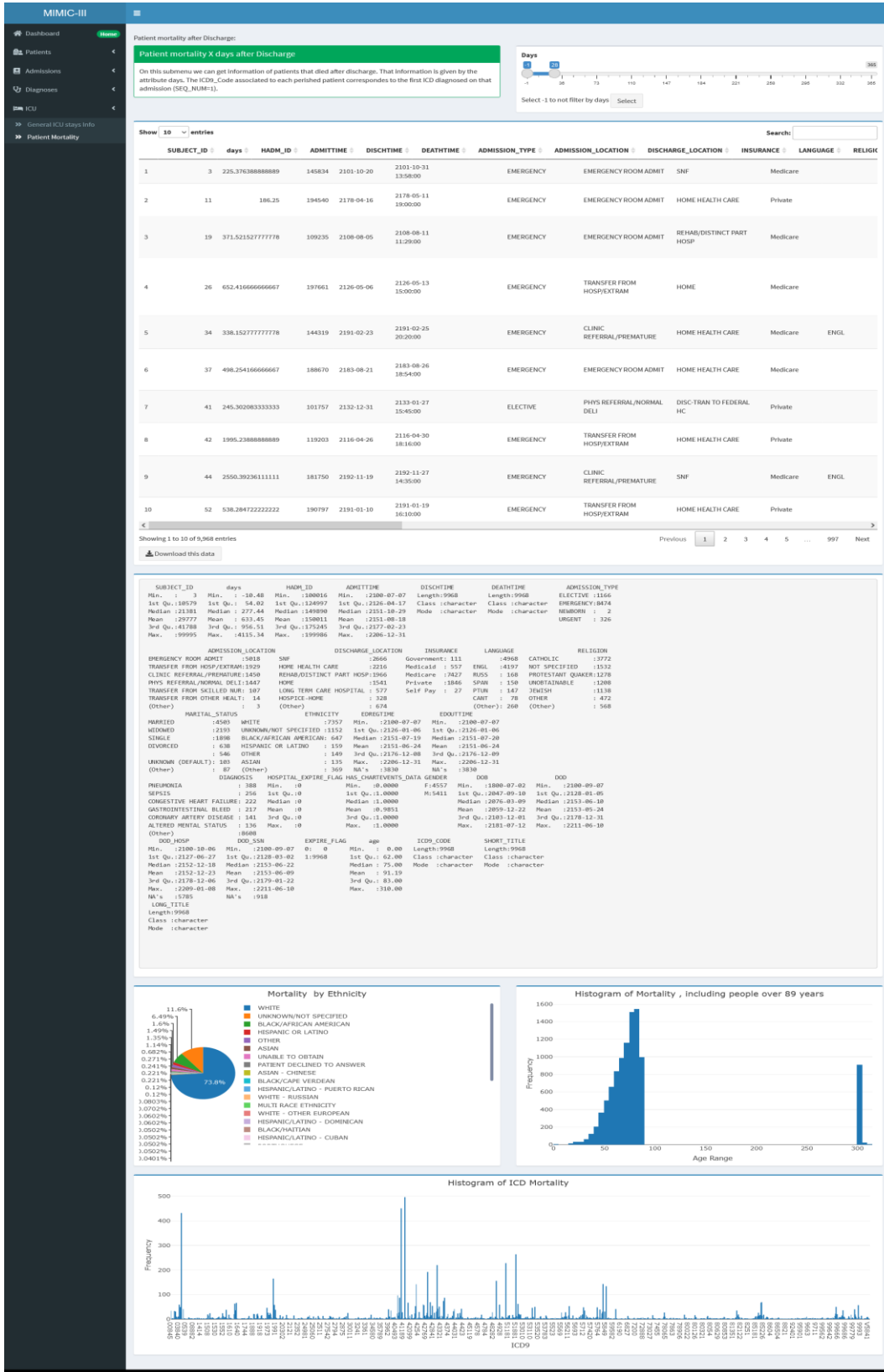


Figura 84 Estudo de mortalidade 28 dias

6. Conclusão

A MIMIC-III é um recurso gratuito, mas de valor imensurável por aquilo que permite e permitirá no futuro estudar, podendo alguns desses estudos serem convertidos em reais avanços na saúde. Porém, embora seja uma fonte de possível conhecimento valioso, ao seu acesso existem algumas barreiras que podem desincentivar a sua utilização.

Existem atualmente poucas ferramentas (e nenhuma acessível online, neste momento) que desempenhem a função deste programa. A larga maioria das ferramentas de visualização de dados orientadas à MIMIC-III focam-se em dados em formato WaveForm.

