PDSP

PROGRAMA DOUTORAL
EM SAUDE PÚBLICA

UNIVERSIDADE DO PORTO
FACULDADE DE MEDICINA

Milton Severo Barros da Silva

Latent models in the development and improvement of tools for health outcomes measurement

Porto - 2012

Milton Severo Barros da Silva

Latent models in the development and improvement of tools for health outcomes measurement

Dissertação de candidatura ao grau de Doutor em Saúde Pública, apresentada à Faculdade de Medicina da Universidade do Porto, realizada sob a orientação científica da Professora Doutora Ana Rita Pires Gaio, do Departamento de Matemática da Faculdade de Ciências da Universidade do Porto e da co-orientadora Professora Doutora Ana Azevedo Cardoso de Oliveira, do Departamento de Epidemiologia, Medicina Preditiva e Saúde Pública da Faculdade de Medicina da Universidade do Porto

Porto 2012

"All models are wrong; some are useful."

Albert Einstein

Este trabalho foi realizado com o apoio financeiro da Fundação para a Ciência e a Tecnologia e da Fundação Astra Zeneca

Porto 2012

Art.º 48º, § 3º - **“A Faculdade não responde pelas doutrinas expendidas na dissertação.”** (Regulamento da Faculdade de Medicina da Universidade do Porto – Decreto-Lei nº 19337 de 29 de Janeiro de 1931)

Corpo Catedrático da Faculdade de Medicina do Porto

Professores Catedráticos Efetivos

Doutor Manuel Alberto Coimbra Sobrinho Simões
Doutor Jorge Manuel Mergulhão Castro Tavares
Doutora Maria Amélia Duarte Ferreira
Doutor José Agostinho Marques Lopes
Doutor Patrício Manuel Vieira Araújo Soares Silva
Doutor Daniel Filipe Lima Moura
Doutor Alberto Manuel Barros da Silva
Doutor José Manuel Lopes Teixeira Amarante
Doutor José Henrique Dias Pinto de Barros
Doutora Maria Fátima Machado Henriques Carneiro
Doutora Isabel Maria Amorim Pereira Ramos
Doutora Deolinda Maria Valente Alves Lima Teixeira
Doutora Maria Dulce Cordeiro Madeira
Doutor Altamiro Manuel Rodrigues Costa Pereira
Doutor Rui Manuel Almeida Mota Cardoso
Doutor António Carlos Freitas Ribeiro Saraiva
Doutor José Carlos Neves da Cunha Areias
Doutor Manuel Jesus Falcão Pestana Vasconcelos
Doutor João Francisco Montenegro Andrade Lima Bernardes
Doutora Maria Leonor Martins Soares David
Doutor Rui Manuel Lopes Nunes
Doutor José Eduardo Torres Eckenroth Guimarães
Doutor Francisco Fernando Rocha Gonçalves
Doutor José Manuel Pereira Dias de Castro Lopes
Doutor Manuel António Caldeira Pais Clemente
Doutor António Albino Coelho Marques Abrantes Teixeira
Doutor Joaquim Adelino Correia Ferreira Leite Moreira

Professores Jubilados ou Aposentados

Doutor Abel José Sampaio da Costa Tavares
Doutor Abel Vitorino Trigo Cabral
Doutor Alexandre Alberto Guerra Sousa Pinto
Doutor Álvaro Jerónimo Leal Machado de Aguiar
Doutor Amândio Gomes Sampaio Tavares
Doutor António Augusto Lopes Vaz
Doutor António Carvalho Almeida Coimbra
Doutor António Fernandes da Fonseca
Doutor António Fernandes Oliveira Barbosa Ribeiro Braga
Doutor António Germano Pina Silva Leal
Doutor António José Pacheco Palha
Doutor António Luís Tomé da Rocha Ribeiro
Doutor António Manuel Sampaio de Araújo Teixeira
Doutor Belmiro dos Santos Patrício
Doutor Cândido Alves Hipólito Reis
Doutor Carlos Rodrigo Magalhães Ramalhão
Doutor Cassiano Pena de Abreu e Lima
Doutor Daniel Santos Pinto Serrão
Doutor Eduardo Jorge Cunha Rodrigues Pereira
Doutor Fernando de Carvalho Cerqueira Magro Ferreira
Doutor Fernando Tavarela Veloso
Doutor Francisco de Sousa Lé
Doutor Henrique José Ferreira Gonçalves Lecour de Menezes
Doutor José Augusto Fleming Torrinha
Doutor José Carvalho de Oliveira
Doutor José Fernando Barros Castro Correia
Doutor José Luís Medina Vieira
Doutor José Manuel Costa Mesquita Guimarães
Doutor Levi Eugénio Ribeiro Guerra
Doutor Luís Alberto Martins Gomes de Almeida
Doutor Manuel Augusto Cardoso de Oliveira
Doutor Manuel Machado Rodrigues Gomes
Doutor Manuel Maria Paula Barbosa
Doutora Maria da Conceição Fernandes Marques Magalhães
Doutora Maria Isabel Amorim de Azevedo
Doutor Mário José Cerqueira Gomes Braga
Doutor Serafim Correia Pinto Guimarães
Doutor Valdemar Miguel Botelho dos Santos Cardoso
Doutor Walter Friedrich Alfred Osswald

Ao abrigo do Art.º 8º do Decreto-Lei nº388/70 fazem parte desta dissertação as seguintes publicações:

I. **Severo M, Gaio R, Lucas R, Barros H.** Assessment of the general public's knowledge about rheumatic diseases: evidence from a Portuguese population-based survey. *BMC Musculoskelet Disord*. 2010;11(1):211.

II. **Severo M, Gaio R, Lourenco P, Alvelos M, Bettencourt P, Azevedo A.** Indirect calibration between clinical observers - application to the New York Heart Association functional classification system. *BMC Res Notes*. 2011;4:276.
   a. Severo M, Gaio AR, Azevedo A. Calibration: effect on the misclassification of NYHA. (*submitted*).

III. **Severo M, Gaio AR, Lourenço P, Alvelos M, Gonçalves A, Lunet N, Bettencourt P, Azevedo A.** Diagnostic value of patterns of symptoms and signs of heart failure: application of latent class analysis with concomitant variables. (**submitted**).
   a. **Severo M, Pereira M, Bettencourt P, Gaio R, Azevedo A.** B-Type Natriuretic Peptide Measured in Serum – Calibration Using Plasma Samples for Research Purposes. *Clinical Laboratory*. 2011; 57(11-12):1015-9.
   b. **Severo M, Lopes C, Lucas R, Barros H.** Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings. *Osteoporos Int*. 2009;20(2):231-7.
   c. **Severo M, Gaio AR**. A simple tool to match a concomitant variable latent class model classification. (*submitted*).

Júri da Prova de Doutoramento

Doutor José Agostinho Marques Lopes (Presidente)
Faculdade de Medicina, Universidade do Porto

Doutor José Henrique Dias Pinto de Barros
Faculdade de Medicina, Universidade do Porto

Doutor José Manuel Gonçalves Dias
Instituto Universitário de Lisboa

Doutora Carla Maria de Moura Lopes
Faculdade de Medicina, Universidade do Porto

Doutor Ana Rita Pires Gaio
Faculdade de Ciências, Universidade do Porto

Doutora Helena Cristina de Matos Canhão
Faculdade de Medicina, Universidade de Lisboa

Doutor Paulo Jorge da Silva Nogueira
Faculdade de Medicina, Universidade de Lisboa

À Fernanda e à Clara

Aos meus Pais

Agradecimentos

Às Professora Doutora Ana Rita Gaio e Professora Doutora Ana Azevedo pela competência com que orientaram esta minha tese e o tempo que generosamente me dedicaram transmitindo-me os melhores e mais úteis ensinamentos, com paciência, lucidez e confiança.

Um sentido agradecimento ao Professor Doutor Henrique Barros pelas suas breves, mas doutas indicações.

À Professora Doutora Carla Lopes por ser a primeira a acreditar nas minhas capacidades e, claro, pelo despertar do interesse na avaliação de questionários e escalas.

À Professora Doutora Maria Amélia Ferreira e ao Professor Doutor Daniel Moura pela oportunidade de trabalhar na área da educação médica.

À Fundação da Ciência e Tecnologia e à Fundação Astra Zeneca pelo apoio financeiro que possibilitou desenvolver este trabalho.

A todo o pessoal do Departamento de Epidemiologia, Medicina Preditiva e Saúde Pública e do Centro de Educação Médica da Faculdade de Medicina da Universidade do Porto

Aos meus familiares que sempre me apoiaram. Aos meus pais, que me deram não somente a vida, mas principalmente a minha educação e condições de estudo.

Eu fortemente agradeço à Fernanda Pereira, por sua extensa paciência, pelo seu amor, por sempre estar disposta a ajudar-me em qualquer situação e principalmente pelo seu apoio que me conforta e me deixa mais forte para superar os desafios a que me proponho.

Contents

Figure List

Chapter 1

Chapter 2

Subchapter 3.1

Subchapter 3.2

Subchapter 3.3

Table list

Introduction

Chapter 1

Chapter 2

Subchapter 2.1

Chapter 3

Subchapter 3.1

Subchapter 3.2

Subchapter 3.3

List of Abbreviations

1-PL - 1-Parameter Logistic

2-PL - 2-Parameter Logistic

AUC – Area Under the (ROC) Curve

BIC - Bayesian Information Criterion

BMI - Body Mass Index

BNP - B-type Natriuretic Peptide

CFI - Comparative Fit Index

EDTA – Ethylene DiamineTetra-acetic Acid

EFA - Exploratory Factor Analysis

FFQ - Food Frequency Questionnaire

GRM - Graded response model

HF - Heart Failure

IOCC - Item Operation Characteristic Curves

IQR - InterQuartile Range

IRT - item Response Theory

LCA - Latent Class Analysis

LCM - Latent Classes Models

ln - Natural Logarithm

LTM - Latent Trait Models

NYHA - New York Heart Association

OR - Odds Ratio

PRI - Population Reference Intake

ROC – Receiver Operating Characteristic Curve

SF36 - Short Form-36

Resumo

Pretendemos com esta tese compreender o papel dos modelos de variáveis latentes para desenvolver e melhorar instrumentos de medição em saúde. Neste contexto, foram desenhados 3 estudos principais:

I. **Severo M, Gaio R, Lucas R, Barros H.** Assessment of the general public's knowledge about rheumatic diseases: evidence from a Portuguese population-based survey. *BMC Musculoskelet Disord*. 2010;11(1):211.

Este estudo teve como objetivo identificar as crenças e os conhecimentos sobre doenças reumáticas numa amostra da população geral e identificar grupos alvo para educação para a saúde, através da aplicação de modelos de variáveis latentes

Participantes foram selecionados durante o seguimento de uma coorte representativa da população adulta do Porto, Portugal; 1626 participantes completaram um questionário que incluíam itens sobre conhecimentos gerais de doenças reumáticas. Modelos de variáveis latentes discretas e contínuas foram usadas para identificar falhas nos conhecimentos e grupos alvo.

O modelo de variáveis latentes contínuas identificou 2 dimensões: um relacionado com crenças (latente 1) e outro relacionado com características, tratamento e impacto das doenças reumáticas (latente 2). O modelo de variáveis latentes com 3 classes refinou estes resultados: A primeira classe apresentava baixa probabilidade de acertar aos itens associados com a primeira latente (média de 39%), a segunda classe apresentou baixa probabilidade de acertar aos itens da segunda dimensão (média de 62%). A terceira classe apresentou uma probabilidade alta de acertar todos os itens (média de 79%).

II. **Severo M, Gaio R, Lourenco P, Alvelos M, Bettencourt P, Azevedo A.** Indirect calibration between clinical observers - application to the New York Heart Association functional classification system. *BMC Res Notes*. 2011;4:276.

Este estudo teve por objetivo calibrar o sistema de classificação NYHA entre diferentes observadores, aspirando aumentar a sua fiabilidade.

Os 265 indivíduos, de um total de 1136 adultos residentes no Porto, Portugal, com idade ≥ 45 anos, que reportaram falta de ar responderam a um questionário com 4 itens para caracterizar a gravidade dos sintomas. O questionário foi aplicado por 7 médicos, que também classificaram a capacidade funcional do indivíduo de acordo com a NYHA. Cada sujeito foi avaliado por um médico. A classificação NYHA pelo foi calibrada com o método concorrente, usando um modelo de traço latente de 1-parâmetro. Discrepâncias entre os observadores foram avaliadas por diferenças nos pontos de corte entre as classes NYHA I-II e II-III no nível da variável de traço latente. A variável de traço latente estimada pelo modelo foi utilizada para prever a classificação NYHA para cada observador.

O nível da variável de traço latente para o primeiro e segundo ponto de corte variou para cada observador de -1,92 a 0,46 e 1,42-2,30, respetivamente. A concordância entre a variável de traço latente estimada e classificação dos observadores NYHA foi de 88% (kappa = 0,61).

a. **Severo M, Gaio AR, Azevedo A.** Calibration: effect on the misclassification of NYHA. (*Submetido*)

Este estudo tem como objetivo mostrar que a calibração do sistema de classificação da NYHA entre diferentes observadores aumenta a sua validade, reduzindo a má classificação.

Os 265 indivíduos, de um total de 1136 adultos residentes no Porto, Portugal, com idade ≥ 45 anos, que reportaram falta de ar responderam a um questionário com 4 itens para caracterizar a gravidade dos sintomas. O questionário foi aplicado por 7 médicos, que também classificaram a capacidade funcional do indivíduo de acordo com a NYHA. Cada sujeito foi avaliado por um médico. A classificação NYHA foi calibrada com o método concorrente, usando um modelo de traço latente de 1-parâmetro.
Comparamos a área sob a curva ROC (AUC) e o valor preditivo da versão calibrada com a versão não calibrada do NYHA para prever a presença de uma série de medidas objetivas do ecocardiograma da função cardíaca.

A AUC mostrou um aumento da capacidade preditiva do NYHA após a calibração, em grande medida pelo aumento da razão de verosimilhança do NYHA I.

III. **Severo M, Gaio AR, Lourenço P, Alvelos M, Gonçalves A, Lunet N, Bettencourt P, Azevedo A.** Diagnostic value of patterns of symptoms and signs of heart failure: application of latent class analysis with concomitant variables. *(Submetido)*.

O propósito dos autores foi identificar padrões de sintomas e sinais baseado em dados recolhidos na prática clinica de rotina, e avaliar o seu valor de diagnóstico, tendo em conta a probabilidade *a priori* de IC.

Baseados num estudo transversal foram avaliados mil e cento e quinze participantes da comunidade com idade≥45 do Porto, Portugal, em 2006-2008. Foram identificados padrões utilizando a análise de classes latentes, usando variáveis concomitantes para prever a classe a que cada participante pertence. Os padrões usaram 11 sintomas e sinais, abrangendo sobrecarga de volume e hipoperfusão. Sexo, idade, educação, obesidade, diabetes e história de enfarte do miocárdio ou IC foram incluídos como variáveis concomitantes.

A solução com 3 padrões foi suportada pelo critério de informação Bayesiano: 10,1% dos participantes apresentavam um padrão com sintomas de falta de ar e sobrecarga de volume (padrão 1), 27,8% apresentavam um padrão caracterizado principalmente por sobrecarga de volume (padrão 2) e 62,1% eram essencialmente assintomáticos (padrão 3); o melhor ajuste do modelo verificou-se quando incluímos as variáveis concomitantes. A razão de verosimilhanças para os padrões 1, 2 e 3 para a disfunção sistólica do ventrículo esquerdo foi 3,4, 1,1 e 0,6, e para a disfunção diastólica do ventrículo esquerdo foi 3,5, 1,4 e 0,5, respetivamente.

a. **Severo M, Pereira M, Bettencourt P, Gaio R, Azevedo A.** B-Type Natriuretic Peptide Measured in Serum – Calibration Using Plasma Samples for Research Purposes. *Clinical Laboratory*. 2011;11

Este estudo teve como objetivo avaliar a precisão das medições de BNP em soro usando um ensaio imunofluorométrico para prever as concentrações de BNP em plasma e classificar os indivíduos de acordo com os pontos de corte habituais.

Foram incluídos 27 indivíduos com idade mínima de 45 anos, participantes de um estudo de coorte Português. Amostras de sangue foram recolhidas em tubos de

plástico de sangue total, contendo um ácido etilenodiaminotetracético para obter plasma ou ativador de coágulo para obter soro. O logaritmo natural de soro BNP foi calibrado com o logaritmo natural de plasma BNP usando uma equação de regressão linear.

Os parâmetros de regressão estimados foram 0,58 (IC 95%: 0,23-0,93) para β_0 e 1,01 (IC 95%: 0,90-1,11) para β_1. A concordância absoluta entre o plasma BNP e que previstos pela equação de acordo com os pontos de corte 30 e 100 pg/mL foram 96,3% (kappa = 0,92) e 96,3% (kappa = 0,91), respetivamente.

b. **Severo M, Lopes C, Lucas R, Barros H.** Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings. *Osteoporos Int*. 2009;20(2):231-7.

Este estudo teve como objetivo o desenvolvimento de uma ferramenta para medir o consumo alimentar de cálcio e de vitamina D em Portugal, e aferir a utilidade de variáveis não alimentares.

Entrevistadores treinados recolheram informação de 2414 adultos da cidade do Porto, Portugal, através de um questionário de frequência alimentar (QFA) semiestruturado. Foram selecionados para ferramenta os alimentos com maior contribuição para o consumo e as variáveis não alimentares (sexo, idade, e índice de massa corporal (IMC)). Diferentes aproximações estatísticas foram usadas para prever o consumo. O gráfico de Bland-Altman foi usado para comparar as previsões da ferramenta com os resultados do QFA.

Os itens selecionados para prever o consumo de cálcio foram o leite (38%), o queijo (12%), o iogurte (10%) e o sexo; para a vitamina D, o peixe gordo (39%), enlatado (9%) e magro (7%), os ovos (5%), a carne vermelha (5%), a idade e o IMC. O gráfico de Bland–Altman mostrou que a média das diferenças foi de 0,0 (limites de concordância= [-220.67; 220.77]) mg/dia e 0,0 (limites de concordância = [-1.03; 1.05])µg/dia, respetivamente para o cálcio e vitamina D.

c. **Severo M, Gaio AR**. A simple tool to match a concomitant variable latent class model classification (*submetido*)

O propósito deste estudo é replicar a classificação obtida pela análise de classes latentes (LCA) com variáveis concomitantes usando ferramentas simples de estatística.

Baseados num estudo transversal foram avaliados mil e cento e quinze participantes da comunidade com idade≥45 do Porto, Portugal, em2006-2008. Foram identificados padrões utilizando a análise de classes latentes, usando variáveis concomitantes para prever a classe a que cada participante pertence. Os padrões usaram 11 sintomas e sinais, abrangendo sobrecarga de volume e hipoperfusão. Sexo, idade, educação, obesidade, diabetes e história de enfarte do miocárdio ou insuficiência cardíaca (IC) foram incluídos como variáveis concomitantes. Análise de classes latentes identificou 3 classes com diferentes perfis clínicos, que chama-mos “padrão de IC sintomático”, “padrão de sobrecarga de volume” e “padrão assintomático”.

Definimos pontuação total como o número de sintomas e sinais observados, e a pontuação concomitante como o número esperado de sintomas e sinais previstos através de uma regressão binomial negativa realizada sobre pontuação total e usando as variáveis concomitantes. Cada pontuação concomitante foi estimada através do sistema de pontos ou uma régua circular baseada na informação das variáveis concomitantes.

A árvore de classificação com a pontuação total e concomitante foi usada para prever classificação dos indivíduos previstos pela análise de classes latentes. A concordância absoluta entre as classes previstas e originais foi de 89,7% (intervalo de confiança de bootstrap 95% (B95%CI) =(88,0; 91,6)) e o respetivo Kappa foi de 0,802 (B95% CI = (0,781; 0,850)).

As principais conclusões da investigação são as seguintes:

1. O uso de modelos latentes aplicados a estas escalas específicas permitiu a identificação de diferentes dimensões e de grupos-alvo relevantes na população geral.

2. Os pontos de corte da classificação da NYHA entre observadores foram bastantes discrepantes e a calibração concorrente através dos modelos de traço latente pode ser usada para calibrar o grande número de observadores na mesma escala, contribuindo para minimizar o problema de fiabilidade da classificação da NYHA. Este tipo de aproximação pode ser útil para minimizar a variabilidade em outras classificações baseadas nas perceções de doentes e/ou médicos.

    a. A metodologia de calibração pode ser útil para melhorar a validade da classificação do NYHA na prática clínica e em contexto de investigação, contribuindo para minimizar a má classificação nesta escala.

3. O uso de variáveis concomitantes pode melhorar o valor de diagnóstico dos padrões de sintomas e sinais e, consequentemente, aumentar a utilidade dos sintomas e sinais tanto para o diagnóstico como em medidas de resposta.

    a. As amostras de soro não podem ser usadas para estimar os valores absolutos de concentração de plasma, mas os valores de BNP medidos em soro e calibrados podem ser usados para classificar de forma correta os indivíduos tendo em conta os pontos de corte habituais.

    b. As equações estimadas pelo melhor modelo de previsão de consumo alimentar de cálcio e de vitamina D permitiram desenvolver um *software* e uma régua circular útil no contexto clínico.

    c. Uma ferramenta simples para o diagnóstico de insuficiência cardíaca foi desenvolvida que permitirá a utilização no contexto clínico dos cuidados primários onde em geral não existem disponíveis de imediato medidas objetivas de função e de estrutura cardíaca.

Abstract

This thesis aimed to understand the role of latent models in the improvement and development of health outcomes measurement. In this context, three main research questions were addressed:

I. **Severo M, Gaio R, Lucas R, Barros H.** Assessment of the general public's knowledge about rheumatic diseases: evidence from a Portuguese population-based survey. *BMC Musculoskelet Disord*. 2010;11(1):211.

This study aimed to identify the incorrect beliefs and common knowledge about rheumatic diseases in a sample of the general population and to identify target groups for health education – application of latent models.

Participants were selected during the follow-up of a representative cohort of adult population of Porto, Portugal; 1626 participants completed a questionnaire that included general knowledge items about rheumatic diseases.
Discrete and continuous latent variable models were used to identify knowledge flaws and the target groups.

A continuous latent variable model identified two dimensions: one related to general beliefs (latent 1) and another concerning characteristics, treatment and impact of rheumatic diseases (latent 2). A 3-class latent variable model refined these results: the first class presented the lowest probabilities of correct answer for items associated with the first latent (mean of 39%), and the second class presented the lowest probabilities of correct answer for items with the second latent (mean of 62%). The third class showed the highest probability of a correct answer for almost all the items (mean of 79%).

II. **Severo M, Gaio R, Lourenco P, Alvelos M, Bettencourt P, Azevedo A.** Indirect calibration between clinical observers - application to the New York Heart Association functional classification system. *BMC Res Notes*. 2011;4:276.

This study aimed to calibrate the NYHA classification system between different observers, aspiring to increase its reliability.

Among 1136 community-dwellers in Porto, Portugal, aged ≥45 years, 265 reporting breathlessness answered a 4-item questionnaire to characterize symptom severity. The questionnaire was administered by 7 physicians who also classified the subject's functional capacity according to NYHA. Each subject was assessed by one physician. We calibrated NYHA classifications by the concurrent method, using 1-parameter latent trait model. Discrepancies between observers were assessed by differences in ability thresholds between NYHA classes I-II and II-III. The ability (standard normal variable) estimated by the model was used to predict the NYHA classification for each observer.

Estimates of the first and second thresholds for each observer ranged from -1.92 to 0.46 and from 1.42 to 2.30 standard deviations of ability, respectively. The agreement between estimated ability and the observers' NYHA classification was 88% (kappa=0.61).

a. **Severo M, Gaio AR Azevedo A.** Calibration: effect on the misclassification of NYHA. (*Submitted*)

Previous studies showed an inter-observer agreement for the NYHA classification of approximately 55%. Thus, the calibration of the NYHA classification system between different observers is expected to increase its validity, reducing misclassification.

At the standardized clinical interview subjects who reported to have breathlessness (n=265) were presented to a 4-item questionnaire on functional capacity to characterize the severity of symptoms. The questionnaire was administered by 7 physicians who also classified the subject's functional capacity according to NYHA. Each subject was assessed by one physician. Calibration of NYHA classification across each set of individuals assessed by each physician was performed by the concurrent method using the four patient items as anchor items. We estimated the area under the ROC curve (AUC) and likelihood ratio using the calibrated and non-calibrated NYHA class I versus II-III to predict the presence of a series of objective structural of functional cardiac abnormalities as assessed by echocardiography at rest.

The area under the ROC curve (AUC) for NYHA class to predict the outcomes considered showed an overall improvement in discrimination of NYHA class after its calibration, largely at the expense of the likelihood ratio of NYHA I.

III. **Severo M, Gaio AR, Lourenço P, Alvelos M, Gonçalves A, Lunet N, Bettencourt P, Azevedo A.** Diagnostic value of patterns of symptoms and signs of heart failure: application of latent class analysis with concomitant variables. Submitted.

The authors aimed to identify patterns of symptoms and signs, based on findings routinely collected in current clinical practice, and to evaluate their diagnostic value, taking into account the a priori likelihood of HF.

Based on the cross-sectional evaluation of 1115 community participants aged ≥45 years from Porto, Portugal, in 2006-2008, patterns were identified by latent class analysis, using concomitant variables to predict class membership. Patterns used eleven symptoms/signs, covering dimensions of congestion and hypoperfusion. Sex, age, education, obesity, diabetes and history of myocardial infarction or HF were included as concomitants.

Bayesian information criteria supported a solution with three patterns: 10.1% of participants followed a pattern with symptoms of troubled breathing and signs of congestion (pattern 1), 27.8% a pattern characterized mainly by signs of congestion (pattern 2) and 62.1% were essentially asymptomatic (pattern 3); model fit was best when including concomitant variables. The likelihood ratio of patterns 1, 2 and 3 for left ventricular systolic dysfunction was 3.4, 1.1 and 0.6, and for left ventricular diastolic dysfunction 3.5, 1.4 and 0.5, respectively.

a. **Severo M, Pereira M, Bettencourt P, Gaio R, Azevedo A.** B-Type Natriuretic Peptide Measured in Serum – Calibration Using Plasma Samples for Research Purposes. *Clinical Laboratory*. 2011;11

This study aimed to evaluate the accuracy of BNP measurements in serum samples using this commercially available immunofluorometric assay to predict the BNP plasma concentration and to classify the individuals using the usual cut-off points.

We enrolled 27 subjects aged at least 45 years, participating in a Portuguese cohort study. Blood samples were collected in plastic whole blood tubes, containing either ethylenediaminetetraacetic acid to obtain plasma or clot

activator to obtain serum. The natural logarithm of serum BNP was calibrated with the natural logarithm of plasma BNP using a linear equation.

The estimated regression parameters were 0.58 (95 % CI: 0.23 - 0.93) for β0 and 1.01 (95 % CI: 0.90 - 1.11) for β1. The absolute agreement between plasma BNP and that predicted by the equation according to the cut-off points 30 and 100 pg/mL were 96.3% (kappa = 0.92) and 96.3% (kappa = 0.91), respectively.

b. **Severo M, Lopes C, Lucas R, Barros H.** Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings. *Osteoporos Int*. 2009;20(2):231-7.

The study aim was to develop a tool to assess the dietary calcium and vitamin D intakes in Portugal, and evaluate the usefulness of non-dietary variables as predictors.

Trained interviewers collected information of 2,414 adults of Porto, Portugal, using a structured questionnaire and a validated semi-quantitative food frequency questionnaire (FFQ). Food items with the highest contribution to the total intake and non-dietary predictors (gender, age and body mass index (BMI)) were selected for the tool. Different statistical approaches were used to predict the intake. A Bland–Altman plot compared the predictions from the tool and the full FFQ.

The items selected to predict intake were milk (38%), cheese (12%), yogurt (10%) and gender for calcium and oily fish (39%), canned fish (9%), white fish (7%), eggs (5%), red meat (5%), age and BMI for vitamin D. The Bland–Altman plot showed that the mean differences were 0.0 (limits of agreement = [-220.67; 220.77]) mg/day and 0.0 (limits of agreement = [-1.03; 1.05]) µg/day, respectively for calcium and vitamin D.

c. **Severo M, Gaio AR**. A simple tool to match a concomitant variable latent class model classification (*submitted*)

The purpose of this study is to mimic the classification obtained from LCA with concomitant variables using a chain of simpler statistical tools.

Data were obtained from a cross-sectional evaluation of 1115 community participants aged≥45 years-old from Porto, Portugal, in 2006-2008. Input variables consisted of eleven symptoms and signs of HF, covering congestion and hypoperfusion. Sex, age, education, diabetes, history of myocardial infarction or heart failure, and obesity were included as concomitant variables. Concomitant variable LCA identified 3 classes with different clinical profiles, which we named "symptomatic heart failure pattern", "congestion pattern" and "no symptoms and signs pattern".

We define the total score to be the number of observed symptoms and signs, and the concomitant score to be the expected number of symptoms and signs predicted by a negative binomial regression performed on the total score and using the concomitant variables. Each concomitant score was then estimated using a point score system or circular ruler based on information of the concomitant variables readily obtained by the practitioner in the office.

A classification tree with total and correspondent concomitant scores as predictors was used to predict membership of subjects in the LCA classes. The absolute agreement between the predicted and the original classes was 89.7% (bootstrap 95% confidence interval (B95%CI) = (88.0, 91.6)) and the respective Kappa was 0.802 (B95% CI = (0.781, 0.850)).

The main conclusions of the present investigation were the following:

1. The use of latent models applied to these specific scales, were able to provide evidence for identification of different dimensions and to identify relevant target groups in the general population.

2. The thresholds of the NYHA classification between observers were very discrepant and that concurrent calibration through latent trait models can be used to calibrate a large number of observers on the same scale. It provides a way to minimize the reliability problem of NYHA classification. This type of approach can be useful to minimize the inter-observer variability in other classifications based on patient's and/or physicians's perception.

    a. The calibration methodology can be useful to improve the validity of NYHA classification in clinical practice and research settings, by increasing the inter-observer reproducibility, and can be used to calibrate a large number of observers on the same scale. It provides a way to minimize the misclassification of NYHA classification.

