Universidade de São Paulo
Faculdade de Saúde Pública

Ciência de dados e políticas públicas de saúde: exemplos práticos.

Joana Raquel Raposo dos Santos

Tese apresentada ao Programa de Pós-Graduação em Saúde Pública da Faculdade de Saúde Pública da Universidade de São Paulo para obtenção do título de doutor em Ciências.

Área de Concentração: Saúde Pública

Orientador: Profº Dr Alexandre Chiavegatto Filho

São Paulo

2020

Ciência de dados e políticas públicas de saúde: exemplos práticos.

Joana Raquel Raposo dos Santos

Tese apresentada ao Programa de Pós-Graduação em Saúde Pública da Faculdade de Saúde Pública da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de Concentração: Saúde Pública

Orientador: Profº Dr Alexandre Chiavegatto Filho

Versão revisada

São Paulo

2020

DEDICATÓRIA

A Deus, pela fé e coragem com que me abençoa todos os dias.

Aos meus pais e irmã, pela paciênca e amor incondicional.

À minha avó que olha por mim.

Ao Rui, pela oportunidade única de aprendizado e crescimento pessoal e académico.

À Hellen, pela amizade desde o primeiro dia.

Ao Prof. Alexandre e outros professores que contribuiram para o meu conhecimento.

Ao Dr. Carlos Dias e à Dra Paula que facilitaram e apoiaram incansavelmente esta jornada.

SANTOS, J. R. R. **Ciência de dados e políticas públicas de saúde: exemplos práticos**. 2020. Tese (Doutorado) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2020.

RESUMO

Introdução: A ciência de dados é uma área do conhecimento impulsionada pela mudança do atual paradigma tecnológico e cientifico, que decorre do aumento do volume de dados, tipo, acesso, armazenamento e desenvolvimento computacional e tecnológico. Esse conhecimento tem permitido importantes avanços em vários setores, mas a contribuição da ciência de dados para as políticas públicas em saúde ainda encontra-se pouco explorada. **Objetivo:** Analisar se técnicas de ciência de dados, como algoritmos preditivos de inteligência artificial (machine learning), técnicas de clusterização de individuos e métodos causais para estudos observacionais podem contribuir para a área das políticas de saúde, identificando grupos-alvo para os quais programas e campanhas possam ser direcionados, permitindo uma alocação mais eficiente de recursos e contribuindo para a elaboração de medidas que auxiliem no desenho e avaliação de políticas públicas de saúde. **Métodos:** Foram utilizados dados do Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS) para caracterização sociodemográfica dos municípios brasileiros, e do Inquérito Nacional de Saúde de Portugal de 2014 para caracterizar a população residente em Portugal. Para a análise preditiva foram utilizados alguns dos algoritmos mais populares de machine learning, como regressão logística penalizada, *random forest*, *gradient boosting trees* e análises de agrupamento com componentes principais. Para a avaliação de um programa público de saúde (Mais Médicos) foram utilizados escores de propensão (*propensity score*) com pareamento. **Resultados:** Foram escritos um total de três artigos científicos, sendo que dois foram publicados e um encontra-se em revisão. O primeiro foi publicado na *International Journal of Public Health*, e trata-se de uma avaliação do Mais Médicos com métodos de escore de propensão. O escore permitiu um pareamento entre unidades municipais (n = 395) com uma boa performance, em que 86 das 97 covariáveis apresentaram um bom balanceamento (medido pela diferença média padronizada, inferior a 25%). O segundo artigo foi publicado na *Health Policy and Technology* e realizou uma análise

de agrupamento de componentes principais para identificar grupos homogêneos entre indivíduos sem plano privado de saúde (n = 12.134). Foram identificados três agrupamentos de indivíduos (indivíduos de meia idade profissionalmente ativos, indivíduos envelhecidos com práticas saudáveis e aqueles psicologicamente vulneráveis), o que pode auxiliar na elaboração de políticas públicas direcionadas. O terceiro artigo encontra-se atualmente em avaliação e realizou uma análise preditiva de inteligência artificial (machine learning) para ausência laboral por motivos de doença com uma amostra populacional do Inquérito Nacional de Saúde (n=6.249), obtendo uma AUC de 0,67 pelo algoritmo de *random forest*. **Conclusão:** A ciência de dados pode ter um papel importante na melhoria da evidência em políticas públicas, especialmente no caso de superar dificuldades de abordagens mais tradicionais, como no estabelecimento de contrafactuais em estudos quase experimentais e por meio da realização de análises preditivas de machine learning para a alocação prioritária de recursos.

Palavras-chave: políticas públicas; ciência de dados: *machine learning.*

SANTOS, J. R. R. **Data science and public policies: pratical examples of application**. 2020. Thesis (Ph. D) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2020.

ABSTRACT

Introduction: Data science is an area of knowledge that has followed the growth of a new technological and scientific paradigm. It results directly from the increase in the volume of data, type, access, storage and from the computational and technological development. This knowledge has allowed important advances in several areas, but the contribution of data science to public health policies is still scarcely explored. **Objective:** To analyze whether the use of data science tools can contribute to improve health policies. In particular, we will identify target groups (using supervised or unsupervised approaches) for which programs and campaigns can be directed, thus contributing to a more efficient allocation of resources and provide evidence that support the design and evaluation of public health programs. **Methods:** We used two different data sources: first, data from the Department of Informatics of the Brazilian Health System of Brasil (DATASUS) was collected to gather information regarding the sociodemographic profile of Brazilian municipalities; and second, the National Health Survey of Portugal in 2014 to gather data regarding Portuguese population. For the statistical analysis, the following algorithms were used: logistic regression, random forest, gradient boosting trees and a cluster analysis in the principal components. To evaluate a large Brazilian health program (Mais Médicos), we applied propensity score matching, and the score was estimated using logistic regression. **Results:** A total of three scientific articles were written. The first was published at the International Journal of Public Health and evaluated the causal effect of the Mais Médicos program. The score resulted in a successful pairing between municipalities (n = 395) of which 86 of the 97 covariates presented good balance (measured by a standardized mean difference lower than 25%). The second was published in Health Policy and Technology and aimed at identifying homogeneous groups among those who did not have a private health plan (n = 12.134). We used a cluster analysis with principal components and found three groups of individuals: professionally active middle aged individuals, healthy elderly individuals, and those psychologically vulnerable. The third

article performed a predictive analysis to identify in advance individuals who are more prone to be absent from work due to illness. We used the National Health Survey (n = 6.249) and a random forest model with an area under the ROC curve of 0.67. **Conclusion:** Data science can play an important role in improving evidence in public policies, namely to overcome difficulties that more traditional approaches are not able to address efficiently. In particular, it can be helpful ine establishing counterfactuals in quasi-experimental studies and performing predictive analyzes for priority allocation of resources.

Keywords: public policies; data science; machine learning

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

ACM: Análise de Correspondência Múltipla

ACP: Analise Componentes Principais

DATASUS: Departamento de Informática do SUS

DID: Diferença nas Diferenças

EUROSTAT: European Statistics

EHIS: European Health Interview Survey

IBGE: Instituto Brasileiro de Geografia e Estatística.

INE: Instituto Nacional Estatística

INS: Inquérito Nacional de Saúde

MBE: Medicina baseada em Evidências

PBE: Política Baseada em Evidências

PMM: Programa Mais Médicos

PS: *Propensity Score* (escore de propensão)

TSE: Tribunal Superior Eleitoral

SUMÁRIO

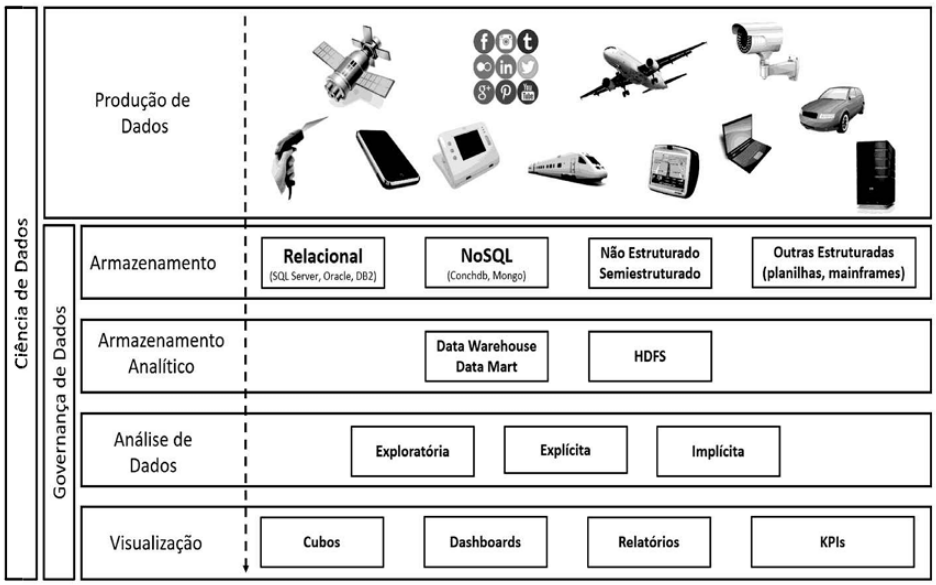
INTRODUÇÃO

Ciência de dados: definição, crescimento e oportunidades

A ciência de dados não é uma área nova, mas o crescimento e a popularização do *big data* (i.e, dados com maior variedade, volume e velocidade crescentes), bem como o avanço tecnológico de armazenamento e processamento de dados, levaram a que a área emergisse e ganhasse expressão, especialmente na última década. A ciência de dados visa manusear e transformar essa quantidade de dados brutos em resultados úteis para auxiliar em decisões.

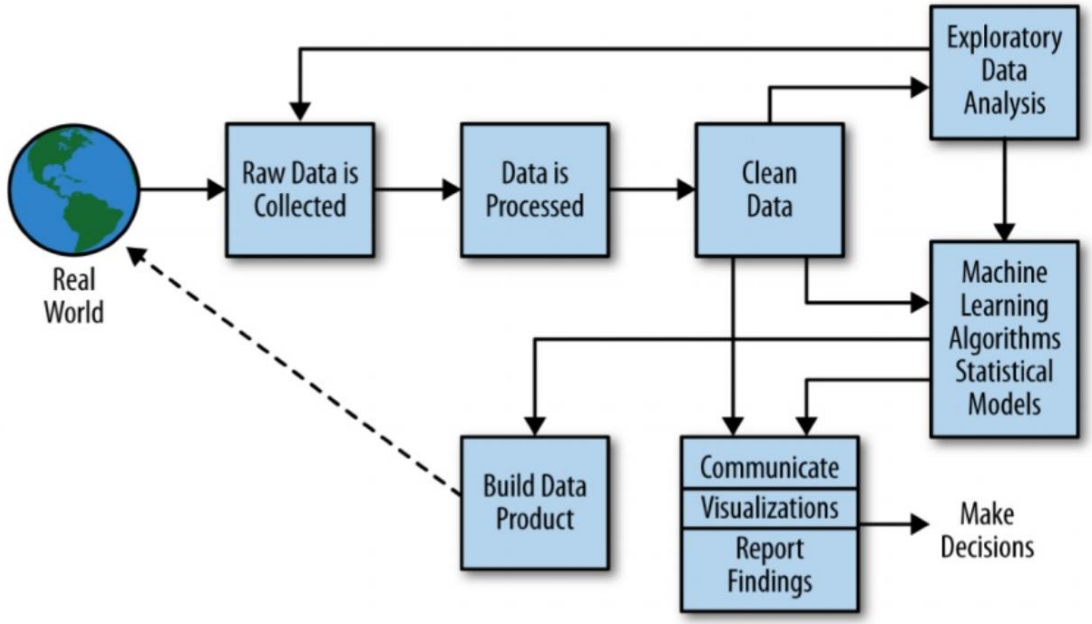
Contudo, a ciência de dados tem gerado alguma controvérsia (SILVER, 2013), especialmente devido à fronteira, aparentemente sutil, com a estatística. Alguns autores têm abordado essa temática e contribuído para o esclarecimento e a diferenciação entre ciência de dados e estatística. No livro “Introdução à Mineração de Dados e Big Data”, Fernando Amaral, professor do Instituto de Matemática e Estatística da USP, definiu ciência de dados como os processos, modelos e tecnologias que estudam os dados durante todo o ciclo de vida, da produção ao descarte (AMARAL, 2015). Nessa definição, a ciência de dados seria mais abrangente que a estatística, sendo que a contempla como uma de suas etapas (análise de dados). Esse conceito torna-se ainda mais claro ao se constatar que os dados podem provir de diversas fontes e assumir diferentes formatos (digitais, analógicos), conforme se observa na Figura 1. O processo como um todo está graficamente representado na Figura 2.

Figura 1 - Panorama da Ciência de Dados.



Fonte: Amaral, F. 2015.

Figura 2- *Descrição de um processo típico de ciência de dados.*



Fonte: Amaral, F. 2015.

Outros autores têm também corroborado essa definição, argumentando que a ciência de dados é mais abrangente que a estatística já que a emergência de novos dados – texto, imagem, sons, vídeo – estruturados ou não, proveniente de

fontes diversas como redes sociais, satélites, dispositivos móveis ou até outros menos complexos, como instrumentos de coleta regular de dados, têm sido considerados ideais para a aplicação de técnicas emergentes de ciência de dados.

É nesse contexto que essa área científica ganha relevância, em conjunto com os algoritmos de *machine learning* enquanto instrumento para responder a uma realidade mais complexa. Na sequência, o aprimoramento desses algoritmos, em conjunto com os dados que se encontram disponíveis em plataformas digitais ou outras formas de armazenamento, têm levado a avanços notórios nos resultados produzidos, divulgados e, por vezes, aplicados. Em suma, a ciência de dados, pela conjuntura favorável em que se move, tem permitido gerar conhecimento que em um outro momento não seria possível.

Atualmente, a ciência de dados tem atuação direta em várias áreas do conhecimento. Porém, o seu desenvolvimento tem sido maior em áreas de mercado privado e ainda é pouco aplicada nas áreas governamentais e sociais, apesar de se reconhecer que a utilização desses dados pode contribuir para avanços significativos no setor público e social.

Alguns exemplos de aplicação incluem *marketing* digital (FERREIRA, 2008), sistemas de recomendação (Netflix e Amazon), reconhecimento de imagens, logística (UPS) (DHL, 2013), saúde (Walgreens), serviços financeiros (por exemplo, serviços de detecção de fraude e risco), comparação de preços (Trivago) e o processo de transformação de dados em conhecimento (MIKE LOUKIDES, 2016).

Machine learning **e a ciência de dados**

O conceito de *machine learning* foi definido por Arthur Samuel no fim da década de 50 como o campo de estudo que permite aos computadores a habilidade de aprenderem sem serem explicitamente programados (SAMUEL, 2000). Esse

conjunto de técnicas permite detectar padrões nos dados e usar novos padrões ainda desconhecidos para predizer fenômenos, ou para realizar outras decisões sob incerteza (MURPHY, 2011). Machine learning é uma ferramenta útil na produção, armazenamento, análise e visualização de dados, contudo muitas vezes é associada apenas à etapa de análise de dados. Por exemplo, um determinado algoritmo pode, numa perspectiva de produção de dados, ser usado na sinalização de uma anomalia em tempo real, com sensores que geram dados minuto a minuto; no armazenamento, uma nuvem pode possuir algoritmos que permitem não só capturar, mas também manter os dados armazenados em segurança; por fim, o aprendizado pode analisar dados de diferentes formatos, estruturados ou não, que vão desde sons e imagens a bancos de dados. Machine learning é, portanto, uma ferramenta que serve à ciência de dados em toda a sua abrangência, e que tem apresentado grande desenvolvimento e suscitado interesse nas últimas décadas.