A única interface que procura desempenhar o mesmo papel já não está disponível (Lee et al, 2015), e fá-lo para uma versão descontinuada da base de dados MIMIC-II. A sua interface era pouco *user-friendly*, embora contivesse ideias interessantes como o estudo de mortalidade dos pacientes 28 dias após receberem alta e utilização de gráficos para mais facilmente apresentar e comparar dados.

Este dashboard feito em R Shiny, inspira-se nos melhores atributos das ferramentas estudadas (tais como, apresentação de informação em gráficos e estudos de mortalidade) e implementa funções em falta referidas nestes trabalhos (Lee et al, 2015), tais como o download dos dados a ser visualizados para o computador do utilizador. Além disso foi tido um cuidado especial no desenho da interface para que fosse fácil de utilizar e coerente ao longo da sua implementação, com um padrão na forma como cada menu e submenu é apresentado.

Utilizando bibliotecas do R, que em muito ajudam a visualização de dados, cumpre o objetivo a que se propõe: ser uma ferramenta de exploração da MIMIC-III que requer zero conhecimentos de programação. Não só podendo ser utilizada por um iniciante na base de dados MIMIC-III, como também, por um investigador que sabe programar e que pretende selecionar ou filtrar um determinado tipo de pacientes com determinadas características para iniciar um estudo ou estudar a viabilidade do mesmo.

6.1 Limitações

Embora esta ferramenta seja capaz de fornecer múltiplas informações relativas aos pacientes da MIMIC-III, possui algumas limitações.

A primeira é não incluir dados da tabela LABEVENTS ou CHARTEVENTS. Esta é uma limitação que a ferramenta apresenta e que seria, sem dúvida, um excelente tópico para trabalho futuro. Não apresenta informações sobre que medicamentos foram administrados aos pacientes, em que quantidades, nem cria gráficos com essa informação.

6.2 Trabalhos futuros

Existem vários caminhos que podem ser seguidos para expandir este trabalho.

A mais imediata seria a adição de dados relativos a dados monitorizados na cama dos pacientes, tais como os que estão presentes nas tabelas INPUTEVENTS_CV e INPUTEVENTS_MV.

A inclusão de dados relativos ao pessoal médico seria também bem-vinda, sendo que para isto não seriam preciso grandes mudanças estruturais no desenho da ferramenta.

A possibilidade de integrar a visualização de dados em formato Waveform, oriundo de MIMIC-III Waveform Database seria uma adição que em muito iria beneficiar esta ferramenta, ampliando assim o seu leque de possíveis utilizações. Poderiam ser criados novos menus onde este tipo de visualização de dados seria implementado.

Outra possibilidade, embora mais complexa, mas não menos interessante seria a utilização de data mining para prever certos acontecimentos, por exemplo: baseado nas características do doente, calcular o número de dias que uma cama de hospital seria necessária. Ou estudar de que forma uma incorreta triagem à entrada do hospital e consequente atraso no internamento pode levar à morte dos pacientes, utilizando técnicas de *data mininig*, elevando assim a fasquia do submenu já criado nesta ferramenta.

7. Bibliografia

Alistair, E. W., Johnson., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Science Data*. <https://www.nature.com/articles/sdata201635.pdf>

Alcaide, D., & Aerts, J. (2021). A visual analytic approach for the identification of ICU patient subpopulations using ICD diagnostic codes. *Peerj Computer Science*, 7, e430. doi: 10.7717/peerj-cs.430 (Alcaide & Aerts, 2021)

Benzon, H., Rathmell, J. P., Wu, C. L., Turk, D. C., & Argoff, C. E. (2008). *Raj's Practical Management of Pain*. Mosby Elsevier.

Cartwright D. J. (2013). ICD-9-CM to ICD-10-CM Codes: What? Why? How?. *Advances in wound care*, 2(10), 588–592. <https://doi.org/10.1089/wound.2013.0478>

Centers for Medicare & Medicaid Services. (2021, Janeiro 12). *Appendix F ICD-9-CM Format*. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/HospitalAppendix_F.pdf

Coad, J., & Coad, N. (2008). Children and young people's preference of thematic design and colour for their hospital environment. *Journal of child health care: for professionals working with children in the hospital and community*, 12(1), 33–48. <https://doi.org/10.1177/1367493507085617>

Coimbra, U. (2022). Anonimização e Pseudonimização. 20 agosto de 2022, de <https://www.uc.pt/protecao-de-dados/protecao-de-dados-pessoais/anonimizacao-e-pseudonimizacao/>