3. The use concomitant variables can improve the diagnostic value of the symptoms and signs patterns and, consequently, improve the usefulness of the symptoms and signs for diagnosis and as an outcome measures.

    a. Serum samples cannot be used to estimate absolute plasma concentrations, but serum BNP values and the calibration equation can be used to classify correctly the individuals with the usual cut-offs.

    b. The equations estimated by the best statistical model to predict the calcium and vitamin D intake allowed for the design of a software and a circular ruler useful in clinical settings.

    c. A simple diagnostic tool for general practitioners and internists was developed. It will enable them to diagnosis heart failure in primary care, where in general objective measures of cardiac structure and function are not available.

Introduction

Every year, new health outcome measures arise in accordance with the most recent scientific developments. Researchers hope the obtained tools will be more reliable, valid, sensitive, and comprehensive than the existent ones, and that therefore will result in minimal response burden. In turn, this raises the need for more comprehensive and accurate evaluation of the psychometric properties of the existing health outcome measures.

A new generation of health outcome tools is being developed based on the principles of latent modeling (1). Latent models express the association between input variables and underlying latent variables (2). Input variables are commonly denoted by *items*, as the items of a questionnaire, survey, examination test… latent variables are unobservable variables, not directly measurable, subjacent to the set of observable items.

Latent models can be divided into two frameworks: latent trait models (LTM), which assume that the latent variables are metrical, and latent class models (LCM) which consider latent variables as categorical (3). Here, metrical means that the variable has realized values in the set of real numbers and may be discrete or continuous. Categorical variables assign individuals to one of a set of categories; they may be unordered or ordered.

In the first situation, the latent variables are often called “abilities” or “traits” while in the second situation the categories of the variables are denoted by “classes”.

In the past decades, the application of latent trait models in health research has increased considerably (4, 5). Examples of this type of analysis can be found in studies of cognitive ability (6), fatigue (7), feelings scale (8), physical functioning (9) or quality of life (1), for example.

In general, LTM assume that a set of J traits is sufficient to “explain”, or account for, both individual performances on the set of items and the interrelationship within all pairs of items. If the set of traits does not exist, it means that the items are all

independent. Whenever J equals 1, *i.e.*, the latent trait is 1-dimensional, the analysis is referred to as item response theory. The more general situation corresponding to J greater than 1 is referred to as multidimensional.

An assumption of LTM is local independence of the items, that is, once the values of the traits are known, the items must be independent. Moreover, if the fitting of the model is good, then item parameters can be estimated independently of the particular sample of individuals that answered the questionnaire (within a linear transformation) (10). This property is known as *item parameter invariance*. In addition, an individual's estimated trait is not dependent upon the particular sample of items chosen from the battery (10). This other property is known as *ability parameter invariance*.

The above features show already that LTM offer more comprehensive and accurate evaluation of the psychometric properties of a given questionnaire over classical measurement techniques [7]. They also allow for an optimal shortening of the questionnaire, when necessary, and for the evaluation of the performance of the reduced measurement. Further, LTM provide an estimate of the reliability of a scale along the whole trait (information function) (11), instead of a single estimate of reliability as we have with Cronbach's alpha, for instance, in the classical measurement theory. The information function provides a graphical representation of the precision of the measurement at each trait value for either an individual test item (item information function, IIF) or the entire test (test information function, TIF). The greater the information present at a given trait value, the more precise and/or reliable the measurement will be at that value (12).

The information function is especially appealing with health outcomes, where it is important to maximize the reliability of a scale at the cut-score, and is very useful in constructing short forms or tailored assessments, ensuring that the selected subset of items provides an adequate reliability.

Another important feature of LTM is that they have a ‘‘built-in’’ linking mechanism (8). Linking is a general term that can be used to refer to both equating and calibration.

LTM define a scale for the underlying latent variable and the items are calibrated with respect to that scale. The linking property of LTM means that, once the items are calibrated for a population (*i.e.*, item parameters are known), comparable scores on a given construct may be calculated for respondents from that population who answered only a subset of the items, without intermediate equating steps. In applications where item parameters are *a priori* not known, the linking of two or more scales is still fairly straightforward provided both forms measure the same construct and there are some overlapping items on the forms. This means that LTM can be used to equate and calibrate a large number of items of different questionnaires; by doing so, we are able to better understand the structure and order of the domain-specific items to each other, as well as the interrelations among items across the ability continuum and to design computer based questionnaires (13) .

Latent class models, also known as analysis of finite mixture models, assume that the latent variable is categorical. This framework brings several statistical advantages over standard classification approaches. Firstly, it allows problems such as the choice of the number of classes and of the classification method to be recast as statistical model choice problems. Secondly, LCM can be potentially improved through the use of concomitant variables, *i.e*., variables that influence the prevalence of classes, thus permitting the identification of more precise categories. Finally, for given values of the response and concomitant variables, posterior class membership probabilities for each individual are produced (14). Examples of health frameworks where latent class analysis was used include drug abuse/dependence (15), alcohol use (16), maternal depression (17) and dietary patterns (18).

There is no obvious choice between LTM and LCM. In the context of disease diagnosis the view that dominates is the categorical one, because it meets clinical needs and allows reporting for health-care planners, while in LTM it is difficult to find natural cut points or thresholds for the traits, reducing its usefulness to provide a classification.

However, LCM ignore possible within-class heterogeneity such as individual differences in severity.

In the following sections we are going to mathematically describe the latent models used in this thesis as well as its applications to health outcomes measurement.

Latent Models

Latent trait models

Latent trait models are a class of latent models for which the latent variables are metrical. In applications to health, they are frequently used in questionnaires, scales, tests,… that have dichotomous or ordinal items.

The most commonly used 1-dimensional LTM for these types of items are presented in table 1. Within this research project, only the 2-parameter logistic model and the graded response model will be used in applications. Therefore, we briefly describe these two methods throughout.

Table 1. Description of several commonly used 1-dimensional LTM

Model	**Item response types**	**Model characteristics**
1-parameter logistic model	Dichotomous	Discrimination power constrained to be equal across items. Thresholds vary across items
2-parameter logistic model	Dichotomous	Discrimination and thresholds vary across items
3-parameter logistic model	Dichotomous	Includes pseudo-guessing parameter, besides unconstrained discrimination and threshold parameters
Graded model	Ordinal	Discrimination and thresholds vary across items
Nominal model	Polytomous	Discrimination and thresholds vary across items
Partial credit model	Polytomous	Discrimination power constrained to be equal across items. Thresholds vary across items.
Rating scale model	Polytomous	Discrimination power and item threshold steps constrained to be equal across items

Two-parameter logistic model

Two-parameter logistic models for dichotomous responses, $\boldsymbol{y}$, assume that the 1-dimensional latent variable, z, is metrical and (standard) normally distributed, the distribution of each item conditional on the latent variable is binomial $B(1, \pi_j(z))$, and the relationships between z and $\boldsymbol{y}$ can be described by an ogive-shaped function called an *item characteristic curve* (ICC),

$$\pi_j(z) = \frac{1}{1+e^{-\beta_{1j}(z-\beta_{0j})}} \qquad (1)$$

where $\pi_j(z) = P(y_j=1|z)$, *i.e.*, the probability of correctly answering item j given the value of z, and β_{j0} and β_{j1} are the parameters of the model to be estimated. This model is similar to a logistic regression; here, however, independent variables are not observable.

For item j, the parameter β_{j0} represents the ability value at which the probability of correctly answering the item is 0.5, and is called *difficulty parameter*. This parameter is expressed on the same scale as the trait. The parameter β_{1j} is called *discrimination parameter* and represents the slope of ICC near the respective difficulty parameter, thus indicating how well an item discriminates individuals with trait value near the difficulty parameter; the higher the value of the discrimination parameter, the higher is its discrimination.

The ICC can also be described by the following parameterization, which has the advantage to immediately allow for the extension of the model to more than one latent variable,

$$\pi_j(z) = \frac{1}{1+e^{-(\beta_{0j}^{'}+\beta_{1j}z)}} \qquad (2)$$

The interpretation of the discrimination parameter remains the same but the (new) difficulty parameter, $\beta_{0j}^{'}$, has a different interpretation from above; it now represents the logit value of an individual at mean value of the trait.

The absolute value of the discrimination parameter describes the relation strength between the correspondent item and the trait. An exact equivalence has been demonstrated between the discrimination parameters of the LTM model and the standardized factor loadings, *i.e.*, correlation coefficient between the latent variable z and the items underlying the continuous variable (3).

$$\beta_{1j}^{*} = \frac{\beta_{1j}}{\left(\beta_{1j}^{2}+1\right)^{1/2}}$$

Graded response model

The graded response models (GRM) are an extension of binary response models appropriate to use when item responses are of ordinal type (19). These models assume that the performance of an individual on the items is explained by only one (standard normal) variable, denote by z as above. In the graded response models, each item is described by a set of curves, called *item operation characteristic curves* (IOCC). The item operation characteristic curve for category k of item j represents the probability of endorsing categories higher than k conditional on the value of the subject's trait. It is given by the equation

$$P(y_j > k) = \frac{1}{1+e^{-\beta_{1j}(z-\beta_{0j(k)})}} \qquad (3)$$

where y_j denotes the response to item j.

Item operation characteristic curves of an item with K categories are characterized by K parameters: the *slope* (discrimination), β_{1j}, which is the same for all categories of the

item, and the *thresholds* (difficulty), $\beta_{0j(k)}$, which are as many as the number of categories minus one. The threshold parameter between two categories represents the ability value at which the probability of indicating the highest of these two or higher is 50%. So, the threshold parameters are expressed in the same scale as the ability. The slope parameter indicates how well an item is able to discriminate individuals with ability values near the respective threshold. The slope parameter may also be interpreted as describing how an item may be related to the ability. The steeper the slope the higher is the item discrimination.

Latent class models

For a fixed number, *K*, of classes, the model assumes that the population density *f* is expressed as a weighted finite sum of *K* component densities, $f_{1,\ldots,} f_k$, with parameters $\theta_{1,..,} \theta_k$, respectively, and each density is identified with a class. For the individual i, let $\boldsymbol{y}_i$ denote the response vector of observations on J variables. In our application, these input variables, also denoted by items or manifest variables, will be either dichotomous or ordinal. The general model is

$$f\left(\mathbf{y}_i \mid \mathbf{x}_i, \psi\right) = \sum_{k=1}^{K} \eta_{k|(\mathbf{x}_i,\alpha)} f_k\left(\mathbf{y}_i \mid \theta_k\right)$$

where η_k is the probability of class k membership, and $\boldsymbol{x}_i$ is a vector of concomitant variables that influences the prevalence of the classes through the parameters α. The vector $\psi=(\alpha, \theta_1,\ldots, \theta_K)$ is the set of model parameters that are to be estimated. The model assumes that $\sum_{k=1}^{K} \eta_k = 1$.

Assuming independence of the coordinate response vectors y_{ij} within each class k, and that the multivariate density f_k is the same across classes, say $f=(f_1,\ldots,f_J)$, the model writes

$$f(\mathbf{y}_i \mid \mathbf{x}_i, \psi) = \sum_{k=1}^{K} \eta_{k|(\mathbf{x}_i,\alpha)} \prod_{j=1}^{J} f_j(\mathbf{y}_{ij} \mid \theta_{jk}) \quad .$$

The (sub)model that takes account of the input variables distribution within each class is denoted by component specific model, and the (sub)model that studies the influence of the concomitant variables on the classes prevalence is called concomitant variable model. The latter model will generally be a multinomial logit model with parameters α.

Individuals are assigned a class according to the standard modal allocation from posterior class memberships.

In this research project, LCM will be considered in two situations: for binary items only, and for items of mixed-mode type. The densities f_k are considered accordingly to the items' type.

Challenges for the application of latent models in health outcomes measurements

The epidemiological transition from infectious to chronic diseases, even in low- and middle income countries, is already well established and is of major relevance to health planning. Musculoskeletal diseases, ischemic heart disease and cerebrovascular disease were amongst the leading causes of morbidity in the world in 2001 (20). Musculoskeletal diseases constitute indeed a major public health challenge for our aging societies (21). Providing the general population and patients with good quality information is an important strategy for the management of chronic diseases. Knowledge leads to changes in attitudes and behaviours, and directly influences health status (22), and adequate information can promote self-management skills necessary for coping with the disease increasing adherence to therapy (23). A recent review (24, 25) identified significant limitations and constraints in measuring osteoporosis

knowledge as a single domain, as it should include multi-dimensional aspects like causes or risk factors, prevention, consequences and treatment.

Heart failure is a complex clinical syndrome resulting from a variety of structural or functional cardiac disorders. The diagnosis of heart failure (HF) requires a compatible clinical syndrome and demonstration of cardiac dysfunction by imaging or functional tests (26, 27). A clinical examination is always the first step in a diagnostic approach to possible HF and further investigation is conditional on initial clinical judgment. However, individual symptoms (such as dyspnoea and fatigue) and signs (*e.g.* third heart sound and evidence of congestion) are generally unreliable and have limited value for diagnosing heart failure (28, 29). Several multidimensional criteria based on symptoms and signs have been developed over decades in an attempt to standardize the clinical assessment of heart failure (30-37). When patients initially labeled as having heart failure are investigated using objective assessment criteria, only around one third are considered to truly have heart failure (38, 39). Obesity, unrecognized myocardial ischaemia or pulmonary disease commonly lead to false positive heart failure diagnoses (38). Additionally, it may be difficult to distinguish pathologic conditions from mere physical deconditioning associated with ageing. Moreover, the varying subjective importance attributed to symptoms justifies a systematic association between reported symptoms and female gender and psychosocial characteristics, both among the healthy and those with cardiac dysfunction. Gender, age, education and obesity are major determinants of symptoms and signs suggestive of heart failure (40), beyond their role as risk factors for heart failure, and may account for false positive and negative classification. Furthermore, the clinical judgment is modified based on the *a priori* likelihood of HF (41), depending mainly on history of HF or myocardial infarction, and on strong risk factors for such conditions.

The New York Heart Association (NYHA) functional classification is one of the steps needed to obtain a good clinical examination of HF. The NYHA classification was designed for clinical assessment of patients by physicians in 4 classes (I, II, III or IV) on the basis of the patient's limitations in physical activities caused by cardiac symptoms (42). The NYHA classification is derived largely by inference from history and/or observation of the patient in certain physical activities, and occasionally by direct or indirect measurement of cardiac function in response to standardized exercises. The class a clinician decides to assign a patient to depends on the clinician´s interpretation of what is "ordinary" physical activity, "slight" and "marked" limitations. This results in a high inter-observer variability. Previous studies showed an inter-observer agreement for the NYHA classification of approximately 55% (43, 44). Consequently the use of NYHA classification as an outcome measure in clinical research is rather poor.

Traditionally, LTM and LCM are used separately in the development of questionnaires and scales. As referred above, there are different advantages in using LTM and LCM to develop health outcomes measurements, thus the combination of both models and consequently the combination of their best features could refine existing instruments; for example, using both models we could obtain simultaneous within class heterogeneity and natural cut points to provide a classification.

Latent models can be useful in the standardization of clinical assessments. Even with training, it may be unrealistic to expect that clinicians will necessarily obtain equivalent clinical examinations, to the extent that inter-observer variability is a negligible issue. LTM have been extensively used for questionnaire and scale standardization but not for clinical examinations; their application in this field could contribute to improve its standardization.

Finally, as referred above, LCM can be potentially improved through the use of concomitant variables. Concomitant variable LCM could account for known determinants of the relevant clinical findings and the a priori probability of the condition. These models could help mimicking the reasoning in clinical diagnosis, which departs from an a priori probability that influences the final clinical conclusion.

Aims

This thesis aimed to understand the role of latent models in the improvement and development of health outcomes measurement. The specific aims were:

1. To identify the incorrect beliefs and common knowledge about rheumatic diseases, from a sample of the general population, and to identify target groups for health education using latent models;

2. To calibrate the NYHA classification system between different observers, aspiring to increase its reliability;

    a. To assess if the calibration of the NYHA classification system between different observers increase its validity, thus reducing misclassification;

3. To identify patterns of symptoms and signs for HF, based on findings routinely collected in current clinical practice, and to evaluate their diagnostic value, taking into account the a priori likelihood of this syndrome;

    a. To evaluate the accuracy of BNP measurements in serum samples using the commercially available immunofluorometric assay to predict the BNP plasma concentration and to classify the individuals using the usual cut-off points;
    b. To develop a tool for the assessment of dietary calcium and vitamin D intakes in Portugal, and to evaluate the usefulness of non-dietary variables as predictors.
    c. To develop a tool that is able to match the classification obtained from concomitant variable LCM using a chain of simpler statistical methods.

Chapters

Chapter 1

Assessment of the general public's knowledge about rheumatic diseases: evidence from a Portuguese population-based survey

Severo et al. BMC Musculoskeletal Disorders 2010, 11:211
http://www.biomedcentral.com/1471-2474/11/211

BMC Musculoskeletal Disorders

RESEARCH ARTICLE Open Access

Assessment of the general public's knowledge about rheumatic diseases: evidence from a Portuguese population-based survey

Milton Severo[1,2*], Rita Gaio[3], Raquel Lucas[1,2], Henrique Barros[1,2]

Abstract

Background: To identify incorrect beliefs and common knowledge about rheumatic diseases in the general population.

Methods: Participants were selected during the follow-up of a representative cohort of adult population of Porto, Portugal; 1626 participants completed a questionnaire that included general knowledge items about rheumatic diseases.
Discrete and continuous latent variable models were used to identify knowledge flaws and the target groups. Odds ratios (OR) estimated by multinomial logistic regression, and 95% confidence intervals (95%CI) were computed to evaluate magnitude of associations.

Results: A continuous latent variable model identified two dimensions: one related to general beliefs (latent 1) and another concerning characteristics, treatment and impact of rheumatic diseases (latent 2). A 3-class latent variable model refined these results: the first class presented the lowest probabilities of correct answer for items associated with the first latent (mean of 39%), and the second class presented the lowest probabilities of correct answer for items with the second latent (mean of 62%). The third class showed the highest probability of a correct answer for almost all the items (mean of 79%). The age and sex standardized prevalence of the classes was 25.7%, 30.8% and 43.5%.
Taking class 2 as reference, class 1 was positively associated with the presence of rheumatic diseases (OR = 2.79; CI95% = (2.10-3.70)), with females (OR = 1.28 CI95% = (0.99-1.67)) and older individuals (OR = 1.04; CI95% = (1.03-1.05)), and was negatively associated with education (OR = 0.84; CI95% = (0.81-0.86)); class 3 was positively associated with education (OR = 1.03; CI95% = (1.00-1.05)) and the presence of rheumatic diseases (OR = 1.29; CI95% = (0.97-1.70)).

Conclusions: There are several knowledge flaws about rheumatic diseases in the general public. One out of four participants considered false general beliefs as true and approximately 30% did not have detailed knowledge on rheumatic disease. Higher education and the presence of disease contributed positively to the overall knowledge. These results suggest some degree of effectiveness of patient education, either conducted by health professionals or self-driven.

Background

Musculoskeletal diseases are among the most prevalent chronic conditions and constitute a major public health challenge for our aging societies [1]. Providing the general population and patients with good quality information is an important strategy in the management of chronic diseases. Knowledge leads to changes in attitudes and behaviours, and directly influences health status [2], and adequate information can promote self-management skills necessary for coping with the disease increasing adherence to therapy [3].

Patient participation in health care has been increasingly advocated: patients should be well informed about diagnosis and prognosis, and involved as fully as possible in disease management, namely in therapeutic decisions. A partnership should be formed between patients and health professionals, especially regarding chronic or life threatening diseases [4]. Involvement in medical

* Correspondence: milton@med.up.pt
[1]Department of Hygiene and Epidemiology, University of Porto Medical School, Porto, Portugal
Full list of author information is available at the end of the article

Severo et al. BMC Musculoskeletal Disorders 2010, 11:211
http://www.biomedcentral.com/1471-2474/11/211

decisions has been positively associated with patient satisfaction with health care [2,5] and improved health outcomes [6]. In rheumatoid arthritis cases, patient education has a positive effect in adherence to treatment, functional disability, global assessment, psychological well-being and depression [7,8].

Several studies showed that requirements for information are associated with patients age and education [9] but the overall level of information about rheumatic diseases is low among patients living with these conditions [9-11]. Although research targeting the general population is scarce, a survey of the Dutch population showed similar results [12] and raised the need for the identification of dimensions involved in knowledge about rheumatic diseases and the quantification of common knowledge in each specific demographic, social or pathology group. If nothing else, such quantification would benefit an education program targeted to musculoskeletal health.

By using a previously developed questionnaire designed to evaluate the overall knowledge level about rheumatic diseases in the general population [12], we aim to identify the incorrect beliefs and common knowledge about rheumatic diseases in a sample of the general population and to identify target groups for health education.

Methods

Participants were selected during the follow-up, conducted in 2005-2008, of a representative cohort of the non-institutionalized adult population of Porto, Portugal - the EpiPorto cohort. Recruitment at baseline was done using random digit dialling [13], selecting a single person over 17 years old in each of the identified households. Trained interviewers collected information, using a standard protocol that comprised multiple exams and questionnaires. Besides questions on social, demographic, clinical and behavioural characteristics participants completed an interviewer-administered questionnaire, comprising 17 statements about rheumatic diseases to be considered true or false. This was the Portuguese version of a Dutch questionnaire designed to evaluate knowledge regarding rheumatic diseases [12].

Cultural validation of the Portuguese version of the scale followed the usual methodology. The first stage consisted of a forward translation completed by 2 independent professional translators, yielding 2 initial Portuguese versions. Translators then synthesized the 2 versions to create a consensus version. Afterwards, 2 different independent translators completed a backward translation. Finally, an expert committee reviewed and compared the final Portuguese translation and the back translations to obtain a final version of the scale.

History of chronic rheumatic disease was self-reported, each individual indicating whether he/she had ever been diagnosed, by a doctor, with rheumatoid arthritis, ankylosing spondylitis, psoriatic arthritis, hand, hip or knee osteoarthritis, osteoporosis or lupus.

At baseline, 2485 participants were recruited, of whom 82 (3.3%) died before follow-up, 199 (8.0%) refused to be re-evaluated and 578 (23.3%) were unreachable by telephone or post. Therefore data from 1626 (65.4%) individuals were available for the present study. They presented a mean age of 58 (±15) years and 9 (±5) years of education; 1014 (62.4%) were women and 528 (32.6%) reported having been diagnosed with at least one rheumatic disease (table 1).

The local ethics committee (Hospital São João) approved the study protocol. All participants gave informed written consent to participate in the study, which was carried out in accordance with the Helsinki Declaration.

Statistical Analysis

Latent variable models were used to identify the incorrect beliefs and common knowledge about rheumatic diseases in the general population and to identify specific target groups.

In the present study, and given the binary structure of the data, two models were used: latent trait models (LTM) and latent class models (LCM).

LTM was used to identify dimensions in knowledge about rheumatic diseases in the general population, thus identifying what we considered to be incorrect beliefs and common knowledge. LTM assume that the performance of an individual while answering the items is

Table 1 Sample characteristics: socio-demographics information and history of rheumatic disease

		N (%)
Gender		
	Women	1014 (62.4)
	Men	612 (37.6)
Self-report Rheumatic Diseases		n (%)
Any		528 (32.6)
Rheumatoid arthritis		48 (3.0)
Ankylosing spondylitis		24 (1.5)
Hand osteoarthritis		211 (13.0)
Hip osteoarthritis		117 (7.2)
Knee osteoarthritis		247 (15.2)
Osteoporosis		265 (16.3)
Other		5 (0.3)
Age (years)		mean ± SD
		58 (14)
Education level (years)		9 (5)

Severo et al. BMC Musculoskeletal Disorders 2010, 11:211
http://www.biomedcentral.com/1471-2474/11/211

explained by one or more (continuous) variables, commonly called "latent variables". LTM is simply a Binary Data Factor Analysis that considers one or more factors.

Interpretation of the model is usually done considering the standardized factor loadings. Each of these expresses the correlation coefficient between the latent variable and an underlying continuous variable obtained from each item [14]. An association is classified as weak if the corresponding standardized loading is less than 0.30, moderate if it is between 0.30 and 0.70, and strong if it is higher than 0.70. Varimax rotation was applied to simplify the standardized factor loadings matrix.

LCA was used to uncover heterogeneous groups of individuals, thereby identifying the target groups. Latent class models (LCM) consider that the performance of an individual on the items is explained by K classes, commonly called "latent classes". Interpretation of the model is usually done by looking at the probabilities of positive response on each item conditional on class membership.

The global *goodness of fit* of the considered latent models was assessed through the likelihood ratio test, via parametric boostrapping - 100 samples - given the sample size and the number of estimated parameters [15]. Marginal *goodness of fit* was also evaluated through residuals inspection. The number of latent variables or classes in the considered LTM or LCM was the smallest providing the best *goodness of fit* to the given data. Correspondence analysis and principal component analysis from the item underlying continuous variables were also applied to confirm that number. Once the latent variables in the LTM were extracted, standardized Cronbach's alpha was estimated from the polychoric correlations between two binary variables [16], inter-item (tetrachoric) correlations mean and item-total biserial correlation coefficient [17] were used to evaluate the internal consistency of the group of items defining each variable.

Odds ratios (OR), estimated by multinomial logistic regression, and their respective 95% confidence intervals (95%CI) were used to measure the magnitude of associations between latent classes and the covariates sex, age, education and self-reported rheumatic diseases.

The distribution of the sample by latent classes was standardized by sex and 10-year age bands according to the 2001 census counts for the city of Porto. Significance level was fixed at 0.05.

Statistical analyses were performed using the software R 2.8.1 [18], and specifically, the ltm and lca command from, respectively, the ltm [19] and e1071 [20] packages.

Results

Among the 1626 participants, 1449 (95.9%) answered all the statements. The proportion of individuals that correctly answered each item ranged from 28% to 93%, corresponding to items 13 and 5, respectively, and the mean of correctly answered items was 10.5 (± 2.3) (table 2). The item-total biserial correlation coefficient computed for each item ranged from 0.31 to 0.60, corresponding to items 5 and 14, respectively. Cronbach's alpha was 0.628 and the inter-item correlation mean was 0.09.

Latent Trait Model

As no prior information on the number of latent variables to be held was available, a one-factor LTM was fit to the 17 items. Seven items showed a moderate-to-strong negative association with the latent variable while four presented a moderate positive association (table 2). A global test of *goodness-of-fit* (G^2 = 1901, $p < 0.01$) enhanced by the inspection of 2 by 2 marginal residuals showed a poor fit of this model.

A two-factor LTM with a varimax rotation presented moderate-to-strong associations between seven items and the first latent variable, and between nine items and the second latent variable. One item showed a weak association with the extracted factors. This model also presented a poor fit (G^2 = 5945, $p < 0.01$) and the inspection of marginal residuals revealed large pairwise residuals for four items (6, 10, 14 and 17) whose statement followed the structure "... is a kind of rheumatic disease." The *goodness-of-fit* was improved after elimination of those items (G^2 = 1901, p = 0.18) (table 2). This final model associates items 2, 3, 4, 12 and 13 with the first latent variable (LT1), and items 1, 5, 7, 8, 9, 15 and 16 with the second one (LT2), providing a standardized alpha of 0.700 and 0.630 and inter-item correlation of 0.32 and 0.20, respectively.

Latent Class Model

A latent class model with three classes was fit (G^2 = 2027.525, $p > 0.99$) to the 13 items considered in the above 2-factor LTM.

The first class presented the lowest probabilities of correct answer for items associated with the first latent (mean of 39%), and the second class presented the lowest probabilities of correct answer for items with the second latent (mean of 62%). The third class showed the highest probability of a correct answer for almost all the items (mean of 79%).

Four hundred and ninety individuals (31.0%) were classified in the first latent class, 443 (28.1%) in the second class and 645 (40.9%) in the third class.

The multinomial logistic regression showed that class, 1 when compared with class 2, was positively associated with the presence of rheumatic disease (OR = 2.79; CI95% = (2.10-3.70)) with female gender (OR = 1.28 CI95% = (0.99-1.67)) and older age (OR = 1.04; CI95% = (1.03-1.05)) and negatively associated with education

Severo et al. BMC Musculoskeletal Disorders 2010, 11:211
http://www.biomedcentral.com/1471-2474/11/211

Table 2 Proportion of correct answers and respective 95% confidence interval (95%CI) for each statement, standardized loadings for the 2-factor latent trait model (LTM) and probability of correct answer in the 2 and 3-classes latent class model (LCM)

			LTM		2-classes LCM		3-classes LCM		
Statement (correct option)	Proportion of correct answers	One Factor	Two factor Model		Class 1 (52%)	Class 2 (48%)	Class 1 (31.1%)	Class 2 (28.1%)	Class 3 (40.9%)
	% (95%CI)	Std.z1	Std. z1	Std. z2	%	%	%	%	%
1. A rheumatic disease is especially characterised by pain and stiffness in muscles and joints (t)	82 (80-84)	0.137	0.109	**0.315**	82	83	**84**	0.74	**88**
2. Rheumatic diseases are only seen in older women (f)	87 (85-89)	**-0.871**	**0.872**	0.000	99	74	68	**92**	**100**
3. In general, rheumatic patients should rest as much as possible and move as little as possible (f)	82 (80-83)	**-0.735**	**0.703**	0.060	95	68	63	**85**	**95**
4. Almost all rheumatic patients will finally end up in a wheelchair (f)	71 (68-73)	**-0.866**	**0.753**	-0.303	97	42	29	**86**	**94**
5. Medications for osteoarthritis cannot cure the disease, but can relieve pain and stiffness (t)	93 (92-95)	**0.529**	-0.153	**0.548**	91	96	**97**	87	**96**
6. Glandular fever is a kind of rheumatic disease (f)*	27 (25-29)	**-0.408**							
7. Rheumatoid arthritis is a rheumatic disease in which the joints are affected with inflammations (t)	89 (87-90)	0.027	0.265	**0.581**	90	88	**90**	79	**96**
8. Affected joints of rheumatic patients can be replaced with artificial joints (t)	54 (52-57)	0.285	-0.089	**0.333**	48	61	**65**	41	**56**
9. Osteoarthritis (wear and tear) is the most common kind of rheumatic disease (t)	88 (87-90)	**0.525**	-0.068	**0.619**	85	92	**95**	72	**97**
10. Multiple sclerosis (MS) is a rheumatic disease (f)*	37 (34-39)	**-0.655**							
11. People can die from the consequences of rheumatic disease (t)	47 (44-49)	0.200	-0.003	0.170	45	47	**49**	38	**51**
12. No kinds of rheumatic diseases can be cured (f)	37 (35-39)	**-0.387**	**0.332**	0.062	45	29	27	**34**	**49**
13. Rheumatoid arthritis is caused by poor diet, and cold and damp weather (f)	28 (25-30)	**-0.680**	**0.529**	-0.099	41	13	08	**34**	**39**
14. Ankylosing spondylitis is a kind of rheumatic disease (t)*	47 (45-50)	0.131							
15. There are more than 100 different kinds of rheumatic diseases (t)	66 (63-68)	**0.391**	-0.054	**0.520**	60	72	**78**	42	**76**
16. About one out of every twenty Portuguese people is being treated for a rheumatic disease (t)	84 (82-86)	**0.613**	-0168	**0.765**	77	92	**97**	59	**95**
17. Fibromyalgia is a rheumatic disease (t)*	38 (36-40)	0.137							

*eliminated from the model

(OR = 0.84; CI95% = (0.81-0.86)). Class 3, compared with class 2, was positively associated with education (OR = 1.03; CI95% = (1.00-1.05)) and with history of rheumatic disease (OR = 1.29; CI95% = (0.97-1.70)) (table 3).