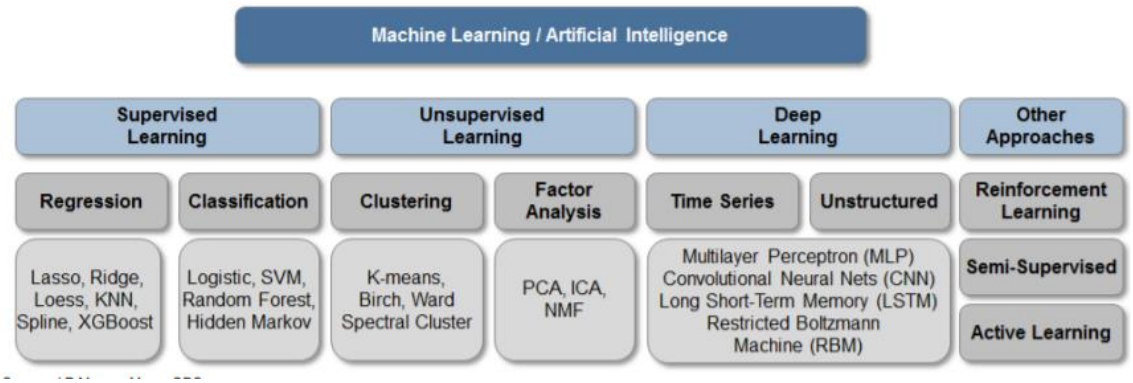
Em um já clássico artigo de Breiman (BREIMAN, 2001a), o autor define as características da modelagem algorítmica, isto é, do machine learning, que o diferencia da estatística. Breiman reforça que a modelação algorítmica tem permitido a evolução de vários campos da ciência, por oposição a áreas em que a estatística tradicional tem sido predominante. Segundo o autor, uma cultura utiliza a geração de modelos estocásticos na compreensão dos fenômenos, enquanto a outra cultura utiliza modelos que procuram a compreensão dos fenômenos sem conhecimento prévio (Breiman, 2001). A predição é, de fato, a finalidade principal da cultura algorítmica para o qual é regra a divisão da amostra (n) em banco de treinamento e banco de teste, com o objetivo de avaliar a performance do modelo (HINDMAN, 2015).

Outra característica é que esses algoritmos têm a vantagem de modelar bem grandes amostras (n) e/ou muitas variáveis preditoras (p). Alguns dos algoritmos de classificação mais populares são modelos logísticos penalizados e Árvores de Classificação e Regressão (CART), inspirados nos modelos de Breiman. Em relação à tarefa de classificação, o objetivo de um modelo supervisionado é construir uma função com base em um banco de treinamento que aprenda as

relações entre variáveis existentes e as observações, de forma a construir um modelo preditivo que, quando aplicado a novas observações, as classifique na sua categoria correta com base nas variáveis utilizadas para o aprendizado (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2006).

Uma divisão clássica de métodos de *machine learning* é a separação entre algoritmos supervisionados e não supervisionados. Recentemente, técnicas mais sofisticadas e flexíveis apareceram como *deep learning* e *reinforment learning* (KOLANOVIC; KRISHNAMACHARI, 2017). (figura 3).

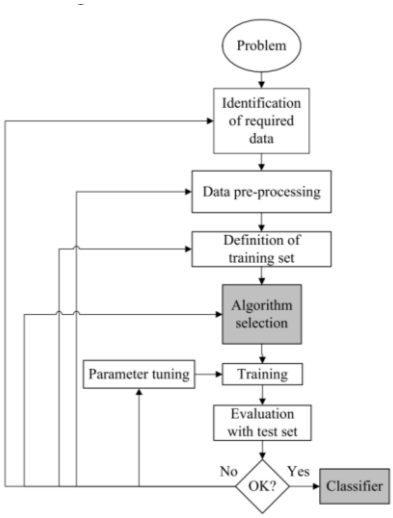
Figura 3-Caracterização de algoritmos utilizados em *machine learning*.



Fonte: KOLANOVIC AND KRISHNAMACHARI, 2017.

Em termos gerais, o que separa esses grupos de algoritmos é a existência de uma variável específica de desfecho para problemas supervisionados. Assim, quando se treina um algoritmo supervisionado, utiliza-se um conjunto de dados que já são mapeados para uma variável desfecho e depois se tenta predizê-la em um novo banco de dados. Algoritmos de classificação e regressão estão enquadrados como supervisionados. O processo encontra-se descrito na figura 4.

Figura 4- Processo para implementação de algoritmos supervisionados.



Fonte: KOTSIANTIS, ZAHARAKIS, AND PINTELAS, 2006.

Por outro lado, algoritmos não supervisionados são treinados com base num conjunto de dados que não tem um desfecho específico a ser predito e em que o interesse é, portanto, descobrir padrões gerais nos dados, como a identificação de clusters e a redução de dimensionalidade. Nesse caso, exemplos de algoritmos não-supervisionados incluem *k means clustering* e a análise de componentes principais.

Em suma, as características de um algoritmo de *machine learning* são as seguintes: a) "*Generalize by learning*", isso é, aprendizado com o banco de treinamento sem que seja necessária a existência de fortes pressupostos a priori; b) no caso de algoritmos supervisionados, o objetivo é a predição em bancos de dados nunca vistos; c) em geral, é mais acurado, automático e rápido que sistemas baseados em regras construídos manualmente; d) ao contrário de modelos paramétricos, que funcionam bem quando já existe algum entendimento sobre a relação entre inputs e a resposta, esses algoritmos permitem encontrar outras relações e assim gerar novas hipóteses; e) lida bem com multidimensionalidade e interações e f) não requer que haja pressupostos sobre a distribuição das variáveis ou linearidade (IZBICKI, 2016).

Alguns autores (COVENEY; DOUGHERTY; HIGHFIELD, 2016) têm exposto desafios decorrentes desses métodos, bem como a necessidade de maior parcimônia na sua utilização. Segundo eles, apesar de esses métodos oferecerem vantagens, não se pode negligenciar o papel que técnicas estatísticas tradicionais têm para garantir bons resultados (COVENEY; DOUGHERTY; HIGHFIELD, 2016). Em particular, características menos positivas de algoritmos de *machine learning* são: a) a performance preditiva depende fortemente do banco de dados utilizado, logo a quantidade e qualidade dos dados é de especial importância; b) alguns algoritmos exigem um certo grau de pré-processamento de dados; c) muitos necessitam que seja feito ajuste de hiperparâmetros para obtenção de modelos com melhor performance; d) é fundamental balancear o trade off entre viés e variância; e) é importante testar vários algoritmos para encontrar o mais adequado para determinada situação (IZBICKI, 2016).

Políticas públicas e ciência de dados

Política pública é um conjunto de metas definido por governantes, com diferentes graus de participação dos governados, com o objetivo de solucionar ou prevenir problemas sociais, que se tornam uma agenda política (PEDROSO; JUHÁSOVÁ; HAMANN, 2019). A política baseada em evidência (PBE) é uma prática que tem reunido consenso relativamente à sua necessidade de implementação junto da população e dos gestores responsáveis por tomarem decisões políticas (RAMOS; SILVA, 2018). A importância da implementação dessa prática no setor público – i.e. do uso da evidência para formular políticas públicas - é fundamental pelo carácter ético de missão pública que é inerente à atividade política. A utilização de evidência no processo de decisão política poderá, assim assegurar, tanto quanto possível, que as decisões irão ao encontro de necessidades reais. A utilização de evidência no processo político pode melhorar todo o sistema, nomeadamente através do apoio na identificação acurada de necessidades de uma população, da otimização de recursos públicos, da gestão de serviços adequados, minimizando efeitos que podem surgir de decisões pouco fundamentadas em ciência.

É nesse contexto que a *ciência de dados,* enquanto disciplina e ferramenta, pode trazer benefícios ao setor social e político. A ciência de dados, aliada ao *big data*, permite analisar dados de várias origens, formatos, estruturas, tamanhos permitindo, assim, uma compreensão da realidade que até recentemente não acontecia de forma sistemática. Se esse fato é verdadeiro em áreas comerciais, financeira e outras, é coerente pensar que a ciência de dados poderia ser igualmente colocada a serviço da política. Tal como CONCOLATO e CHEN (2017) escrevem:

Apesar da [ciência de dados] não estar muito desenvolvida teoricamente numa ótica científica, há uma oportunidade imensa de crescimento. A ciência de dados tem provado ser determinante na indústria tecnológica devido à necessidade crescente de compreender a quantidade de dados sendo produzida e a precisar de análise. Em suma, dados estão por todo o lado e em vários formatos.Os cientistas estão usando estatisitica e técnicas de análise de inteligência artificial como machine learning para compreender conjuntos massivos de dados e naturalmente, a sua tentativa de compreender relações entre os bancos de dados.

[traduzido pela autora]

“Even though [Data science] is not much developed in theory from a scientific perspective for Data Science, there is still great opportunity for tremendous growth. Data Science is proving to be of paramount importance to the IT industry due to the increased need for understanding the insurmountable amount of data being produced and in need of analysis. In short, data is everywhere with various formats. Scientists are currently using statistical and AI analysis techniques like machine learning methods to understand massive sets of data, and naturally, they attempt to find relationships among datasets.”

(CONCOLATO; CHEN, 2017)

Simultaneamente, as políticas atuais de transparência e de acessibilidade aos dados, estimuladas pelo Banco Mundial e pela Comissão Européia também deixam antever a importância que a ciência de dados tem hoje e terá no futuro. Muitos desses dados são relativos a populações para as quais algoritmos mais sofisticados, naturalmente ligados à ciência de dados, podem trazer *insights* diferentes e gerar evidência para apoio à formulação de políticas. Porém, é importante considerar o impacto que a junção de todas essas fontes pode ter, e o avanço e aprimoramento que pode resultar para a PBE.

Recentemente, Susan Athey do Instituto de Políticas Públicas de Stanford tem desenvolvido e melhorado algoritmos de machine learning que predizem com boa performance os grupos que podem se beneficiar mais de certas políticas (Stanford, 2018). A mesma pesquisadora refere ainda que algorimos de machine learning podem apoiar situações em que se pretende estudar causalidade e efeitos de políticas públicas, na ausência do contrafatual[1]. (ATHEY, 2015)

O departamento de ciência de dados da Universidade de Chicago tem dinamizado a articulação entre a ciência de dados e o bem social, em inglês, *Data science for social good* (UNIVERSITY OF CHICAGO, 2017) . Nesse âmbito, têm surgido vários projetos que visam demonstrar a aplicabilidade da ciência de dados para o aprimoramento de políticas, como por exemplo, para reduzir casos de corrupção num departamento público de *procurement* através de um sinalizador de anomalias, a sinalização de desigualdade de acesso a transportes

[1] O contrafatual é a situação que se verificaria se o programa/intervenção não tivesse sido aplicado, pelo que é desconhecida.

públicos, entre outros (UNIVERSITY OF CHICAGO, 2017). Por fim, salienta-se também a utilidade e complementariedade desses algoritmos, desenvolvidos no contexto de ciência de dados, para aprimorar métodos econométricos cujo objetivo é a avaliação de uma política, como no cálculo de escores de propensão. Esses escores costumam derivar de regressões logísticas, mas quando o número de covariáveis é elevado, traduzindo, portanto, uma realidade complexa que se deseja captar, é possível que outros algoritmos consigam estimar de forma mais acurada essa probabilidade. (LINDEN; YARNOLD, 2016)

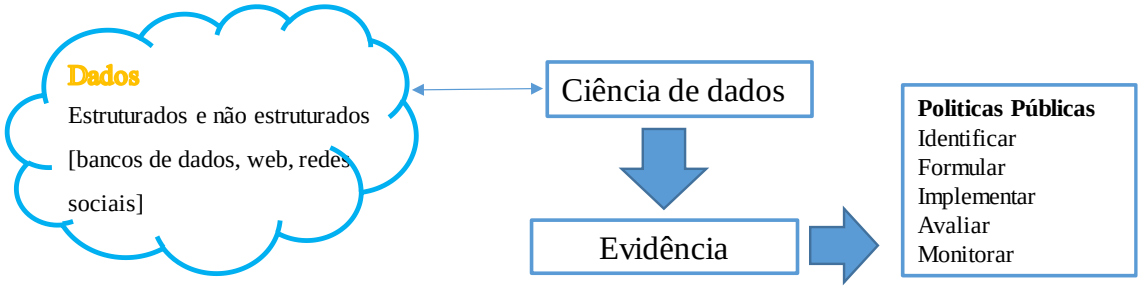
O *Berkman Kleiman Center Collection* enumera outras potencialidades ciência de dados no campo social, mas alerta para a importância de uma estrutura de governança ética global, bem como a importância de se prestar apoio aos gestores responsáveis pela tomada de decisão em políticas públicas para que sejam incluídos e estimulados a usar a evidência – nomeadamente decorrente da ciência de dados – nas suas políticas.

A esse propósito, importa salientar a vantagem comparativa que a ciência de dados oferece. Em uma revisão alusiva aos 50 anos de ciência de dados, Donoho aponta como uma das áreas de atuação da ciência de dados, a visualização e apresentação (DONOHO, 2017). Apoiados por pacotes estatísticos criados para esse objetivo, a ciência de dados atribui grande importância à visualização, seja através de gráficos mais simples, seja através de *dashboards* ou representações multidimensionais. Nesse sentido, essa característica não pode ser desconsiderada no contexto do envolvimento dos agentes políticos.

Por fim, a possibilidade de realizar predições é uma característica dos algoritmos utilizados em machine learning. Apesar da descrição e compreensão também serem uma capacidade desses algoritmos, a predição e a aprendizagem com os dados (*learning from data*) ocupam um papel predominante na utilização desses algoritmos. De fato, é nesse aspecto preditivo que mais se distanciam de abordagens tradicionais e da estatística, em que a compreensão do fenômeno em questão é central. Um exemplo concreto dessa tarefa preditiva está presente

no artigo de Pan, em que foi utilizado um algoritmo para identificar grávidas com alto risco e priorizar as mais carentes, para atribuição de subsídios, de forma a melhorar a otimização de recursos. (Pan & et al., 2017)

Figura 5 - *Relação entre a ciência de dados e as políticas públicas.*



Fonte: A própria.

Fontes de dados no setor público e ciência de dados

Relativamente ao potencial das fontes de dados, MICHIE (1991) menciona que:

"*Estes bancos de dados fornecem o material bruto na oferta de informação mas têm sido largamente impenetráveis como fontes de conhecimento especializado. Técnicas de computação podem agora ser usadas, contudo, em articulação com métodos estatísticos bem estabelecidos, para gerar classificadores estruturados em regras que não só performam bem na classificação de novas amostras mas também possuem uma boa qualidade em estruturas que visam explicação.*"

[traduzido pela autora]

"[…] these databases provide the raw material for Information supply but have been largely impenetrable as potential sources of expert knowledge. Computer-oriented techniques can now be used, however, in integration with established methods from classical statistics to generate rule structured classifiers which not only make a better job of classifying new data sampled from the same source but also possess the quality of clear explanatory structure."

(MICHIE, 1991)

A coleta de dados sobre a população que governa é da responsabilidade do Estado, que só assim poderá analisar dados sobre seus cidadãos nas mais diferentes esferas demográfica, de saúde e social. Os censos, por exemplo, são

uma das fontes mais usadas, além dos registos de nascimentos e óbitos. Outras fontes são os Inquéritos de Saúde que são primariamente usados para observar e monitorar o estado de saúde e doença da população, desde que coletados sistematicamente e em períodos regulares. Tal como os Inquéritos Nacionais de Saúde (INS), e, no âmbito da saúde, os sistemas de vigilância de doenças, também têm um papel importante na função que o Estado tem em assegurar a saúde da sua população. Outros sistemas e registos de dados são utilizados pelo Estado, desde que possibilitem construir e melhorar as suas funções básicas de governação. Entre eles encontram-se também os registos de programas e intervenções de políticas que permitem monitorar ou avaliar a sua evolução ou performance. Neste trabalho, foram utilizados os registros do Ministério da Saúde (MS) brasileiro e do Departamento de Informática do Sistema Único de Saúde (DATASUS) que possuem dados ao nível municipal sobre prevalências, incidências, entre outros. A lei brasileira também permite o acesso a dados mais confidenciais e que podem ser requeridos ao MS.

A seguir são caracterizados primeiro os Inquéritos Nacionais de Saúde Português e em seguida são descritas as áreas que foram identificadas por terem potencial e que foram eventualmente testadas para confirmar a sua importância para o planejamento em saúde pública. Em segundo lugar, foi caracterizados o DATASUS, que é a fonte de dados utilizada para o primeiro objetivo.

Inquéritos Nacionais de Saúde

Inquérito Nacional de Saúde Português

Os Inquéritos Nacionais de Saúde (INS) são realizados em Portugal desde a década de 80 e são instrumentos valiosos para o planejamento e monitoramento da saúde (DIAS, 2009). A coleta de informação relativa à condição de saúde, acesso e utilização de serviços de saúde e ainda características

sociodemográficas de indivíduos que são convidados aleatoriamente a participar, constitui um repositório de informações sobre o perfil de uma população que precisa ser mais explorado. Analisar e reportar o que esses dados mostram é fundamental para um processo que não se esgota na observação e monitoramento, mas que deve contribuir para a formulação de ações, programas ou políticas ajustadas ao perfil dessa população. Outro ponto positivo desses Inquéritos refere-se ao número de variáveis e número de indivíduos, e que oferecem uma caracterização detalhada de cada indivíduo. A importância desses Inquéritos é reconhecida mundialmente, com vários países conduzindo esses estudos de forma periódica: nos EUA, em vários países europeus com supervisão do EUROSTAT, no Brasil, entre outros.