Plotly Technologies Inc. (2015). Collaborative data science. From <https://plot.ly>

Data Flair. (2022). *Object Oriented Programming (OOP) in R | Create R Objects & Classes*. <https://data-flair.training/blogs/object-oriented-programming-in-r/>

Do Not Pay Quick Reference Card. (2017). *Social Security Death Master File*. *Iperia*, 112-248. https://fiscal.treasury.gov/files/dnp/20170804QRCDMF_FINALv2.pdf

Festag, S., & Spreckelsen, C. (2019). A Visualisation and Extraction Tool for Time Series in the MIMIC III Database. *Studies in health technology and informatics*, 267, 134–141. <https://doi.org/10.3233/SHTI190817> (Festag & Spreckelsen, 2019)

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101 (23), 215–e220. <https://doi.org/10.13026/C2XW26>

Hirsch, J. A., Nicola, G., McGinty, G., Liu, R. W., Barr, R. M., Chittle, M. D., & Manchikanti, L. (2016). ICD-10: History and Context. *AJNR. American journal of neuroradiology*, 37(4), 596–599. <https://doi.org/10.3174/ajnr.A4696>

Inc, R. (2020). Easy web applications in R. 29 de Janeiro de 2022, de <https://shiny.rstudio.com/>

Johnson, A., Pollard, T., Blundell, J., Gow, B., Paris, N., & Dang, K. et al. (2022). MIT-LCP/mimic-code: MIMIC Code v2.2.1. 28 de agosto de 2022, de <https://zenodo.org/record/6818823>

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). *MIMIC-III, a freely accessible critical care database*. *Science Data* 3. <https://doi.org/10.1038/sdata.2016.35>

Knowles, J. (2020). Convenience Functions for Education Data [R package eeptools version 1.2.4]. 29 de abril de 2022, de <https://cran.r-project.org/web/packages/eeptools/index.html>

Lee, J., Ribey, E. & Wallace, J.R. A web-based data visualization tool for the MIMIC-II database. *BMC Med Inform Decis Mak* 16, 15 (2015). <https://doi.org/10.1186/s12911-016-0256-9>

LightWAVE 0.71. (2022). 28 julho 2022, from <https://physionet.org/lightwave/>

MIT Critical Data. (2016). Secondary Analysis of Electronic Health Records. doi: 10.1007/978-3-319-43742-2

Massachusetts Institute of Technology Cambridge. (2017, junho 30). *R Language Definition*. http://web.mit.edu/~r/current/arch/amd64_linux26/lib/R/doc/manual/R-lang.pdf

Medical Information Mart for Intensive Care. (2021, Junho 26). *MIMIC-III documentation*.
<https://mimic.mit.edu/docs/iii/>

Miles, J. (2017). Reconstruction of the MIMIC-III Database for Data Analytics. Retrieved from http://pi.cs.oswego.edu/~jmiles3/mimic/Miles-MIMIC-Project_report.pdf

MIMIC II Clinical Database - Demo. (2016). 28 de abril de 2022, from <https://archive.physionet.org/mimic2/demo/> ("MIMIC II Clinical Database - Demo", 2016)

National Cancer Institute. (2018, Dezembro 3). *What Is the ICD?*.
<https://training.seer.cancer.gov/icd10cm/intro.html>

National Technical Information Service. (2017). *Limited Access Death Master File Download*. <https://dmf.ntis.gov/>

Neamatullah, I., Douglass, M.M., Lehman, L.H., Reisner, A., Villarroel, M., Long, W. J., Sxolovits, P., Moody, G. B., Mark, R. G., & Clifford G. D. (2008). Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*, 8(32), 1-17.
<https://doi.org/10.1186/1472-6947-8-32>

PhysioBank ATM. (2022). 27 de setembro de 2022, de <https://archive.physionet.org/cgi-bin/atm/ATM>

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austris. <https://www.R-project.org/>.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA
URL <http://www.rstudio.com/>.

Shiny - Create a page with fluid layout — fluidPage. (2020). Retrieved 16 April 2022, from <https://shiny.rstudio.com/reference/shiny/latest/fluidPage.html>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., & François, R. et al. (2022). Welcome to the Tidyverse. 28 julho 2022, de <https://cran.r-project.org/web/packages/tidyverse/citation.html>

Xie, Y. (2017). Shiny - Using selectize input. maio 2022, de <https://shiny.rstudio.com/articles/selectize.html>