After adjustment for all variables, age and gender effect were attenuated, while education and the history of rheumatic disease as the major determinants.

The age and sex standardized prevalence of latent classes were 25.7%, 30.8% and 43.5% in classes 1, 2 and 3, respectively.

Discussion

This survey revealed limited knowledge regarding rheumatic diseases at the general population level: there were difficulties regarding the identification of whether diseases where rheumatic (ankylosing spondylitis and fibromyalgia) or not (glandular fever and multiple sclerosis), and more than fifty percent believed that people with rheumatic diseases cannot be cured and cannot die from those illnesses (table 2). The latter finding is similar to that reported in other studies: in Canada [21] a study on women aged 65-90 years showed that only 36% agreed that health problems caused by osteoporosis can be life-threatening and another study carried out in US adults [22] found that only 63% correctly answer "false" to the statement "No medications can treat osteoporosis".

Considering these results, it is important to separate these knowledge domains in order to identify possible

Severo et al. BMC Musculoskeletal Disorders 2010, 11:211
http://www.biomedcentral.com/1471-2474/11/211

Table 3 Multinomial logistic regression model for latent classes by gender, age, education level, and self-report rheumatic diseases

		Class 3	Class 1	Class 3	Class 1
		Crude OR (95CI%)	Crude OR (95CI%)	OR** (95CI%)	OR** (95CI%)
Sex					
	Women	1.41 (1.10-1.80)	1.28 (0.99-1.67)	1.32 (1.02-1.72)	0.92 (0.68-1.25)
	Men	1	1	1	1
Age (years)		1.00 (0.99-1.00)	1.04 (1.03-1.05)	1.00 (0.99-1.01)	1.01 (1.00-1.02)
Education level (years)		1.03 (1.00-1.05)	0.84 (0.81-0.86)	1.03 (1.00-1.06)	0.86 (0.83-0.89)
Self-reported Rheumatic Pathologies					
	None	1	1	1	1
	At least one	1.29 (0.97-1.70)	2.79 (2.10-3.70)	1.34 (0.98-1.85)	1.77 (1.27-2.47)

*latent class 2 as reference; **adjusted for all variables

knowledge flaws and understand educational needs. Summarizing rheumatic diseases knowledge as a single domain is very limited. A recent review [21,23] identified significant limitations and constraints in measuring osteoporosis knowledge as a single domain, as it should include multi-dimensional aspects like causes or risk factors, prevention, consequences and treatment. A similar situation holds for the present questionnaire when one tries to summarize rheumatic diseases knowledge after a single value obtained from the 17 statements. The *goodness-of-fit* test suggest that the 2-factor LTM (13 items) is the best solution. The first latent was associated with the following statements, which probably represent wrong general beliefs: rheumatic diseases are more frequent in older women, rheumatoid arthritis is caused by poor diet, cold and damp weather, rheumatic patients should rest and move as little as possible and rheumatic diseases cannot be cured, and all rheumatic patients end up in wheelchairs. Items about aetiology, treatment and impact of the rheumatic diseases were related with second latent - which reveals specific knowledge.

The 3-class LCM showed that 25.7% agreed with the false general beliefs but did have specific knowledge (Class 1), 30.8% did not agree with the general beliefs and did not have specific knowledge (Class 2) and 43.5% did not agree with general beliefs and had specific knowledge (Class 3). Overall this suggests that almost 60% of the individuals had some flaws in their overall knowledge about rheumatic diseases.

Considering this relationship between the 3-class LCM and 2-factor LTM (Figure 1), we expected that a history of at least one rheumatic disease was the major determinant of the first class, since the prevalence of rheumatic diseases is higher in women and in older individuals, and we also expected that this association would be extended to sex and age; for the second class, the expected major determinant would be education; and for the third class the presence of both characteristics: rheumatic disease history and higher educational level. The multinomial logistic regression confirmed these expectations: class 1 when compared with class 2 was positively associated with the history of rheumatic disease, female sex and old age and negatively associated with education; and class 3 when compared with class 2 was positively associated with education and the history of rheumatic disease. As expected, after adjustment for the presence of disease and education, the effect of age and sex was attenuated. Therefore the major determinants of knowledge were education and the presence of rheumatic diseases. Although we did not measure the effectiveness of patient education, these results suggest some degree of effectiveness of such, either conducted by health professionals or self-driven.

The effectiveness of futures educational programs about rheumatic diseases directed to general population/patient population might be improved by targeting the eldest and low educated fraction of the population to counteract wrong general beliefs. As reported in other studies education is not one programme, but a strategy that is tailored to each population, the programme should remodel the interpretative structures of individuals because providing educational information, by itself, has no beneficial impact [24].

There are a number of limitations to this study. First, this sample of the study was significantly older and had higher frequency of women when compared with census counts for the city of Porto, which could lead to a selection bias. However we have tried to minimize this by estimating the age and sex standardized prevalence of latent classes. Additionally, we used only a pool of 13 items to indentify incorrect beliefs and common knowledge and targets groups in the overall knowledge about rheumatic diseases. This is somehow limited, as the moderated alpha shows, considering that we are trying to measure a multi-dimensional concept, with a multitude of possible items.

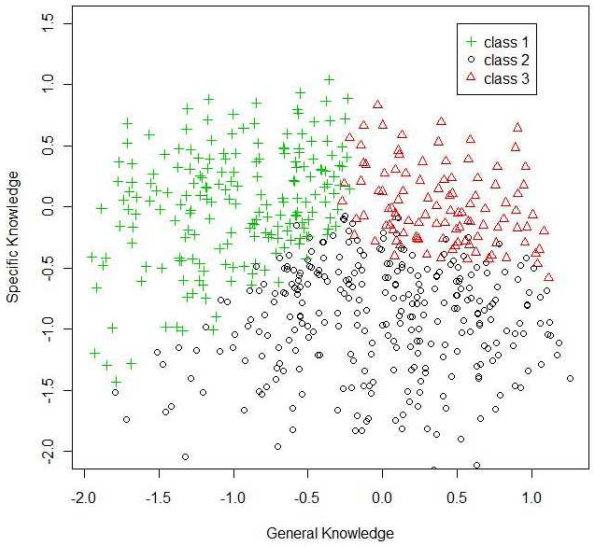

Figure 1 The latent classes of the 3-latent class model allocated in the 2 dimensions (general and specific knowledge) of 2-factor latent trait model (13 items).

Conclusions

The use of latent models applied to this specific scale, we were able to provide evidence for identification of different knowledge domains regarding rheumatic diseases in the general population. Additionally, this method was instrumental to identify relevant target groups for educational programmes.

This study showed that there are several knowledge flaws about rheumatic diseases in the general population. One out of four considered the false general beliefs as true and approximately 30% did not have detailed knowledge on rheumatic disease. Higher education and the presence of disease contributed positively to the overall knowledge. However there is a major flaw in identifying what is and what is not a rheumatic disease in the general population.

Author details
[1]Department of Hygiene and Epidemiology, University of Porto Medical School, Porto, Portugal. [2]Institute of Public Health of the University of Porto, Porto, Portugal. [3]Department of Mathematics, University of Porto Science School; Center of Mathematics of the University of Porto.

Authors' contributions
MS participated in the study design, performed the statistical analysis and helped to draft the manuscript. RG performed the statistical analysis and helped to draft the manuscript. RL and HB participated in the study design and helped to draft the manuscript.
All authors read and approved the final manuscript

Competing interests
The authors declare that they have no competing interests.

Received: 25 November 2009 Accepted: 16 September 2010
Published: 16 September 2010

References

1. Hazes JM, Woolf AD: The bone and joint decade 2000-2010. *The Journal of rheumatology* 2000, 27(1):1-3.
2. Brekke M, Hjortdahl P, Kvien TK: Involvement and satisfaction: a Norwegian study of health care among 1,024 patients with rheumatoid arthritis and 1,509 patients with chronic noninflammatory musculoskeletal pain. *Arthritis and rheumatism* 2001, 45(1):8-15.
3. Taal E, Rasker JJ, Wiegman O: Group education for rheumatoid arthritis patients. *Semin Arthritis Rheum* 1997, 26(6):805-816.
4. Charles C, Whelan T, Gafni A: What do we mean by partnership in making decisions about treatment? *BMJ (Clinical research ed)* 1999, 319(7212):780-782.
5. Kjeken I, Dagfinrud H, Mowinckel P, Uhlig T, Kvien TK, Finset A: Rheumatology care: involvement in medical decisions, received information, satisfaction with care, and unmet health care needs in patients with rheumatoid arthritis and ankylosing spondylitis. *Arthritis and rheumatism* 2006, 55(3):394-401.
6. Stewart M, Brown JB, Donner A, McWhinney IR, Oates J, Weston WW, Jordan J: The impact of patient-centered care on outcomes. *The Journal of family practice* 2000, 49(9):796-804.
7. Riemsma RP, Taal E, Kirwan JR, Rasker JJ: Systematic review of rheumatoid arthritis patient education. *Arthritis and rheumatism* 2004, 51(6):1045-1059.

8. Hill J, Bird H, Johnson S: **Effect of patient education on adherence to drug treatment for rheumatoid arthritis: a randomised controlled trial.** *Annals of the rheumatic diseases* 2001, **60(9)**:869-875.
9. Neame R, Hammond A, Deighton C: **Need for information and for involvement in decision making among patients with rheumatoid arthritis: a questionnaire survey.** *Arthritis and rheumatism* 2005, **53(2)**:249-255.
10. Hennell SL, Brownsell C, Dawson JK: **Development, validation and use of a patient knowledge questionnaire (PKQ) for patients with early rheumatoid arthritis.** *Rheumatology (Oxford, England)* 2004, **43(4)**:467-471.
11. Hill J, Bird HA, Hopkins R, Lawton C, Wright V: **The development and use of Patient Knowledge Questionnaire in rheumatoid arthritis.** *British journal of rheumatology* 1991, **30(1)**:45-49.
12. Wardt EM, Taal E, Rasker JJ: **The general public's knowledge and perceptions about rheumatic diseases.** *Annals of the rheumatic diseases* 2000, **59(1)**:32-38.
13. Ramos E, Lopes C, Barros H: **Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction.** *Annals of epidemiology* 2004, **14(6)**:437-441.
14. Bartholomew D, Steele F, Moustaki I, Galbraith J: **The Analysis and Interpretation of Multivariate Data for Social Scientists.** *Chapman & Hall/CRC* 2002.
15. Bartholomew DJ, Knott M: **Latent Variable Models and Factor Analysis: Hodder Arnold.** 1999.
16. Cronbach L: **Coefficient alpha and the internal structure of tests.** *Psychometrika* 1951, **16**:297-333.
17. Drasgow F: **Polychoric and polyserial correlations.**Edited by: Kotz S, Johnson N 1986, 68-74.
18. **R: A Language and Environment for Statistical Computing.** Vienna, Austria: {R Development Core Team} 2008.
19. Rizopoulos D: **ltm: An R package for latent variable modeling and item response theory analyses.** *Journal of Statistical Software* 2006, **17(5)**:1-25.
20. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A: **Misc Functions of the Department of Statistics (e1071).** *TU Wien. R package version* 2005, **1**:5-7.
21. Cadarette SM, Gignac MA, Beaton DE, Jaglal SB, Hawker GA: **Psychometric properties of the "Osteoporosis and You" questionnaire: osteoporosis knowledge deficits among older community-dwelling women.** *Osteoporos Int* 2007, **18(7)**:981-989.
22. Solomon DH, Finkelstein JS, Polinski JM, Arnold M, Licari A, Cabral D, Canning C, Avorn J, Katz JN: **A randomized controlled trial of mailed osteoporosis education to older adults.** *Osteoporos Int* 2006, **17(5)**:760-767.
23. Werner P: **Knowledge about osteoporosis: assessment, correlates and outcomes.** *Osteoporos Int* 2005, **16(2)**:115-127.
24. Ramos-Remus C, Salcedo-Rocha AL, Prieto-Parra RE, Galvan-Villegas F: **How important is patient education?** *Baillieres Best Pract Res Clin Rheumatol* 2000, **14(4)**:689-703.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1471-2474/11/211/prepub

doi:10.1186/1471-2474-11-211
Cite this article as: Severo *et al.*: Assessment of the general public's knowledge about rheumatic diseases: evidence from a Portuguese population-based survey. *BMC Musculoskeletal Disorders* 2010 11:211.

Chapter 2

Indirect calibration between clinical observers - application to the New York Heart Association functional classification system

Severo et al. BMC Research Notes 2011, 4:276
http://www.biomedcentral.com/1756-0500/4/276

BMC Research Notes

RESEARCH ARTICLE Open Access

Indirect calibration between clinical observers - application to the New York Heart Association functional classification system

Milton Severo[1,2*], Rita Gaio[3], Patrícia Lourenço[4], Margarida Alvelos[4], Paulo Bettencourt[4] and Ana Azevedo[1,2,4]

Abstract

Background: Previous studies showed an inter-observer agreement for the NYHA classification of approximately 55%. The aim of this study was to calibrate the New York Heart Association (NYHA) classification system between observers, increasing its reliability.

Results: Among 1136 community-dwellers in Porto, Portugal, aged ≥ 45 years, 265 reporting breathlessness answered a 4-item questionnaire to characterize symptom severity. The questionnaire was administered by 7 physicians who also classified the subject's functional capacity according to NYHA. Each subject was assessed by one physician. We calibrated NYHA classifications by the concurrent method, using 1-parameter logistic graded response model. Discrepancies between observers were assessed by differences in ability thresholds between NYHA classes I-II and II-III. The ability estimated by the model was used to predict the NYHA classification for each observer.

Estimates of the first and second thresholds for each observer ranged from -1.92 to 0.46 and from 1.42 to 2.30, respectively. The agreement between estimated ability and the observers' NYHA classification was 88% (kappa = 0.61).

Conclusions: The study objectively indicates the main reason why several studies have reported low inter-observer is the existence of discrepant thresholds between observers in the definition of NYHA classes. The concurrent method can be used to minimize the reliability problem of NYHA classification.

Keywords: dyspnea, physical exertion, questionnaires, New York Heart Association, calibration, reliability, equating

Background

The New York Heart Association (NYHA) functional classification was originally conceptualized and described in 1928 and most recently updated in 1994 as a method of assessing functional disability induced by cardiac diseases in patients encountered in clinical practice [1]. The NYHA system was designed for clinical assessment of patients by physicians in 4 classes (I, II, III or IV) on the basis of the patient's limitations in physical activities caused by cardiac symptoms. The NYHA classification is derived largely by inference from history and/or observation of the patient in certain physical activities, and occasionally by direct or indirect measurement of cardiac function in response to standardized exercises. There was an attempt to increase the objectivity of the NYHA classification by adding an objective assessment, based on measurements such as electrocardiogram, stress test, X-ray and echocardiogram. Despite this attempt, the NYHA classification remains essentially subjective [2]. The class a clinician decides to assign a patient to depends on the clinician's interpretation of what is "ordinary" physical activity, "slight" and "marked" limitations. This results in a high inter-observer variability. Previous studies showed an inter-observer agreement for the NYHA classification of approximately 55% [3,4]. Consequently the use of NYHA classification as an outcome measure in clinical research is rather poor. However this classification system has been widely used in

* Correspondence: milton@med.up.pt
[1]Department of Clinical Epidemiology, Predictive Medicine and Public Health, University of Porto Medical School, Porto, Portugal
Full list of author information is available at the end of the article

clinical epidemiology studies as an inclusion criterion and also as an outcome measure [2]. It is also used in routine clinical practice.

The aim of this study was to calibrate the NYHA classification system between different observers, aspiring to increase its reliability, by quantifying the discrepancy in thresholds in functional capacity that lead an observer to assign a NYHA class to a patient.

Methods

Participants were selected within the first follow-up of a cohort, representative at baseline of the non-institutionalized adult population of Porto, Portugal - the EPIPorto cohort study. At baseline, households were selected by random digit dialling [5]. After the identification of a household, permanent residents were characterized according to age and gender, and one individual aged 18 years or older was randomly selected and invited to visit our department for an interview and physical examination. If there was a refusal, replacement was not allowed within the same household.

Trained interviewers collected information, using a standard protocol that comprised questions on social, demographic, clinical and behavioural characteristics. At baseline, 2485 participants were recruited. Between October 2006 and July 2008, all participants aged ≥ 45 years were eligible to a systematic evaluation, at our department, of measures of cardiac structure and function, which included a cardiovascular clinical history and physical examination, and a transthoracic echocardiogram.

Among 2048 eligible to this study, 134 (6.5%) had died, 198 (9.7%) refused to be re-evaluated and 580 (28.3%) were lost to follow up (unreachable by telephone or post). Therefore 1136 (55.4%) individuals aged ≥ 45 years were assessed by 8 physicians experienced in the management of heart failure patients.

At the standardized clinical interview applied by these physicians, subjects who reported to have breathlessness (n = 265; 23.3%) were presented to a 4-item questionnaire on functional capacity to characterize the severity of symptoms: 1) whether breathlessness is felt when walking on steep plane, horizontal plane or at rest; 2) distance walked until perception of breathlessness; 3) sets of stairs (10-15 steps) climbed until perception of breathlessness; 4) whether mild, moderate or intense efforts are necessary to elicit breathlessness. These will hereafter be referred to as "anchor items".

The same physician administered the questionnaire and classified the subject's functional capacity using the NYHA classification. This classification, defined by each physician for each subject, will hereafter be referred to as "target items". The assessment of the NYHA classification was carried out after the administration of the 4 anchor items. NYHA class IV was aggregated to class NYHA III because only one individual was classified in NYHA IV.

The Medical Outcomes Study Short Form-36 (SF36) was used to assess health-related quality of life [6]. The scale had been previously translated and the adapted Portuguese version was validated [7]; each sub-domain of the SF-36 is scored from 0 to 100, with increasing values representing better health. Participants completed a physical activity questionnaire designed to estimate usual individual daily energy expenditure, focused on the activity in the past year. Time spent in a variety of activities per day, including work, transport to and from work, household chores, sports, sedentary leisure time and sleep, was self-reported and activity intensity categorized as very light, light, moderate and heavy with a corresponding average of 1.5, 2.5, 5.0 and 7.0 METs respectively, where one MET is equal to the energy expended at the basal metabolic rate or at rest [8]. A severity scale was applied to measure fatigue [9], with increasing values representing higher severity.

The local ethics committee (Hospital São João) approved the study and participants provided written informed consent.

Statistical analysis

Different correlation coefficients were used to evaluate the magnitude of the association between anchor items and the target items (NYHA classifications): correlations between two (artificial) ordinal variables were evaluated through polychoric correlations, and between interval and (artificial) ordinal variables through polyserial correlations.

Exploratory factor analyses (weighted least square) on the 4 ordinal anchor items combined with each target item was used to evaluate homogeneity (i.e., to confirm there was a single latent variable) of the items and the Cronbach's alpha was used to measure the reliability [10]. The global goodness of fit of the underlying structure with 1 factor was evaluated using the comparative fit index (CFI) recommended when N < 250 [11].

The convergent and divergent validity of the 4 anchor items was assessed through the correlation between the questionnaire's raw score and the 4 physical dimensions of the health-related quality of life scale SF36 (physical function, role physical, bodily pain and general health perception), a scale for fatigue and daily physical activity. The raw score was estimated by the sum of all anchor items.

Calibration

Each set of individuals assessed by each physician was considered as a group. Calibration of NYHA classification across different groups was performed by the concurrent method. Concurrent calibration involves

Severo et al. BMC Research Notes 2011, 4:276
http://www.biomedcentral.com/1756-0500/4/276

estimating item and ability parameters in all groups simultaneously, i.e., by combining data from these distinct groups. Items not taken by one of the groups are treated as either not reached or missing [12]. Given the ordinal nature of the items, this is a particular use of the 1-dimensional logistic graded response model (GRM) from item response theory (IRT). Fit of the model was based on approximate marginal Maximum Likelihood. The four patient items were used as anchor items and the 7 obtained NYHA classifications as target items (observer 3 NYHA classification was eliminated for the GRM and dyspnea item was aggregated in two classes 0 vs. 1 and 2 because of the small sample size).

Exploratory factor analysis (EFA) supported that only 1 dimension was reflected in the ordinal items. Thus, 1-dimensional logistic graded response models (GRM) from item response theory (IRT) were used [13]. These models assume that the performance of an individual on the items is explained by only one (standard normal) variable, commonly called "ability". "Ability" is the term that denotes the unobserved hypothetical variable (a latent trait) subjacent to graded response models. In our study, ability refers to the functional capacity of the subject that we are trying to characterize. Higher ability values represent worse functional capacity (more severe symptoms). In the graded response models, each item is described by a set of curves, item operation characteristic curves (IOCC). The item operation characteristic curves for category k represent the probability of endorsing categories higher than k conditional on subject's ability.

The item operation characteristic curves of an item are characterized by several parameters: the slope (discrimination), which is the same for all categories, and the thresholds (difficulty), which are as many as the number of categories minus one. For example, one item with 3 categories has 3 category characteristic curves, one slope and two thresholds: t_1 to define I versus II-IV and t_2 to define I-II versus III-IV.

The threshold parameter between two categories represents the ability value at which the probability of indicating the highest of these two or higher is 50%. So, the threshold parameters are expressed in the same scale as the ability. The slope parameter indicates how well an item is able to discriminate individuals with ability values near the respective threshold. The slope parameter may also be interpreted as describing how an item may be related to the ability. The steeper the slope the higher is the item discrimination. We fitted a 1-parameter logistic (1-PL) GRM assuming a unique slope (discrimination parameter) for all items.

Quality of the calibration

The thresholds estimated for each observer were used as ability cut-off points to predict the observed NYHA classifications, this procedure permitted to assess the ability fit with the target items and the agreement between observers.

In the first case NYHA predictions were sample-specific, i.e., the NYHA predictions were estimated separately for each sample assessed by each of the observers and compared with the observed NYHA classifications.

In the second case NYHA prediction were not sample-specific, i.e., all individuals were classify using the thresholds estimated for each observer regardless of the observer that assessed each individual and compare with each other.

The agreement was assessed with both the absolute agreement and the Cohen's weighted kappa coefficient. Guidelines for interpreting kappa statistics suggest that values between 0.81-1.00 indicate almost perfect agreement, 0.61-0.80 substantial agreement, 0.41-0.60 moderate agreement, 0.21-0.40 fair agreement, and values less than 0.21 are poor or slight agreement [14].

Statistical analyses were performed using the software R 2.12.1 [15], and specifically, the ltm [16] and plink packages and the Mplus software [17].

Results

The number of individuals assessed by each observer ranged from 10 (3.7%) to 80 (30.4%). The participants were similar by each observer in terms of sex, education and clinical history, systolic blood pressure but showed significant differences in age, body mass index and diastolic blood pressure (Table 1). The NYHA classification showed significant differences by observer, with the prevalence of class I, II and III/IV ranging from 9.3 to 58.8%, 29.2 to 83.3% and 4.8% to 20.0%, respectively. Missing data for the anchor items was equal to or less than 3.0% for all items with the exception of the item 2, 12%. The distribution of NYHA classification in the sample was 85 (33.3%), 147 (57.6%) and 23 (9.0%) for class I, II and III-IV, respectively.

Homogeneity

The polychoric correlations between each item and the NYHA classification of all observers were positive and statistically significant (Table 2).

Exploratory factor analysis conducted separately for the 7 observers combined with the 4 anchor items revealed a first factor that accounted more than 70% of the variance, and the first eigenvalue was 3.4 times larger than the second eigenvalue. The fit index met the criteria to support the 1 factor structure, the CFI ranged from 0.934 to 1.00 with the exception of observer 1 than obtained a value of 0.687. The Cronbach's alpha ranged from 0.607 and 0.839 for the 4 common items combined with each observer NYHA classification (Table 3).

Table 1 Characteristics of the study sample by observers

Observer	total	1	2	3	4	5	6	7	8	P
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	
Total	265 (100)	21 (7.7)	61 (22.5)	10 (3.7)	80 (30.4)	21 (7.7)	18 (6.6)	24 (8.9)	28 (10.3)	< 0.001
Men	68 (25.7)	7 (33.3)	21 (34.4)	5 (50.0)	14 (17.5)	5 (23.8)	4 (22.2)	7 (29.2)	5 (17.9)	0.180
History of myocardial infarction	19 (7.2)	1 (4.8)	4 (6.6)	1 (10.0)	7 (8.7)	1 (5.0)	0 (0.0)	3 (12.5)	2 (7.1)	0.870
History of angina	32 (12.1)	3 (14.3)	12 (19.2)	1 (10.0)	8 (10.0)	1 (4.8)	1 (5.9)	3 (12.5)	3 (10.7)	0.606
History of heart failure	43 (16.3)	3 (14.3)	13 (21.3)	1 (10.0)	15 (18.8)	1 (4.8)	1 (5.9)	5 (20.8)	4 (14.3)	0.582
Left ventricular systolic dysfunction	**18 (7.1)**	**1 (4.8)**	**6 (10.2)**	**0 (0.0)**	**5 (6.6)**	**0 (0.0)**	**0 (0.0)**	**3 (13.0)**	**3 (13.0)**	**0.571**
NYHA classification										
I	85 (33.3)	10 (47.6)	26 (43.3)	3 (30.0)	7 (9.3)	10 (58.8)	2 (11.1)	15 (62.5)	12 (42.9)	< 0.001
II	147 (57.6)	10 (47.6)	29 (48.3)	5 (50.0)	61 (81.3)	5 (29.4)	15 (83.3)	7 (29.2)	14 (50.0)	
III and IV	23 (9.0)	1 (4.8)	5 (8.3)	2 (20.0)	7 (9.3)	2 (11.8)	1 (5.6)	2 (8.3)	2 (7.1)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	
Age (years)	65.8 (9.7)	67.5 (9.1)	67.6 (10.3)	66.8 (10.5)	65.6 (9.1)	68.6 (7.9)	62.1 (7.9)	60.0 (8.7)	67.1 (10.3)	0.021
Systolic blood pressure (mmHg)	136 (21)	128 (19)	137 (23)	142 (15)	135 (21)	142 (22)	136 (23)	138 (20)	137 (21)	0.574
Diastolic blood pressure (mmHg)	80 (12)	81 (10)	79 (13)	84 (13)	77 (10)	77 (12)	88 (15)	82 (10)	81 (14)	0.025
Body mass index (kg/m^2)	30 (5.5)	31.4 (7.1)	28.7 (4.5)	28.8 (6.3)	29.4 (4.7)	27.9 (4.6)	33.0 (6.1)	32.9 (5.9)	31.2 (6.4)	0.001
	Med (IQR)	Med (IQR)	Med (IQR)	Med (IQR)	Med (IQR)	Med (IQR)	Med (IQR)	Med (IQR)	Med (IQR)	
Education (years)	4 (5)	4 (4)	4 (3)	7 (7)	4 (5)	4 (7)	5 (3)	4 (6)	4 (0)	0.368

SD: standard deviation.
Med: median.
IQR: interquartile range.

Validity of the anchor items

The raw score on the 4-item questionnaire showed a positive correlation with NYHA classification (Table 2). The raw score showed a moderate negative correlation with the 4 physical dimensions of SF36, a positive correlation with the severity of fatigue and no association with total physical activity. The correlations between NYHA classification and SF36, fatigue and physical activity were similar to the ones obtained with the raw score of the questionnaire (Table 4).

Concurrent calibration

Inspection of the thresholds between classes I-II and II-III provided information about the ability extremes (Table 5). Estimates for the first threshold ranged from -1.92 to 0.69 (median = -0.10) standard deviations of ability, and for the second threshold ranged from 0.27 to 2.26 (median = 1.69). Observers 4 (t1 = -1.92) and 7 (t1 = 0.43) showed the lowest and highest first threshold, respectively, while for the second threshold it was observer 4 (t2 = 1.24) and 8 (t2 = 2.26), respectively (Figure 1). "Effort" was the anchor item whose first threshold was closest to the median of the observers' first threshold, that is, this anchor item is considered to distinguish classes I and II. The "effort" anchor item second threshold was closest to the median of the observers' second thresholds, that is, these anchor items are considered to distinguish classes II and III.

The results of the calibration with the 1-PL graded response model showed that the observers and the patient anchor items showed a high discrimination (β = 2.27, standard error = 0.176).

Quality of the calibration

The agreement between the NYHA classification according to the thresholds estimated for each observer for the ability and the observers' NYHA classification observed (target items) ranged from 76 to 89% with a median of 88%, the weighted Kappa ranged from 0.42 to 0.83 with median of 0.61 (table 6). This means that after taking into account the discrepancies in thresholds between observers, their NYHA classification is well predicted by the ability, with a substantial agreement.