Em Portugal, o primeiro INS data de 1987, com um intervalo de aproximadamente 10 anos entre os inquéritos realizados: 1987, 1995, 2005 e 2014. Desde 2005, porém, os INS são supervisionados pelo EUROSTAT e foram integrados no Inquérito de Saúde Europeu (EHIS). Essa modalidade, regulamentada pelo regulamento UE 141/2013 prevê que o Inquérito seja implementado em todos os países incluídos no regulamento, de acordo com as diretrizes de qualidade e confidencialidade e acesso previstas e dentro da calendarização. O documento legal estabelece ainda quais as perguntas que são obrigatórias e comuns a todos os Estados Participantes - designado módulo europeu - mas deixa a cada Estado a possibilidade de adicionar perguntas que possam ser relevantes ao nível nacional.

O INS 2014 ocorreu em parceria entre o Instituto de Estatística e o Instituto Nacional de Saúde (INSA) de Portugal. O mesmo envolveu cerca de 200 questões/variáveis (INE, 2018), nas seguintes áreas: estado de saúde, cuidados e determinantes de saúde, com questões integradas no módulo europeu, entre outras questões relativas à saúde reprodutiva, de incapacidade de longa duração, consumo alimentar e satisfação com a vida. Essas questões podem ser consultadas no documento metodológico (INE 2018) ou no site do EUROSTAT (EUROSTAT, 2019). O INS é um estudo transversal, de base populacional, com uma amostra representativa, selecionada de forma probabilística em multi-

etapas por NUTS II[2]. O INS 2014 incidiu sobre toda a população maior de 15 anos, não institucionalizada, residente em Portugal e envolveu uma amostra de mais de 18.000 participantes. (INE, 2018). A coleta de dados ocorreu entre setembro e dezembro de 2014. Os dados são confidenciais e anônimos, mas podem ser solicitados online.

Dado o número de questões colocadas e a amostra, é possível extrair conhecimentos relevantes sobre a população. De fato, o INS proporciona caracterizações muito completas sobre os indivíduos, e, se desejável, com inferência para a população, e que pode apoiar no planejamento em saúde. Nesse sentido, as duas temáticas selecionadas para estudo foram a predição de absenteísmo laboral e a caracterização de população sem plano privado de saúde. Em seguida, será apresentado o fundamento teórico para essas temáticas. Ainda, é importante mencionar que essa escolha se deveu à disponibilidade de dados e ao que era possível e relevante dentre o banco que poderia contribuir para a temática do estudo.

O absenteísmo laboral e predição

A ausência de trabalho atribuído a doenças e acidentes é uma preocupação importante da sociedade pois representa um alto ônus econômico (KOCAKULAH et al., 2016). À medida que a expectativa de vida aumenta e a incidência das doenças crônicas cresce, pode-se esperar um percentual maior de incapacidade no curto e no longo prazo (KOCAKULAH et al., 2016). À medida que as reformas

---

2 NUTS é o acrônimo **“**Nomenclatura das Unidades Territoriais para Fins Estatísticos”, sistema hierárquico de divisão do território em regiões. Essa nomenclatura foi criada pelo Eurostat no início dos anos 1970, visando a harmonização das estatísticas dos vários países em termos de recolha, compilação e divulgação de estatísticas regionais. A nomenclatura subdivide-se em 3 níveis (NUTS I, NUTS II, NUTS III), definidos de acordo com critérios populacionais, administrativos e geográficos.

públicas que impulsionam a aposentadoria para idades posteriores estão aumentando em todo o mundo, o absenteísmo do trabalho também tende a aumentar e isso é especialmente desafiador em países com altos níveis de proteção social, uma vez que está associado a pensões relacionadas a doenças e invalidez (LUND et al., 2008). Alguns estudos realizaram predições de funcionários em risco de ausência por doença (AIRAKSINEN et al., 2018) (BOOT et al., 2017) No entanto, nenhum foi realizado com dados coletados por uma pesquisa de saúde nacional e representativa. Isso é importante porque as pesquisas nacionais de saúde podem trazer informações detalhadas sobre a condição de saúde, acesso e utilização, ao contrário de fontes mais restritas, como dados administrativos das empresas. Além disso, prever quem é mais propenso a se ausentar do trabalho pode ser valioso já que identificar antecipadamente esses trabalhadores permite que os tomadores de decisão de serviços sociais e empresas planejem e apoiem esses indivíduos com ações específicas, além de alocar recursos com maior eficiência.

Planos privados de saúde e estratégias direcionadas para melhoria da saúde

A conceção de medidas e ações políticas que visam agir sobre grupos-alvo muito específicos pode ter um papel importante na gestão pública porque abordagens universais podem ser menos eficazes e mais caras (MCLAREN; PETIT, 2018). Além disso, dada a escassez dos recursos, é fundamental identificar e caracterizar grupos mais vulneráveis na população de forma a dotá-los dos bens necessários para apoiar em questões essenciais como a saúde. Entre esses grupos está a população que não tem seguro privado de saúde. Um estudo realizado no Brasil mostrou que o perfil socioeconômico de indivíduos que possuem um seguro de saúde privado é diferente daqueles que não têm (VIACAVA; SOUZA-JÚNIOR; SZWARCWALD, 2005), levantando questões

sobre quem são esses indivíduos, uma vez que estes últimos são mais vulneráveis a problemas de saúde. De fato, a caracterização dessa população pode trazer dados importantes para o seu estudo e desenho de intervenções mais eficazes. Informações nacionais sobre essa população podem ser encontradas em bancos de dados com variáveis demográficas, socioeconômicas e relacionadas à saúde. Em particular, as pesquisas nacionais de saúde (NHS) não apenas fornecem amostras representativas, mas também fornecem um grande conjunto de variáveis que são úteis para caracterizar indivíduos em termos demográficos, sociais, de percepção e status de saúde, acesso a cuidados de saúde, práticas de conhecimento em cuidados de saúde e uso de cuidados de saúde.

DATASUS

O DATASUS é o Departamento de Informática do Sistema Único de Saúde (SUS) Brasileiro. Os bancos de dados podem ser acessados online e compreendem estatísticas vitais (mortalidade e nascidos vivos), dados epidemiológicos e de mortalidade, morbidade, incapacidade, acesso a serviços, qualidade da atenção, condições de vida e fatores ambientais, assistência à saúde da população, os cadastros da Rede Assistencial, o cadastro dos estabelecimentos de saúde, além de informações sobre recursos financeiros e informações demográficas e socioeconômicas.

O objetivo desse sistema é disponibilizar informações que podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão baseadas em evidências e elaboração de programas de ações de saúde (DATASUS, 2018a).

O DATASUS disponibiliza tanto dados individuais quanto agregados. Os dados encontram-se disponíveis online mas caso não se verifique alguma informação, o usuário pode solicitá-la, sob a Lei de Acesso à Informação (LAI) n° 12.527, de

18 de novembro de 2011, que assegura o direito fundamental de acesso às informações produzidas ou armazenadas por órgãos e entidades da União, Estados, Distrito Federal e Municípios.

Assim, o Departamento de Informática do SUS é uma ferramenta de grande potencial em ciência de dados na medida em que possui um conjunto de dados muito vasto, coletado consistentemente há décadas, de forma mais ou menos sistemática, em especial em relação aos municípios brasileiros.

Avaliação de políticas de Saúde - O Programa de Saúde Mais Médicos

O planejamento, monitoramento e avaliação são fases características do processo das políticas públicas. Em qualquer uma das fases, a evidência é necessária pois contribui para processos mais rigorosos e tomadas de decisão mais transparentes. Assim, a política de saúde brasileira Mais Médicos (PMM), pela dimensão e impacto esperados, bem como pela quantidade de dados disponíveis para estudá-la e acesso aos mesmos pelo DATASUS, reúne várias condições necessárias para gerar evidência de qualidade e apoiar na sua avaliação.

O Programa Mais Médicos (PMM) é um programa lançado em 2013 para dar resposta à baixa taxa de médicos por habitante no Brasil, especialmente em zonas mais carentes como o Norte e o Nordeste (2,7, 1,09 e 1,3 médicos por 1000 habitantes) (SCHEFFER, 2015). É importante mencionar que alguns municípios não tinham qualquer médico antes do início do Programa.

Essa política, instituída pela Lei No 12.871, de 22 de outubro de 2013, teve como objetivo principal aumentar o número de médicos que trabalham em áreas remotas e nas periferias das grandes cidades. Além da alocação de médicos nas localidades mais carentes, o PMM apresentou outros objetivos, nomeadamente melhorar a formação médica e investir em infraestrutura de atenção primária

(Presidência da República 2013). Para participar do PMM, os municípios deveriam ser considerados elegíveis de acordo com as necessidades locais, e as áreas dentro de um limiar de pobreza predefinido foram consideradas prioritárias (PRESIDÊNCIA DA REPUBLICA, 2013) (CONASS, 2013).

Segundo o Ministério da Saúde do Brasil, em 2015, mais de 4.000 municípios receberam médicos do PMM (DATASUS, 2018b). O fornecimento de médicos para as áreas mais remotas foi a principal prioridade do Programa e ganhou visibilidade pública devido à contratação de médicos estrangeiros, principalmente de origem cubana, para trabalhar em áreas remotas. Até 2016, cerca de 17.000 médicos haviam sido alocados em todo o Brasil, com uma porcentagem equilibrada de médicos cubanos e brasileiros (47% e 46%, respectivamente) (PAHO/WHO, 2018).

Os elevados custos associados ao programa aumentaram a pressão por medições independentes de seu impacto na saúde (MINISTERIO SAUDE, 2013). Além disso, a medição desses efeitos toma um papel fundamental para se realizarem os ajustamentos necessários.

Porém, medir o impacto de um experimento social ou de um programa social, envolve muitas dificuldades especialmente porque medir o efeito decorrente de um programa pressupõe a atribuição desse efeito ao Programa. Inferir a causalidade requer cautela, especialmente na área social, e é necessário um desenho de estudo robusto para assegurar atribuição de efeitos ao Programa. A randomização do programa (ou tratamento) e a existência de grupos comparáveis antes do tratamento são etapas fundamentais, para, de certa forma, simular a presença de um experimento clínico. No âmbito social, essa exigência envolve muitos obstáculos levando ao surgimento de métodos que tentam simular um experimento social aleatório entre grupos semelhantes no início da intervenção. Entre outros, a técnica de escores de propensão para pareamento tem sido utilizada com esse propósito. Trata-se de um método que estima um escore que deve refletir a probabilidade de uma unidade ser selecionada para tratamento, com base nas covariáveis observadas antes do

experimento e de forma a parear essas unidades pela sua semelhança. Essa probabilidade é frequentemente calculada por meio de uma regressão logística.

Apesar de tradicionalmente este escore ser feito com regressão logística, alguns autores têm argumentado que há vantagem de utilizar algoritmos de machine learning na estimação do escore de propensão (LINDEN; YARNOLD, 2016). Inclusive, alguns autores têm demonstrado que os algoritmos de classificação tiveram uma boa performance na estimativa do escore (ELIZABETH A. STUART, BRIAN K. LEE, AND JUSTIN LESSLER, 2010).

OBJETIVO GERAL

Nesse contexto, o presente projeto propõe analisar se o uso da ciência de dados, em especial de algoritmos de machine learning, pode contribuir em alguma medida para a área das políticas de saúde da seguinte forma: 1) identificando grupos-alvo (utilizando abordagem supervisionada ou não supervisionada) para os quais programas e campanhas possam ser direcionados, visando a alocação mais eficiente de recursos, e 2) contribuindo para a elaboração de medidas que auxiliem no desenho ou avaliação de programas.

MÉTODOS

Esta seção tem como objetivo apresentar e descrever que métodos foram utilizados, bem como pontos relativos ao delineamento do estudo, fonte de dados, população e variáveis utilizadas em cada um dos artigos e que ainda não tinham sido abordados.

ARTIGO 1

Delineamento Estudo quase experimental, antes e depois.

Fonte IBGE, TSE, DATASUS e Ministério da Saúde.

Variáveis [em anexo no artigo]

Métodos: Escore de propensão seguido de um *matching* ou pareamento entre municípios pelo mesmo escore com estimativa da diferença nos indicadores de saúde selecionados, antes e depois do PMM e entre o grupo que recebeu o Programa e não recebeu (diferença nas diferenças).

Foi utilizada a regressão logística para estimar um escore de propensão para pareamento – *propensity score matching*. Esse método, característico para avaliar programas de saúde - requer a estimativa de um escore, que, neste projeto, foi inicialmente calculado por meio de um algoritmo de árvore. De fato, o objetivo inicial seria utilizar um modelo de machine learning em combinação com o pareamento, porém essa não se mostrou a melhor estratégia, pelo que foi usada uma regressão logística. De notar que nesta seção será abordado em maior detalhe o *propensity score matching* com diferença nas diferenças.

O escore de propensão para pareamento ou *propensity score matching* é uma técnica econométrica que visa aferir um escore que reflita a probabilidade de atribuição de um tratamento a uma unidade de observação. Esse escore tem fundamento na teoria causal de Rubin (RUBIN, 1974), em que avaliar o efeito ou

impacto de um programa ou tratamento é um desafio porque atribuir um efeito (causalidade) a um programa requer que se saiba o que teria acontecido a uma mesma unidade se não tivesse recebido esse programa (contrafatual). Outra premissa fundamental para avaliar o tratamento é garantir que este tenha sido distribuído de forma aleatória, garantindo que a probabilidade de receber o tratamento seja a mesma para todas as unidades de observação do estudo. Assim, a alocação do tratamento seria o elemento diferenciador das unidades que recebem ou não o tratamento (RUBIN, 1974).

Propensity score matching

O *propensity score matching* surge como forma de superação dessas dificuldades. Para cada unidade de observação, é estimado um escore de propensão, tradicionalmente, utilizando um modelo de regressão logística (LEITE, 2017). Essa regressão deve ser ajustada para todas as covariáveis observadas que possam ter interferido na alocação do tratamento (por exemplo, casos que foram atribuidos subsidios para participar em um programa ou experimento social) e nos indicadores de saúde relativamente aos quais se mede o efeito. Pensando-se em um programa de saúde em que há dois grupos de mulheres participantes que receberam ou não um subsídio para ir ao médico durante o pré-natal, e em que se quer saber se este subsídio teve efeito no número de óbitos fetais, por exemplo. Nesse caso, deve-se ajustar pelas variáveis de confundimento que são aquelas que terão tido um efeito direto na mortalidade (efeito em que queremos medir se o tratamento/subsidio teve impacto) e na probabilidade de se ter recebido o subsidio, isto é, o tratamento. Seriam, nesse exemplo, o grupo etário, entre outras. Porém, é preciso cuidado em não adicionar covariáveis que sejam anteriores à variável resposta, que, no caso acima, seria por exemplo, a entrada em funcionamento de um serviço de saúde que tenha ocorrido já durante o Programa. O motivo é que este poderá ter levado a que estas mulheres tenham melhor assistência e, consequentemente,

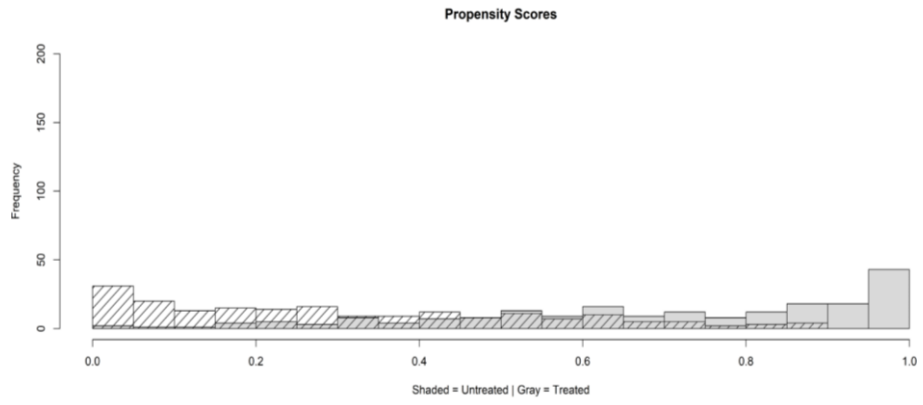
menor chance de desefecho de óbito fetal porque tiveram, entretanto, acesso a este serviços de saúde.