The agreement between observers predicted classifications for all individuals according to the thresholds estimated for each observer for the ability ranged from 30 to 97% with a median of 65%, the weighted Kappa ranged from 0.00 to 0.94 with median of 0.21. This means

Table 2 Score of each anchor item, the distribution of the items and the polychoric correlation of each item with NYHA classification

	Raw Score	N (%)	r+
Total		**N = 265**	
Do you usually have breathlessness or difficulty breathing? ("dyspnea")		N = 263 (99.2)	
Yes, when walking on steep plane	0	185 (70.3)	0.57
Yes, when walking on the horizontal plane	1	68 (25.9)	
Yes, even at rest	2	10 (3.8)	
If yes, how long can you walk before you have to stop? ("distance")*		N = 232 (87.5)	
0-100 metres	2	97 (41.8)	0.33
101-500 metres	1	86 (37.1)	
501-2500 metres	0	49 (21.1)	
If yes, after how many sets of stairs (10-15 steps) do you have to stop? ("stairs")*		N = 258 (97.3)	
1 set	2	93 (36.0)	0.66
2 sets	1	75 (29.1)	
3 or more sets	0	90 (34.9)	
If yes, in your view, what level of effort induces breathlessness? ("effort")		N = 257 (97.0)	
Great efforts	0	105 (40.9)	0.67
Average efforts	1	89 (34.6)	
Small efforts	2	63 (24.5)	
		Median (IQR)	
Raw score (0-8)		4 (2-6)	0.62

* inverse order for the final score
+ polychoric correlation between each item and NYHA classification of all observers
IQR: interquartile range.

that without taking into account the discrepancies in thresholds between observers, the agreement between NYHA observers classification is fair.

Discussion

Several studies have shown that the NYHA classification is valid but not reproducible [2,4], and associated with symptom burden, quality of life, exercise capacity, and increased risk of ischemic stroke [18-20]. Nevertheless, the NYHA classification was originally designed as a clinical, not a research tool. Although much has been written regarding the limitations of the NYHA of classification as an outcome measure [21], investigators continue to use it in clinical research. The popularity of the NYHA classification system is based on its simplicity [4]. Any system that might replace it should be more accurate without being more complex. So the aim of this study was not to build a new system but to improve the NYHA system. To do so, we used IRT models to equate and calibrate a large number of observers on the

Table 3 Exploratory factor analysis and internal consistency conducted separately for the 7 observers NYHA classification (target items) and combined with the 4 anchor items

Item	Eigenvalue1	Eigenvalue2	CFI[1]	Alpha Cronbach
Observer 1	3.261	1.856	0.687	0.799
Observer 2	3.383	0.929	0.988	0.778
Observer 3	—	—	—	—
Observer 4	3.754	0.803	0.992	0.790
Observer 5	3.972	0.864	0.987	0.839
Observer 6	3.056	1.406	0.934	0.607
Observer 7	3.363	1.178	0.984	0.732
Observer 8	3.368	0.863	1.000	0.791

[1]Comparative Fit Index.

Table 4 Correlation between the raw score (sum of 4 items) and NYHA with fatigue scale, the daily physical activity, the 4 physical sub-dimensions (physical function, role physic, pain and health perception) and the general physical function of Short Form 36

	Raw Score	NYHA
Fatigue scale	0.36	0.42
Total physical activity (mets)	-0.02	-0.02
General physical health (SF36)	-0.40	-0.40
Physical function	-0.33	-0.36
Role physical	-0.27	-0.30
Bodily pain	-0.35	-0.38
General health perception	-0.47	-0.44

Severo et al. BMC Research Notes 2011, 4:276
http://www.biomedcentral.com/1756-0500/4/276

Table 5 One-dimensional 2 parameter logistic graded response model with equal discrimination parameters across items

	Threshold 1[1] t_1 (se)	Threshold 2[2] t_2 (se)	Item Discrimination[3] B (se)
"Dyspnea"	0.688 (0.106)	—	2.268 (0.176)
"Distance"	-0.349 (0.104)	0.278 (0.094)	2.268 (0.176)
"Stairs"	-0.493 (0.102)	0.459 (0.390)	2.268 (0.176)
"Effort"	-0.316 (0.099)	0.865 (0.123)	2.268 (0.176)
Observer 1	-0.549 (0.283)	1.503 (0.287)	2.268 (0.176)
Observer 2	-0.099 (0.168)	1.692 (0.631)	2.268 (0.176)
Observer 3	—	—	—
Observer 4	-1.920 (0.253)	1.420 (1.237)	2.268 (0.176)
Observer 5	0.331 (0.312)	1.671 (0.309)	2.268 (0.176)
Observer 6	-1.211 (0.449)	2.178 (0.728)	2.268 (0.176)
Observer 7	0.430 (0.269)	1.804 (3.418)	2.268 (0.176)
Observer 8	0.339 (0.247)	2.263 (0.305)	2.268 (0.176)

[1]Threshold 1 - level of ability above which 50% of subjects were NYHA class II-III

[2]Threshold 2 - level of ability above which 50% of subjects were NYHA class III

[3]Item discrimination - represents the slope of the item characteristic curves at the value of the threshold and indicates the extent to which the item is related to the ability

se: standard error.

same scale; by doing so, we were able to identify observers with lower and higher thresholds for classification, as well as to understand the relations with anchor items across the ability continuum, and to improve the NYHA classification system.

The present study objectively indicates the main reason why several studies have reported low inter-observer reliability and, consequently, the limited usefulness of the NYHA classification as an outcome measure. The main reason is the existence of discrepant thresholds

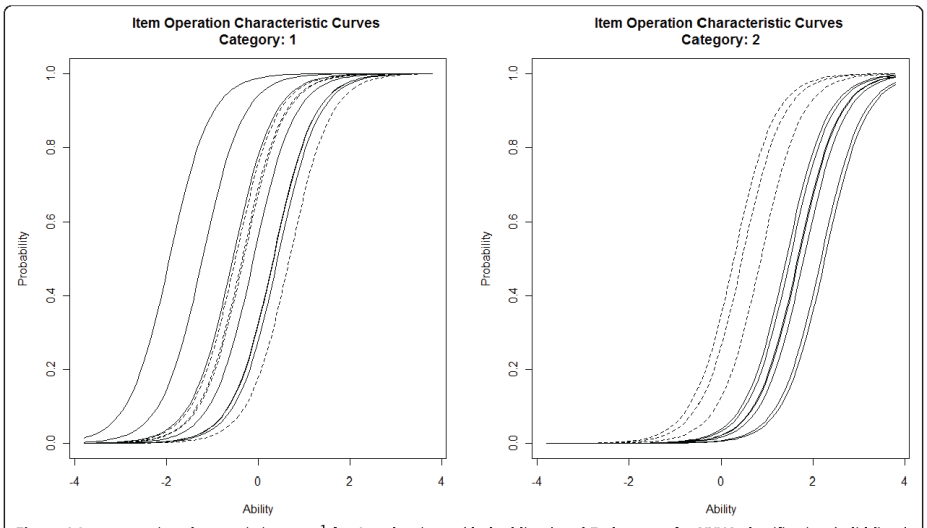

Figure 1 Item operation characteristic curves[1] for 4 anchor items (dashed lines) and 7 observers for NYHA classification (solid lines). [1]Item operation characteristic curves (IOCC) for category k represent the probability of endorsing categories higher than k conditional on subject's ability.

Table 6 Agreement between the observers and between the observers and the ability estimated by the concurrent calibration

	Observer 1	Observer 2	Observer 4	Observer 5	Observer 6	Observer 7	Observer 8	Ability
Observer 1	.....	82.89[1]	70.72[1]	65.40[1]	72.24[1]	59.32[1]	62.36[1]	0.76[1]
Observer 2	0.68[2]	—	53.61[1]	82.51[1]	58.17[1]	76.43[1]	79.47[1]	0.88[1]
Observer 4	0.20[2]	0.09[2]	.....	36.12[1]	90.87[1]	30.04[1]	33.08[1]	0.87[1]
Observer 5	0.43[2]	0.69[2]	0.06[2]	—	40.68[1]	93.92[1]	96.96[1]	0.88[1]
Observer 6	0.21[2]	0.12[2]	0.00[2]	0.06[2]	—	38.4[1]	42.21[1]	0.89[1]
Observer 7	0.33[2]	0.57[2]	0.01[2]	0.87[2]	0.05[2]	.....	96.2[1]	0.88[1]
Observer 8	0.36[2]	0.62[2]	0.00[2]	0.94[2]	0.06[2]	0.92[2]	—	0.79[1]
Ability	0.56[2]	0.80[2]	0.42[2]	0.83[2]	0.47[2]	0.77[2]	0.61[2]	.....

[1]Upper triangle shows the % of absolute agreement
[2]Lower triangle shows the weighted kappa.

between observers in the definition of NYHA class I, II and III individuals. Although the observers in study were experienced physicians well trained in the management of heart failure, there were still discrepancies between their (subjective) evaluations.

The focus should therefore be on the identification of differences between the evaluations of the observers and on the calibration of those classifications.

Although intra-observer reliability is more important to interpret changes in NYHA class in the individual patient who is assessed repeatedly by the same physician, inter-observer variability is of special concern when patients are assessed by different physicians. This is particularly important, in practice, in unscheduled visits to the clinic or the emergency department, where patients are not assessed by their usual attendant. These unscheduled visits are usually due to worsening symptoms and an increase in NYHA class, in comparison with the previous clinical state, is used as a criterion for clinical decisions such as hospital admission and intensity of therapy adjustment such as use of intravenous medication.

Therefore, in each setting the NYHA classification is to be used, it would be useful to identify the differences between the assessments of the observers and calibrate their classifications. For the calibration with the IRT methodology to be possible, a set of anchor items is needed. These items should be reliable and valid. In this sample, the 4 anchor items combined with each target item showed good homogeneity (strong first factor) and reliability (alpha > 0.61). Furthermore, these items showed content validity on the basis of a previous study [3], which concluded that the self-reported distance (70%) and difficulty in climbing stairs (60%) were the items more commonly used by senior cardiologists and trainees in cardiology to classify patients in NYHA classes. Our study showed that these anchor items had a strong association with the NYHA classification and that had a similar association with scales that measure related constructs. So these results confirm the reliability of the anchor items and their validity to assess the same construct as the NYHA classification.

The improvement in the absolute agreement (65% to 88%) between the ability scale predictions of the NYHA classification between observers and the ability scale predictions of the NYHA classification with the observers' NYHA classifications observed, show how the subjectivity of the thresholds can affect the reliability of the NYHA classification. At the same time this improvement confirms the quality of the calibration obtained.

The calibration methodology can be useful to improve the reliability between observers in clinical practice and research settings. In clinical practice it is possible to use the anchor items' relations with ability to explain the differences between observers and give guidelines to improve the inter-observers reliability. For example, if we wanted to calibrate the threshold between NYHA I and II for all observers, we would advise all observers to use endorsement of the second category of the "Effort" item for the definition of class NYHA II. Similarly if we wanted to calibrate the threshold between NYHA II and III we would advise all observers to use endorsement of at least the third category of the "effort" item. In research settings the ability scale, defined using both the anchor items and an operator's classification, can be used as a refined NYHA classification, independently of the subjectivity of the observers.

The major limitation of this study is its small size. Whereas the minimum number of individuals required to properly fit a 1-PL model is 200 [22], only slightly less than the 263 individuals assessed here, a proper 2-parameter logistic (2-PL) GRM allowing the slope to vary among the items would require a larger sample size. An inadequate sample size would be expected to yield unstable item parameters and higher standard errors, which was the case in our study.

In the present study, each individual was assessed by only one observer, opposed to the ideal situation where that individual would be assessed by all observers. We

do not think of this as a limitation. When we compared the individuals assessed by each of the observers there were no statistically significant differences in sex, clinical history, systolic blood pressure, education and left ventricular systolic dysfunction; only age, body mass index and diastolic blood pressure showed small differences. Consequently, overall the individuals that each observer assessed were very similar. On the other hand, the anchor items were related to each observer's NYHA classification. So even if the sample assessed by each observer was very discrepant, the anchor items would guarantee a good calibration. Therefore we are confident that this limitation did not have a major impact on the results.

The anchor items proposed to calibrate the NYHA classifications are not assumed to be the gold standard and are not intended replace the NYHA classification by themselves. The study only validated these anchor items against the NYHA classifications, supporting that they could be used to calibrate different observers in using NYHA classification. We do not intend to question the validity of either the anchor items or NYHA classification to measure true functional capacity, in which case we would need to confront each of them with quantitative measures of functional capacity like the 6-minute walk test or a cardiopulmonary exercise test with measurement of oxygen consumption.

Self-reported distance is a subjective measure and many factors influence a patient's answer, including psychosocial factors and perceptions of distance. Patients' ability to estimate 100 m, 500 m and 2500 m distance was shown to be poor [3]. However, the use of additional anchor items is expected to attenuate the impact of this potential error in each of them.

The physicians were aware of patients' responses to the 4-anchor items. It is therefore possible that this fact influenced their ratings and thus violated the assumption of local independence of the statistical model. Separate calibration with the mean/mean method [23] was use as sensitivity analysis (data not shown) and the results obtained were similar to the concurrent analysis, also there were no significant differences between the observed and expected frequencies of items for the 7 observers models and only one pair of anchor items in 1 out of the 7 observers graded response model (observer 2) showed local dependencies.

The generalisation of the calibration method proposed is limited by the lack of individuals classified as NHYA class IV.

Conclusions

In conclusion, this study showed that the thresholds of the NYHA classification between observers were very discrepant and that concurrent calibration through IRT models can be used to calibrate a large number of observers on the same scale. It provides a way to minimize the reliability problem of NYHA classification. This type of approach can be useful to minimize the inter-observer variability in other classifications based on patient's and/or physicians's perception.

Author details

[1]Department of Clinical Epidemiology, Predictive Medicine and Public Health, University of Porto Medical School, Porto, Portugal. [2]Institute of Public Health of the University of Porto, Porto, Portugal. [3]Department of Pure Mathematics, University of Porto Science School, Portugal. [4]Heart Failure Clinic, Department of Internal Medicine, Hospital S. João, Porto, Portugal.

Authors' contributions

MS participated in the study design, performed the statistical analysis and helped to draft the manuscript. RG performed the statistical analysis and helped to draft the manuscript. PL, MA, PB and AA participated in the study design and helped to draft the manuscript.
All authors read and approved the final manuscript

Competing interests

The authors declare that they have no competing interests.

Received: 14 June 2011 Accepted: 3 August 2011
Published: 3 August 2011

References

1. *Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels*. 9, revised edition. Little, Brown; 1994.
2. Bennett JA, Riegel B, Bittner V, Nichols J: **Validity and reliability of the NYHA classes for measuring research outcomes in patients with cardiac disease.** *Heart Lung* 2002, **31**:262-270.
3. Raphael C, Briscoe C, Davies J, Ian Whinnett Z, Manisty C, Sutton R, Mayet J, Francis DP: **Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure.** *Heart* 2007, **93**:476-482.
4. Goldman L, Hashimoto B, Cook EF, Loscalzo A: **Comparative reproducibility and validity of systems for assessing cardiovascular functional class: advantages of a new specific activity scale.** *Circulation* 1981, **64**:1227-1234.
5. Ramos E, Lopes C, Barros H: **Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction.** *Ann Epidemiol* 2004, **14**:437-441.
6. McHorney CA, Ware JE Jr, Raczek AE: **The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs.** *Med Care* 1993, **31**:247-263.
7. Severo M, Santos AC, Lopes C, Barros H: **Reliability and validity in measuring physical and mental health construct of the Portuguese version of MOS SF-36.** *Acta Med Port* 2006, **19**:281-287.
8. Ainsworth BE, Haskell WL, Leon AS, Jacobs DR Jr, Montoye HJ, Sallis JF, Paffenbarger RS Jr: **Compendium of physical activities: classification of energy costs of human physical activities.** *Med Sci Sports Exerc* 1993, **25**:71-80.
9. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD: **The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus.** *Arch Neurol* 1989, **46**:1121-1123.
10. Cortina JM: **What Is Coefficient Alpha? An Examination of Theory and Applications.** *Journal of applied psychology* 1993, **78**:98-98.
11. Hu L, Bentler PM: **Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives.** *Structural Equation Modeling: A Multidisciplinary Journal* 1999, **6**:1-55.
12. McHorney CA, Cohen AS: **Equating health status measures with item response theory: illustrations with functional status items.** *Med Care* 2000, **38**:II43-59.
13. Samejima F: **Graded response model.** *Handbook of modern item response theory* 1997, 85-100.

Subchapter 2.1

Calibration: effect on the misclassification of NYHA

Calibration: effect on the misclassification of NYHA

Milton Severo, MSc[1,2], Rita Gaio, PhD[3], Ana Azevedo, PhD[1,2,4]

(1) Department of Clinical Epidemiology, Predictive Medicine and Public Health, University of Porto Medical School, Porto, Portugal.

(2) Institute of Public Health of the University of Porto, Porto, Portugal.

(3) Department of Mathematics, University of Porto Science School;

(4) Heart Failure Clinic, Department of Internal Medicine, Centro Hospitalar São João, Porto, Portugal.

MS: milton@med.up.pt
RG: argaio@fc.up.pt
AA: anazev@med.up.pt

Correspondence:
Milton Severo
Departamento de Epidemiologia Clínica, Medicina Preditiva e Saúde Pública
Faculdade de Medicina do Porto
Alameda Prof. Hernâni Monteiro
4200-319 Porto
milton@med.up.pt
Tel +351225513652
Fax: +351225513653

Abstract

Objective: The calibration of the NYHA classification system between different observers is expected to increase its validity, reducing misclassification.

Study Design and Setting: At the standardized clinical interview subjects who reported to have breathlessness (n=265) were presented to a 4-item questionnaire on functional capacity. The questionnaire was administered by 7 physicians who also classified the subject's functional capacity according to NYHA. Calibration of NYHA classification across each set of individuals assessed by each physician was performed by the concurrent method using the four patient items as anchor items. We estimated the area under the ROC curve (AUC) and likelihood ratio using the calibrated and non-calibrated NYHA class I *versus* II-III to predict the presence of a series of objective structural of functional cardiac abnormalities as assessed by echocardiography at rest.

Results: The area under the ROC curve (AUC) for NYHA class to predict the outcomes considered showed an overall improvement in discrimination of NYHA class after its calibration, largely at the expense of the likelihood ratio of NYHA I.

Conclusion: The calibration methodology can be useful to improve the validity of NYHA classification in clinical practice and research settings. It provides a way to minimize the misclassification of NYHA classification.

Keywords: Dyspnea; Heart failure; New York Heart Association; Calibration; Misclassification.

Background

The New York Heart Association (NYHA) system was designed for translation of the clinical assessment of patients by physicians into 4 functional classes on the basis of the patient's limitations in physical activities caused by cardiac symptoms. The class a clinician decides to assign a patient to depends on the clinician´s interpretation of what is "ordinary" physical activity, "slight" and "marked" limitations. This results in a high inter-observer variability. Previous studies showed an inter-observer agreement for the NYHA classification of approximately 55%[1]. Thus, the calibration of the NYHA classification system between different observers is expected to increase its validity, reducing misclassification. Although much has been written regarding the limitations of the NYHA class as an outcome measure[2], it is still used in clinical research and practice. The popularity of the NYHA classification system is based on its simplicity.

Methods

This study is based on a follow-up visit of the EPIPorto cohort[3], which at baseline (1999-2003) was representative of the non-institutionalized adult population of Porto, Portugal. Between October 2006 and July 2008, all participants aged ≥ 45 years were eligible to a systematic evaluation of cardiac structure and function, which included a cardiovascular clinical history and physical examination, and a transthoracic M-mode and bi-dimensional echocardiogram. The Ethics Committee of Hospital de São João approved the study and participants provided written informed consent.

Among 1914 participants in the age range of interest, 198 (10.3%) refused to be re-evaluated and 580 (30.3%) were lost to follow up (unreachable by telephone or post).

Therefore 1136 (59.3%) individuals aged ≥45 years were assessed by 8 physicians experienced in the management of heart failure patients. At the standardized clinical interview subjects who reported to have breathlessness (n=265; 23.3%) were presented to a 4-item questionnaire on functional capacity to characterize the severity of symptoms: 1) whether breathlessness is felt when walking on steep plane, horizontal plane or at rest; 2) distance walked until perception of breathlessness; 3) sets of stairs (10-15 steps) climbed until perception of breathlessness; 4) whether mild, moderate or intense efforts are necessary to elicit breathlessness. The same physician administered the questionnaire and classified the subject's functional capacity using the NYHA classification. Each participant was assessed by only one observer. As expected given the community base of the study, there were no participants in class IV and very few in class III which were therefore aggregated with class II.

Calibration of NYHA classification across each set of individuals assessed by each physician was performed by the concurrent method using the four patient items as anchor items [4]. We estimated the likelihood ratio and predictive value using the calibrated and non-calibrated NYHA class I *versus* II-III to predict the presence of a series of objective structural of functional cardiac abnormalities as assessed by echocardiography at rest.

Left ventricular dilatation was considered when end diastolic left ventricular diameter was larger than 58 mm in men and 52 mm in women and left ventricular systolic dysfunction was defined by an ejection fraction below 45%, assessed by Simpson's method, or by visual estimate. Left ventricular hypertrophy was defined as left ventricular mass index> 110 g/m^2 in women and 125 g/m^2 in men and left atrium dilation was defined as left atrial volume index> 40ml/m^2. Valvular abnormalities were considered when moderate or severe. Diastolic dysfunction was defined according to the European Society of Cardiology guidelines for diastolic heart failure, based on tissue Doppler assessment of e' and e, atrial fibrillation, B-type natriuretic peptide and age[5].

4

Results:

The prevalence of NYHA II-III in the sample was 170 (66.7%) and 143 (56.1%) without and with calibration, respectively. From the non-calibrated to the calibrated NYHA class, 16 and 43 individuals changed from I to II-III and from II-III to I, respectively. Almost half of this sample of community participants reporting troubled breathing had at least one of the objective cardiac structural or functional abnormalities considered.

In general, the likelihood ratios showed that NYHA class changed pre- to posttest probability of cardiac abnormalities very little (higher than 0.3 for NYHA I and below 1.5 for NYHA II-III). This was expected, considering that many of these abnormalities are known to be asymptomatic and the symptoms are unspecific (table 1).

With regard to the effect of the calibration, the area under the ROC curve (AUC) for NYHA class to predict the outcomes considered showed an overall improvement in discrimination of NYHA class after its calibration (table 1), largely at the expense of the likelihood ratio of NYHA I.

Conclusion

These results objectively indicate that the calibration can improve the diagnostic value of the NYHA classification and, consequently, improve the usefulness of the NYHA classification as an outcome measure. Although the observers in study were experienced physicians well trained in the management of heart failure, there were still discrepancies between their (subjective) classifications; we expect that the calibration should have a more marked effect in less trained physicians.

A major limitation of this study is the lack of individuals classified as NHYA class III and IV. However, the discrepancies between observers are expected to be smaller (thus the effects

of calibration also smaller) for higher functional classes, particularly for dyspnea at rest. Also, the setting in which these participants were inquired does not reproduce clinical practice triggered by symptoms that lead the patient to seek care; this context probably explains the large proportion of NYHA I after people having reported breathlessness.

In conclusion, this study showed that the calibration methodology can be useful to improve the validity of NYHA classification in clinical practice and research settings, by increasing the inter-observer reproducibility, and can be used to calibrate a large number of observers on the same scale. It provides a way to minimize the misclassification of NYHA classification. This type of approach can be useful to minimize the misclassification in other classifications based on patient's and/or physicians' perception.

Competing interests

The authors declare that they have no competing interests.

References

1. Raphael C, Briscoe C, Davies J, Ian Whinnett Z, Manisty C, Sutton R, Mayet J, Francis DP: **Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure**. *Heart* 2007, **93**(4):476-482.
2. Tedesco C, Manning S, Lindsay R, Alexander C, Owen R, Smucker ML: **Functional assessment of elderly patients after percutaneous aortic balloon valvuloplasty: New York Heart Association classification versus functional status questionnaire**. *Heart Lung* 1990, **19**(2):118-125.
3. Ramos E, Lopes C, Barros H: **Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction**. *Annals of epidemiology* 2004, **14**(6):437-441.
4. Severo M, Gaio R, Lourenco P, Alvelos M, Bettencourt P, Azevedo A: **Indirect calibration between clinical observers - application to the New York Heart Association functional classification system**. *BMC Res Notes* 2011, **4**:276.
5. Paulus WJ, Tschope C, Sanderson JE, Rusconi C, Flachskampf FA, Rademakers FE, Marino P, Smiseth OA, De Keulenaer G, Leite-Moreira AF *et al*: **How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology**. *Eur Heart J* 2007, **28**(20):2539-2550.

Table 1. Likelihood ratio and predictive value of NYHA I vs. NYHA II-III for several cardiac structural and functional parameters.

		Calibrated NYHA class			Non-calibrated NYHA class		
		I	II-III		I	II-III	
		N=112 (43.9%)	N=143 (56.1%)		N=85 (33.3%)	N=170 (66.7%)	
	N (%)	LR (-PV)	LR (+PV)	AUC	LR (-PV)	LR (+PV)	AUC
Outcomes							
Left ventricular dilation	28 (11.0)	0.54 (93.8%)	1.39 (14.7%)	0.645	0.51 (94.1%)	1.27 (13.5%)	0.593
Left ventricular systolic dysfunction	18 (7.4)	0.48 (96.3%)	1.44 (10.3%)	0.656	0.46 (96.4%)	1.30 (9.4%)	0.588
Left ventricular hypertrophy	59 (23.1)	0.29 (92.0%)	1.79 (35.0%)	0.702	0.66 (83.5%)	1.19 (26.5%)	0.600
Diastolic dysfunction	67 (27.6)	0.48 (84.5%)	1.58 (37.6%)	0.652	0.65 (80.2%)	1.21 (31.5%)	0.602
Left atrium dilation	61 (24.6)	0.68 (81.8%)	1.29 (29.7%)	0.592	0.91 (77.1%)	1.05 (25.5%)	0.542
Valvular disease	16 (6.3)	0.27 (98.2%)	1.63 (10.0%)	0.699	0.36 (97.6%)	1.34 (8.3%)	0.627
Any of the above	111 (45.7)	0.51 (70.1%)	1.65 (58.1%)	0.635	0.74 (61.7%)	1.16 (49.4%)	0.572

AUC, area under the ROC curve; LR, likelihood ratio; NYHA, New York Heart Association; PV, predictive value

Chapter 3

Diagnostic value of patterns of symptoms and signs of heart failure: application of latent class analysis with concomitant variables.

Title: Diagnostic value of patterns of symptoms and signs of heart failure: application of latent class analysis with concomitant variables

Milton Severo, MSc,[a,b] Rita Gaio, PhD,[c,d] Patrícia Lourenço, MD, MSc,[e] Margarida Alvelos, MD,[e] Alexandra Gonçalves, MD,[f] Nuno Lunet, PharmD, MPH, PhD,[a,b] Paulo Bettencourt, MD, PhD,[e] Ana Azevedo, MD, PhD[a,b,e]

(a) Department of Clinical Epidemiology, Predictive Medicine and Public Health, University of Porto Medical School, Porto, Portugal.

(b) Institute of Public Health of the University of Porto, Porto, Portugal.

(c) Department of Mathematics, University of Porto Science School.

(d) Mathematics Center, University of Porto

(e) Heart Failure Clinic, Department of Internal Medicine, Hospital São João, Porto, Portugal.

(f) Department of Cardiology, Hospital São João, Porto, Portugal

Correspondence:
Milton Severo
Departamento de Epidemiologia Clínica, Medicina Preditiva e Saúde Pública
Faculdade de Medicina do Porto
Alameda Prof. Hernâni Monteiro
4200-319 Porto
milton@med.up.pt
Tel +351225513652
Fax: +351225513653

ABSTRACT

Background: The diagnosis of heart failure (HF) requires a compatible clinical syndrome and demonstration of cardiac dysfunction by imaging or functional tests. Since individual symptoms and signs are generally unreliable and have limited value for diagnosing HF, the authors aimed to identify patterns of symptoms and signs, based on findings routinely collected in current clinical practice, and to evaluate their diagnostic value, taking into account the *a priori* likelihood of HF.

Methods: Based on the cross-sectional evaluation of 1115 community participants aged ≥45 years from Porto, Portugal, in 2006-2008, patterns were identified by latent class analysis, using concomitant variables to predict class membership. Patterns used eleven symptoms/signs, covering dimensions of congestion and hypoperfusion. Sex, age, education, obesity, diabetes and history of myocardial infarction or HF were included as concomitants.

Results: Bayesian information criteria supported a solution with three patterns: 10.1% of participants followed a pattern with symptoms of troubled breathing and signs of congestion (pattern 1), 27.8% a pattern characterized mainly by signs of congestion (pattern 2) and 62.1% were essentially asymptomatic (pattern 3); model fit was best when including concomitant variables. The likelihood ratio of patterns 1, 2 and 3 for left ventricular systolic dysfunction was 3.4, 1.1 and 0.6, and for left ventricular diastolic dysfunction 3.5, 1.4 and 0.5, respectively.

Conclusions: The use concomitant variables can improve the diagnostic value of the symptoms and signs patterns and, consequently, improve the usefulness of the symptoms and signs for diagnosis and as an outcome measures.

Key words: Heart failure; Latent class models; Classification; Diagnosis; Concomitant Variables; Cardiac Dysfunction.