A probabilidade predita do score é dado pelo logaritmo da regressão, e é dada pela fórmula:

$$P\ (p) = \log(\frac{p}{1-p})$$

Em que *p* é a probabilidade de sucesso de um determinado evento, podendo assumir valores no intervalo [0,1]. O evento é, nesse trabalho, receber o programa Mais Médicos. Como já foi referido anteriormente, a estimativa de uma probabilidade pode também ser realizada através de algoritmos de machine learning, mais sofisticados e com menos pressupostos que modelos logísticos, por exemplo. A distribuição gráfica do resultado dessa probabilidade, geralmente feita por um histograma ou boxplot de acordo com a alocação efetiva do programa (isto é, do grupo de controle e de tratamento) é muito útil para verificar se existe área suficientemente sobreposta para pareamento das unidades.

Figura 6 - Exemplo de um histograma de *propensity scores*.

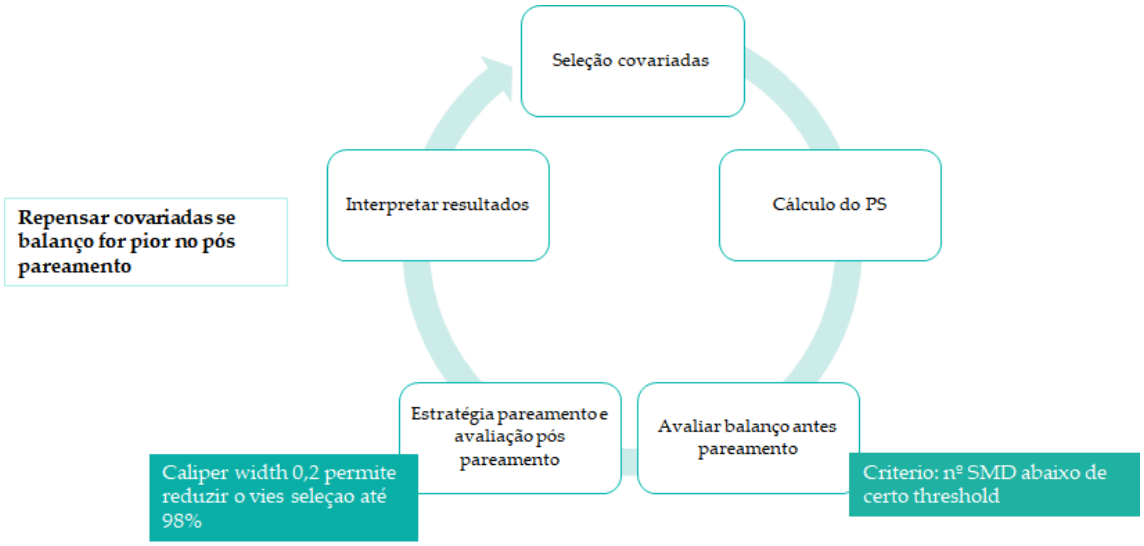


Fonte: A própria.

Nota: Este histograma reflete uma distribuição dos escores de propensão para o grupo de controle, isto é, não receber o programa (0) e tratamento, isto é, receber o programa (1).

Após a estimativa de uma regressão, acontece que as unidades de observação, divididas entre grupo de controle e de tratamento já têm um escore e é com base no mesmo que elas são pareadas. A forma de pareamento pode então, ser escolhida de acordo com diferentes técnicas, como a técnica de vizinho mais próximo, a escolha de uma distância (designado de *caliper*) correspondente a um valor estabelecido do desvio padrão do escore, entre outras. Essa escolha irá impactar no pareamento e nos resultados. A escolha de até 0.2 desvio padrão do escore de propensão para cada unidade tem sido documentado como a melhor técnica, na medida em que permite reduzir o viés na escolha da unidade para pareamento, minimizando portanto o risco de se parearem municípios com escores muito diferentes (AUSTIN, 2011). Porém, essa técnica só pode ser aplicada quando há uma área de sobreposição suficientemente grande e escores muito parecidos nos dois grupos. A figura abaixo representa o ciclo desse processo.

Figura 7 - Processo caracterizador de um escore de propensão para pareamento.



Fonte: A própria.

Nota: SMD: Standardized means difference é calculada da seguinte forma:

$$d = \frac{(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}}$$

Em que d é SMD, $\bar{x}$ é a média da covariável e s^2 a variância para cada grupo.

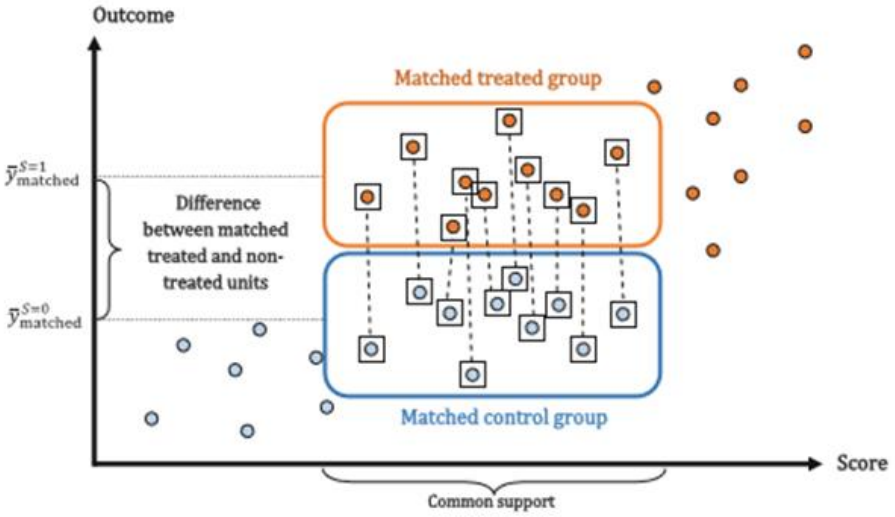
Nota: O fluxograma apresenta o processo do escore de propensão ao pareamento e relativamente ao qual apresentamos um exemplo prático. Partimos de uma situação em que temos um tratamento cujo efeito queremos medir e dois grupos, um que recebeu e outro não. Por exemplo, dois grupos de mulheres, um grupo recebe o tratamento/programa – que assumimos ser um subsídio que promova o número de consultas prenatais - e o outro não, e temos interesse em saber o efeito deste tratamento em um indicador de saúde, como a mortalidade perinatal e neonatal. O primeiro passo é selecionar as covariáveis que possam ter efeito na alocação do tratamento e na mortalidade – idade, escolaridade, renda, entre outros - e assegurar que não incluimos nenhuma covariada que possa ter ocorrido depois deste subsidio e que possa, simultaneamente, ter interferido na melhoria da mortalidade (por exemplo, número de consultas prenatais que cada mulher fez após a construção de um posto de saúde). De seguida, estas covariavéis servirão para estimar um escore, isto é uma probabilidade, com uma regressão logística e cada mulher terá um probabilidade de receber o tratamento e que será entre 0 e 1, sendo que é conhecido o grupo de quem recebeu e não. O package R *Matching* permite depois fazer o pareamento destas mulheres e avaliar, para cada grupo de quem recebeu e não o tratamento se estes grupos eram diferentes (avaliando o p-valor se era significante ou não, bem como o *standardized mean difference* (SMD) ou diferença média estandardizada) antes e depois do pareamento. Ou seja, queremos que, após o pareamento, tenhamos pareamentos que deixam de ser significantes entre grupos e valores de SMD de 0.25, sendo que 0.10 é o ideal. Estas referências podem ser encontradas em Austin (AUSTIN, 2009). Ainda, no pareamento podemos definir a distância a que queremos que duas mulheres, uma que recebeu o subsidio e outra que não recebeu, sejam pareadas tendo por base o seu escore e eventualmente, decidir descartar ou não, e repôr ou não. A fig. 7 mostra este processo muito claramente. No nosso exemplo, imagine-se uma mulher A que recebeu o Tratamento e B e C não. A tem um escore de 0.459, B de 0.458 e C de 0.454. A distância – ou caliper – definida foi de 0.02. Então, o algoritmo vai parear A com B. C, porque tem uma distância além de 0.02 será descartada.

A medição do efeito pode ser feita comparando a variável em que há interesse - a mortalidade, por exemplo - nos tratados (*Average Treatment on the Treated*) em cada unidade relativamente aos não tratados ou, olhando para o efeito médio

na população (*Average Treatment Effect*). O ATT foi a opção para o presente estudo, já que o objetivo foi avaliar o efeito causal, esperado, nos municípios que receberam o programa e não na população em geral, para o qual são necessários suposições mais difíceis de cumprir no trabalho (SHADISH, 2010).

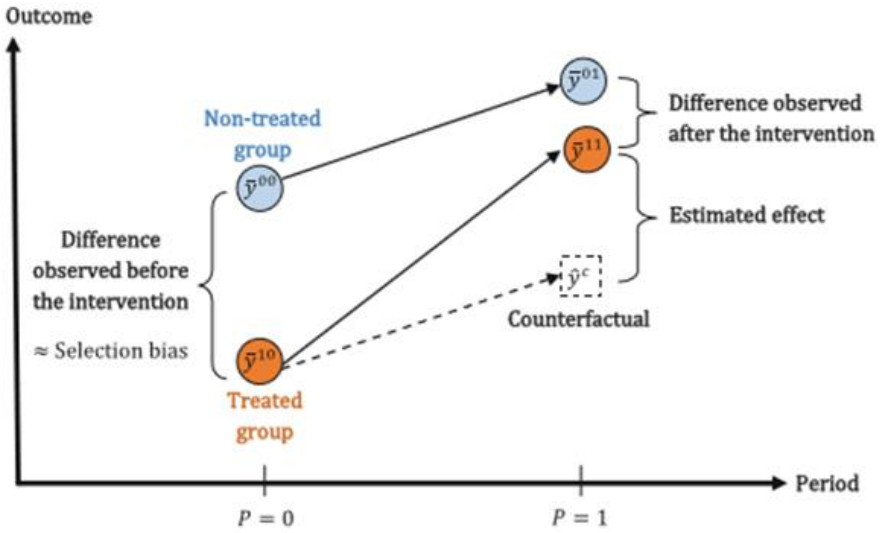
Como referido anteriormente, o efeito de um programa é medido em uma variável de interesse estabelecida logo no início do desenho de estudo e que diz respeito aos indicadores nos quais se deseja ver o efeito. Neste trabalho foram definidas várias, como serão vistas adiante: mortalidade geral e infantil, porcentagem de nascidos-vidos com peso baixo ao nascer e consultas de pré-natal, entre outras. Essa variável pode corresponder a um valor em um ponto no tempo (um determinado mês, ano) ou corresponder a um período de tempo mais longo. No presente caso, foi analisada a diferença na variável resposta em dois pontos no tempo, que correspondem a um momento antes do programa e depois do programa, resultando em duas diferenças para a variável resposta: uma para o grupo de tratamento e outra para o grupo de controle. No final, foi realizada uma diferença das diferenças (*Difference in Differences* ou DID) (JOSSELIN; LE MAUX, 2017) que é uma dupla subtração entre as variáveis resposta para cada um dos grupos, antes e depois do programa. O pareamento serve para estabelecer as unidades mais parecidas no momento anterior ao programa e assim observar a diferença na variável resposta, conforme se ilustra na imagem a seguir:

Figura 8 - Representação gráfica de um escore de propensão para pareamento.



Fonte: JOSSELIN; LE MAUX, 2017.

Figura 9 - Representação gráfica de um avaliação antes e depois de uma intervenção.



Fonte: JOSSELIN; LE MAUX, 2017.

ARTIGO 2.

Delineamento Estudo transversal.

Fonte INS 2014.

Variáveis [em anexo artigo 3]

Métodos: Análise de Agrupamentos Hierárquica nos componentes principais. Foi utilizado um algoritmo não supervisionado, no caso uma análise de agrupamento, após uma análise de correspondência múltipla (ACM). Esse método é designado de análise hierárquica de agrupamento nos componentes principais. Neste trabalho, para descrever o método foi realizada uma explicação sucinta, separada, das duas estratégias que compõem uma análise hierárquica de agrupamento nos componentes principais: uma análise de correspondência múltipla e uma análise de agrupamento hierárquica ascendente. Posteriormente, explicamos de que forma essas duas técnicas se conjugam para dar corpo à análise hierárquica de agrupamento nos componentes principais.

Agrupamento hierárquico ascendente nas componentes principais

Essa técnica foi utilizada no terceiro artigo e decorre da aplicação de uma análise de agrupamento após se ter usado a analise multivariada para a identificação inicial dos agrupamentos. O uso simultâneo dessas técnicas permite análises de agrupamento mais robustas: por um lado, a ACM permite a descrição dos dados e a Análise Hierárquica permite ultrapassar dificuldades do uso exclusivo da ACM, nomeadamente as que passam pela interpretação de proximidades geradas por fatores para além do plano principal e pela compressão excessiva e deformação dos dados. É importante ainda referir como se articulam essas duas técnicas para identificação dos *clusters*. Como referido anteriormente, a ACM permite sumarizar as categorias e os indivíduos em coordenadas que são dispostas no espaço multidimensional. Por sua vez, essas coordenadas

contribuem para a construção dos componentes principais, sendo que a contribuição de uma categoria para um eixo depende da sua representação no gráfico e está relacionada com o desvio esperado do ponto em relação à média dos pontos dessa mesma categoria. Esse desvio é tanto maior quanto maior o desvio do valor esperado, logo, a sua inércia é maior, contribuindo mais para o chi-quadrado, o qual por sua vez, também mede as associações entre as categorias e distância entre individuos. Os componentes principais respondem pela maioria do desvio dos valores esperados nos dados. A cada componente corresponde uma proporção da inércia, isto é, variabilidade, explicada pelos pontos que o compõem, em que cada componente ou eixo é inferior a 1. A soma das inércias, explicadas por cada componente, totalizam 1. Quando aplicados aos resultados de uma análise de cluster, os componentes principais são usados para agrupar os indivíduos. A coordenada fatorial de cada indivíduo se torna um valor em uma variável (o eixo), e os indivíduos com maiores semelhanças nas coordenadas fatoriais são os que têm mais probabilidade de serem agrupados em um cluster ou adicionados a um cluster em uma determinada etapa. A qualidade da classificação resultante depende de quão compactos ou homogêneos os clusters são e de quão bem eles são separados um do outro.

Análise de Correspondência Múltipla (ACM)

A Análise de Correspondência Múltipla (ACM) é uma técnica de redução de dimensionalidade, que caracteriza-se pela capacidade de sumarizar ou projetar um grande número de atributos num espaço de menor dimensão, mas que represente e mantenha a informação presente no conjunto de dados (MIGUEL A. CARREIRA-PERPINAN, 2014). Num banco de dados, por exemplo, podemos verificar a existência de um elevado número de atributos/variáveis correlacionados e que pode prejudicar a visualização e análise dos dados. Inclusive, em termos computacionais, o aumento de atributos faz com que os algoritmos tenham que processar um volume de dados muito maior, o que se torna computacionalmente custoso em termos de tempo. Porém, para o

tratamento e análise desse banco de dados, muitas vezes é importante sumarizar sem perder a informação presente no mesmo. Nesse contexto, a redução de dimensionalidade é um procedimento que permite representar em um número de dimensões menor que a dimensão incial do conjunto dos dados, mantendo a informação constante nesses dados.

Existem diferentes métodos para reduzir dimensionalidade e que dependem do tipo de variáveis presente no banco de dados. Esses métodos funcionam identificando conjuntos de variáveis não correlacionadas entre si que explicam a maior parte da variabilidade dos dados, de forma a que o produto final seja uma matriz de rank menor que permite explicar os dados originais e reconstrui-los de forma o mais aproximada possível. A redução de dimensionalidade pode ser utilizada enquanto técnica de pré-processamento para implementar posteriormente uma tarefa de *machine learning* como por exemplo classificação, pois espera-se que quando o espaço de características contém somente as características mais salientes, o classificador seja mais rápido. A redução de dimensionalidade permite visualizar estes dados em menos dimensões e podemos observar os padrões de forma mais clara.

A Análise de Correspondência Múltipla (ACM) é uma técnica de redução da dimensionalidade utilizada em bancos de dados categóricos, opondo-se, nesse aspecto, à Análise de Componentes Principais (ACP), utilizadas principalmente em bancos com dados contínuos. A ACM permite visualizar associações entre linhas e colunas de uma matriz de dados. Essa visualização gráfica tem um importante papel no desenvolvimento de *insights* sobre a estrutura dos dados. Uma das principais vantagens é a capacidade de descrever graficamente dados de pesquisa de grandes bancos de dados em poucas dimensões. De fato, esse tipo de técnica permite: 1) sumarizar e descrever a informação mais importante de um conjunto de dados 2) oferecer uma representação geométrica dos dados e 3) facilitar a interpretação e observar associações entre variáveis (HJELLBREKKE, 2007).