INTRODUCTION

Heart failure is a complex clinical syndrome resulting from a variety of structural or functional cardiac disorders. The diagnosis of heart failure (HF) requires a compatible clinical syndrome and demonstration of cardiac dysfunction by imaging or functional tests[1,2]. A clinical examination is always the first step in a diagnostic approach to possible HF and further investigation is conditional on initial clinical judgment. However, individual symptoms (such as dyspnoea and fatigue) and signs (*e.g.* third heart sound and evidence of congestion) are generally unreliable and have limited value for diagnosing heart failure[3,4]. Several multidimensional criteria based on symptoms and signs have been developed over decades in an attempt to standardize the clinical assessment of heart failure [5-12]. When patients initially labeled as having heart failure are investigated using objective assessment criteria, only around one third are considered to truly have heart failure [13,14]. Obesity, unrecognized myocardial ischaemia or pulmonary disease commonly lead to false positive heart failure diagnoses [13]. Additionally, it may be difficult to distinguish pathologic conditions from mere physical deconditioning associated with ageing. Moreover, the varying subjective importance attributed to symptoms justifies a systematic association between reported symptoms and female gender and psychosocial characteristics, both among the healthy and those with cardiac dysfunction. Gender, age, education and obesity are major determinants of symptoms and signs suggestive of heart failure [15], beyond their role as risk factors for heart failure, and may account for false positive and negative classification. Furthermore, the clinical judgment is modified based on the *a*

priori likelihood of HF[16], depending mainly on a history of HF or myocardial infarction, and on strong risk factors for such conditions.

Systolic and diastolic heart failure are very similar at the bedside though very different when one considers cardiac structure and function, and response to therapy. About half of patients with symptomatic heart failure have preserved left ventricular systolic function [17]. Heart failure with preserved systolic function has long been mainly an exclusion diagnosis, but the most recent guidelines require evidence of left ventricular diastolic dysfunction [18]. The prevalence of heart failure with preserved ejection fraction increased over the last years [19], in part as a consequence of increasing acknowledgment of its importance and of improvements in the ability to recognize it. Until a recent past, most clinical and epidemiologic studies considered only HF with left ventricular systolic dysfunction, and community-based studies or consecutive series of patients with any kind of HF are likely to currently yield scenarios of diagnostic reasoning and validity that contrast with past cohorts.

The aims of this study were to identify patterns of symptoms and signs of heart failure, based on findings routinely collected in current clinical practice, including concomitant variables to predict the pattern membership; and to evaluate the diagnostic value of different patterns.

METHODS

Study design and sample selection

Participants were selected within the first follow-up of a cohort, representative at baseline of the non-institutionalized adult population of Porto, Portugal – the EPIPorto cohort study. In 1999-2003, cohort assembly was done by random digit dialing, using households as the sampling frame, followed by random selection of one person aged 18 years or older in each household. Refusals were not substituted within the same household. The proportion of participation was 70%[20]. At baseline, 2485 participants were recruited. Between October 2006 and July 2008, participants aged 45 years or over were eligible to a systematic evaluation of parameters of cardiac structure and function, which included a cardiovascular clinical history and physical examination, and a bidimensional transthoracic echocardiogram. Among 2048 cohort members that would be in the eligible age range at this time, 134 (6.5%) had died, 198 (9.7%) refused to be re-evaluated and 580 (28.3%) were lost to follow up (unreachable by telephone or post), and 21 (1.0%) had missing values in at least one of the variables used in the present analysis. Therefore, 1115 (54.6%) individuals aged 45 years or over were analyzed to develop the new epidemiologic classification scheme for clinical HF, with mean (standard deviation) follow-up period of 7 (2.7) years. When comparing their baseline characteristics with the 933 cohort members of the same age range who were not included in the present analysis, participants were significantly younger [mean (standard deviation) age: 57 (10) vs. 59 (13) years, p<0.001], had a higher level of education (median: 7 vs. 4 years, p<0.001), and a lower prevalence of obesity (22.5% vs. 28.9%, p=0.001), hypertension (55.3% vs. 65.2%, p<0.001) and previous myocardial infarction (3.5% vs. 8.4%, p=0.09), while there were no gender differences (men: 39.0% vs. 37.4%, p=0.456).

5

This investigation conformed to the principles expressed in the Declaration of Helsinki. The local ethics committee approved the study and participants provided written informed consent.

Data collection and variables definition

A structured questionnaire was applied by non-physician trained interviewers to obtain data on sociodemographic characteristics and lifestyles. Clinical history and physical examination were performed by physicians experienced in the management of heart failure patients.

The New York Heart Association (NYHA) classification across each set of individuals assessed by each physician was calibrated to increase the inter-observer reproducibility[21].

Echocardiograms were performed by trained cardiologists using the same equipment (Sonos 5500 Phillips), following a standardized protocol. All measures resulted from averaging three observations.

Obesity was defined as body mass index higher than or equal to 30 kg/m^2 [22]. History of myocardial infarction was defined as self-reported medical diagnosis of myocardial infarction or history of coronary artery bypass graft.

An overnight fasting venous blood sample was withdrawn with plasma or serum samples and used for B-type natriuretic peptide (BNP) measurement using a commercially available immunofluorimetric assay (Triage BNP Test, BIOSITE diagnostics, San Diego, CA, USA). A establish an equation was the use for the estimation of the BNP concentration in plasma when only serum is available[23]. BNP

values were available for 630 participants and, for analysis, they were dichotomized by the cut-off point 100 pg/mL, previously established for the diagnosis of heart failure [24].

Left ventricular dilatation was considered when end diastolic left ventricular diameter was larger than 58 mm in men and 52 mm in women and left ventricular systolic dysfunction was defined by an ejection fraction below 45%, assessed by Simpson's method, or by visual estimate. In 29 subjects (2.6%) it was not possible to quantify the ejection fraction due to poor acoustic window.

Left ventricular hypertrophy was defined as left ventricular mass index> 110 g/m^2 in women and 125 g/m^2 in men, and left atrium dilation was defined as left atrial volume index> 40ml/m^2 [25]. Valvular abnormalities were considered when moderate or severe. Diastolic dysfunction was defined according to the European Society of Cardiology guidelines for diastolic heart failure [18].

Statistical analysis

Latent class analysis

Latent class analysis (LCA) is used to uncover distinct groups of individuals from a sample (patterns), homogeneous within the group, considering that the performance of an individual in a set of items is explained by a categorical latent variable with K classes, commonly called "latent classes". Interpretation of the model is usually based on item profiles in each category, obtained from the probabilities of endorsing each item response, conditional on class membership.

In this study, the number of latent classes (patterns) was defined according to the bayesian information criterion (BIC). Starting from one single class and increasing

one class at each step, the best solution was identified when the increase in the number of classes did not lead to a decrease in BIC.

LCA used eleven symptoms and signs to define a syndrome suggestive of HF or important for differential diagnosis, including dyspnoea, orthopnoea, nocturnal paroxysmal dyspnoea, fatigue, self-percepted and clinically confirmed edema, hepatojugular reflux or jugular venous distension, pulmonary rales, heart murmur, trophic signs of chronic venous insufficiency and visible varicose veins.

The items selection was based on their clinical relevance for the definition of HF and prevalence [1,2,18]. Other relevant signs and symptoms such as third heart sound (0.8%), heart rate higher than 120 beats per minute (0.1%) and hepatomegaly (1.5%) were not taken into account because they occurred in less than 2% of the study sample.

In LCA, factors which are known to have a large impact on the prevalence of symptoms and signs suggestive of heart failure were used as concomitant variables, namely sex, age, education, obesity, diabetes and history of myocardial infarction or heart failure. Other relevant concomitant variables, namely smoking, alcohol intake, hypertension and history of valvular diseases were not included because they did not show a significant effect in the patterns in this sample.

In LCA, concomitant variables are covariates considered in the process of formation of the latent classes, by the multinomial regression of latent classes on concomitant variables, to allow for different contributions of the items to define the classes for different levels of concomitants [26].

All LCA models were fitted using MPlus (version 5.2; Muthen & Muthen).

8

Patterns' diagnostic value

To assess the diagnostic value of the defined patterns, we estimated the likelihood ratio and predictive value of patterns of symptoms and signs, with and without concomitant variables, to predict the presence of a series of outcomes, corresponding to objective structural of functional cardiac abnormalities as assessed by echocardiography at rest. The likelihood ratio measures the ratio between the prevalence of each pattern in subjects with and without the outcome. The predictive value is the *a posteriori* probability of the outcome, conditional on the clinical pattern.

The diagnostic value of high BNP (BNP ≥100 pg/mL) was evaluated in a subgroup of the study sample with BNP measured in blood collected at the time of clinical and echocardiographic examination (n=630). In untreated subjects, a concentration of BNP under 100 pg/mL has high negative predictive value and makes HF an unlikely diagnosis[1].

RESULTS

Patterns of symptoms and signs of heart failure

Relying only on signs and symptoms, the increase in log likelihood values leveled off when increasing from two to three and BIC reached its optimum value at three classes, supporting preference for a 3-class solution. The inclusion of

concomitant variables led to an improvement (decrease) in BIC values in all tested models and the 3-class model was again the best solution according to BIC (Table 1).

The final model with concomitant variables had the following item profiles: class 1 had high probabilities for all 11 items (symptomatic heart failure pattern), class 2 had high probability for volume overload and lower probability for troubled breathing (congestion pattern) and class 3 had low endorsement probabilities for all items (no symptoms and signs pattern) (Table 2).

The estimated prevalence of classes 1, 2 and 3 without concomitant variables was 9.6%, 19.2% and 71.1%; the estimated prevalence of classes 1, 2 and 3 with concomitant variables was 10.1%, 27.8% and 62.1%, respectively. When considering gender, age, education, obesity, diabetes and history of myocardial infarction or heart failure, the discrimination to distinguish a third class increased mainly as a result of the reclassification of around a quarter of participants initially classified as non cases into class 2, supporting the importance of including concomitant variables when judging the value of symptoms and signs of HF.

Taking class 3 as reference, class 1 was positively associated with age (OR=1.07 per year), obesity (OR=6.00), diabetes (OR=2.33) and history of myocardial infarction or heart failure (OR=12.94), and negatively associated with male sex (OR=0.11) and education (OR=0.80 per year); class 2 was positively associated with age (OR=1.12 per year) and obesity (OR=3.34). All these associations were statistically significant (Table 2).

Diagnostic value of clinical patterns

The prevalence of left ventricular systolic dysfunction and left ventricular dilation was lower than 5%, the prevalence of diastolic dysfunction, left ventricular hypertrophy, and left atrial dilation varied between 12 and 17%, valvular disease affected 2.8% of the sample, and almost 30% had any of the former abnormalities on echocardiogram.

In general, the likelihood ratios showed that the patterns without concomitant variables changed pre- to posttest probability of cardiac abnormalities very little (minimum 0.6 for pattern 3 and maximum 4.1 for pattern 1). The area under the ROC curve (AUC) for symptoms and signs patterns to predict the outcomes considered showed an overall improvement in discrimination of symptoms and signs patterns after the use of concomitant variables, largely at the expense of the likelihood ratio of pattern 3, whose value for exclusion of HF increased (Table 3). The negative likelihood ratios of pattern 3 were better than each individual symptoms and signs.

Pattern 1 is 3-fold more likely and pattern 3 is 5-fold less likely in subjects with BNP above vs. below 100 pg/mL (Table 3), resulting in safe exclusion of high BNP in pattern 3 defined with concomitant variables.

DISCUSSION

In this study, the authors succeeded in identifying three patterns of syndromic aggregation of symptoms and signs for heart failure, based on findings routinely collected in current clinical practice, by the application of latent class analysis with concomitant variables to account for known determinants of the relevant clinical

11

findings and the *a priori* probability of the condition. These models could be useful to standardize and quantify the probabilistic reasoning in clinical diagnosis, upon which decisions of further investigation and even treatment need to be made.

In general, the likelihood ratios showed that the symptoms and signs patterns generated only relatively small changes from pre- to posttest probability of cardiac abnormalities. This was expected, considering that many of these abnormalities are known to be asymptomatic in a large proportion of patients for a long time and the symptoms are unspecific, and is compatible with previous quantifications of the value of symptoms and signs [27]. The patterns showed a small diagnostic value and a high value to exclude a high BNP value, an established biomarker for the diagnosis of heart failure.

The LCA is a probabilistic approach to disease classification which allows the identification of more precise categories of disease conditions [28]. It has been used to validate diagnostic tests in the absence of a perfect reference standard [28], which is the situation for HF. A new feature of this study is that for the first time, to our knowledge, these classifications integrated factors which have a large impact on the prevalence of symptoms and signs suggestive of HF. The novelty in our application is that class probabilities are adjusted for concomitant variables. Specifically, the model estimates the increase or decrease in class probabilities for individuals conditional to the respective concomitant variables pattern, contributing to increased discrimination and a decrease in the number of false negatives and false positives. The inclusion of these variables improved model fit. The area under the ROC curve (AUC) for symptoms and signs patterns to predict the outcomes considered showed an overall improvement in discrimination of symptoms and signs patterns after the use of concomitant variables,

largely at the expense of the likelihood ratio of pattern 3 (no symptoms and signs) to exclude cardiac abnormalities. The inclusion of concomitant variables helped to refine class 3 (reclassifying individuals from class 3 to class 2, thus decreasing the false negatives). These results objectively indicate that the use concomitant variables can improve the diagnostic value of the symptoms and signs patterns and, consequently, improve the usefulness of the symptoms and signs for diagnosis and as an outcome measures. The potential for application in other settings of complex diagnoses is very high.

In patterns with concomitant variables, class 1 was more prevalent in women, individuals with history of myocardial infarction or heart failure, diabetic and obese individuals, increased with age and decreased with education, while class 2 only increased with age and obesity. These associations reflect the influence of gender and education on subjective importance attributed to symptoms [15], likely a psycho-social effect, while age, obesity, diabetes, and history of myocardial infarction or heart failure are biologically associated with decreasing cardiac function even at asymptomatic stages.

The patterns identified by this methodological approach depend on the type of population being studied. In this study, in a sample of the general population, three patterns were identified: symptomatic HF pattern, symptoms and signs of congestion and no symptoms and signs. In a previous study, using a similar approach in subjects discharged after myocardial infarction or acute heart failure, the authors were able to distinguish different patterns (non-cases, heart failure and advanced heart failure) [12]. The patterns identified in the current study are appropriate for diagnosis in the general population only. The low prevalence of advanced and severe HF cases is a limitation of

this study and could have underestimated the discriminative capacity of this set of items. Also, the prevalence of more specific symptoms and signs such as a third heart sound [29] was too low in this sample to be able to take them into account, suggesting that the proposed patterns are likely to be more sensitive but less specific than previously available scores such as the Framingham criteria, supporting their usefulness as potential screening tools or initial clinical investigation that do not aim to replace full investigation in the clinical setting. A valid assessment of diastolic dysfunction, currently recommended for the diagnosis of HF [18], using up-to-date technology, is a major advantage of this study. The prevalence of hypertension is very large in this population[30], while this is a low coronary heart disease risk country, expectedly increasing the burden of diastolic abnormalities. Since mild diastolic changes have more questionable clinical significance[31], though discrimination of such mild cases would require technological means that were unavailable in our study such as study of pulmonary veins flow[18], this could have contributed to a less favorable performance of the clinical diagnosis in this setting.

The role of BNP testing is clearly defined for diagnosing patients with suspected heart failure [32]. Ventricular wall stretch is the major determinant of increased BNP concentrations and peptide levels have limited accuracy in differentiating heart failure with left ventricular systolic dysfunction or with preserved ejection fraction [33]. The high sensitivity of the model with concomitant variables is supported by the virtually null prevalence of high BNP among subjects classified as no symptoms and signs pattern. However, a large proportion of participants classified in classes 1 and 2 still did not have BNP above 100 pg/mL, arguing in favor of the likely existence of false positives in these patterns.

The fact that validation by comparison with echocardiographic parameters and BNP values was performed using the same sample in which the LCA models were fit is a limitation of this study. However, all clinical and echocardiographic data were collected blinded to each other and to BNP values, preventing an artificially high correlation between the LCA and both cardiac abnormalities and the biomarker, due to the subjective and observer-dependent nature of the items being assessed.

Future developments of this research work aim at translating the validated patterns into a classification score, using an approach for making complex statistical models useful to practitioners and researchers, such as the a circular ruler[34] or points system[35], which was used for example to develop the widely used Framingham risk scores. Use of this tool will allow the identification of high-risk candidates for heart failure who are likely to have a substantial yield of positive findings when tested for objective measures of cardiac dysfunction in clinical practice, as well as to confidently exclude heart failure in others, thus orienting the clinical investigation in alternative directions. Such a tool could also increase discrimination and decrease the number of false negatives and false positives in epidemiological studies on HF, in which subjects are classified depending on a set of systematically collected data without the integrated view of one clinician to weigh the whole complex picture of a case.

Disclosures: The authors declare that they have no competing interests.

Table 1. Latent class analysis for heart failure symptoms and signs, with and without concomitant variables (sex, age, education, obesity, diabetes mellitus, and history of myocardial infarction or heart failure), in the general population aged ≥ 45 years, Porto, Portugal, 2006-2008.

	Symptoms and signs				**Symptoms and signs with concomitant variables**[2]			
Number of classes[1]	Log L	Number of parameters	BIC	p[3]	Log L	Number of parameters	BIC	p[3]
LCA								
1 class	-5060.161	14	10218					
2 classes	-4727.150	29	9657	<0.001	-4596.351	35	9438	<0.001
3 classes*	**-4613.189**	**44**	**9535**	**<0.001**	**-4438.071**	**56**	**9269**	**0.003**
4 classes	-4575.405	59	9564	0.966	-4382.080	77	9304	0.770

BIC, Bayesian information criteria; HF, heart failure; LCA, latent class analysis; Log L, log likelihood.
[1] The bold font denotes the best models according to lowest BIC.
[2] A putative role of concomitant variables only exists in models with at least two latent classes, in which concomitants can influence the classification in different groups.
[3]Lo-Mendell-Rubin likelihood ratio test of model fit to quantify the likelihood that the data can be described by a model with one-less class.

Table 2. Marginal percentage of subjects with each symptom and sign in each assigned latent class (pattern), with and without including concomitant variables (sex, age, education, obesity, diabetes, and history of myocardial infarction or heart failure) to predict class membership, in the general population aged ≥ 45 years, Porto, Portugal, 2006-2008.

		Pattern of symptoms and signs			Pattern of symptoms and signs with concomitant variables		
	Total	Class 1 9.6	Class 2 19.2	Class 3 71.1	Class 1 10.1	Class 2 27.8	Class 3 62.1
Dyspnoea							
NYHA I	86.9	18.3	91.4	95.5	20.9	90.6	96.9
NYHA II	9.2	45.2	8.0	4.3	45.1	9.0	2.9
NYHA III	3.9	36.4	0.7	0.2	33.9	0.4	0.2
Fatigue							
No	81.7	18.5	88.5	88.6	23.1	89.6	88.5
Yes	18.3	81.5	11.5	11.4	76.9	10.4	11.5
Orthopnoea							
No	91.6	41.4	98.1	96.7	44.9	97.5	96.9
Yes, 1 pillow	3.8	19.2	0.0	2.8	19.5	0.2	2.7
Yes, 2 or more pillows	4.7	39.4	1.9	0.5	35.6	2.2	0.4
Nocturnal paroxysmal dyspnea							
No	93.8	62.2	100.0	96.3	63.0	100	96.2
Yes	6.2	37.8	0.0	3.7	37.0	0.0	3.8
Heart murmur							
No	93.9	87.6	91.1	95.8	86.1	90.6	97.0
Yes	6.1	12.4	8.9	4.2	13.9	9.4	3.0
Pulmonary rales							
No	91.9	78.0	85.9	96.0	80.3	84.9	97.2
Yes	8.1	22.0	14.1	4.0	19.7	15.1	2.8
Hepatojugular reflux or jugular venous distension							
No	90.1	76.0	79.7	95.7	79.2	79.0	97.2
Yes	9.9	24.0	20.3	4.3	20.8	21.0	2.8
Lower limb oedema at the end of the day (symptom)							
No	73.2	35.3	55.5	84.6	35.0	61.0	85.5
Yes	26.8	64.7	44.5	15.4	65.0	39.0	14.5
Lower limb oedema (physical examination)							
No	83.7	55.4	62.3	95.2	56.2	67.1	96.3
Ankle	14.3	36.3	34.1	4.3	36.3	29.5	3.3
Up to the knee	2.0	8.4	3.6	0.5	7.5	3.4	0.4
Trophic signs of chronic venous insufficiency							
No	83.8	67.7	50.2	97.6	68.0	60.4	97.3
Yes	16.2	32.3	49.8	2.4	32.0	39.6	2.7
Visible varicose veins							
No	58.5	44.3	15.1	75.5	40.2	23.5	77.9
Yes	41.5	55.7	84.9	24.5	59.8	76.5	22.1
Concomitant variables					**Multinomial logistic regression**		
					OR (95%CI)	OR (95%CI)	
Sex (male)					0.11 (0.04, 0.32)	0.54 (0.28, 1.06)	Ref*
Age (per year)					1.07 (1.03, 1.12)	1.12 (1.09, 1.16)	Ref*
Education (per year)					0.80 (0.70, 0.91)	0.99 (0.94, 1.04)	Ref*
Obesity					6.00 (2.85, 12.64)	3.34 (1.47, 7.59)	Ref*
Diabetes					2.33 (1.07, 5.08)	0.61 (0.28, 1.32)	Ref*
History of myocardial infarction or heart failure					12.94 (4.95, 33.80)	1.60 (0.61, 4.17)	Ref*

95%CI, 95% confidence interval; HF, heart failure; LCA, latent class analysis; NYHA, New York Heart Association; OR, odds ratio.

* LCA - class 3 taken as reference.

Table 3. Likelihood ratio and predictive value (%) of patterns of symptoms and signs with and without concomitant variables, for the presence of objective cardiac structural and functional parameters. Area under the ROC curve for the classification with and without concomitants.

		Pattern of symptoms and signs				Pattern of symptoms and signs with concomitant variables			
		1	2	3		1	2	3	
	Outcomes prevalence N (%)	LR (PV)	LR (PV)	LR (PV)	AUC	LR (PV)	LR (PV)	LR (PV)	AUC
Outcomes									
Left ventricular systolic dysfunction	39 (3.5)	3.6 (11.5)	1.1 (3.9)	0.6 (2.3)	64.5	3.4 (11.1)	1.1 (3.7)	0.6 (2.2)	65.1
Left ventricular dilation	55 (4.9)	4.1 (16.8)	1.2 (5.6)	0.6 (2.8)	67.4	4.4 (17.7)	1.0 (4.9)	0.5 (2.5)	69.5
Diastolic dysfunction	161 (14.5)	3.6 (38.0)	1.6 (21.9)	0.6 (9.8)	65.2	3.6 (38.1)	1.7 (22.5)	0.5 (7.6)	69.3
Left ventricular hypertrophy	145 (12.9)	3.4 (33.6)	1.3 (16.4)	0.7 (9.4)	63.0	3.5 (34.5)	1.4 (17.8)	0.5 (7.4)	67.5
Left atrium dilation	185 (16.6)	2.2 (30.8)	1.5 (23.8)	0.7 (13.1)	60.0	2.4 (33.0)	1.5 (23.0)	0.6 (11.5)	62.9
Valvular disease	31 (2.8)	4.0 (10.6)	0.7 (1.9)	0.7 (2.1)	63.0	3.4 (9.0)	1.5 (4.3)	0.4 (1.2)	71.1
Any of the above	322 (29.6)	2.6 (52.4)	1.5 (38.5)	0.8 (24.2)	60.0	3.1 (56.9)	1.6 (39.9)	0.6 (20.7)	64.5
High BNP(≥100 pg/mL)	40 (6.3)	3.6 (19.4)	1.8 (10.7)	0.4 (2.6)	72.9	3.4 (18.2)	1.7 (9.9)	0.2 (1.7)	75.1

*AUC, area under the curve; LR, likelihood ratio; PV, predictive value; ROC, receiver operating characteristic.

REFERENCES

1. Dickstein K, Cohen-Solal A, Filippatos G, McMurray JJ, Ponikowski P, Poole-Wilson PA, Stromberg A, van Veldhuisen DJ, Atar D, Hoes AW, Keren A, Mebazaa A, Nieminen M, Priori SG, Swedberg K, Vahanian A, Camm J, De Caterina R, Dean V, Funck-Brentano C, Hellemans I, Kristensen SD, McGregor K, Sechtem U, Silber S, Tendera M, Widimsky P, Zamorano JL. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). *Eur Heart J* 2008;29:2388-442.

2. Hunt SA. ACC/AHA 2005 guideline update for the diagnosis and management of chronic heart failure in the adult: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure). *J Am Coll Cardiol* 2005;46:e1-82.

3. Hobbs FD, Doust J, Mant J, Cowie MR. Heart failure: Diagnosis of heart failure in primary care. *Heart*;96:1773-7.

4. Mant J, Doust J, Roalfe A, Barton P, Cowie MR, Glasziou P, Mant D, McManus RJ, Holder R, Deeks J, Fletcher K, Qume M, Sohanpal S, Sanders S, Hobbs FD. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. *Health Technol Assess* 2009;13:1-207, iii.

5. Mosterd A, Deckers JW, Hoes AW, Nederpel A, Smeets A, Linker DT, Grobbee DE. Classification of heart failure in population based research: an assessment of six heart failure scores. *Eur J Epidemiol* 1997;13:491-502.

6. McKee PA, Castelli WP, McNamara PM, Kannel WB. The natural history of congestive heart failure: the Framingham study. *N Engl J Med* 1971;285:1441-6.

7. Eriksson H, Caidahl K, Larsson B, Ohlson LO, Welin L, Wilhelmsen L, Svardsudd K. Cardiac and pulmonary causes of dyspnoea--validation of a scoring test for clinical-epidemiological use: the Study of Men Born in 1913. *Eur Heart J* 1987;8:1007-14.

8. Walma EP, Hoes AW, Prins A, Boukes FS, van der Does E. Withdrawing long-term diuretic therapy in the elderly: a study in general practice in The Netherlands. *Fam Med* 1993;25:661-4.

9. Schocken DD, Arrieta MI, Leaverton PE, Ross EA. Prevalence and mortality rate of congestive heart failure in the United States. *J Am Coll Cardiol* 1992;20:301-6.

10. Gheorghiade M, Beller GA. Effects of discontinuing maintenance digoxin therapy in patients with ischemic heart disease and congestive heart failure in sinus rhythm. *Am J Cardiol* 1983;51:1243-50.

11. Carlson KJ, Lee DC, Goroll AH, Leahy M, Johnson RA. An analysis of physicians' reasons for prescribing long-term digitalis therapy in outpatients. *J Chronic Dis* 1985;38:733-9.

12. Kim J, Jacobs DR, Jr., Luepker RV, Shahar E, Margolis KL, Becker MP. Prognostic value of a novel classification scheme for heart failure: the Minnesota Heart Failure Criteria. *Am J Epidemiol* 2006;164:184-93.

13. Remes J, Miettinen H, Reunanen A, Pyorala K. Validity of clinical diagnosis of heart failure in primary health care. *Eur Heart J* 1991;12:315-21.

14. Wheeldon NM, MacDonald TM, Flucker CJ, McKendrick AD, McDevitt DG, Struthers AD. Echocardiography in chronic heart failure in the community. *QJM* 1993;86:17.
15. Azevedo A, Bettencourt P, Pimenta J, Frioes F, Abreu-Lima C, Hense HW, Barros H. Clinical syndrome suggestive of heart failure is frequently attributable to non-cardiac disorders--population-based study. *Eur J Heart Fail* 2007;9:391-6.
16. Bianchi MT, Alexander BM. Evidence based diagnosis: does the language reflect the theory? *BMJ* 2006;333:442-5.
17. Petrie M, McMurray J. Changes in notions about heart failure. *Lancet* 2001;358:432-4.
18. Paulus WJ, Tschope C, Sanderson JE, Rusconi C, Flachskampf FA, Rademakers FE, Marino P, Smiseth OA, De Keulenaer G, Leite-Moreira AF, Borbely A, Edes I, Handoko ML, Heymans S, Pezzali N, Pieske B, Dickstein K, Fraser AG, Brutsaert DL. How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology. *Eur Heart J* 2007;28:2539-50.
19. Owan TE, Hodge DO, Herges RM, Jacobsen SJ, Roger VL, Redfield MM. Trends in prevalence and outcome of heart failure with preserved ejection fraction. *N Engl J Med* 2006;355:251-9.
20. Ramos E, Lopes C, Barros H. Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction. *Ann Epidemiol* 2004;14:437-41.
21. Severo M, Gaio R, Lourenco P, Alvelos M, Bettencourt P, Azevedo A. Indirect calibration between clinical observers - application to the New York Heart Association functional classification system. *BMC Res Notes* 2011;4:276.
22. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: executive summary. Expert Panel on the Identification, Evaluation, and Treatment of Overweight in Adults. *Am J Clin Nutr* 1998;68:899-917.
23. Severo M, Pereira M, Bettencourt P, Gaio R, Azevedo A. B-type natriuretic peptide measured in serum--calibration using plasma samples for research purposes. *Clin Lab* 2011;57:1015-9.
24. Zaphiriou A, Robb S, Murray-Thomas T, Mendez G, Fox K, McDonagh T, Hardman SM, Dargie HJ, Cowie MR. The diagnostic accuracy of plasma BNP and NTproBNP in patients referred from primary care with suspected heart failure: results of the UK natriuretic peptide study. *Eur J Heart Fail* 2005;7:537-41.
25. Kuch B, Schunkert H, Muscholl M, Doring A, von Scheidt W, Hense HW. [Distribution, determinants and reference values of left ventricular parameters in the general population--results of the MONICA/KORA echocardiography studies]. *Gesundheitswesen* 2005;67 Suppl 1:S68-73.
26. Vermunt JK, Magidson J. Latent class cluster analysis. *Applied latent class analysis* 2002:89–106.
27. Wang CS, FitzGerald JM, Schulzer M, Mak E, Ayas NT. Does this dyspneic patient in the emergency department have congestive heart failure? *JAMA: the journal of the American Medical Association* 2005;294:1944-1956.
28. Guggenmoos-Holzmann I, van Houwelingen HC. The (in)validity of sensitivity and specificity. *Stat Med* 2000;19:1783-92.
29. Wang CS, FitzGerald JM, Schulzer M, Mak E, Ayas NT. Does this dyspneic patient in the emergency department have congestive heart failure? *JAMA: the journal of the American Medical Association* 2005;294:1944.

30. Pereira M, Carreira H, Vales C, Rocha V, Azevedo A, Lunet N. Trends in hypertension prevalence (1990-2005) and mean blood pressure (1975-2005) in Portugal: a systematic review. *Blood Press* 2012.
31. Goncalves A, Almeida PB, Lourenco P, Alvelos M, Betrencourt P, Azevedo A. Clinical significance of impaired relaxation pattern in middle-aged and elderly adults in the general population. *Rev Port Cardiol* 2010;29:1799-806.
32. Bettencourt PM. Clinical usefulness of B-type natriuretic peptide measurement: present and future perspectives. *Heart* 2005;91:1489-94.
33. Maisel AS, McCord J, Nowak RM, Hollander JE, Wu AH, Duc P, Omland T, Storrow AB, Krishnaswamy P, Abraham WT, Clopton P, Steg G, Aumont MC, Westheim A, Knudsen CW, Perez A, Kamin R, Kazanegra R, Herrmann HC, McCullough PA. Bedside B-Type natriuretic peptide in the emergency diagnosis of heart failure with reduced or preserved ejection fraction. Results from the Breathing Not Properly Multinational Study. *J Am Coll Cardiol* 2003;41:2010-7.
34. Severo M, Lopes C, Lucas R, Barros H. Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings. *Osteoporos Int* 2009;20:231-7.
35. Sullivan LM, Massaro JM, D'Agostino RB, Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med* 2004;23:1631-60.

Subchapter 3.1

B-Type Natriuretic Peptide Measured in Serum – Calibration Using Plasma Samples for Research Purposes.

Clin. Lab. 2011;57:XXX-XXX

SHORT COMMUNICATION

B-Type Natriuretic Peptide Measured in Serum – Calibration Using Plasma Samples for Research Purposes

M. SEVERO [1,2], M. PEREIRA [1,2], P. BETTENCOURT [3], R. GAIO [4,5], A. AZEVEDO [1,2]

[1] *Department of Hygiene and Epidemiology, University of Porto Medical School, Porto, Portugal*
[2] *Institute of Public Health of the University of Porto, Porto, Portugal*
[3] *Department of Internal Medicine, Hospital de S. João and University of Porto Medical School, Porto, Portugal*
[4] *Department of Mathematics, University of Porto Science School, Porto, Portugal*
[5] *Center for Mathematics, University of Porto, Porto, Portugal*

SUMMARY

Background: **We aimed to establish an equation for the estimation of the BNP concentration in plasma when only serum is available.**
Methods: **We enrolled 27 subjects aged at least 45 years, participating in a Portuguese cohort study. Blood samples were collected in plastic whole blood tubes, containing either ethylenediaminetetraacetic acid to obtain plasma or clot activator to obtain serum. The natural logarithm of serum BNP was calibrated with the natural logarithm of plasma BNP using a linear equation.**
Results: **The estimated regression parameters were 0.58 (95 % CI: 0.23 - 0.93) for β0 and 1.01 (95 % CI: 0.90 - 1.11) for β1. The absolute agreement between plasma BNP and that predicted by the equation according to the cut-off points 30 and 100 pg/mL were 96.3% (kappa = 0.92) and 96.3% (kappa = 0.91), respectively.**
Conclusions: **Serum samples cannot be used to estimate absolute plasma concentrations, but serum BNP values and the calibration equation can be used to classify correctly the individuals with the usual cut-offs.**
(Clin. Lab. 2011;57:XXX-XXX)

KEY WORDS

B-type natriuretic peptide; calibration; plasma; serum

INTRODUCTION

B-type natriuretic peptide (BNP) is secreted from the heart in response to pressure or volume overload. BNP levels increase when heart failure (HF) develops or worsens [1].
HF patients, even when stable, have higher BNP levels than persons with normal heart function.
In untreated subjects, a concentration of BNP under 100 pg/mL has high negative predictive value and makes HF an unlikely diagnosis [2].
The increase of BNP plasma concentration with the decline of heart function makes the blood concentration measurements of this peptide useful for clinical and research purposes [3].

In research projects, it is usual to store biological samples for later use. The commercially available immunefluorometric assay (Triage BNP Test, BIOSITE diagnostics, San Diego, CA, USA) recommends the use of whole blood or plasma specimens using EDTA as the anticoagulant to measure BNP. According to the manufacturer's recommendations, it should not be used to measure the BNP concentration in serum.
Thus, the purpose of this study was to evaluate the accuracy of BNP measurements in serum samples using this commercially available immunofluorometric assay to predict the BNP plasma concentration and to classify the individuals using the usual cut-off points. For that, we compared BNP plasma levels with BNP serum levels, in samples frozen at −20 °C, without protease inhibitors, using a sample of community subjects that had both serum and plasma samples stored as part of a population-based survey.

Short Communication accepted May 20, 2011

MATERIAL AND METHODS

Among participants in a health and nutrition survey of the adult population of Porto, Portugal [4], we enrolled all 27 subjects aged at least 45 years for whom plasma and serum samples had been collected and were still available. In this convenience sample, the mean (standard deviation) age was 67 (9) years, 14 (52%) were women and the mean (standard deviation) body mass index was 28.3 (5.0) kg/m^2. Blood samples were collected in plastic whole blood tubes (BD Vacutainer Systems, Plymouth, UK).

For plasma samples the tubes contained ethylenediaminetetraacetic acid (EDTA), whereas for serum samples the tubes contained clot activator (silica particles that coat the walls of the tube). Both were centrifuged within 2 hours of collection at 4000 rpm (2000 x g) for 10 minutes at room temperature. Aliquots of plasma and serum were separated and both were stored at −20 °C until measurement. Measurements were made at the same time, using the same protocol, after thawing the samples at room temperature for 15 minutes. BNP was measured using a commercially available immunofluorometric assay (Triage BNP Test, BIOSITE Diagnostics, San Diego, CA, USA). The reagents necessary to run the test are murine BNP monoclonal antibodies and BNP polyclonal antibodies labeled with a fluorescent dye. The analytical measuring range of the test varies from 5 pg/mL to 5000 pg/mL. The reported coefficient of variation is 9.9 % for mean BNP plasma concentration 71.3 pg/mL, 12.0 % for 629.9 pg/mL, and 12.2 % for 4087.9 pg/mL. For statistical analysis, BNP values below the detection threshold (5.0 pg/mL) were assumed as 2.5 pg/mL.

The distribution of BNP values was summarized as a median (interquartile range (IQR)). The fractional polynomials [5] were used to equate, thus calibrating the serum with plasma BNP values.

Correlations were performed with Spearman's correlation coefficient and the agreement was assessed with Bland-Altman plots and Cohen's kappa. Guidelines for interpreting kappa statistics suggest that values between 0.81 - 1.00 indicate almost perfect agreement, 0.61 - 0.80 substantial agreement, 0.41 - 0.60 moderate agreement, 0.21 - 0.40 fair agreement, and values less than 0.21 are poor or slight agreement [6].

The performance of the calibration equation was evaluated by leave-one-out cross-validation. Briefly, the equation was trained on 27 minus 1 individuals, and the trained equation was then used to test the individual that had been left out. This process was repeated until every individual in the dataset had been used once as an unseen test individual. The agreement estimated by leave-one-out cross-validation was then compared with the one estimated using the whole sample to evaluate possible over-fitting.

RESULTS

The median time of storage was 42 months (IQR = 1). In plasma, the BNP concentration ranged from minimum value of 2.5 pg/mL to a maximum value 635.0 pg/mL with median value of 21.0 pg/mL (IQR = 133.12). In serum, the BNP concentration ranged from 2.50 pg/mL to 440.0 pg/mL with median value of 10.6 pg/mL (IQR = 49.5). BNP concentration was significantly higher in plasma than in serum ($p < 0.001$). The median value of the difference between plasma and serum concentration was 11.7 pg/mL (IQR = 40.3) However, Spearman's correlation coefficient between the plasma and serum BNP concentrations was 0.96 (Figure 1). So we used fractional polynomials to equate, thus calibrating the serum with plasma BNP. Considering that the distribution of BNP concentration is skewed, we calibrated the natural logarithm (ln) of plasma BNP. The estimated equation was the following:

$$\ln(BNPplasma) \sim \beta_0 + \beta_1 \ln(BNPSerum)$$

$$\ln(BNPplasma) \sim 0.58 + 1.01 \times \ln(BNPSerum)$$

The estimated regression parameters were 0.58 [95 % confidence interval (95 % CI): 0.23 - 0.93] for β_0 and 1.01 (95 % CI: 0.90 - 1.11) for β_1. The difference between the ln plasma BNP and ln plasma BNP predicted by the equation showed a mean difference of 0 and limits of agreement ranging from -0.74 to 0.74 (Figure 2). Thus, the ratio between plasma BNP and plasma BNP predicted by the equation showed mean ratio of 100% and limits of agreement ranging between 48 % and 210 %. According to the cut points 30 and 100 pg/mL, the absolute agreement between the plasma BNP and the BNP predicted by the equation was 96.3 % (kappa = 0.92; 95 % CI: 0.78 - 1.00) and 96.3 % (Kappa = 0.91; 95 % CI: 0.74 - 1.00), respectively (Table 1).

The leave-one-out cross-validation showed that the limits of agreement between the ln plasma BNP and ln plasma BNP predicted by the calibration equation increased slightly, ranging from -0.80 to 0.79, and the accuracy and the kappa coefficient were exactly the same.

Table 1. Agreement between plasma BNP and that predicted by the calibration equation according to the cut-off points 30 and 100 pg/mL.

Predicted plasma BNP using serum samples			
Plasma BNP	<30 pg/mL	≥30 pg/mL	Absolute Agreement (kappa)
<30 pg/mL	14 (51.9%)	0 (0%)	96.3% (0.92)
≥30 pg/mL	1 (3.7%)	12 (44.4%)	
	<100 pg	≥100 pg	
<100 pg/mL	18 (66.7%)	1 (3.7%)	96.3% (0.91)
≥100 pg/mL	0 (0.0%)	8 (29.6%)	

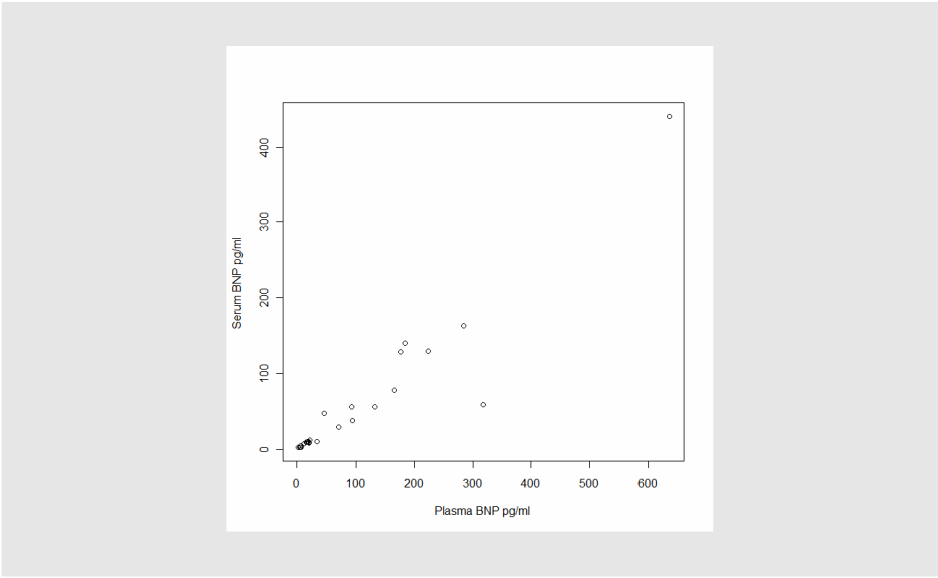

Figure 1. BNP concentration in serum versus BNP concentration in plasma.

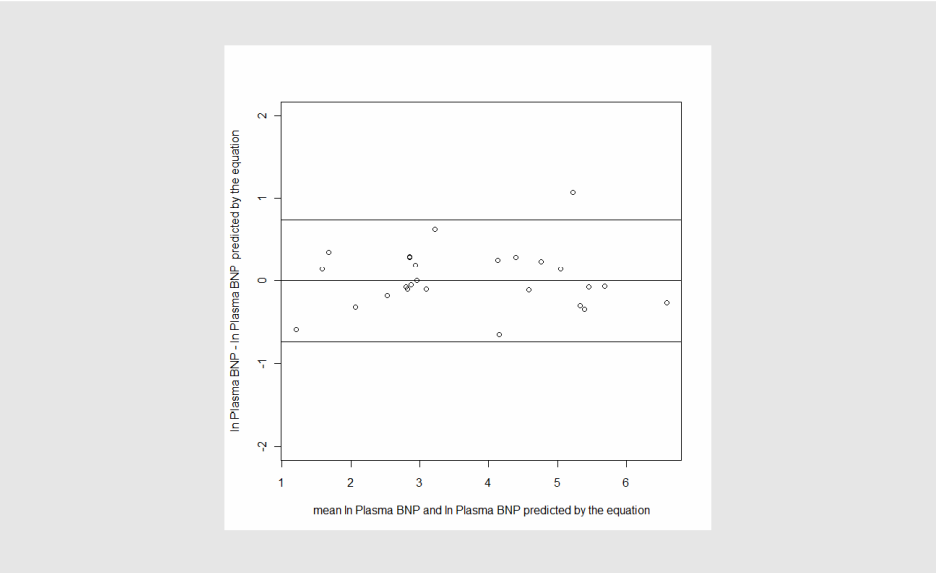

Figure 2. Bland-Altman plot for the difference between ln of Plasma BNP and ln of Plasma BNP predicted by the calibration equation.

DISCUSSION

Our study lends support to the usefulness of BNP measurements in serum samples to classify community individuals with the usual plasma BNP cut-offs [2,7] when only serum samples are stored.

Serum BNP concentration was lower than plasma concentration, suggesting that the method used to extract clotting factors influences the concentration of BNP. This result and the strong correlation coefficient between the plasma and serum BNP concentration confirm the validity of establishing an equation that will allow estimation of the BNP value in plasma when only serum is available.

After equating, no significant difference between the predicted plasma BNP using serum samples and the plasma BNP values was found, although the limits of agreement showed 2-fold differences in relation to the absolute plasma concentration.

This shows that serum samples cannot be used to estimate absolute plasma concentrations. However, the agreement was excellent when we use the BNP values predicted by the equation to classify the individuals with the usual cut-off points, showing the usefulness of BNP measurements in serum samples to classify the individuals.

Previous studies showed that the storage of samples at –20 °C or at –80 °C did not prevent BNP degradation, especially for high BNP values [8,9]. Thus, we cannot exclude that BNP degradation throughout storing time is different for the plasma and serum samples.

Consequently, the major limitation of this study was that we did not evaluate if the equation calibrates equally well at different measurement times. We acknowledge that, given this limitation, our results apply to serum stored in the same conditions as we describe and for the same time length, but cannot be generalized to other conditions.

Another possible limitation is that the BNP distribution varies across populations, which could limit the generalizability of the agreement, namely for patient populations. Still, our confidence in the generalization of the calibration equation is enhanced since the cross-validation showed that the agreement was almost the same.

In conclusion, the serum samples cannot be used to estimate absolute plasma concentration values, but we can use serum BNP values and the equation to classify the community individuals with the usual cut-offs.

Declaration of Interest:
The authors declare that they have no competing interests.

References:

1. Suttner SW, Boldt J. Natriuretic peptide system: physiology and clinical utility. Curr Opin Crit Care. 2004;10(5):336-41.
2. Dickstein KA, Cohen-Solal G, Filippatos, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). Eur Heart J 2008; 29(19):2388-442.
3. Bettencourt PM, Clinical usefulness of B-type natriuretic peptide measurement: present and future perspectives. Heart 2005; 91(11):1489-94.
4. Ramos E, Lopes C, Barros H. Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction. Ann Epidemiol 2004;14(6):437-41.
5. Royston P. Ambler G. Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 1999;28(5):964-74.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-74.
7. Zaphiriou A, Robb S, Murray-Thomas T, et al. The diagnostic accuracy of plasma BNP and NTproBNP in patients referred from primary care with suspected heart failure: results of the UK natriuretic peptide study. Eur J Heart Fail 2005;7(4):537-41.
8. Pereira M, Azevedo A, Severo M, et al. Long-term stability of endogenous B-type natriuretic peptide during storage at -20 degrees C for later measurement with Biosite Triage assay. Clin Biochem 2007;40(15):1104-7.
9. Pereira M, Azevedo A, Severo M, et al. Long-term stability of endogenous B-type natriuretic peptide after storage at -20 degrees C or -80 degrees C. Clin Chem Lab Med 2008;46(8):1171-4.

Correspondence:
Milton Severo
Serviço de Higiene e Epidemiologia
Faculdade de Medicina do Porto
Alameda Prof. Hernâni Monteiro
4200-319 Porto
Tel.: +351225513652
Fax: +351225513653
E-mail: milton@med.up.pt

Subchapter 3.2

Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings.

Osteoporos Int (2009) 20:231–237
DOI 10.1007/s00198-008-0647-6

ORIGINAL ARTICLE

Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings

M. Severo · C. Lopes · R. Lucas · H. Barros

Received: 3 December 2007 / Accepted: 28 April 2008 / Published online: 5 June 2008

Abstract
Summary The study aim was to develop a tool (software and ruler) to assess the dietary calcium and vitamin D intakes in Portugal, and evaluate the usefulness of non-dietary variables as intake predictors. Our findings indicated that is possible to estimate both using three and six food items, respectively, and non-dietary predictors.
Introduction The study aim was to develop a tool to assess the dietary calcium and vitamin D intakes in Portugal, and evaluate the usefulness of non-dietary variables as predictors.
Methods Trained interviewers collected information of 2,414 adults of Porto, Portugal, using a structured questionnaire and a validated semi-quantitative food frequency questionnaire (FFQ). Food items with the highest contribution to the total intake and non-dietary predictors (gender, age and body mass index (BMI)) were selected for the tool. Different statistical approaches were used to predict the intake. A Bland–Altman plot compared the predictions from the tool and the full FFQ.
Results The items selected to predict intake were milk (38%), cheese (12%), yogurt (10%) and gender for calcium and oily fish (39%), canned fish (9%), white fish (7%), eggs (5%), red meat (5%), age and BMI for vitamin D. The Bland–Altman plot showed that the mean differences were 0.0 (limits of agreement = [-220.67; 220.77]) mg/day and 0.0 (limits of agreement = [-1.03; 1.05]) μg/day, respectively for calcium and vitamin D.
Conclusion The equations estimated by the best statistical model to predict the calcium and vitamin D intake allowed for the design of a software and a circular ruler useful in clinical settings.

M. Severo (✉) · C. Lopes · R. Lucas · H. Barros
Department of Hygiene and Epidemiology,
University of Porto Medical School,
Alameda Prof. Hernani Monteiro,
4200-319 Porto, Portugal
e-mail: milton@med.up.pt

Keywords Brief food frequency questionnaire · Calcium · Osteoporosis · Vitamin D

Introduction

Osteoporosis is the main cause of fractures in middle-age men and women. The lifetime risk of fracture, regardless of anatomical location, is almost 40% for white women and 13% for white men from age 50 onward [1].

A study in Portugal including 5,964 women aged 20 and 89 years showed a 10.1% age-standardized prevalence of osteoporosis [2], while another study found a 2% prevalence in men [3].

Appropriate doses of calcium and vitamin D intake were shown to be pharmacologically active, safe, and effective for the prevention and treatment of osteoporotic fractures [4, 5].

The population reference intake (PRI) for calcium defined by the European Scientific Committee on Food was 700 mg calcium per day for adults [6]. In North America the adequate intakes were defined between 1000 to 1200 mg of calcium per day for adults [7]. The current allowance of vitamin D recommended by most European countries is 5 μg/day for adults but 10 μg/day for individuals older than 60–65 years. The recommended PRI, by the European Scientific Committee on Food, is 0–10 μg/day and 10 μg/day for people aged 18 to 64 and over 65 years old, respectively [8]. The adequate intakes were defined as 5 μg/day [7]; however the 2005 Dietary Guidelines for Americans recommend consuming extra vitamin D (25 μg/day) in populations at increased risk of

deficiency, such as older adults, people with darker skin, and exposed to insufficient UV radiation [9].

Large-scale epidemiological studies typically use food frequency questionnaires (FFQs) to assess nutrient intake [10]. For practical reasons, the use of brief FFQs can be more efficient in clinical settings. These brief tools are simpler to complete and less expensive than the larger ones [11, 12]. Several studies have developed and validated short versions of FFQ to assess calcium [13–16] and vitamin D intakes individually [17]. Also, brief FFQ computerized versions were designed to assess calcium intake in Belgian women [16].

The most common approach to choose a list of food items for a brief FFQ, designed to assess a specific nutrient, is based on the identification of the items with the highest levels of that particular nutrient or those foods and beverages that explain most of the total nutrient intake assessed by a comprehensive FFQ. Then, linear regression models are commonly used to estimate absolute specific nutrient dietary intake, based only on the food items selected [15]. However, other statistical approaches may be used, such as other generalized linear models, classification and regression trees or neural networks. Also, previous studies included only dietary variables in the models regardless of possible non-dietary predictors of intake.

Our study aim was to develop a quick and easy tool to assess, in the busy clinical setting, the absolute dietary calcium and vitamin D intakes regardless of supplementation, in Portugal. Additionally, the usefulness of non-dietary variables, namely gender, age and body mass index (BMI) as intake predictors was evaluated.

Participants and methods

As previously described elsewhere, non-institutionalized inhabitants of Porto, Portugal, older than 18 years, were selected using random digit dialling, during the assembling of the EpiPorto cohort [18]. In brief, after the identification of a household, permanent residents were characterized according to age and gender, and one individual was selected by simple random sampling and invited to visit our department for interview and physical examination. If there was a refusal, replacement was not allowed. The participation proportion was estimated in 70% [19]. The São João University Hospital Ethics Committee approved the study and all participants gave written informed consent.

Trained interviewers collected information using a structured questionnaire, which comprised socio-demographic data, personal and family medical history and behavioural characteristics. Diet was recorded using a semi-quantitative FFQ designed according to Willet and colleagues [10] and adapted and validated for the Portuguese population [20, 21]. The FFQ used had been previously validated by comparison with four 7-day food records (each one in a different season of the year). The FFQ referred to the previous 12 months and comprised 86 food items or beverages categories and a frequency section with nine possible responses ranging from never to six or more times per day. Participants were also asked to specify any items not included in the list mentioned by the interviewer. For each food item, in addition to the average frequency of consumption, a predetermined portion size based on a photograph manual with three portion sizes (i.e., small, medium, or large) was chosen by participants. Nutrient intake data were obtained by multiplying the frequency of consumption of each food item by the nutrient content of the standard portion size in grams (g). The software Food Processor Plus® (ESHA Research, Salem, Oregon) was used. The database underlying this software is based on nutritional values published by the US Department of Agriculture. In the present study nutritional values for typical national foods and recipes were computed and considered for the calculation of the intake, using the Portuguese Tables of Food Composition [22, 23]. The contents in nutrients of the different food items were estimated after cooking and processing.

Among the 2,488 participants, included in the evaluation, we excluded 56 subjects who scored less than 24 in the Mini Mental State Examination [24], six who were not capable of answering the FFQ for reasons other than cognitive impairment and 12 that refused to complete the questionnaire. A sample of 2414 participants (61.7% women) remained for analysis, with a mean (standard deviation) age of 52.4 (sd=15.3) years and 8.7 (sd=5.1) years of education.

Tool development

The first step was to describe the distribution of calcium and vitamin D intake, Mann-Whitney test was used to compare two independent samples, Spearman rank correlation test was use to evaluate the correlation and Kolmogorov–Smirnov to test the distribution of both nutrients.

The next step was to identify all foods items with the highest contribution to the total intake of calcium and vitamin D, assessed by the full FFQ, without considering nutrient intakes from supplements. In order to determine how many food items should be retained to the tool, we plotted the contribution of each food item in descending order and look for a "big gap" or an "elbow" after achieving 50% of the total contribution, separately for calcium and vitamin D. Gender, age and body mass index were added to selected food items as predictors.

Different statistical approaches were used to choose the model that best predicted the intake of calcium and

vitamin D. Five predictive models were tested. The first set of models estimated the intake as continuous variable–generalized linear models (GLM) with link function identity and regression trees; the second group predicted categories of intake (tertiles)–multinomial regression, classification regression and neural networks.

The prediction reliability of each model was evaluated by tenfold cross validation [25]: the sample was divided in ten partitions, the model was constructed with nine out of the ten partitions, and the proportion of error of the predictions in the subset left out (prediction error) was calculated. In this study an error was defined as a prediction value classified in a different tertile from the true value estimated by the full FFQ. This procedure was repeated for every subset. The final prediction error of each model was estimated by the mean prediction error off al ten subsets.

Wilcoxon test was use to compare the prediction error distribution (distribution of the prediction error estimated in each subset) between models. The prediction error distribution of the GLM model (standard model) was compared with distribution of all other models.

After choosing the final predicted model, was measured the effect of gender, age and BMI as non-dietary predictors of intake for the food items not selected to integrate the model. Likelihood ratio test was used to measure the effect of the inclusion of non-dietary predictors in the model by comparing the model with and without the non-dietary predictors. The Wald's test was use to measure the effect of each non-dietary predictor. The agreement between the predictions from the model selected and the FFQ was evaluated by kappa and Bland–Altman plots. The model selected to predict the calcium and vitamin D intake was use to design software and a circular ruler useful in clinical settings. Data was analyzed using the software R 2.4.0.

Results

Calcium and vitamin D Intake

Both nutrients followed a gamma distribution. The median (interquartile range (IQR)) calcium intake was 871.5 (IQR= 543.6) mg/day. Women and younger subjects with lower body mass index had higher intake levels (Table 1). The total mean intake of vitamin D was 3.1 (IQR=2.3) μg/day and older people had lower intake levels (Table 1).

Table 1 Calcium (mg/day) and vitamin D (μg/day) intake by gender, body mass index (BMI) and age

	1st Quartile	Median	3rd Quartile	rho Spearman	P
Calcium					
	647.3	871.5	1191.0		
Gender					
Female	659.8	918.0	1247.0		<0.001
Male	634.2	814.2	1095.4		
BMI (kg/m2)					
[15–25)	670.4	920.0	1236.0	-0.10	<0.001
[25–30)	663.7	861.0	1173.5		
[30–52]	599.8	788.2	1127.3		
Age (years)					
[18–30)	760.5	1060.2	1338.5	-0.11	<0.001
[30–45)	663.7	869.0	1198.3		
[45–65)	639.8	845.5	1178.4		
[65–92]	617.4	839.8	1129.4		
Vitamin D					
	2.4	3.1	4.7		
Gender					
Female	2.4	3.1	4.7		0.375
Male	2.4	3.2	4.8		
BMI (kg/m2)					
[15–25)	2.5	3.2	4.7	-0.04	0.062
[25–30)	2.4	3.1	4.7		
[30–52]	2.4	3.2	4.8		
Age (years)					
[18–30)	2.5	3.3	4.8	-0.09	
[30–45)	2.5	3.1	4.7		<0.001
[45–65)	2.5	3.2	4.8		
[65–92]	2.1	2.8	4.5		

Identification of food items with the higher contribution to calcium and vitamin D intake

The foods items with the highest contribution to calcium intake were: semi-skimmed milk (27.0%), cheese (11.5%), yogurt (9.6%) and skimmed milk (8.9%). For the intake vitamin D, contributors were oily fish (38.6%), canned fish (9.1%), white fish (7.4%), eggs (4.9%), red meat (4.9%) and cod fish (4.7%) (Table 2).

Evaluation of the different calcium and vitamin D intake prediction models

The models were constructed using the food items with the highest percent contribution and including also gender, age and body mass index as indirect indicators of calcium and vitamin D intakes from the remaining foods and beverages.

Calcium intake predictive models considered three items: total milk (aggregating semi-skimmed milk, skimmed milk and whole milk), yogurt and cheese.

The mean of the prediction error in tertiles of intake for ten subsets were 17.1% for generalized linear model, 25.7% for regression trees, 17.1% for multinomial regression, 22.9% for classification trees and 18.3% for neural networks. The GLM distribution of the prediction error was compared with the other models, and were obtain significant differences with regression trees (p=0.001), classification trees (p=0.002) and neural networks (p=0.048)

Table 2 The 10 food items with the highest contribution to calcium and vitamin D intake

	Calcium		Vitamin D	
1	Semi-skimmed milk	27.0%	Oily fish	38.6%
2	Cheese	11.5%	Canned fish	9.1%
3	Yogurt	9.6%	White fish	7.4%
4	Skimmed milk	8.9%	Eggs	4.9%
5	White bread	5.1%	Red meat	4.9%
6	Soup	3.7%	Cod fish	4.8%
7	Wheat bread	2.5%	Cereals	3.7%
8	Dry beans and pies	2.5%	Semi-skimmed milk	3.1%
9	Whole milk	1.8%	Croquette, rissole and similars	2.9%
10	Orange	1.8%	Cheese	2.6%

models, and was not found with multinomial regression (p=0.312). The same indirect indicators were considered for vitamin D, the food items selected were: oily fish, canned fish, white fish, eggs, red meat and cod fish. The mean of the prediction error in tertiles of intake for ten subsets were: 18.4% for generalised linear models, 25.3% for regression trees, 18.1% for multinomial regression, 22.1% for classification trees and 20.4% for neural networks. As with calcium, significant differences were found with regression trees (p=0.003), classification trees (p=0.001) and neural networks (p=0.021) model and with multinomial regression model were not found (p=0.799), when compared with the generalised linear model.

The prediction of the final model

For both nutrients the final model chosen was GLM with gamma distribution (Table 3). After adjusting for the selected food items men had higher calcium intake levels than women from other food items (β=40.6 mg/day, p<0.001). Younger subjects (β=-0.007 μg/day, p<0.001) with lower body mass index (β=-0.015 μg/day, p<0.001) had higher vitamin D intake levels from other food items. Likelihood Ratio Test, compared the model only with dietary predictors with the model with dietary and non-dietary predictors and showed that there were significant differences between both models respectively for calcium (p<0.001) and vitamin D (p<0.001).