De forma a elucidar sobre o processo de ACM, é apresentado a seguir um exemplo concreto, adaptado de (NASCIMENTO, 2011) que visa explicar a transformação de uma matriz indicadora para uma matriz de correspondência e perfis de linha e/ou coluna, perfis médios e massas. Imagine-se a seguinte base de dados, em que temos na linha observações, ou indivíduos, e nas colunas as categorias. Em suma, tem-se uma tabela dummy (Fig. 18), em que nas marginais das linhas e das colunas são obtidos o total:

Figura 10 - Matriz Indicadora do exemplo.

OBSERVAÇÕES	Variáveis e suas categorias								Total de Linha
	V1		V2		V3		V4		
	S	N	S	N	S	N	S	N	
O1	0	1	1	0	1	0	0	1	4
O2	1	0	0	1	0	1	0	1	4
O3	0	1	1	0	0	1	0	1	4
O4	0	1	1	0	1	0	0	1	4
O5	1	0	0	1	1	0	1	0	4
Total de coluna	2	3	3	2	3	2	1	4	20

Fonte: NASCIMENTO, 2011.

De seguida esta matriz é transformada em uma *Matriz de Correspondência* (Fig. 19) em que,

$$f_{ij} = \frac{a_{ij}}{T}$$

f_{ij} corresponde à frequência relativa da resposta da observação *i a uma categoria j;*

a_{ij} corresponde à resposta da observação i a uma categoria j podendo assumir 0 ou 1;

T o número total de amostras.

Figura 11 - Matriz Correspondência do exemplo.

OBSERVAÇÕES	Variáveis e suas categorias								Total de Linha (r_i)
	V1		V2		V3		V4		
	S	N	S	N	S	N	S	N	
O1	0	0,05	0,05	0	0,05	0	0	0,05	0,2
O2	0,05	0	0	0,05	0	0,05	0	0,05	0,2
O3	0	0,05	0,05	0	0	0,05	0	0,05	0,2
O4	0	0,05	0,05	0	0,05	0	0	0,05	0,2
O5	0,05	0	0	0,05	0,05	0	0,05	0	0,2
Total de coluna (c_j)	0,1	0,15	0,15	0,1	0,15	0,1	0,05	0,2	1,0

Fonte: NASCIMENTO, 2011.

De seguida, é possivel construir a Matriz de Correspondência padronizada (Fig. 20):

$$g_{ij} = \frac{f_{ij}}{\sqrt{r_i}\sqrt{c_j}}$$

Onde:

f_{ij} é a frequência relativa padronizada da resposta da observação i a categoria j;

r_i é a frequência relativa marginal da linha i;

c_j é a frequência relativa marginal da coluna j.

Figura 12 - Matriz Correspondencia Padronizada.

OBSERVAÇÕES	Variáveis e suas categorias							
	V1		V2		V3		V4	
	S	N	S	N	S	N	S	N
O1	0	0,289	0,289	0	0,289	0	0	0,250
O2	0,354	0	0	0,354	0	0,354	0	0,250
O3	0	0,289	0,289	0	0	0,354	0	0,250
O4	0	0,289	0,289	0	0,289	0	0	0,250
O5	0,354	0	0	0,354	0,289	0	0,500	0

Fonte: NASCIMENTO, 2011.

A matriz de correspondência padronizada dará origem a uma matriz de autovetores e autovalores em que as suas colunas são as dimensões. A matriz de autovetores (Fig. 21) é a matriz de coordenadas de coluna e a matriz de autovalores (Fig. 22) permitirá construir a matriz de valores singulares (Fig. 23). A construção da matriz de autovetores decorre da decomposição por Valores Singulares da Matriz de Correspondência paronizada.

Figura 13 - Matriz de autovetores e Coordenada de colunas.

Colunas	Autovetores							
V1S	-0,316	0,475	-0,141	-0,237	-0,018	0,680	-0,193	0,314
V1N	-0,388	-0,386	0,115	0,190	-0,019	0,193	0,680	0,388
V2S	-0,388	-0,386	0,115	0,190	-0,019	-0,193	-0,680	0,388
V2N	-0,316	0,475	-0,141	-0,237	-0,018	-0,680	0,193	0,314
V3S	-0,387	0,000	0,542	-0,328	0,565	0,000	0,000	-0,360
V3N	-0,316	0,001	-0,664	0,397	0,466	0,000	0,000	-0,289
V4S	-0,223	0,448	0,399	0,667	-0,300	0,000	0,000	-0,237
V4N	-0,447	-0,224	-0,200	-0,325	-0,610	0,000	0,000	-0,482

Fonte: NASCIMENTO, 2011.

Figura 14 - Matriz autovalores.

1,000	0	0	0	0	0	0	0
0	0,625	0	0	0	0	0	0
0	0	0,321	0	0	0	0	0
0	0	0	0,054	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Fonte: NASCIMENTO, 2011.

A inércia, ou variância da nuvem, é representada pela matriz de valores singulares, em que:

$$\sigma_{ij} = \sqrt{\lambda_{ij}}$$

E $\sqrt{\lambda_{ij}}$ é o autovalor.

Figura 15 - Matriz de valores singulares.

1,000	0	0	0	0	0	0	0
0	0,791	0	0	0	0	0	0
0	0	0,567	0	0	0	0	0
0	0	0	0,232	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Fonte: NASCIMENTO, 2011.

Os valores singulares são utilizados para explicar a inércia de cada dimensão e para calcular as *coordenadas de linha e principais dos pontos* (Fig. 24).

Figura 16 - Coordenadas de linha.

Linhas	Dimensões							
	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8
O1	-0,448	-0,353	0,305	-0,285	0	0	0	0
O2	-0,447	0,355	-0,679	-0,468	0	0	0	0
O3	-0,448	-0,352	-0,386	0,729	0	0	0	0
O4	-0,448	-0,353	0,305	-0,285	0	0	0	0
O5	-0,447	0,708	0,452	0,306	0	0	0	0

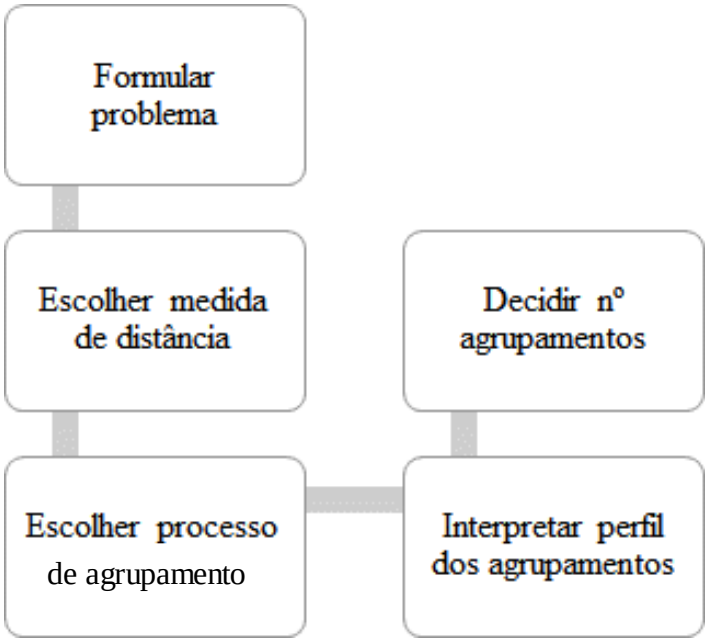
Fonte: NASCIMENTO, 2011.

Estas coordenadas são, então, posicionadas num Mapa de Correspondência.

Análise de Agrupamento

Como referido anteriormente, algoritmos não supervisionados são caracterizados por qualquer função que, tendo um conjunto de preditores, a variável resposta, *y* é desconhecida. A análise de *clusters*, ou análise de agrupamentos, tem como finalidade construir grupos *(clusters)* com base nas características que as observações possuem. Essa técnica classifica os objetos de modo que cada um seja semelhante aos outros dentro do seu grupo. Assim, os agrupamentos resultantes devem exibir elevada homogeneidade interna e elevada heterogeneidade externa. Uma das finalidades da utilização da análise de agrupamentos inclui a redução de dados por meio da redução de informação de uma população inteira, ou de uma amostra, para a informação de subgrupos específicos e menores (HAIR, JR, RE ANDERSON, 2005). As etapas de uma análise de agrupamentos estão descrita na figura abaixo.

Figura 17 - Processo de Análise de Agrupamento**.**



Fonte: A própria.

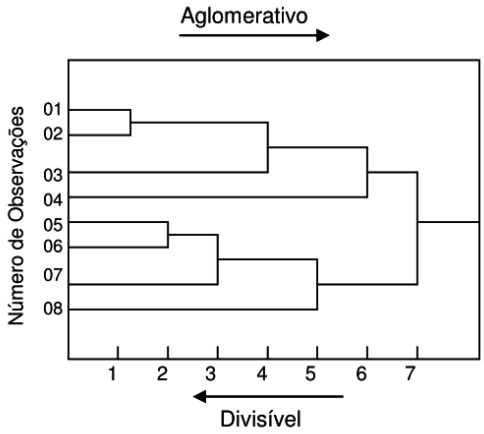
A figura 25 descreve as principais etapas numa análise de agrupamento. Após formular a pergunta de pesquisa, que deve ter em conta a natureza descritiva desse tipo de análise, o pesquisador deverá identificar qual a medida de distância e qual a estratégia que irá utilizar para agrupar as observações (HAIR, JR, RE ANDERSON, 2005). Em seguida, são apresentadas de forma sumária as decisões tomadas nessa análise relativamente aos dois pontos descritos.

Estratégia de agrupamento

No contexto das análises de agrupamento, existem duas estratégias de agrupamento possiveis: (1) método não hierárquico por repartição e (2) *clustering* hierárquico aglomerativo e/ou divisivo (Fig. 20). Em (1) um conjunto de observações é dividido em uma partição simples de *k* grupos, em que *k* corresponde a um valor selecionado pelo pesquisador. Neste, o algoritmo mais utilizado é o *k means,* e que se traduz num centróide que é o valor médio estimado a partir das variáveis do grupo (HAIR, JR, RE ANDERSON, 2005).

O agrupamento herárquico, por sua vez, pode ser aglomerativo ou divisivo. No primeiro, os elementos, individuais, são sucessivamente aglomerados, terminando no nó que corresponde ao momento em que já não existe nenhum objeto pendente. Os elementos mais similares são fusionados primeiramente. No cluster divisivo, o processo funciona de forma contrária, isto é, os conjuntos de elementos são separados progressivamente, até que se chegue ao nível dos sub-conjuntos menores. Ambos os tipos de agrupamento hierárquico estão descritos na figura 26.

Figura 18 - Dendrograma.



Fonte: NASCIMENTO, 2011.

Legenda: A imagem é um dendrograma, também designado de árvore hierárquica. O dendrograma, como se observa é um gráfico em forma de árvore em que o nível de similaridade é indicado pelo eixo horizontal e em que, as linhas que partem dos elementos – na figura, presentes no eixo vertical – têm uma altura correspondente ao nível em que os elementos foram considerados semelhantes. A imagem mostra ainda a direção do processo de agrupamento, segundo ser divisivo ou aglomerativo.

Medidas de Proximidade

As medidas de distância, ou proximidade, são a etapa seguinte na construção do agrupamento e são calculadas para todos os pares dos objetos assim

determinando os elementos selecionados na formação do *cluster*. Quando os elementos são agrupados com base em medidas de dissimilaridade, quanto menor for o valor mais parecidos são os elementos. Uma das medidas mais utilizadas é a Euclidiana (DE), definida pela fórmula:

$$\mathrm{DE} = \sum_{n=1}^{p} \left(x_{ij} + x_{i'j}\right)^2$$

Onde x_{ij} é a j-ésima caracteristica do i-ésimo indivíduo e $x_{i'j}$ é a j-ésima caracteristica do i-ésimo indivíduo, em que a posição de qualquer um destes pontos em um p –espaço Euclideano - é um vetor. Quanto mais próximo de zero mais similares eles são. O processo encontra-se na Fig.27.

Figura 19 - Matriz de proximidade de distâncias Euclidianas entre observações.

	Observação						
	A	B	C	D	E	F	G
A	-						
B	3.162	-					
C	5.099	2.00	-				
D	5.099	2.828	2.000	-			
E	5.000	2.236	2.236	4.123	-		
F	6.403	3.606	3.000	5.000	1.414	-	
G	3.606	2.236	3.606	5.000	2.000	3.162	-

Fonte: Própria.

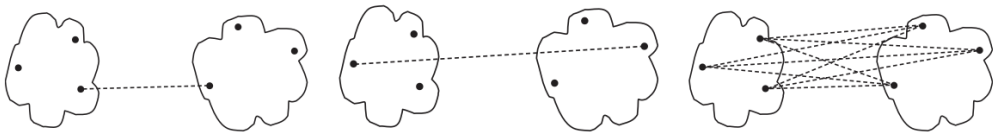
Legenda: A imagem é uma matriz com os valores de distâncias Euclideanas entre observações, isto é, de dissimilaridade. A uma menor distância corresponde menor dissimilaridade, pelo que o par de elementos E e F apreseta o menor valor e são os mais parecidos. No figura 23 apresentamos estes elementos dispostos num espaço.

Técnicas de aglomeração

As técnicas aglomerativas dizem respeito àquelas que serão utilizadas para desenvolver agregados e são utilizadas no conjunto dos métodos hierárquicos.

As técnicas utilizadas envolvem diferentes tipos de ligação: simples, completa, média, centróide e Ward (HAIR, JR, RE ANDERSON, 2005) (fig. 24).

Figura 20 - Possibilidades de técnicas de ligação em análise de agrupamentos.



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Fonte: BRIGITTE LE ROUX, 2010.

O primeiro (a) – é designado de ligação simples ou método do vizinho mais próximo – e utiliza a distância mínima entre dois itens e os coloca no primeiro agrupamento. Em seguida, a próxima distância mais curta é calculada e um terceiro item é agregado aos dois iniciais e assim sucessivamente.

O segundo (b) – ligação completa – é semelhante ao primeiro, mas baseia-se na máxima distância entre dois itens em grupos diferentes. Esse método nem sempre funciona bem em todas as situações, principalmente quando os grupos tendem a ser alongados, ou seja, quando a forma dos clusters é mais oval.

No método da ligação média (c), a distância entre dois grupos é calculada pela média entre todos os pares de itens pertencentes aos grupos. Esse método tende a combinar agregados com pequena variação interna.

O método do centróide utiliza como medida de similaridade entre dois agrupamentos a distância entre seus centróides, definidos como os valores médios dos elementos sobre as variáveis em estudo. Deste modo, a cada adição de um novo elemento ao grupo um novo centróide é computado. Assim, a distância entre dois grupos é igual à distância entre seus centróides.

Finalmente, o método de Ward ou também conhecido por método de variância mínima, utiliza uma abordagem diferente dos demais. Esse método visa

minimizar a perda associada a cada agrupamento. A perda é quantificada pela diferença entre a soma dos erros quadráticos de cada padrão e a média do agrupamento em que se encontra. A soma dos erros quadráticos é definido como: (HAIR, JR, RE ANDERSON, 2005):

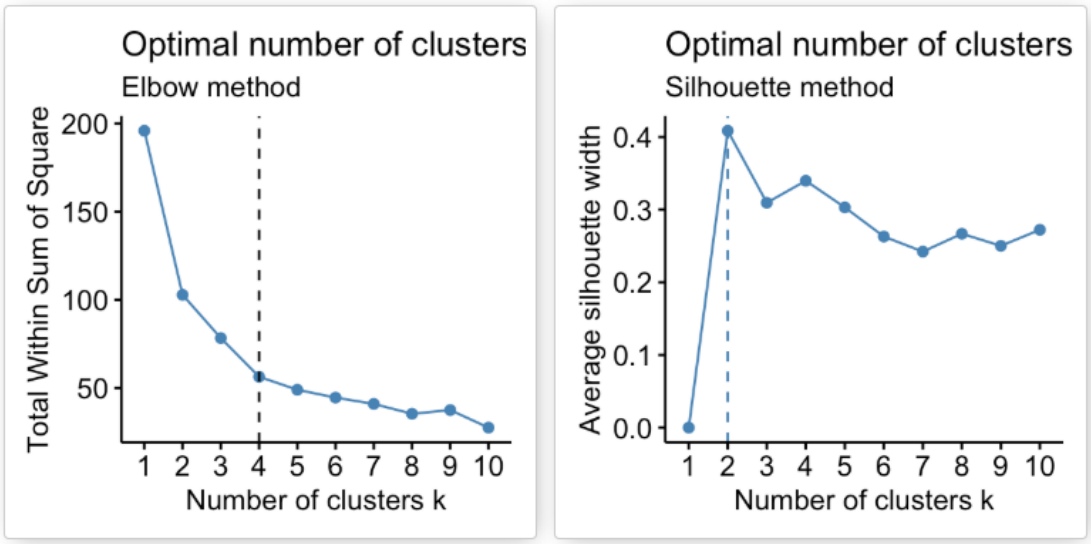
$$ESS_k = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2$$

Onde k é o agrupamento em questão, n é o total de elementos do agrupamento k e x_i é o i-ésimo elemento do agrupamento k.