The full FFQ and the final model predictions of calcium intake were split into two categories according to the PRI and then compared. Eighty nine percent of the subjects were correctly classified (Table 4) and kappa was 0.75 (95%CI: 0.72–0.78). The vitamin D intake was divided into two categories (0 to 5 μg and over 5 μg) according the recommendation of most European countries intake, and 94% of the subjects were correctly classified (Table 4) and kappa was 0.83 (95%CI: 0.80–0.86).

A Bland–Altman plot showed that the mean differences between Full FFQ and the predicted model selected were 0.0 with 95% limits of agreement between −220.7 to 220.8 mg/day for calcium and 0.0 with limits of agreement between −1.0 to 1.0 μg/day for vitamin D (Fig. 1).

Tool for assessing the calcium and vitamin D intake

Using these results, one could design and build software and a circular ruler to estimate the calcium and vitamin D intake.

Software

The software version was divided in two parts: one were the user insert the gender, age, height and weight and a second were he chooses the frequency of the food items selected for the brief tool of intake (milk, cheese and yogurt for calcium and oily fish, canned fish, white fish, eggs and red meat for vitamin D). The frequency was composed of nine possible responses ranging from never to six or more times per day as the original full FFQ. After the introduction of this information the software uses the GLM equations estimated to calculate the calcium and vitamin D individual intake.

Circular ruler

The circular ruler was divided in two parts: one to measure the calcium intake and another to measure the vitamin D intake (Fig. 2). Each one was composed with four circumferences. In the first part we transformed the intake between 0 to 3000 mg/day of calcium in 0 to 360°. In the second part we transform 0 to 12 μg/day of vitamin D in 0 to 360°. Next, we calculate for each food item selected and

Table 3 Generalized linear model with gamma distribution (link function identity) for calcium and vitamin D intake

Calcium			Vitamin D		
	Coefficients	p		Coefficients	p
Intercept	360.590	<0.001	Intercept	1.748	<0.001
Milk	1.169	<0.001	Oily fish	0.076	<0.001
Cheese	8.026	<0.001	Canned fish	0.073	<0.001
Yogurt	1.783	<0.001	White fish	0.009	<0.001
—			Cod fish	0.009	<0.001
—			Eggs	0.024	<0.001
—			Red meat	0.002	<0.001
Gender (male)	40.592	<0.001	Gender (male)	0.004	0.841
Age	-0.101	0.533	Age	-0.007	<0.001
BMI	-0.893	0.053	BMI	-0.015	<0.001

Table 4 Comparison between the full FFQ and the GLM model predictions of calcium and vitamin D intake

	Calcium model (mg/day)	[173–700] n (%)	(700–2632] n (%)	Total (%)
Calcium FFQ (mg)	[173–700]	602 (25.3)	135 (5.7)	737 (31.0)
	(700–2632]	116 (4.9)	1522 (64.1)	1638 (69.0)
	Total (%)	718 (30.2)	1657 (69.8)	
	Vitamin D model (μg/day)	[0.4–5.0] n (%)	(5.0–11.9] n (%)	Total (%)
Vitamin D FFQ (μg)	[0.4–5.0]	1802 (75.9)	70 (2.9)	1872 (78.9)
	(5.0–11.9]	63 (2.7)	439 (18.5)	502 (21.1)
	Total (%)	1865 (78.6)	509 (21.4)	

frequency per day the corresponding calcium and vitamin D intake using the equations estimated by the GLM and transform the intake in degrees.

In the first circumference for the calcium ruler we sign the calcium intake in degrees for nine categories of frequency of milk intake, in the second the intake of yogurt, in the third the intake of cheese and in the last circumference was added the gender and the constant contribution to the intake.

In the same way for vitamin D, we sign to the first circumference the intake of vitamin D in degrees for the nine categories of oily fish and canned fished aggregated because the coefficients in the GLM equations were similar (0.076 and 0.073, respectively), in the second white fish and cod fish aggregated by the same reason (0.009 in both), in the third the intake of eggs and in the last circumference the red meat and the constant.

It was possible then to add each food item contribution and estimate the calcium and vitamin D intake. Example: to know the daily calcium intake of a subject who has a daily average intake of two cups of milk, one 125 g yogurt, and a weekly intake of three 30 g slices of cheese, we would match the zero of the inner circle (yogurt circle) with the daily frequency of two cups of milk, then match the zero of the next inner circle (cheese circle) with the 1 yogurt daily frequency and do the same with the last inner circle (sex circle) and match with 3/7 of daily cheese slices, finally we look at the result in grams of calcium in the outer disk, according to sex.

Discussion

This study identified the food items that contributed with the greatest proportion of calcium and vitamin D intake in a population-based survey. The results confirm that the food items with higher contribution to calcium intake were milk, cheese and yogurt; while for vitamin D they were oily fish, canned fish, white fish, eggs and red meat. These food items are the same as other countries [26, 27].

As a consequence milk, yogurt and cheese were included in the brief tool for the estimation of calcium intake. Oily

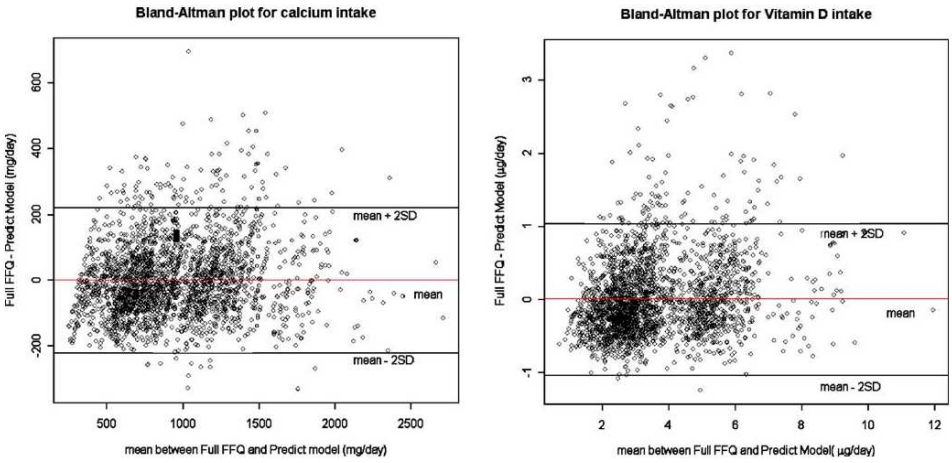

Fig. 1 A Bland–Altman plot between the full FFQ and the GLM model predictions for calcium and vitamin D intake

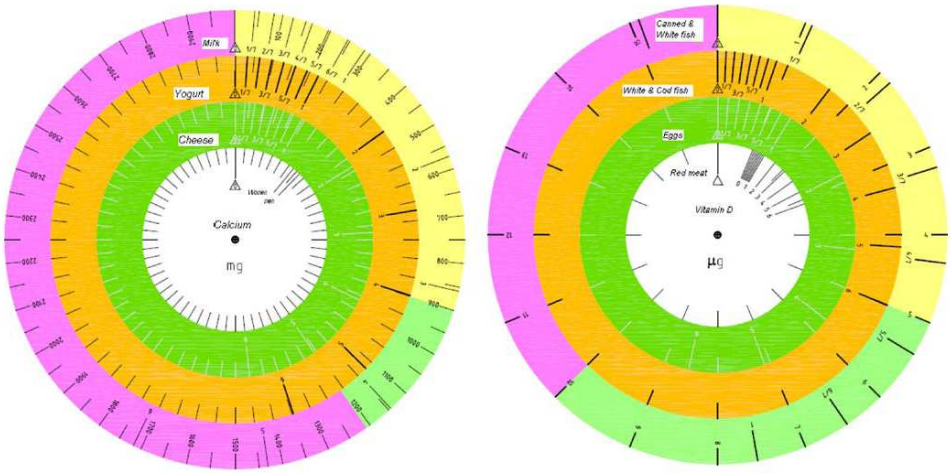

Fig. 2 Circular ruler to estimate the calcium and vitamin D intake

fish, canned fish, white fish, cod fish, eggs and red meat were used in the measurement of vitamin D intake.

Our study showed that men had on average higher intakes of calcium after adjusted for milk, cheese and yogurt consumption and the vitamin D intake decreased with increasing age and body mass index after adjusting the food items with the highest contribution (Table 3). This shows that is possible to assess the intake of calcium and vitamin D among the other food items using non-dietary predictors, namely gender, age and body mass index. The use of indirect indicators to estimate the intake of calcium and vitamin D could be advantageous to brief tool designed to be applied in the general population [28].

The generalization linear model and the multinomial showed the lowest prediction error mean; however the first has the advantage of predicting the quantity of intake, while the second only provides the category of intake. Additionally, when the distribution of the prediction error for the generalized linear regression model was compared with the other statistical approaches, the distribution was significantly lower or similar. This means that the use of regression models to estimate absolute specific nutrient dietary intake [15] is a good statistical approach, moreover when the model is generated from a nutrient with normal distribution or when the sample from which the model is generated is large. Conversely, in cases where the distribution is not normal (gamma distribution) and the sample is small the estimate of the individual intake, and to a lesser extent the population intake, may be affected.

Using this study, one could design and build software and a circular ruler to estimate calcium and vitamin D intakes (Fig. 2), which are efficient in suppressing the time and resource limitations in clinical settings. The circular ruler advantage is that is possible to use when no computer is available or among people without special computer skills. This can be important when the use of computer is not generalized. An disadvantaged is that was not possible to construct a version with all significant variables for vitamin D, that is the ruler did not had in count the effect of age and body mass index. In terms fitness, the percentage of individuals correctly classified were 94%, the kappa was 0.83 and the mean difference between the model only with the food items predictions and the full FFQ was 0.0 with limits of agreement between −1.1 to 1.1 µg/day for vitamin D.

Although calcium supplementation up to 1500 mg/day and vitamin D up to 25 µg (1000 IU) has been regarded as safe [29] the excess of these nutrients may be related with some health problems [30]; therefore, this brief tool may have practical clinical relevance in the decision of prescribing supplementation or a rich diet in calcium and vitamin D. Nevertheless, while relevant, food consumption is a part of the daily supply of vitamin D; sunlight exposure is the most relevant source of the hormone, so the decision of prescribing supplementation or a rich diet in calcium and vitamin D should have this in account.

However, this study presents some limitations: we did not administer an abbreviated instrument consisting only with the items selected, which could have artificially inflated the level of agreement; and the major sources of calcium and vitamin D can vary across different subpopulations, which could limited the generalization of the prediction model. Still our confidence in the generalization

of the model is enhanced since the studied population is substantially heterogeneous in age, gender, education. And on another hand the cross-validation showed prediction error did not increase significantly in any one of the subsets ([12.6–21.1] and [14.8–21.8], for calcium and vitamin D).

Our findings indicate that the brief tool to estimate the calcium and vitamin D using three and six food items and non-dietary predictors, respectively, was comparable with the full 86-item FFQ. The software or circular ruler may then provide an efficient way to assess calcium and vitamin D dietary intake for the general population in epidemiological surveys or in clinical settings, by means of a quick an easy tool, and taking into consideration that is not possible to measure the intake of this two nutrients directly, this study provides a tool useful for health professionals in the prevention and management of osteoporosis.

Acknowledgements The authors wish to thanks the contribution of Fernanda Pereira and José Relvas in the circular ruler conception.

Source of funding Supported by Fundação para a Ciência e Tecnologia, POCI/SAUESP/61160/2004

Conflicts of interest None

Reference

1. Melton LJ 3rd, Chrischilles EA, Cooper C, Lane AW, Riggs BL (1992) Perspective. How many women have osteoporosis? J Bone Miner Res 7:1005–1010
2. Araújo D, Pereira J, Barros H (1997) Osteoporose em mulheres portuguesas. Acta Reum Port 82:7–13 (in Portuguese)
3. Silva J, Carrapito H, Reis P (1999) Diagnóstico densitométrico de osteoporose: critérios de referência na população portuguesa. Acta Reum Port 22:9–18 (in Portuguese)
4. Boonen S, Rizzoli R, Meunier PJ, Stone M, Nuki G, Syversen U, Lehtonen-Veromaa M, Lips P, Johnell O, Reginster JY (2004) The need for clinical guidance in the use of calcium and vitamin D in the management of osteoporosis: a consensus report. Osteoporos Int 15:511–519
5. Boonen S, Vanderschueren D, Haentjens P, Lips P (2006) Calcium and vitamin D in the prevention and treatment of osteoporosis-a clinical update. J Intern Med 259:539–552
6. Opinion of the scientific committee on food on the tolerable upper intake level of calcium. In. European Commission Scientific Committee on Food
7. Vedral J (1997) Dietary reference intakes: for calcium phosphorus magnesium vitamin d and fluoride. National Academy Press
8. Opinion of the scientific committee on food on the tolerable upper intake level of vitamin d. In. European Commission Scientific Committee on Food
9. Johnson MA, Kimlin MG (2006) Vitamin D, aging, and the 2005 Dietary Guidelines for Americans. Nutr Rev 64:410–421
10. Willett W (1998) Food frequency methods. Nutritional Epidemiology
11. Andersen LF, Johansson L, Solvoll K (2002) Usefulness of a short food frequency questionnaire for screening of low intake of fruit and vegetable and for intake of fat. Eur J Public Health 12:208–213
12. Byers T, Marshall J, Fiedler R (1985) Assessing nutrient intake with an abbreviated dietary interview. Am J Epidemiol 122:41–50
13. Sebring NG, Denkinger BI, Menzie CM, Yanoff LB, Parikh SJ, Yanovski JA (2007) Validation of three food frequency questionnaires to assess dietary calcium intake in adults. J Am Diet Assoc 107:752–759
14. Cummings SR, Block G, McHenry K, Baron RB (1987) Evaluation of two food frequency methods of measuring dietary calcium intake. Am J Epidemiol 126:796–802
15. Blalock SJ, Currey SS, DeVellis RF, Anderson JJ, Gold DT, Dooley MA (1998) Using a short food frequency questionnaire to estimate dietary calcium consumption: a tool for patient education. Arthritis Care Res 11:479–484
16. Matthys C, Pynaert I, Roe M, Fairweather-Tait SJ, Heath AL, De Henauw S (2004) Validity and reproducibility of a computerised tool for assessing the iron, calcium and vitamin C intake of Belgian women. Eur J Clin Nutr 58:1297–1305
17. Blalock SJ, Norton LL, Patel RA, Cabral K, Thomas CL (2003) Development and assessment of a short instrument for assessing dietary intakes of calcium and vitamin D. J Am Pharm Assoc 43:685–693
18. Santos AC, Ebrahim S, Barros H (2007) Alcohol intake, smoking, sleeping hours, physical activity and the metabolic syndrome. Prev Med 44:328–334
19. Ramos E, Lopes C, Barros H (2004) Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction. Ann Epidemiol 14:437–441
20. Lopes C, Aro A, Azevedo A, Ramos E, Barros H (2007) Intake and adipose tissue composition of fatty acids and risk of myocardial infarction in a male Portuguese community sample. J Am Diet Assoc 107:276–286
21. Lopes C (2000) Reproducibility and validity of semi-quantitative food frequency questionnaire. In Diet and Myocardial Infarction: A Community-Based Case–Control Study. In. University of Porto, Porto. (in Portuguese)
22. Ferreira F, Graça M (1985) Tabela da composição dos alimentos portugueses. Instituto Nacional de Saúde Dr. Ricardo Jorge. (in Portuguese)
23. Tabela da composição dos alimentos. Instituto Nacional de Saúde Dr. Ricardo Jorge. (in Portuguese)
24. Folstein MF, Folstein SE, McHugh PR (1975) "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12:189–198
25. Efron B, Tibshirani R (1997) Improvements on cross-validation: The.632+ bootstrap method. J Am Stat Assoc 92:548–560
26. Fleming KH, Heimbach JT (1994) Consumption of calcium in the U.S.: food sources and intake levels. J Nutr 124:1426S–1430S
27. Nakamura K, Nashimoto M, Okuda Y (2002) Fish as a major source of vitamin d in the japanese diet. Nutrition 18:415–416
28. Magkos F, Manios Y, Babaroutsi E, Sidossis LS (2006) Development and validation of a food frequency questionnaire for assessing dietary calcium intake in the general population. Osteoporos Int 17:304–312
29. Francis RM (2008) What do we currently know about nutrition and bone health in relation to United Kingdom public health policy with particular reference to calcium and vitamin D? Br J Nutr 99:155–159
30. Jackson RD, LaCroix AZ, Gass M, Wallace RB, Robbins J, Lewis CE, Bassford T, Beresford SA, Black HR, Blanchette P, Bonds DE, Brunner RL, Brzyski RG, Caan B, Cauley JA, Chlebowski RT, Cummings SR, Granek I, Hays J, Heiss G, Hendrix SL, Howard BV, Hsia J, Hubbell FA, Johnson KC, Judd H, Kotchen JM, Kuller LH, Langer RD, Lasser NL, Limacher MC, Ludlam S, Manson JE, Margolis KL, McGowan J, Ockene JK, O'Sullivan MJ, Phillips L, Prentice RL, Sarto GE, Stefanick ML, Van Horn L, Wactawski-Wende J, Whitlock E, Anderson GL, Assaf AR, Barad D (2006) Calcium plus vitamin D supplementation and the risk of fractures. N Engl J Med 354:669–683

Subchapter 3.3

A simple tool to match a concomitant variable latent class model classification

Title: A simple tool to match a concomitant variable latent class model classification

Milton Severo,[1,2] A. Rita Gaio, [3,4]

(1) Department of Hygiene and Epidemiology, University of Porto Medical School, Porto, Portugal.

(2) Institute of Public Health of the University of Porto, Porto, Portugal.

(3) Department of Mathematics, University of Porto Science School;

(4) Centre of Mathematics, University of Porto, Porto, Portugal

Correspondence:

Milton Severo

Serviço de Higiene e Epidemiologia

Faculdade de Medicina do Porto

Alameda Prof. Hernâni Monteiro

4200-319 Porto

milton@med.up.pt

Tel +351225513652

Fax: +351225513653

ABSTRACT

Concomitant variable latent class analysis is a multivariate statistical procedure for the identification of underlying categorical structures. In spite of its precision, it is fairly complex to provide general practitioners with at-hand effective tools. The purpose of this study is to reproduce the classification obtained from a concomitant variable latent class model using very simple algebraic calculations. We design a device for the heart failure diagnosis based solely on findings that are routinely collected in current clinical practice and that requires almost no calculations for the general practitioner.

key words: heart failure; concomitant variable latent class model; point system; circular ruler.

1. Introduction

Latent class analysis (LCA), also known as analysis of finite mixture models, is a classification method that can be helpful in disease diagnosis, amongst several other applications. It allows for the standardization and quantification of the probabilistic reasoning in clinical diagnosis, upon which decisions of further investigation and even treatment need to be made. LCA can be potentially improved through the use of concomitant variables, i.e., variables that influence the prevalence of the classes, thus permitting the identification of more precise categories. Latent class modeling brings several statistical advantages over standard classification approaches: it allows problems such as choice of the number of classes and of the classification method to be recast as statistical model choice problems; it allows for patterns prevalence to depend on a set of concomitant variables; and it produces posterior class membership probabilities for each individual, given values of the input and concomitant variables.

Concomitant variable latent class analysis has been employed to validate diagnostic tests in the absence of a perfect reference standard (1), which is the situation, for example, of heart failure (HF). This is a complex clinical syndrome resulting from a variety of structural or functional cardiac disorders. A clinical examination is always the first step in a diagnostic approach to possible HF and further investigation requires demonstration of cardiac dysfunction by imaging or functional tests. Individual symptoms (such as dyspnoea and fatigue) and signs (*e.g.* third heart sound and evidence of congestion) are generally unreliable and have limited value for diagnosing heart failure (2, 3). To avoid the need for expensive clinical measurements and as an

attempt to standardize the clinical assessment of HF, several multidimensional criteria based solely on symptoms and signs have been developed over decades (4-11). All but one of these studies defined HF directly as a combination of one or more symptoms and signs according to medical expertise; the criterion discrimination assessment consisted then of a cross validation of the criterion results with a classification obtained from an external exhaustive diagnosis. The remaining study used latent class analysis (without concomitant variables) but the approach did not succeed in popularity, most probably due to its statistical complexity, as far as application was concerned. Therefore there is yet no established instrument that can provide general practitioners with a quick, cheap, easy and effective diagnosis of HF. For practical purposes, simplicity may be better than complexity, even when simplicity results in some relative weakness. Clinical practitioners who are not research scientists may have difficulties in the implementation of complex statistical models in their practice routine.

A recent study conducted by the authors (12) succeeded in identifying three patterns of syndromic aggregation of symptoms and signs for heart failure, based on findings routinely collected in current clinical practice. These patterns were identified by a concomitant variable latent class model; its latent part accounted for known determinants of the relevant clinical findings, and the *a priori* probabilities were described by individuals risk profile. The model was validated against objective cardiac structural and functional parameters. Nevertheless, clinical application of the developed methodology requires statistical modeling and computer resources.

The purpose of this study is to reproduce the classification obtained from a concomitant variable latent class model using very simple algebraic calculations, and

to design a device for the heart failure diagnosis based solely on findings that are routinely collected in current clinical practice and that requires almost no calculations for the general practitioner.

2. **Concomitant variable latent class models**

For a fixed number, *K*, of classes, the model assumes that the population density *f* is expressed as a weighted finite sum of *K* component densities, $f_1,\ldots, f_k$, with parameters $\vartheta_1,.., \vartheta_k$, respectively, and each density is identified with a class. For the individual i, let $\boldsymbol{y}_i$ denote the response vector of observations on J variables. In our application, these input variables, also denoted by items or manifest variables, will be either dichotomous or ordinal. The general model is

$$f(\mathbf{y}_i \mid \mathbf{x}_i, \psi) = \sum_{k=1}^{K} \eta_{k|(\mathbf{x}_i,\alpha)} f_k(\mathbf{y}_i \mid \theta_k)$$

where η_k is the probability of class k membership, and $\boldsymbol{x}_i$ is a vector of concomitant variables that influences the prevalence of the classes through the parameters α. The vector $\psi=(\alpha, \vartheta_1,\ldots, \vartheta_K)$ is the set of model parameters that are to be estimated. The model assumes that $\sum_{k=1}^{K} \eta_k = 1$.

Assuming independence of the coordinate response vectors y_{ij} within each class k, and that the multivariate density f_k is the same across classes, say $f=(f_1,\ldots,f_J)$, the model writes

$$f(\mathbf{y}_i \mid \mathbf{x}_i, \psi) = \sum_{k=1}^{K} \eta_{k|(\mathbf{x}_i,\alpha)} \prod_{j=1}^{J} f_j(\mathbf{y}_{ij} \mid \theta_{jk})$$

The (sub)model that takes account of the input variables distribution within each class is denoted by component specific model, and the (sub)model that studies the influence of the concomitant variables on the classes prevalence is called concomitant variable model. The latter model will generally be a multinomial logit model with parameters α.

Individuals are assigned a class according to the standard modal allocation from posterior class memberships. The *a posteriori* probabilities are given by Bayes theorem:

$$P(k \mid y_i, x_i, \psi) = \eta_{k|(\mathbf{x}_i,\alpha)} \prod_{j=1}^{J} f_j(\mathbf{y}_{ij} \mid \theta_{jk}) \Big/ \sum_{k=1}^{K} \eta_{k|(\mathbf{x}_i,\alpha)} \prod_{j=1}^{J} f_j(\mathbf{y}_{ij} \mid \theta_{jk})$$

Design of the classification tool

The construction of a classification tool that could faithfully reproduce the classification of the model we have developed (12) was performed in three steps: the first two corresponded to the two sub-models of the concomitant variable latent class analysis, and the last step linked the previous two steps.

Latent class models are within the framework of latent models. The latter express the association between input variables and underlying latent variables. Roughly speaking,

latent models formulate a generalized linear model whose independent variable is the latent variable and the dependent variables are the items. Depending on the type of the prior distribution of the latent variable, the models are called latent class models (LCM) or latent trait models (LTM), for the types categorical or metrical, respectively.

As the *a priori* distribution of the latent variable has been shown to be essentially arbitrary, and its choice to be largely a convention, assuming a continuous *a priori* distribution will not change the fitting of the model (13).

There is no obvious choice between LTM and LCM. In the context of disease diagnosis the view that dominates is the categorical one, because it meets clinical needs and allows reporting for health-care planners. Moreover, LTM does not provide a classification, as it is difficult to find natural cut points or thresholds for the traits.

In the case of binary input data with a unique continuous latent variable, it is known that a sufficient statistic for the distribution of the latent variable is given by a linear combination of the item response variables, with weights equal to the slope coefficients estimated by the latent trait model (13). For individual i, we then have

$$T_i = \sum_{J=1}^{J} \alpha_j y_i$$

which is known as the score of the individual.

This property is lost when we move to ordinal data (graded response model), but it continues to hold for nominal data (generalized partial credit model), where the sufficient statistics are weighted sums of the responses with weights equal to the slope coefficient for each item and category. The score for individual i is then

$$T_i = \sum_{j=1}^{J} \sum_{s=0}^{m_j - 1} \alpha_{j(s)} y_{i,j(s)}$$

where m_j is the number of categories in item j and $y_{i,j(s)}$ takes the value 1 when the individual responds to the s^{th} category of item j and 0 otherwise. If $\alpha_{j(s)} = \alpha \cdot s$ for $s=0,..,m_j-1$ and $j=1,\ldots J$, i.e., all item slopes are equal and, for each item, the slopes of the several categories are proportional to sequential numbers, the previous score is

$$T_i = \sum_{j=1}^{J} \sum_{s=0}^{m_j - 1} \alpha\, s\, y_{i,j(s)}$$

Assume moreover that all components of the response vector are of dichotomous or ordinal type. Then the sum of the values each individual take on the items, denoted by **total score**, is a surrogate measure of that individual's value on the latent variable, as far as the component specific model is concerned.

In addition, we use a generalized linear model to extract an expected total score from the concomitant variables, which we call **concomitant score**, and think of it as a surrogate measure of the concomitant variables influence on the total score.

A classification tree with total and concomitant scores as predictors is then used to obtain the final classification.

Application to heart failure diagnosis

Study design and sample selection

Participants were selected within the first follow-up of a cohort, representative at baseline of the non-institutionalized adult population of Porto, Portugal – the EPIPorto cohort study. In 1999-2003, cohort assembly was done by random digit dialing, using households as the sampling frame, followed by random selection of one person aged 18 years or older in each household. Refusals were not substituted within the same household. The proportion of participation was 70%(14). At baseline, 2485 participants were recruited. Between October 2006 and July 2008, participants aged 45 years or over were eligible to a systematic evaluation of parameters of cardiac structure and function, which included a cardiovascular clinical history and physical examination, and a bidimensional transthoracic echocardiogram. Among 2048 cohort members that would be in the eligible age range at this time, 134 (6.5%) had died, 198 (9.7%) refused to be re-evaluated and 580 (28.3%) were lost to follow up (unreachable by telephone or post), and 21 (1.0%) had missing values in at least one of the variables used in the present analysis. Therefore, 1115 (54.6%) individuals aged 45 years or over were analyzed to develop the new epidemiologic classification scheme for clinical HF, with mean (standard deviation) follow-up period of 7 (2.7) years.

Total and concomitant scores

The input variables consisted of eleven symptoms and signs to define a syndrome suggestive of HF or important for differential diagnosis, including dyspnoea, orthopnoea, nocturnal paroxysmal dyspnoea, fatigue, self-percepted and clinically confirmed edema, hepatojugular reflux or jugular venous distension, pulmonary rales, heart murmur, trophic signs of chronic venous insufficiency and visible varicose veins.

Sex, age, education, diabetes, history of myocardial infarction or heart failure, and obesity were included as concomitant variables. The number of latent classes was chosen according to the Bayesian Information Criterion (BIC). Concomitant variable LCA identified 3 classes with different clinical profiles, which we named "symptomatic heart failure pattern", "congestion pattern" and "no symptoms and signs pattern". Associations with the concomitant variables were statistically significant.

The assumptions enumerated above were confirmed through exploratory factor analysis and generalized partial credit models (GPCM).

Exploratory factor analysis supported the existence of only 2 dimensions: hypoperfusion and congestion (table 1). We considered the fitting of the 1-factor GPCM for each dimension, with and without constrained discrimination coefficients. The obtained BIC values suggested that the 1-factor solution with constrained coefficients was preferred to the 1-factor solution without constraints.

We therefore defined the total score to be the number of observed symptoms and signs, and the concomitant score to be the expected number of symptoms and signs predicted by a negative binomial regression model performed on the total score and using the concomitant variables as covariates (table 2). Separate models were developed for each dimension. The first model showed that the expected number of symptoms and signs of hypoperfusion was positively associated with obesity, diabetes and history myocardial infarction or heart failure, and negatively associated with male gender and education; the second model showed that the expected number of symptoms and signs of congestion were positively associated with age, obesity and history myocardial infarction or heart failure and negatively associated with male gender and education.

A classification tree with total and correspondent concomitant scores as predictors was used to predict membership of subjects in the LCA classes. The classification tree for the patterns of heart failure, Figure 1, had seven terminal nodes and used three covariates (hypoperfusion and congestion total scores, and congestion concomitant score).

Agreement among classifications was assessed with the absolute agreement and Cohen's kappa statistics, via boostrapping - 100 samples – within the whole estimation process. The absolute agreement between the predicted and the original classes was 89.7% (bootstrap 95% confidence interval (B95%CI) = (88.0, 91.6)) and the respective Kappa statistic was 0.802 (B95% CI = (0.781, 0.850)).

Each concomitant score was then estimated using two methods: a circular ruler(15), and the point score system(16), based on the information of the concomitant variables readily obtained by the practitioner in the office.

Congestion circular ruler

The concomitant variables were organized into meaningful categories, a reference value (e.g. mid-point) for each category was determined, and a reference concomitant profile was designed by choosing a base category for each concomitant variable. The base category was the category assigned 0 points in the scoring system. We therefore considered 45-54 years of age, 0-4 years of education completed, female with no obesity, no diabetes and no history of myocardial infarction and heart failure as the reference concomitant profile. Less healthy concomitant states are assigned positive points so that a high total conveys a high expected number of symptoms.