Seleção número de agrupamentos

Por fim, a última decisão na construção de agrupamentos, ou seja, o número de agrupamentos, também conhecida pela regra da parada. Existem algumas técnicas que permitem verificar o número de agrupamentos, como por exemplo a regra do cotovelo ou silhueta (fig.29) (ALBOUKADEL KASSAMBARA, 2019), ou, ainda, o dendrograma que, num cluster hierárquico, permite identitifcar o número de agrupamentos pelo tamanho das linhas verticais, como explicamos anteriormente neste trabalho. Contudo, qualquer uma delas não dispensa o olhar critico do pesquisador quando o algoritmo “propõe” um número de agrupamentos. Qualquer uma das técnicas usadas é, apenas, uma representação gráfica resultante das opções tomadas anteriormente, ou seja, elas mostram em que número de agrupamentos (eixo horizontal da fig.29) se dá um decréscimo na métrica escolhida (eixo y) que já não justifica acrescentar outro cluster. Veja-se que nós queremos que esse valor (a escolha de adicionar outro cluster) seja ótimo, ou seja, que maximize a distância entre clusters e minimize a distância inter-clusters (HUSSON, F. JOSSE, J, PAGÉS, 2008).

Figura 21 - Técnica de cotovelo para escolha do número ótimo de clusters, de acordo com a métrica de proximidade entre observações escolhida.



Fonte: ALBOUKADEL KASSAMBARA, 2019.

ARTIGO 3.

Delineamento: Estudo transversal.

Fonte: INS 2014.

Variáveis [em anexo no artigo 1]

Métodos Foram utilizados algoritmos supervisionados como regressão logística penalizada e algoritmos baseados em árvores de classificação, como random forests e boosting trees.

Regressão logística

A regressão logística é uma técnica estatística que tem como objetivo produzir um modelo que permita a predição de uma variável categórica a partir de uma série de variáveis explicativas. Apesar de se tratar de uma técnica utilizada tradicionalmente em estatística, também é utilizada enquanto algoritmo de

machine learning. Esse modelo permite estimar uma probabilidade em que a variável resposta é modelada por meio da distribuição binomial e que tem como parâmetro, *p*, a probabilidade de ocorrência de uma classe especifica (SANTOS, 2018). A probabilidade estimada deve estar no intervalo [0,1] e apresentar relação direta com os preditores mensurados para modelar a relação entre a probabilidade de determinada resposta e os preditores. Trata-se de um método de predição que não assume pressupostos sobre a distribuição dos dados (POHAR; BLAS; TURK, 2004) pelo que é amplamente usado em predição.

O ajuste do modelo logístico pode ser realizado a partir da aplicação do método da máxima verossimilhança, que estima valores para o vetor de parâmetros β. Essas estimativas corresponderão ao *logit* que depois de modelado resultará em uma probabilidade predita para cada observação do conjunto de treinamento, mais próxima da verdadeira classe a qual a observação pertence. Para uma resposta dicotômica, rotulada com 0s e 1s, se $P(x)$ é a probabilidade associada à presença de determinada resposta, a chance desse desfecho ocorrer será:

$$\frac{Pr(Y=1|X=x)}{Pr(Y=0|X=x)} = \frac{p(X)}{1-p(X)} = \exp(\beta_0 + \Sigma_j \beta_j x_j)$$

A fronteira de decisão linear do modelo de regressão logística estará relacionada à escolha de um ponto de corte para $P(x)$. Por exemplo, se o ponto de corte escolhido for $P(x) = 0.5$, indivíduos com $P(x) > 0.5$ serão classificados como um evento (presença de determinada resposta) e aqueles com $P(x) < 0.5$ como um não evento (ausência de determinada resposta). Logo, $P(x) = 0.5$ é a opção inicial para definir a fronteira de decisão para o modelo. É a partir da aplicação deste ponto de corte que faz a regressão logística um algoritmo de classificação em machine learning na medida em que se torna uma decisão categórica binária.

Métodos Ensemble – Random forests

Os métodos *ensemble* – ou métodos de aprendizado por agrupamento - nos quais se inserem as florestas aleatórias (random forests), são modelos não paramétricos que funcionam a partir de uma coleção de hipóteses que no final são combinadas, ao contrário dos métodos tradicionais em que apenas um resultado é obtido. No caso das florestas aleatórias, é criada uma combinação (*ensemble*) de árvores de decisão, que visa melhorar a acurácia e estabilidade do modelo. É importante esclarecer que a árvore de decisão funciona através de particionamentos recursivos no espaço das covariáveis, em que cada particionamento se designa de nó e cada resultado é uma folha (IZBICKI, 2016). A seleção das covariáveis e ponto de corte que darão origem a um nó da árvore é geralmente feito por via do critério de impureza, que se traduz em duas métricas no caso da classificação: índice de Gini e entropia e que visam obter o maior ganho de informação em determinada combinação covariáveis.

O algoritmo de floresta aleatória, contudo, faz essa seleção em diferentes subconjuntos de preditores permitindo modificar a criação de cada árvore, para que essas sejam diferentes umas das outras e também para reduzir a correlação entre as árvores (SANTOS, 2018). Mais especificamente, são definidos inicialmente o número de covariáveis m utilizadas para fazer a divisão em cada nó, em que $m < d$. Essas m covariáveis são escolhidas aleatoriamente dentre as covariáveis originais d, e a cada nó criado, um novo subconjunto de covariáveis é sorteado. O valor de m pode ser escolhido por validação cruzada, mas em geral $m = \sqrt{d}$, leva a uma boa performance. O crescimento de cada árvore em uma floresta aleatória não é apenas baseado em subconjuntos de variáveis explicativas em cada nó, mas também por meio um subconjunto aleatório da amostra – em que são sorteadas observações e colunas – e são agregadas as previsões individuais para cada árvore e que irão contribuir para a previsão final. Esse processo é denominado de bagging ou bootstrap aggregation (HASTIE et al., 2013). Nesse processo de treinamento/teste de cada subconjunto, o erro para cada subconjunto ou árvore é testado numa parte da

amostra que foi deixada de lado nessa mesma árvore, e se chama de out-of-bag (OOB).

Os hiperparâmetros de uma floresta aleatória podem ser ajustados, entre eles o número de observações no conjunto de treinamento, o número de árvores na floresta e número de covariáveis utilizados para construir cada árvore.

Algumas das vantagens desse método são a capacidade de lidar com correlações e bancos de dados em que há poucas observações e muitos preditores e interações. Outra vantagem é a possibilidade de identificar e listar as variáveis preditoras mais importantes, designado de *feature importance*. Este processo de seleção é feito a partir de diferentes métricas que consideram, por exemplo, o desempenho do modelo nas observações OOB (KUHN; JOHNSON, 2013) ou o índice de Gini (KUHN; JOHNSON, 2013).

Métodos ensemble – Gradient Boosting

Boosting é um algoritmo de predição pertencente à categoria das árvores de decisão. Esse algoritmo consiste na construção de árvores sequenciais, em que cada uma é aprimorada com base nos resíduos da anterior (CHENG LI, 2013) permitindo uma boa performance em vários modelos preditivos. De uma forma bem intuitiva, tenta-se ensinar um modelo a predizer valores Y na forma $h(x) = Y$ e minimizando o resíduo deste modelo. Esse modelo fornecerá informações sobre o resíduo e que serão contempladas no modelo (ou árvore) seguinte (LEG/UFPR, 2016). Ou seja,

$$resíduo = g(x) + residuo2$$

Em que $g(x)$ é uma função não comtemplada em $h(x)$ mas vai ajudar a explicar Y, ou seja,

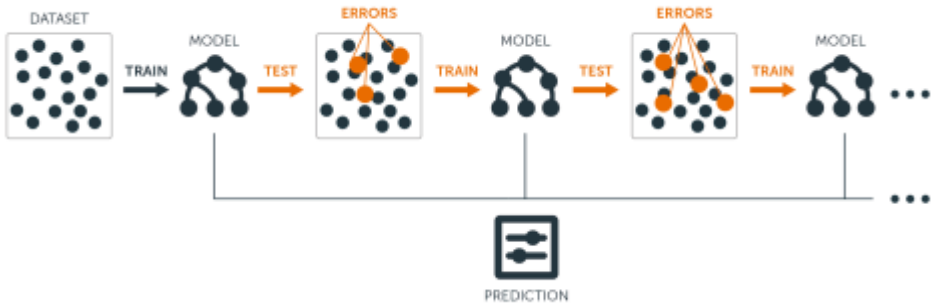
$$Y = h(x) + g(x) + residuo2$$

A árvore daqui resultante combinará essas duas funções de forma a que:

$$h\,(x)^2 = h(x)^1 + g(x)$$

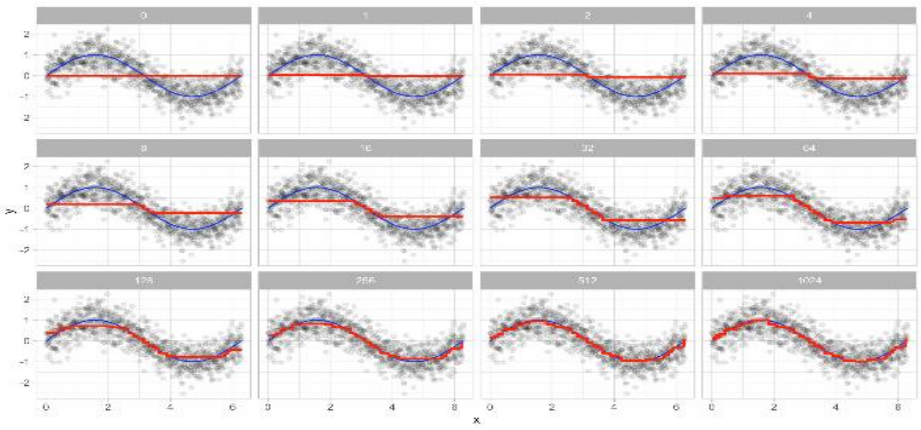
a cada interação. Esse processo decorre de forma sucessiva até que algum hiperparâmetro indique o fim, ou seja, o processo ocorre pelo número de iterações programadas. O modelo final é o resultado do conjunto aditivo de árvores geradas a partir do aprimoramento dos erros das anteriores:

Figura 22 -Caracterização do processo da técnica Boosting.



Fonte: BOEHKME, 2018.

Figura 23 - Caracterização do processo da técnica Boosting.



Fonte: BOEHKME, 2018.

Notas: Nesta imagem são ilustrados os passos para melhoria da performance do modelo através do ajuste dos modelos aos resíduos anteriores, de forma sequencial como explicado anteriormente.

Avaliação da performance e limitações

A avaliação de um modelo de machine learning é feita no conjunto de teste. A performance do modelo, no caso de se tratar de um modelo de classificação supervisionado, pode ser avaliada mediante diversas métricas. Porém, dependendo do objetivo da pesquisa e do banco que foi utilizado, pode ser interessante e necessário utilizar várias opções, pois algumas podem esconder fragilidades, como conjuntos com dados desbalanceados, em que o modelo pode ter sido treinado com base em poucas observações de uma determinada classe. Então, o ideal será sempre analisar e contrapor várias métricas. Para algoritmos supervisionados de classificação, as métricas mais utilizadas *são*:

a) Acurácia, ou taxa de erro, que corresponde à taxa de predições corretas ou incorretas realizado por um modelo. Trata-se do desempenho geral do modelo e é dada pela equação:

$$\boldsymbol{Acurácia} = \frac{\boldsymbol{Verdadeiros\ Positivos} + \boldsymbol{Verdadeiros\ Negativos}}{\boldsymbol{Total}}$$

b) Sensibilidade (ou *recall*) ou seja, a proporção de casos positivos classificados como positivos pelo modelo:

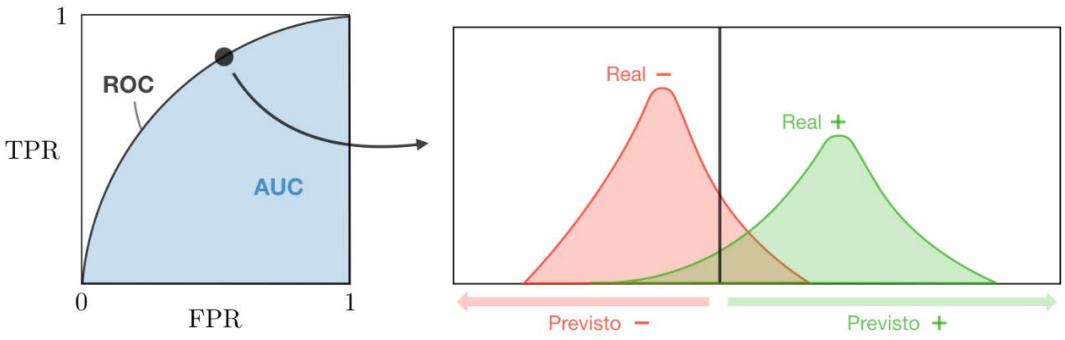
$$\boldsymbol{Sensibilidade} = \frac{\boldsymbol{Verdadeiros\ positivos}}{\boldsymbol{Verdadeiros\ positivos} + \boldsymbol{Falsos\ Negativos}}$$

c) Especificidade, que corresponde à proporção de casos negativos classificados como negativos pelo modelo:

$$\boldsymbol{Especificidade} = \frac{\boldsymbol{Verdadeiros\ negativos}}{\boldsymbol{Verdadeiros\ negativos} + \boldsymbol{Falsos\ positivos}}$$

d) ROC e AUC - A curva de operação do receptor, também chamada de ROC (*Receiver Operating Characteristic*) é uma representação gráfica que ilustra a peformance de um modelo de classificação binário. Ela representa um valor contínuo no intervalo [0,1] e que corresponde à área abaixo da curva (AUC). Essa área é obtida pela razão de Verdadeiros Positivos ou Sensibilidade (TPR, ou *True Positive Rate* em inglês) e pela razão de Falsos Positivos ou Especificidade (FPR, ou *False Positive Rate*). Existem muitas opções de ponte de corte definidas pelo pesquisador para estabelecer um limiar de decisão relativamente ao que será contabilizado como resposta ausente e/ou presente. Assim, para cada ponto de corte podem ser calculados valores de sensibilidade e especificidade, que serão dispostos em um gráfico denominado curva ROC, que apresenta no eixo das ordenadas os valores de sensibilidade e nas abscissas o complemento da especificidade, ou seja, o valor (1-especificidade).

Figura 24 - Ilustração de uma curva ROC na figura a esquerda e de um hipotético ponto de corte na curva, com as curvas de sensibilidade e especificidade à direita que correspondem ao ponto de corte apresentado na curva.



Fonte: AMIDI, 2017.

Nota: Na figura à direita observamos o ponto de corte situado entre duas curvas, a de especificidade e sensibilidade. Quando o ponto de corte (isto é, a linha vertical) se move para a

esquerda, a área da curva da sensibilidade aumenta, aumentando a identificação de verdadeiros positivos, à custa do aumento também de falsos positivos. Assim, o ponto de corte ótimo é aquele que permitiria a maximização dessa área (AUC), representado pela imagem à esquerda.

Outro instrumento muito útil para avaliar a performance do modelo é a Matriz de Confusão. A seguinte imagem revela como se deve interpretar e as métricas que podem ser retiradas da mesma:

Figura 25 - Matriz de confusão para avaliação de uma análise preditiva de classificação binária.

	CLASSE PREVISTA	
CLASSE REAL	VP Verdadeiro Positivo	FN Falso Negativo Erro tipo 2
	FP Falso Positivo Erro tipo 1	VN Verdadeiro negativo

Fonte: A própria.

No canto superior esquerdo estão 1) os Verdadeiros Positivos (VP) que correspondem às observações que eram casos e foram corretamente identificados como tal; 2) Os Falsos Positivos (FP), no canto inferior esquerdo, e que corresponde às observações que não eram casos mas foram preditos como tal 3) Verdadeiros negativos (VN), corretamente preditos como negativos e 4) Falsos Negativos (FN), preditos como negativos mas sendo positivos.

Porém, na presença de dados desbalanceados, isto é, quando o número de observações é baixo relativamente às outras classes, o modelo pode apresentar uma boa acurácia mas pode esconder uma baixa performance na identificação de classes com frequência baixa. Nesse caso, pode ser necessário olhar para algumas das seguintes métricas (FERRI; HERNÁNDEZ-ORALLO; MODROIU, 2009):

a) Precisão: É a probabilidade das classificações positivas serem realmente positivas:

$$Precisão = \frac{Verdadeiros\ Postivos}{Verdadeiros\ Positivos + Falsos\ Positivos}$$

b) F1 *score*: Essa métrica combina precisão e recall de modo a trazer um número único que indique a qualidade geral do modelo e trabalha bem até com conjuntos de dados que possuem classes desproporcionais.

$$F1\ score = \frac{2 * precisão * recall}{precisão + recall}$$

c) Cohen Kappa: Essa métrica também é útil para classes desbalanceadas e corresponde:

$$\frac{Acurácia\ observada - acurácia\ esperada}{1 - acurácia\ esperada}.$$

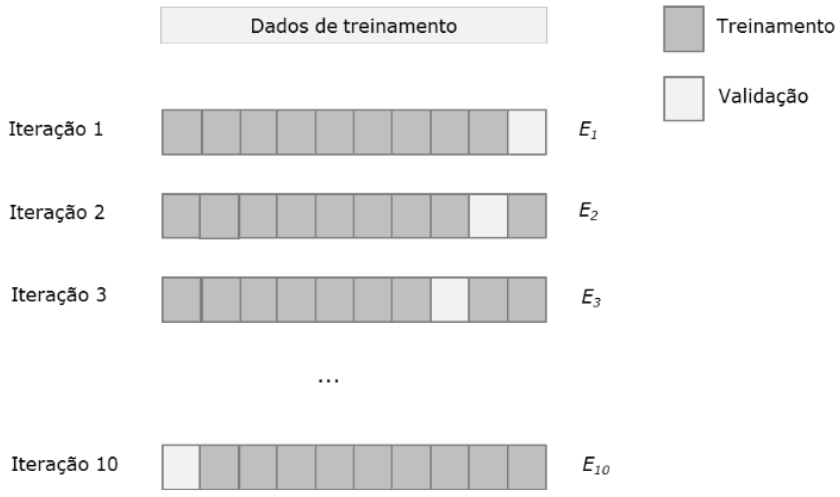
Em que valores ≤ 0 indicam sem concordância, 0.01–0.20 pouco a nenhuma, 0.21–0.40 como razoável, 0.41– 0.60 moderada, 0.61–0.80 substancial e 0.81–1.00 como quase perfeito. (LANDIS, 1977)

Umas das técnicas mais usadas em *machine learning* que permite melhorar a performance de um modelo é a validação cruzada (KUHN; JOHNSON, 2013). Às vezes, a par do pré-processamento dos dados - padronizar, aferir e recategorizar os dados originais – existem técnicas mais específicas para tentar melhorar os resultados obtidos pelo algoritmo (KUHN; JOHNSON, 2013), nomeadamente a validação cruzada que foi usada neste artigo.

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados (KOHAVI, 1995) sendo muito útil quando o objetivo é a predição. Caracteriza-se pelo particionamento do conjunto de dados de forma aleatória, em subconjuntos mutualmente exclusivos, e posteriormente, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos

(dados de validação ou de teste) são utilizados para a validação do modelo. Embora existam diversas formas de realizar o particionamento dos dados, as três mais utilizadas são: o método holdout, isto é, particionar o banco em duas partes, uma para treinamento e teste, o k-fold e o leave-one-out (KOHAVI, 1995). A validação cruzada permite ajustar os hiperparâmetros ainda no conjunto de treinamento, para que o erro estimado no conjunto teste seja melhor e permita a generalização a amostras novas, melhorando, assim a performance do algoritmo. Outro aspeto importante é que a validação cruzada permite minimizar a instabilidade do algoritmo, a qual decorre do fato de que pequenas alterações nos dados de treinamento podem levar a classificadores diferentes e mudanças grandes na precisão da classificação. Essa instabilidade advém da seleção aleatória das unidades para serem treinadas e é uma característica de alguns algoritmos (KUHN; JOHNSON, 2013). Adicionalmente, quando num algoritmo de classificação, a variável resposta, y, apresenta desbalanceamento em relação à proporção de eventos, é necessário que no momento do particionamento o pesquisador leve esse aspecto em consideração.

Figura 26 - Validação cruzada de 10 partes (k fold).



Fonte: SANTOS, 2018.

Legenda: Os dados são divididos em 10 partes. O modelo é treinado em 9 partes e testado na 10ª parte. Este processo é repetido *n* vezes. O resultado final é a média dos erros de cada iteração.

RESULTADOS

Os resultados da tese estão apresentados sob a forma de três artigos científicos apresentados abaixo.

ARTIGO 1

[Publicado na *International Journal of Public Health*]

TITLE: Assessing the impact of a doctor in remote areas of Brazil.

Santos, J[1] [2], Santos, H[1] , Matias-Dias, C [2] , Filho, AC [1]

[1] Faculty of Public Health - University of São Paulo, São Paulo, Brazil

[2] Department of Epidemiology, Institute of Public Health Dr Ricardo Jorge, Lisbon, Portugal

ABSTRACT

Objective: The More Doctors Program (MDP) is an ongoing Brazilian policy that aims to improve healthcare by providing physicians to the most vulnerable municipalities. We aimed to measure the impact of MDP in mortality and infant mortality rate, the proportion of livebirths with low weight, prenatal appointments, childbirths at 1st and 5th minute Apgar, public health investment and immunization in Brazil. Methods: Municipal health indicators were collected before and after the intervention (2012 and 2015). Effects were measured by applying propensity score matching (PSM) with difference-in-differences (DiD). Results: Our findings show that infant mortality presented the highest improvement during the period (a decrease of 11 infant deaths per 1,000 live births, $p<0.01$). A significant effect, albeit smaller, was also found for the age-standardized total mortality (a decrease of 5 deaths per 10,000 residents), proportion of children with Apgar score lower

than 8 in the 5th minute and children with low birth weight. Conclusion: MDP contributed to improve important health indicators, highlighting the importance of a doctor in remote areas of Brazil.

INTRODUCTION

Brazil has a historical shortage and uneven distribution of physicians. In 2013, the country had 2 physicians per 1,000 inhabitants (Scheffer 2015) but the Southeast presented over twice more physicians per capita than the North and Northeast (2.7, 1.09 and 1.3, respectively). To address this problem, in 2013 the Brazilian government, with close support from the Pan American Health Organization (WHO/PAHO) (PAHO/WHO 2018), introduced the More Doctors Program (MDP).

The main objective of the program was to reduce inequalities in access to primary healthcare in Brazil. The MDP entailed a total of 8 main objectives, such as increasing the number of physicians in rural and remote areas, improving medical training and investing in primary care infrastructure (Presidência da Republica 2013). To participate in the MDP, municipalities must be considered eligible according to local necessities, and areas within a pre-defined poverty threshold were considered priority (CONASS 2013) (Presidência da Republica 2013).

According to the Brazilian Ministry of Health, in 2015 more than 4.000 municipalities received doctors from the MDP (DATASUS 2018). The provision of doctors to the most remote areas was the main priority of the Program and gained public visibility due to the hiring of foreign physicians, mostly of Cuban origin, to work in remote areas. Until 2016, around 17.000 physicians had been allocated all over Brazil, with a balanced percentage of Cuban and Brazilian doctors (47% and 46%, respectively). (PAHO/WHO 2018)

The large costs associated with the program increased the pressure for independent measurements of its health impact (Ministerio Saude 2013). The impact of a program can be measured by determining if a potential change can

be specifically attributed to the program (Josselin and Le Maux 2017). Assessing impact requires the existence of a counterfactual (Peixoto 2012) that can frequently be approximated by randomizing the intervention, which was not the case for the MDP.

A conceptual framework was proposed by Rubin in 1974 to overcome this constraint (Rubin 1974). Rubin´s causal model (Holland 1986) (Shadish 2010) lies on the following assumptions 1) each individual i has a potential outcome Y_i^1 associated with participating in the treatment (Zi =1) and a potential outcome Y_i^0 if not participating (Zi=0) with the effect being the difference between the outcome with and without treatment, for the same individual; and 2) simulating randomization in observational studies is possible, under particular assumptions, through the assignment of a propensity score which is the conditional probability of being assigned a particular treatment given a vector of observed covariates. (Rosenbaum and Rubin 1983). As a result, one is able to measure the Average Treatment Effect on the Treated (ATT), i.e. the difference between the expected values of the potential outcomes of treated individuals (Leite 2017). An important assumption of this framework is ignorability, or “no hidden bias”, meaning that the assignment of the treatment can be assumed to be random based on all observable characteristics of the observations under study (Stuart 2010). Although with different objectives, several studies have been carried out under this conceptual framework (Filho et al. 2012) (Gebel and Voßemer 2014) (Jonk et al. 2015). All of these three studies employed a propensity score matching approach to measure the effect of a program in health indicators, when the study design was not experimental and random allocation was not possible. Our study used a similar technique to assess the effect of the MDP. Our analysis applied a similar technique to assess the effect of the MDP, followed by a difference-in-differences (DiD) approach, a statistical technique used to estimate effects of a policy before and after the intervention between treatment and control groups (Josselin and Le Maux 2017). In short, a major strength of combining these approaches is that, under these assumptions, it is possible to estimate the impact of a program on a population even without a randomized experiment. By

matching units that were similar before the program, the PSM reduces the treatment assignment bias and mimics randomization.

Before the Program, some municipalities in Brazil did not have a doctor. It is expected that one doctor contributes more to improve health outcomes in these areas than in places that already have a few doctors. The objective of this study is then to measure the impact of the More Doctors Program in 2015 in these municipalities for the following health outcomes: crude and age standardized mortality rate, infant mortality rate, percentage of livebirths with low weight, percentage of pregnancies with least 7 prenatal appointments, percentage of newborns with 1st minute Apgar less than 8 and percentage of newborns with 5th minute Apgar less than 8, per capita public health investment and immunization rate in municipalities of Brazil.

METHODS

We analyzed the municipalities of Brazil that did not have a physician in 2012, the year before the start of MDP (n=395). For this purpose, the ratio of physicians per capita was estimated, with data collected from DATASUS (DATASUS 2018). Covariates were collected from IBGE, TSE (Electoral Superior Court) and Ministry of Health (MOH) (Electronic supplement 1). Information on the provision of physicians for each municipality and cycle of allocation was provided by the Ministry of Health (Federal 2018). Rates and proportions for the outcomes of interest for each municipality were estimated: crude and age standardized mortality, infant mortality rate (IMR), the proportion of births with at least 7 antenatal visits, proportion of births with low weight (under 2.5kg), the proportion of births with Apgar score below eight at first (Apgar 1) and fifth minute (Apgar 5) and the rate of mortality due to undetermined causes. Variables were measured before the intervention to avoid endogeneity.

To build the propensity score we used a set of covariates (Electronic supplement 1) related to sociodemographic and health access characteristics of municipalities as long as they are conceptualized to be prior to the MDP.

Nonetheless, covariates with zero variance were removed (Garrido et al. 2014) but correlated variables were kept. We used these variables to identify comparable municipalities with and without the intervention (i.e. participating or not in the MDP) using propensity matching approach. A logistic regression was performed to estimate the propensity score. We then inspected the common support area visually and treated municipalities were matched to similar controls on the propensity score with a caliper of 0.1 standard deviation from the PS (with replacement) (Austin 2011a) (Austin 2011b).

We then analyzed the health outcomes from the two group of municipalities in two time points, in the year immediately before the program started in 2012 (when not available, data was collected from 2010, the year of the last census); and and then after the beginning of the program (2015 – the most recent data available). We applied a difference-in-differences analysis (Gebel and Voßemer 2014) which consists of a double subtraction between the outcomes before and after the intervention and between the intervention and control groups (Jonk et al. 2015) (Katchova 2013)

Covariate balance before and after the propensity matching approach was assessed by the number of covariates with standard mean difference less than 25%, as previously used in the literature (Elizabeth A. Stuart, Brian K. Lee, and Finbarr P. Leacy 2013) (Leite 2017) (Austin 2009). Assessment of PS balance was measured by taking into account whether the number of balanced covariates improved after matching and later checked with p-value of t-test to verify if groups after matching were different from each other. The package *matching* for R was used for the analyzes. (Sekhon 2018)

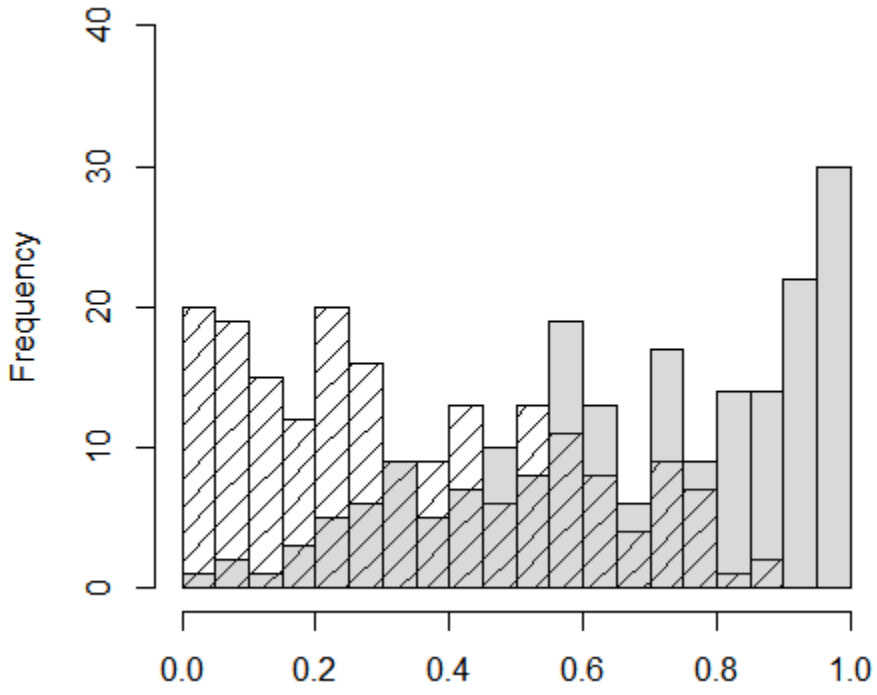
RESULTS

Our sample was composed of 395 municipalities that did not have a physician in 2012. From these, 194 did not receive a physician and 201 did until May 2014 (corresponding to the 5th cycle of the program). Of the municipalities without a

physician, 31% were in the south, 27% in the northeast, 26% in the southeast, 10% in the north and 7% in the center-west region.

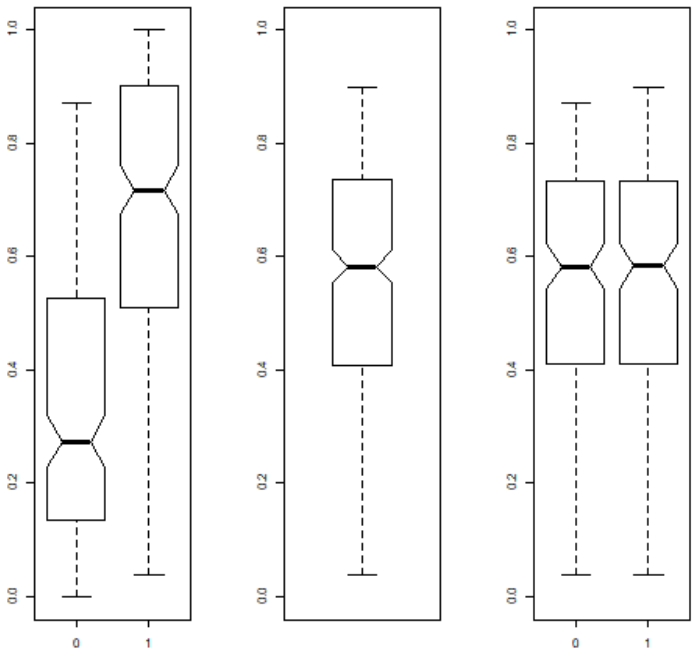
The distribution of the results of the propensity scores for the two groups (with and without MDP) is presented in Figure 1. It is possible to identify an interesting overlap of treated and untreated municipalities, which allows for pairwise comparisons. Figure 2 presents the distribution of propensity score before and after matching, for all municipalities and according to having, or not, received the MDP. The complete numerical distribution of the covariates for the two groups before and after matching, as well as their standardized means difference (in %), is presented in Electronic Supplement 2. We obtained 163 pairs within this caliper width and municipalities that were not within this exact caliper width of distance from another municipality with opposite treatment were dropped from further analyzes.