Subsequently we computed how far each category of each concomitant variable was from the base category in terms of regression units. Then we defined the constant for the points system, or the number of regression units that would correspond to one point. Here, we let constant reflect the increase associated with a 5-years increase in age. Points were rounded to the nearest integer. The concomitant scores could now be estimated using a simple point score system based on the information readily obtained by a practitioner in the office (table 3). For example, a subject with a profile given by 50 years of age, 10 years of education, female, obese, no diabetes, and no history of myocardial infarction or heart failure, would receive a total score of 3 points that corresponded to 1.2 points in the congestion concomitant score.

Conclusion

In this study, the authors succeeded in deriving a tool that will enable general practitioners and internists to improve the diagnosis of heart failure in primary care, where in general objective measures of cardiac structure and function are not available. The trade-off between model simplicity (points system/circular ruler and classification trees) and predictive accuracy (latent class analysis with concomitant variables) seemed to compensate. Arguably, the primary benefit of an integrated points system/circular ruler and classification trees is their simplicity and in this case the predictive accuracy was not compromised by the model simplicity.

Our mimicked concomitant variable latent class model was based on two main ideas: that the sum of the values each individual take on the items, denoted by total score, is a surrogate measure of that individual's value on the component specific model; and

that the expected total score obtained from the concomitant variables, denoted by concomitant score, is as a surrogate measure of that individual's value on the concomitant variable model.

Exploratory analysis applied to these specific symptoms and signs provided evidence for the existence of two dimensions and showed that each symptom and sign within each factor could be given a similar weight (similar discrimination). This suggested simply the sum of adequate symptoms and signs. These results confirmed that it was possible to use two total scores, as surrogate measures of the individual's value on the component specific model.

The Framingham Heart Study has been a leader in the development and dissemination of a simple mathematical scheme, which they called a points system, for making regression models useful to practitioners (16). The system is easy to use, it does not require a calculator or computer and it simplifies the estimation of complex statistical models. As the multinomial regression in the concomitant variable model was replaced by a negative binomial regression model, a similar points system was also used in this work.

The excellent agreement between the classification obtained from the concomitant variable latent class model and the classification obtained from our tool evidenced the usefulness of our mimicked approach. Nevertheless, this agreement can be overestimated once the tool was applied to only one sample of individuals.

Moreover, the estimated regression coefficients and tree thresholds can vary across different subpopulations. We have used bootstrapping to circumvent this question; results showed a small bias for the estimates.

In conclusion, the combination of a point system/circular ruler with classification trees allowed the simplification of a complex classification system like LCA with concomitant variables, and constructed an easy to use tool to improve the diagnosis of HF in clinical settings.

Table 1. Exploratory factor analysis for symptom and signs of heart failure

Symptom and signs of heart failure (categories)	Factor 1	Factor 2
Dyspnoea (NYHA I, II and III/VI)	0.848	0.110
Fatigue (no| yes)	0.752	-0.032
Orthopnoea (no| yes, 1 pillow| yes, 2 or more pillows)	0.871	-0.014
Nocturnal paroxysmal dyspnea (no| yes)	0.771	-0.113
Heart murmur (no| yes)	0.169	0.211
Pulmonary rales (no| yes)	0.214	0.355
Hepatojugular reflux or jugular venous distension (no| yes)	0.137	0.403
Lower limb oedema at the end of the day - symptom (no| yes)	0.338	0.353
Lower limb oedema - physical examination (no |ankle | Up to the knee)	0.256	0.557
Trophic signs of chronic venous insufficiency (no| yes)	-0.028	0.828
Visible varicose vein (no| yes)	-0.105	0.740
Fit results		
Comparative Fit Index	0.967	
Root Mean Square Error of Approximation	0.046	

Table 2. Estimated parameters of the Negative Binomial models used to obtain the expected number of symptoms and signs of hypoperfusion and expected number of symptoms and signs of congestion dimension, conditional on the covariates pattern (defined by gender, age, education, obesity, diabetes and history myocardial infarction or heart failure).

Covariates	**Model 1** **B (95%IC)**	**Model 2** **B (95%IC)**
Intercept	-0.466 (-1.365, 0.433)	-1.932 (-2.387, -1.481)
Age (per 10 years)	0.022 (-0.098, 0.142)	0.340 (0.280, 0.400)
Education (per 4 years completed)	-0.320 (-0.434, -0.209)	-0.077 (-0.130, -0.024)
Male Sex	-0.644 (-0.910, -0.382)	-0.305 (-0.431, -0.181)
Obesity	0.463 (0.214, 0.711)	0.378 (0.253, 0.502)
Diabetes	0.554 (0.255, 0.855)	0.096 (-0.062, 0.250)
History myocardial infarction or heart failure	1.067 (0.753, 1.387)	0.284 (0.116, 0.449)

Model 1: expected number of symptoms and signs of hypoperfusion;
Model 2: expected number of symptoms and signs of congestion;

Table 3. Points system for the congestion concomitant score

Congestion concomitant Score	-2	-1	0	+1	+2	+4	+6	+8
Age (years)			45-54		55-64	65-74	75-84	85-94
Education (years)	13-25	5-12	0-4					
Gender	Male		Female					
Obesity			No		Yes			
Diabetes			No	Yes				
History myocardial infarction or heart failure			No		Yes			

Points	Expected number of congestion symptoms and signs
-4	0.4
-3	0.5
-2	0.5
-1	0.6
0	0.7
1	0.9
2	1.1
3	1.2
4	1.5
5	1.8
6	2.1
7	2.5
8	2.9
9	3.5
10	4.1
11	4.9
12	5.8
13	6.8

		Hypoperfusion total score		
		0-1	2	3-8
Congestion total score	Congestion Conc. Score			
0	0-1.7			
	1.7-8			
1	0-1.7			
	1.7-8			
2-6	0-8			

Figure 1. Classification trees for the patterns symptoms and signs of heart failure estimated by latent class analysis with concomitant variables (symptomatic heart failure pattern: red; congestion pattern: yellow and no symptoms and signs pattern).

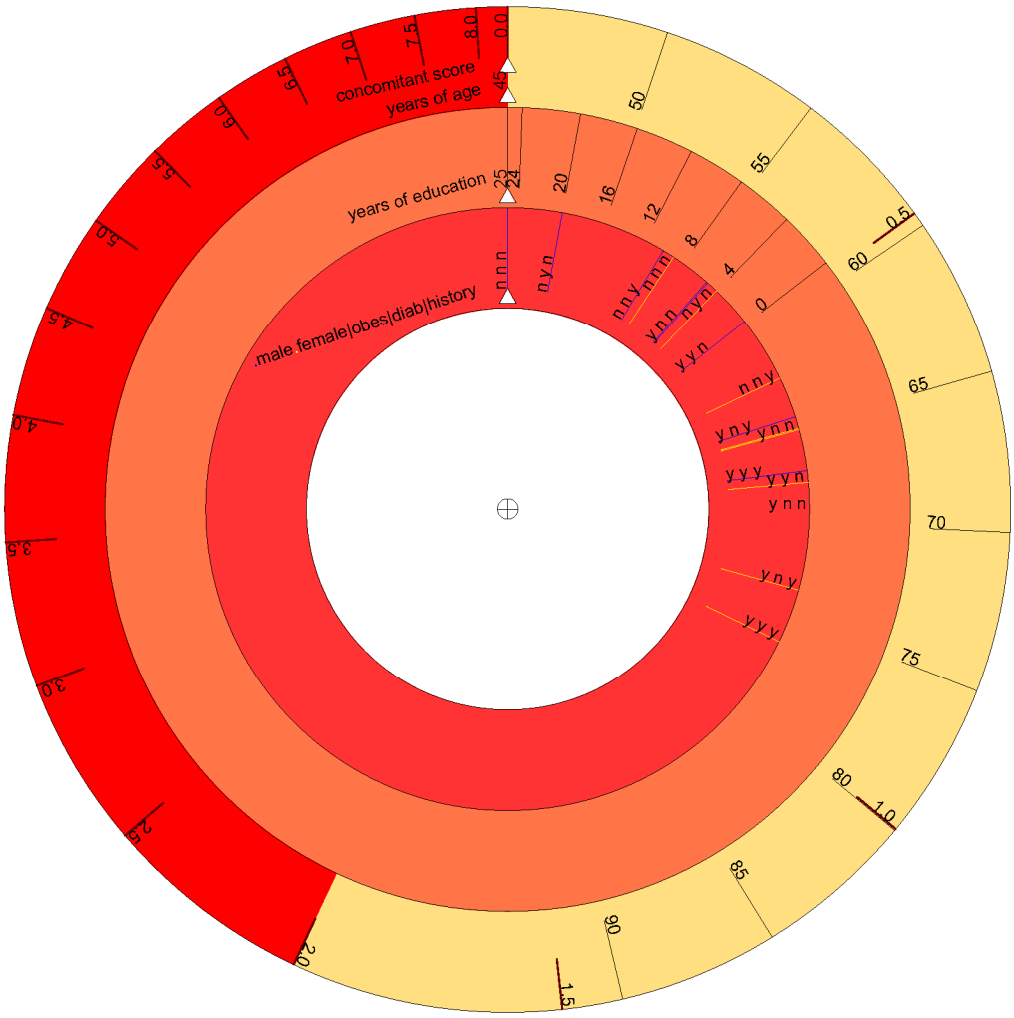

Figure 2. Circular ruler to estimate the congestion concomitant score

References

1. Guggenmoos-Holzmann I, van Houwelingen HC. The (in)validity of sensitivity and specificity. Stat Med 2000;19:1783-92.
2. Hobbs FD, Doust J, Mant J, et al. Heart failure: Diagnosis of heart failure in primary care. Heart;96:1773-7.
3. Mant J, Doust J, Roalfe A, et al. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. Health Technol Assess 2009;13:1-207, iii.
4. Mosterd A, Deckers JW, Hoes AW, et al. Classification of heart failure in population based research: an assessment of six heart failure scores. Eur J Epidemiol 1997;13:491-502.
5. McKee PA, Castelli WP, McNamara PM, et al. The natural history of congestive heart failure: the Framingham study. N Engl J Med 1971;285:1441-6.
6. Eriksson H, Caidahl K, Larsson B, et al. Cardiac and pulmonary causes of dyspnoea--validation of a scoring test for clinical-epidemiological use: the Study of Men Born in 1913. Eur Heart J 1987;8:1007-14.
7. Walma EP, Hoes AW, Prins A, et al. Withdrawing long-term diuretic therapy in the elderly: a study in general practice in The Netherlands. Fam Med 1993;25:661-4.
8. Schocken DD, Arrieta MI, Leaverton PE, et al. Prevalence and mortality rate of congestive heart failure in the United States. J Am Coll Cardiol 1992;20:301-6.
9. Gheorghiade M, Beller GA. Effects of discontinuing maintenance digoxin therapy in patients with ischemic heart disease and congestive heart failure in sinus rhythm. Am J Cardiol 1983;51:1243-50.
10. Carlson KJ, Lee DC, Goroll AH, et al. An analysis of physicians' reasons for prescribing long-term digitalis therapy in outpatients. J Chronic Dis 1985;38:733-9.
11. Kim J, Jacobs DR, Jr., Luepker RV, et al. Prognostic value of a novel classification scheme for heart failure: the Minnesota Heart Failure Criteria. Am J Epidemiol 2006;164:184-93.
12. Severo M, Gaio R, Pimenta J, et al. Prognostic value of a novel classification scheme of clinical symptoms and signs of heart failure adjusted for major confounders. J Epidemiol Community Health 2011;65:A15-A.
13. Bartholomew DJ, Knott M. Latent Variable Models and Factor Analysis. Hodder Arnold, 1999.
14. Ramos E, Lopes C, Barros H. Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction. Ann Epidemiol 2004;14:437-41.
15. Severo M, Lopes C, Lucas R, et al. Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings. Osteoporos Int 2009;20:231-7.
16. Sullivan LM, Massaro JM. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. Statistics in medicine 2004;23:1631-60.

Discussion

This thesis aimed to understand the role of latent models in the improvement and development of health outcomes measurement. In this context, three main research questions were addressed:

1. The refinement of a knowledge questionnaire about rheumatic diseases directed towards the general population, the identification of beliefs and knowledge about these diseases, and the detection of target groups for health education.

2. The calibration of the NYHA classification system between different observers, aspiring to increase its reliability and validity.

3. The role of concomitant variable latent class models in the diagnosis of HF, and how to translate the resulting classification to the clinical practice.

Latent models, both trait and class, provided the evidence base for the identification of knowledge domains in the general population regarding rheumatic diseases. Additionally, this method was instrumental to identify relevant target groups for educational programmes.

The first identified latent trait was associated with the following statements, which probably represent wrong general beliefs: “rheumatic diseases are more frequent in older women”, “rheumatoid arthritis is caused by poor diet”, “cold and damp weather”, “rheumatic patients should rest and move as little as possible” and “rheumatic diseases cannot be cured”, and “all rheumatic patients end up in wheelchairs”. Items about aetiology, treatment and impact of rheumatic diseases were related with the second indentified latent trait, which was therefore denoted specific knowledge.

The latent 3-class model showed that 25.7% of the individuals agreed with the false general beliefs but did have specific knowledge about the diseases, 30.8% did not agree with the general beliefs and did not have specific knowledge, and 43.5% did not agree with general beliefs and had specific knowledge.

Overall this meant that almost 60% of the individuals had some gaps in the overall knowledge about rheumatic diseases. There were difficulties regarding the identification of whether diseases where rheumatic (ankylosing spondylitis and fibromyalgia) or not (glandular fever and multiple sclerosis), and more than fifty percent believed that people with rheumatic diseases cannot be cured and cannot die from those illnesses. The latter finding is similar to that reported by others: a study with women aged 65-90 years in Canada (24) showed that only 36% agreed that health problems caused by osteoporosis can be life-threatening and another study carried out in US adults (45) found that only 63% correctly respond “false” to the statement “no medications can treat osteoporosis”.

This study confirmed that it is important that educational programs about rheumatic diseases should be centred in the eldest and low educated individuals, in order to counteract wrong general beliefs together with the obvious improvement of the identification of the different rheumatic diseases. Specific education programs for health professionals could also improve the communication of the specific knowledge to patients and their relatives.

The application of LTM to the NYHA classification has objectively indicated the main reason why several studies have reported low inter-observer reliability and consequent limited usefulness of that classification as an outcome measure. The main point was the existence of discrepant thresholds between observers in the definition of NYHA class I, II and III. Although the observers in the study were experienced physicians with training in the management of heart failure, there were still discrepancies between their (subjective) evaluations.

Intra-observer reliability is very important to interpret changes in NYHA class in the individual patient who is assessed repeatedly by the same physician. Nevertheless inter-observer variability should be of special concern when patients are assessed by different physicians. This is particularly important, in practice, in unscheduled visits to the clinic or the emergency department, where patients are not assessed by their usual attendant. These unscheduled visits are usually due to worsening symptoms and an increase in the NYHA class, in

comparison with the previous clinical state, is used as a criterion for clinical decisions such as hospital admission and intensity of therapy adjustment such as the use of intravenous medication.

Therefore, in each setting the NYHA classification was to be used, it would be useful to identify the differences between the observers' assessments and to calibrate their classifications.

The absolute agreement between the predictions of the NYHA classification across observers was 65%. The same statistic was improved to 88% after comparison of the trait predictions with the observer classification, for each observer. This showed how the subjectivity of the thresholds affected the reliability of the NYHA classification. At the same time this improvement confirmed the quality of the obtained calibration.

Calibration methodology can be useful to improve the reliability between observers in clinical practice and research settings. In clinical practice, it is possible to use the relation between *anchor* items and ability to standardize the classification among observers, and to give guidelines to improve the inter-observers reliability. In the research setting, the scale defined by the anchor items and the operators classification can be used as a standardized NYHA classification that minimizes the subjectivity of each observer classification.

In order to evaluate the validity of the NYHA calibration, we assessed its predictive value with respect to a series of objective structural of functional cardiac abnormalities as assessed by echocardiography at rest.

The discrimination power of the calibrated NYHA was shown to be better than its correspondent non-calibrated scale, largely at the expense of the likelihood ratio of NYHA I.

The major limitation of this calibration study is its small size. Whereas it was slightly larger than the minimum number required to properly fit a 1-PL latent trait model, it was not large enough to allow for the application of a 2-PL model (11); indeed we obtained unstable item parameters and large standard errors.

Each individual was assessed by one observer only, opposed to the ideal situation where that individual should be assessed by all observers.

We do not think of this as a limitation. First, there were no statistically significant differences among the groups evaluated by each observer regarding sex, clinical history, systolic blood pressure, education and left ventricular systolic dysfunction; only age, body mass index and diastolic blood pressure showed significant small differences. On the other hand, the anchor items were related to each observer's NYHA classification; so even if there were an observer group that was discrepant from the others, the anchor items would guarantee a good calibration. Therefore we are confident that this apparent limitation did not have a major impact on the results.

The study addressing the last research topic succeeded in identifying three patterns of syndromic aggregation of symptoms and signs for HF. Based on findings routinely collected in current clinical practice, we applied concomitant variable LCM to account for known determinants of the relevant clinical findings and the *a priori* probability of the condition. These models were shown to be useful to standardize and quantify the probabilistic reasoning in clinical diagnosis, upon which decisions of further investigation and even treatment need to be made.

Most of the obtained likelihood ratios were small, showing that the patterns generated relatively small changes from pre- to post-test probability of cardiac abnormalities. On one hand, this was expected, considering that many of these abnormalities are known to be asymptomatic in a large proportion of patients for a long time, and the symptoms are unspecific. On the other hand, it is compatible with previous quantifications of the diagnostic value of symptoms and signs (46).

The concentration level of B-type natriuretic peptide is an established biomarker for the diagnosis of HF. The obtained patterns showed a good diagnostic performance for exclusion of high BNP values. However these concentrations were not measured according to the manufacturer's recommendations. The manufacturer's recommends the use of whole blood or plasma specimens to measure BNP and we have measured it in serum. Nevertheless, in a

pilot study we established a calibration equation to allow for the use of serum BNP and for the classification of the community individuals with the usual cut-offs.

For the first time, to our knowledge, the obtained HF classification integrated factors that have a large impact on the prevalence of symptoms and signs suggestive of HF. The novelty in our application, with respect to previous HF classifications, is that class probabilities are adjusted for concomitant variables. Specifically, the model estimates the increase or decrease in class probabilities for individuals conditional on the respective concomitant variables values, contributing to an increase in the discrimination and to a decrease in the number of false negatives and false positives. The inclusion of these variables improved the fitting of the model.

The discrimination power of the concomitant variable LCM was shown to be better than that of its correspondent LCM, largely at the expense of the likelihood ratio of pattern 3 (no symptoms and signs) to exclude cardiac abnormalities. These results objectively indicate that the use of concomitant variables can improve the diagnostic value of the symptoms and signs patterns and, consequently, improve the usefulness of the symptoms and signs for diagnosis and as an outcome measure. The potential for application in other settings of complex diagnoses is very high.

The patterns identified by this methodological approach depend on the type of population being studied. In this study, in a sample of the general population, three patterns were identified: “symptomatic HF”, “symptoms and signs of congestion” and “no symptoms and signs”. In a previous study, using a similar approach in subjects discharged after myocardial infarction or acute heart failure, the authors were able to distinguish different patterns (non-cases, heart failure and advanced heart failure) (37). The patterns identified in the current study are appropriate for diagnosis in the general population only.

The low prevalence of advanced and severe HF cases is a limitation of this study and could have underestimated the discriminative capacity of this set of items. Also, the prevalence of more specific symptoms and signs such as a third heart sound (47) was too low in this sample to be able to take them into account. The proposed patterns are likely to be more

sensitive but less specific than previously available scores such as the Framingham criteria, supporting their usefulness as potential screening tools or initial clinical investigation that do not aim to replace full investigation in the clinical setting.

Clinical practitioners who are not research scientists may have difficulties in the implementation of complex statistical models in their practice routine, such as concomitant variable latent class analysis. The present study aims at translating the validated patterns into a classification score, using an approach for making complex statistical models useful to practitioners and researchers, such as a circular ruler (48) or points system (49), which was used for example to develop the widely used Framingham risk scores. The use of this tool will allow the identification of high-risk candidates for heart failure who are likely to have a substantial yield of positive findings when tested for objective measures of cardiac dysfunction in clinical practice. In addition, it will confidently exclude heart failure in others, thus orienting the clinical investigation towards alternative directions. Such a tool could also increase the discrimination and decrease the number of false negatives and false positives in epidemiological studies on HF.

The study succeeded in deriving a tool that will enable general practitioners and internists to diagnose heart failure in primary care, where in general objective measures of cardiac structure and function are not available. The followed approach was similar to that of a previous one, where we designed a tool to estimate the calcium and vitamin D intakes, from a simplification of a complex statistical model. The circular ruler main advantage is that it can be used when no computer is available or among people without special computer skills.

In the HF context, the tool integrated a circular ruler and a table on the back. There was a small trade-off between model simplicity (circular ruler and table) and predictive accuracy (latent class analysis with concomitant variables). The quality of the agreement was excellent. Nevertheless, this agreement can be overestimated once the tool was applied to only one sample of individuals.

Moreover, the estimated regression coefficients and tree thresholds can vary across different subpopulations. We have used bootstrapping to circumvent this question; results showed a small bias for the estimates.

Conclusions

The application of latent trait and latent class methodologies provided the evidence base for the identification of knowledge domains in the general population regarding rheumatic diseases. The method was instrumental to identify relevant target groups for educational programmes. It was shown that there were several knowledge flaws about rheumatic diseases. One out of four individuals considered false general beliefs as true, and approximately 30% did not have detailed knowledge on rheumatic diseases. Higher education and the presence of disease contributed positively to the overall knowledge.

Latent models can also be useful in the standardization of clinical assessments, namely in the identification of different thresholds for different observers. In the NYHA classification, those cut-offs tend to be discrepant as there is some subjectivity inherent to the classification. Concurrent calibration through latent trait models can be used to link a large number of observers to the same scale. It provides a way to maximize the reliability and to minimize the misclassification of a classification.

Classical latent class analysis can be adapted to include effects on the prevalence of the different patterns, in a methodology called concomitant variable latent class analysis. These models can be useful to standardize and quantify the probabilistic reasoning in clinical diagnosis, upon which decisions of further investigation and even treatment need to be made.

In the heart failure framework, concomitant variables not directly related to clinical findings were used to account for *a priori* probabilities of the condition. We succeeded in identifying three patterns of syndromic aggregation of symptoms and signs for heart failure, based on findings routinely collected in current clinical practice. Relatively to the classical approach, validity was improved.

A simple diagnostic tool for general practitioners and internists was developed. It will enable them to diagnosis heart failure in primary care, where in general objective measures of cardiac structure and function are not available.

References

1. **Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45:S22-31.**
2. **Reeve BB. An introduction to modern measurement theory. National Cancer Inst 2002.**
3. **Bartholomew DJ, Knott M. Latent Variable Models and Factor Analysis. Hodder Arnold, 1999.**
4. **Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol 2009;5:27-48.**
5. **Thomas ML. The value of item response theory in clinical assessment: a review. Assessment 2010;18:291-307.**
6. **Crane PK, Narasimhalu K, Gibbons LE, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol 2008;61:1018-27 e9.**
7. **Lai JS, Cook K, Stone A, et al. Classical test theory and item response theory/Rasch model to assess differences between patient-reported fatigue using 7-day and 4-week recall periods. J Clin Epidemiol 2009;62:991-7.**
8. **Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16 Suppl 1:5-18.**
9. **Hays RD, Liu H, Spritzer K, et al. Item response theory analyses of physical functioning items in the medical outcomes study. Med Care 2007;45:S32-8.**
10. **Hambleton RK. Principles and selected applications of item response theory. 1989.**
11. **Downing SM. Item response theory: applications of modern test theory in medical education. Med Educ 2003;37:739-45.**
12. **De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 2010;44:109-17.**
13. **McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. Med Care 2000;38:II43-59.**
14. **Vermunt JK, Magidson J. Latent class cluster analysis. Applied latent class analysis 2002:89-106.**
15. **Agrawal A, Lynskey MT, Madden PA, et al. A latent class analysis of illicit drug abuse/dependence: results from the National Epidemiological Survey on Alcohol and Related Conditions. Addiction 2007;102:94-104.**
16. **Sacco P, Bucholz KK, Spitznagel EL. Alcohol use among older adults in the National Epidemiologic Survey on Alcohol and Related Conditions: a latent class analysis. J Stud Alcohol Drugs 2009;70:829-38.**
17. **Campbell SB, Morgan-Lopez AA, Cox MJ, et al. A latent class analysis of maternal depressive symptoms over 12 years and offspring adjustment in adolescence. J Abnorm Psychol 2009;118:479-93.**
18. **Fahey MT, Thane CW, Bramwell GD, et al. Conditional Gaussian mixture modelling for dietary pattern analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2007;170:149-66.**
19. **Samejima F. Graded response model. Handbook of modern item response theory 1997:85-100.**
20. **Lopez AD, Mathers CD, Ezzati M, et al. Measuring the global burden of disease and risk factors, 1990â€“2001. Global burden of disease and risk factors 2006:1.**

21. **Hazes JM, Woolf AD. The bone and joint decade 2000-2010. J Rheumatol 2000;27:1-3.**
22. **Brekke M, Hjortdahl P, Kvien TK. Involvement and satisfaction: a Norwegian study of health care among 1,024 patients with rheumatoid arthritis and 1,509 patients with chronic noninflammatory musculoskeletal pain. Arthritis Rheum 2001;45:8-15.**
23. **Taal E, Rasker JJ, Wiegman O. Group education for rheumatoid arthritis patients. Semin Arthritis Rheum 1997;26:805-16.**
24. **Cadarette SM, Gignac MA, Beaton DE, et al. Psychometric properties of the "Osteoporosis and You" questionnaire: osteoporosis knowledge deficits among older community-dwelling women. Osteoporos Int 2007;18:981-9.**
25. **Werner P. Knowledge about osteoporosis: assessment, correlates and outcomes. Osteoporos Int 2005;16:115-27.**
26. **Dickstein K, Cohen-Solal A, Filippatos G, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). Eur Heart J 2008;29:2388-442.**
27. **Hunt SA. ACC/AHA 2005 guideline update for the diagnosis and management of chronic heart failure in the adult: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure). J Am Coll Cardiol 2005;46:e1-82.**
28. **Hobbs FD, Doust J, Mant J, et al. Heart failure: Diagnosis of heart failure in primary care. Heart;96:1773-7.**
29. **Mant J, Doust J, Roalfe A, et al. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. Health Technol Assess 2009;13:1-207, iii.**
30. **Mosterd A, Deckers JW, Hoes AW, et al. Classification of heart failure in population based research: an assessment of six heart failure scores. Eur J Epidemiol 1997;13:491-502.**
31. **McKee PA, Castelli WP, McNamara PM, et al. The natural history of congestive heart failure: the Framingham study. N Engl J Med 1971;285:1441-6.**
32. **Eriksson H, Caidahl K, Larsson B, et al. Cardiac and pulmonary causes of dyspnoea--validation of a scoring test for clinical-epidemiological use: the Study of Men Born in 1913. Eur Heart J 1987;8:1007-14.**
33. **Walma EP, Hoes AW, Prins A, et al. Withdrawing long-term diuretic therapy in the elderly: a study in general practice in The Netherlands. Fam Med 1993;25:661-4.**
34. **Schocken DD, Arrieta MI, Leaverton PE, et al. Prevalence and mortality rate of congestive heart failure in the United States. J Am Coll Cardiol 1992;20:301-6.**
35. **Gheorghiade M, Beller GA. Effects of discontinuing maintenance digoxin therapy in patients with ischemic heart disease and congestive heart failure in sinus rhythm. Am J Cardiol 1983;51:1243-50.**
36. **Carlson KJ, Lee DC, Goroll AH, et al. An analysis of physicians' reasons for prescribing long-term digitalis therapy in outpatients. J Chronic Dis 1985;38:733-9.**

37. Kim J, Jacobs DR, Jr., Luepker RV, et al. Prognostic value of a novel classification scheme for heart failure: the Minnesota Heart Failure Criteria. Am J Epidemiol 2006;164:184-93.
38. Remes J, Miettinen H, Reunanen A, et al. Validity of clinical diagnosis of heart failure in primary health care. Eur Heart J 1991;12:315-21.
39. Wheeldon NM, MacDonald TM, Flucker CJ, et al. Echocardiography in chronic heart failure in the community. QJM 1993;86:17.
40. Azevedo A, Bettencourt P, Pimenta J, et al. Clinical syndrome suggestive of heart failure is frequently attributable to non-cardiac disorders--population-based study. Eur J Heart Fail 2007;9:391-6.
41. Bianchi MT, Alexander BM. Evidence based diagnosis: does the language reflect the theory? BMJ 2006;333:442-5.
42. Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels. Little, Brown
1994.
43. Raphael C, Briscoe C, Davies J, et al. Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure. Heart 2007;93:476-82.
44. Goldman L, Hashimoto B, Cook EF, et al. Comparative reproducibility and validity of systems for assessing cardiovascular functional class: advantages of a new specific activity scale. Circulation 1981;64:1227-34.
45. Solomon DH, Finkelstein JS, Polinski JM, et al. A randomized controlled trial of mailed osteoporosis education to older adults. Osteoporos Int 2006;17:760-7.
46. Wang CS, FitzGerald JM, Schulzer M, et al. Does this dyspneic patient in the emergency department have congestive heart failure? JAMA: the journal of the American Medical Association 2005;294:1944-56.
47. Wang CS, FitzGerald JM, Schulzer M, et al. Does this dyspneic patient in the emergency department have congestive heart failure? JAMA: the journal of the American Medical Association 2005;294:1944.
48. Severo M, Lopes C, Lucas R, et al. Development of a tool for the assessment of calcium and vitamin D intakes in clinical settings. Osteoporos Int 2009;20:231-7.
49. Sullivan LM, Massaro JM, D'Agostino RB, Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med 2004;23:1631-60.