Figure 1 Propensity score values distribution for treated and untreated Brazilian municipalities, 2012.



Legend: Shaded bars correspond to municipalities without the More Doctors Program and grey to municipalities with the More Doctors Program.

Figure 2 Boxplot of propensity scores distribution before and after matching for Brazilian municipalities, 2012.



Legend: Boxplot in the left corresponds to propensity scores before matching and according to municipalities that received the More Doctors Program (1) or not the More Doctors Program (0); boxplot in the middle corresponds to propensity scores after matching in all municipalities and the boxplot in the right corresponds to propensity scores distribution according to municipalities that received (1) or not (0) the Program, after matching.

Electronic Supplement 2 shows that before matching, out of 97 covariates, 86 presented a SMD of less than 25%, and 75 had a significant t-test p value. After matching, 94 presented a SMD of less than 25%, and 92 covariates increased their t-test p-value, indicating better balance between treated and untreated.

Table 1 presents the effect of the program on health outcomes after differences-in-differences. There was a significant effect in the age-standardized mortality rate, with a decrease of 5 deaths per 10,000 inhabitants between 2012 and 2015 (p=0.01), a decrease in the IMR (p<0.01) of 11 deaths per 1000 live births, a decrease in the proportion of newborns with low birth weight and with Apgar at 5th minute lower than 8 (p<0.01) and in the proportion of children with low birth weight (p=0.01). However for the last two outcomes, the decrease was very small.

Table 1. Estimated effects of the More Doctors Program for health outcomes after difference-in-differences propensity score matching, Brazil, 2012-2015.

Health outcome	Effect estimates		
	Effect estimates after centering and scaling	**Effect estimates in absolute values**	**P value (t-test)**
Crude mortality rate	-0.119	-0.255	0.344
Age-standardized mortality rate	-0.284	-0.511	0.01*
Infant mortality rate	-0.399	-11.871	0.001*
Percentage of livebirths with low weight	-0.307	-0.023	0.01*
Percentage of livebirths with at least7 prenatal appointments	-0.009	-0.001	0.944
Public health investment per capita	0.056	0.284	0.626
Percentage of childbirths with 1st minute Apgar less than 8	0.036	0.004	0.800
Percentage of childbirths with 5th minute Apgar less than 8	-0.307	-0.012	0.01*
Immunization rate	0.170	7.553	0.260
Mortality rate by undetermined causes	0.046	76.541	0.744

*significant p-value

DISCUSSION

We found that the MDP improved some health indicators in municipalities that joined the program when compared to those that did not after 2 years of participation. Propensity score matching was successful in improving the balance for most of the covariates. After matching, 97.9% were considered to be well balanced between the two groups (intervention and control), meaning that the intervention and control municipalities, matched to measure the impact of the Program, were very similar at baseline.

Our study found that the highest effect of the MDP was in decreasing the infant mortality rate in 11 deaths per 1,000 livebirths. A previous study examined the effects of the MDP in infant health and found an increase in the number of prenatal visits (Carrilllo and Feres 2017). However, no significant effect was found for infant mortality between treated and untreated areas. This study differs from ours due to the fact that our sample is comprised of 395 municipalities purposively selected to include only municipalities that before treatment had no physician. We used this approach because the introduction of one physician can have a different effect for municipalities that did not previously have a physician, in relation to those that previously already had a few. We also included a wider range of control variables, such as hospital and ambulatory care access.

There is some previous evidence for the association between the presence of a physician density and IMR (Farahani et al. 2009) (Anand and Bärnighausen 2004). One study, from Shi et al (SHI et al., 2004) found the supply of primary care physicians, especially of family doctors, was significantly associated with lower infant mortality and reduced rates of low birth weight. Another study, performed from 2005-2012 in municipalities in Brazil also found a negative relationship between density of primary care physicians and infant mortality (RUSSO et al., 2019).

It is well established that physicians play an important role in disease prevention especially for maternal and child health (Cutler et al. 2011), which are easier and less expensive to prevent. Although our analyzes were not able to assess the pathway through which the IMR drop occurred, a recent study (Liebert and Mäder 2016) found that physicians can prevent diseases by providing essential information to the community. In particular, a doctor can influence neonatal and post natal deaths by providing information to mothers regarding sanitary practices that prevent infants´ deaths, such as poor nutrition, smoking, low birth weight, infectious and sexually transmitted diseases. (SHI et al., 2004) In the case of our study, these are likely to have an immediate and direct effect in the evolution of pregnancy and the health of the newborn.

We also found a significant improvement in other maternal and child health indicators, namely a decrease in the proportion of children with Apgar less than 8 on the 5th minute and the proportion of newborns with low birth weight, though the magnitude of these effects were very small (0.01 and 0.02 respectively). Another study from Bladimir and Feres (Carrilllo and Feres 2017) found that although the Program increased the number of physicians in treated areas and the number of prenatal care visits by 10%, the effects in low birth weight had coefficients of very small magnitude, finding a significant improvement only in areas with high social spending in education.

Regarding the decrease in age-standardized mortality rate, previous studies using different methodologies also found a significant decrease in the mortality rate of municipalities that joined the MDP (Bastos et al. 2015). Bastos et al. found a decrease of 0.538 deaths per 1,000 inhabitants, and Santos (Fernanda dos Santos 2018) a decrease of 0.235 for the group of municipalities among the 20% poverty risk.

Other outcomes such as the proportion of pregnant women with at least seven antenatal visits, public health investment in health in the municipality, proportion of newborns with Apgar lower than 8 in the 1st minute, immunization coverage and mortality rate due to undetermined causes were not statistically significant.

This is line with both studies mentioned before, namely for pregnant women with antenatal care visits (Fernanda dos Santos 2018) and immunization (Carrilllo and Feres 2017)

This study has a few limitations. Data for the allocation of doctors was given by the MoH and the original data presented some duplicates regarding the allocation of doctors according to cycles. This information was necessary to identify municipalities with and without the MDP, and although we performed the necessary corrections there might still be residual bias. Another limitation is the fact that 2015 is very close to the start date of the program so effects may still be very small at this point. Finally, although we included a large number of covariates to assess the propensity score (a total of 97) there may be still some remaining bias regarding unobserved variables.

Our study found that the MDP contributed to improve important health indicators in Brazilian municipalities, especially infant mortality. Other indicators such as age-standardized mortality and the percentage of children with low birth weight also improved, but to a lesser extent. These results highlight the importance of a doctor in bringing health to communities in remote areas of Brazil.

REFERENCES

Anand S, Bärnighausen T (2004) Human resources and health outcomes : cross-country. Lancet 365:1603–1609. doi: 10.1016/S0140-6736(04)17313-3

Austin P (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 3385–3397. doi: 10.1002/sim

Austin PC (2011a) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 46:399–424. doi: 10.1080/00273171.2011.568786

Austin PC (2011b) Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat 10:150–161. doi: 10.1002/pst.433

Bastos SQ, Rodrigues LC, Morais B, et al (2015) Impacto do Programa Mais Médicos nas Taxas de Mortalidade. CBFM, XXIII, Porto Alegre, RS

Carrilllo B, Feres J (2017) More Doctors, Better Health? Evidence from a Physician Distribution Policy

CONASS (2013) Programa Mais Médicos-Nota Tecnica 23/2013

Cutler D, Deaton A, Lleras-muney A (2011) The Determinants of Mortality. 20:97–120

DATASUS (2018) DATASUS. http://datasus.saude.gov.br/. Accessed 7 Jul 2017

Elizabeth A. Stuart, Brian K. Lee, and Finbarr P. Leacy S (2013) Prognostic score–based balance measures for propensity score methods in comparative effectiveness research. J Clin Epidemiol 66:1–14. doi: 10.1016/j.jclinepi.2013.01.013.Prognostic

Farahani M, Subramanian S V, Canning D, Aging P on the GD of (2009) Short and long-term relationship between physician density on infant mortality: a longitudinal econometric analysis. PGDA Work Pap No 49

Federal G (2018) Mais Medicos. In: Gov. Fed. http://maismedicos.gov.br/. Accessed 7 Jul 2017

Fernanda dos Santos (2018) Programa Mais Médicos Uma avaliaçao de impacto sobre indicadores no brasil. UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

Filho ADPC, Kawachi I, Gotlieb SLD (2012) Propensity score matching approach to test the association of income inequality and mortality in São Paulo, Brazil. J Epidemiol Community Health 66:14–17. doi: 10.1136/jech.2010.108852

Garrido MM, Kelley AS, Paris J, et al (2014) Methods for constructing and assessing propensity scores. Health Serv Res 49:1701–1720. doi: 10.1111/1475-6773.12182

Gebel M, Voßemer J (2014) The impact of employment transitions on health in Germany. Adifference-in-differences propensity score matching approach. Soc Sci Med 108:128–136. doi: 10.1016/j.socscimed.2014.02.039

Holland PW (1986) Statistics and Causal Inference: Rejoinder. J Am Stat Assoc 81:968. doi: 10.2307/2289069

Jonk Y, Lawson K, O'Connor H, et al (2015) How effective is health coaching in reducing health services expenditures? Med Care 53:133–140. doi: 10.1097/MLR.0000000000000287

Josselin J-M, Le Maux B (2017) Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education. In: Springer (ed) Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education., 1st 2017. Springer International Publishing, pp 489–531

Katchova, A Propensity Score Matching. In: Econom. Acad. https://sites.google.com/site/econometricsacademy/econometrics-models/propensity-score-matching. Accessed 24 Jan 2017

Leite W (2017) Practical propensity score methods using R. SAGE Publications, USA

Liebert H, Mäder B (2016) Marginal effects of physician coverage on infant and disease mortality. York

Ministerio Saude (2013) TC 80: Cooperação Técnica entre o Ministério da Saúde e a Organização Pan-Americana da Saúde - Projeto Acesso da população Brasileira à Atenção Básica em Saúde

PAHO/WHO (2018) Ministério da Saúde reforça cooperação com a OPAS para continuidade do Mais Médicos. http://portalms.saude.gov.br/noticias/agencia-saude/42756-ministerio-da-saude-reforca-cooperacao-com-a-opas-para-continuidade-do-mais-medicos

Peixoto B et al (2012) Avaliação Economica de projetos sociais. Fundaçao Itaú

Presidência da Republica (2013) LEI Nº 12.871, DE 22 DE OUTUBRO DE 2013.

Rosenbaum PR, Rubin DB (1983) Biometrika Trust The Central Role of the Propensity Score in Observational Studies for Causal Effects The central role of the propensity score in observational studies for causal effects. Source: Biometrika Biometrika 70:41–55. doi: 10.2307/2335942

Rubin DB (1974) Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. J Educ Psychol 66:688–701

Russo LX, Scott A, Sivey P, Dias J (2019) Primary care physicians and infant mortality: Evidence from Brazil. PLoS One 14:1-16. Doi:10.1371/journal.pone.0217614

Scheffer M et al (2015) Demografia médica no Brasil 2015. Departamento de Medicina Preventiva, Faculdade de Medicina da USP. Conselho Regional de Medicina do Estado de São Paulo. Conselho Federal de Medicina., São Paulo

Sekhon JS (2018) Multivariate and Propensity Score Matching Software for Causal Inference. http://sekhon.berkeley.edu/matching. Accessed 17 Aug 2018

Shadish WR (2010) Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings. Psychol Methods 15:3–17. doi: 10.1037/a0015916

Shi L, Macinko J, Starfield B, et al (2004) Primary care, infant mortality, and low birth weight in the states of the USA. J Epidemiol Community Health 58:374–380. doi: 10.1136/jech.2003.013078

Stuart EA (2010) Matching methods for causal inference: A review and a look forward. Stat Sci 25:1–21. doi: 10.1214/09-STS313

ES 1 List of covariates used to build the propensity score, source of collection and year, for municipalities.

1. Socio Demographic	**SOURCE**	**Year the data refers to.**
Number of residents	IBGE[1]	2010
Life expectancy	IBGE	2010
Older individuals (%)	IBGE	2010
Dependants (%)	IBGE	2010
Women (%)	IBGE	2010
Livebirths per capita 2010	IBGE	2010
Married people (%)	IBGE	2010
Evangelicals (%)	IBGE	2010
Disabled (%)	IBGE	2010
Populational density	IBGE	2010
Households in urban areas (%)	IBGE	2010
Households with refrigerator (%)	IBGE	2010
Households with computer (%)	IBGE	2010
Households with car (%)	IBGE	2010
Household density (%)	IBGE	2010
Household with electricity (%)	IBGE	2010
Households with urban afforestation (%)	IBGE	2010
Households in urban areas, with paved roads (%)	IBGE	2010
White individuals (%)	IBGE	2010
Literacy rate (%)	IBGE	2010
Higher education completed (%)	IBGE	2010
Secondary school completed (%)	IBGE	2010
Individuals living in the municipality for less than 10 years (%)	IBGE	2010
Average income	IBGE	2010
Unemployment rate	IBGE	2010
Child labour (%)	IBGE	2010
Retirees (%)	IBGE	2010
Individuals working more than 45 hour per week (%)	IBGE	2010
Low income children (%)	IBGE	2010
GDP per capita	IBGE	2010
Gini index	IBGE	2010
Coverage of families with public health care (%)	DATASUS[2]	2012

Individuals working in a different municipality (%)	IBGE	2010
Hospital beds per capita	DATASUS	2012
Political party of the mayor belonging to the federal government coalition in 2010	TSE[3]	2010
Region (Northeast, North, Southeast, Center-West and South)	IBGE	2010
2.Health Access [a]		
Hospital admission rate by material collection (puncture/biopsy)	DATASUS	2012
Hospital admission rate by emergency appointment	DATASUS	2012
Hospital admission rate by treatment of infectious diseases and parasites	DATASUS	2012
Hospital admission rate by treatment of blood and other immunity system diseases	DATASUS	2012
Hospital admission rate by treatment of metabolic and nutritional diseases	DATASUS	2012
Hospital admission rate by treatment of nervous system related diseases	DATASUS	2012
Hospital admission rate by treatment of cardiovascular diseases	DATASUS	2012
Hospital admission rate by treatment of digestive system related diseases	DATASUS	2012
Hospital admission rate by treatment of skin related diseases	DATASUS	2012
Hospital admission rate by treatment of musculoskeletal system related diseases	DATASUS	2012
Hospital admission rate by treatment of pregnancy, childbirth and puerperium related diseases	DATASUS	2012
Hospital admission rate by treatment of patients under long term care	DATASUS	2012
Hospital admission rate by treatment of ear and respiratory tract diseases	DATASUS	2012
Hospital admission rate by treatment of the genitourinary system diseases	DATASUS	2012
Hospital admission rate by treatment of affections occurred during neonatal period	DATASUS	2012
Hospital admission rate by treatment of mental diseases	DATASUS	2012
Hospital admission rate by procedure, chemotherapy	DATASUS	2012
Hospital admission rate by oncology (general)	DATASUS	2012
Hospital admission rate by nephrology (general)	DATASUS	2012
Hospital admission rate by trauma and injuries	DATASUS	2012
Hospital admission rate by intoxications and poisoning	DATASUS	2012
Hospital admission rate by complications due to health procedures	DATASUS	2012
Hospital admission rate by labor and childbirth	DATASUS	2012
Hospital admission rate by type of admissions and procedures - Total	DATASUS	2012
Outpatient care admissions rate by type of attendance, appointment	DATASUS	2012
Outpatient care admissions rate by type of attendance, emergency	DATASUS	2012
Outpatient care admissions rate by type of attendance: not informed	DATASUS	2012
Outpatient care admissions rate by type of attendance, total	DATASUS	2012
Outpatient care admissions rate by procedure, diagnosis	DATASUS	2012

Outpatient care admissions rate by procedure, clinical	DATASUS	2012
Outpatient care admissions rate by procedure, surgery	DATASUS	2012
Outpatient care admissions rate by procedure, organ transplantation	DATASUS	2012
Outpatient care admissions rate by procedure, for medication dispensary	DATASUS	2012
Outpatient care admissions rate by procedure, orthoses, prostheses and related materials	DATASUS	2012
Hospital admissions rate by procedure, clinical	DATASUS	2012
Hospital admissions rate by procedure, surgery	DATASUS	2012
Hospital morbidity rate according to ICD 10, Chapter I	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter II	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter III	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter IV	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter V	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter VI	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter VII	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter VIII	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter IX	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter X	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XI	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XII	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XIII	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XIV	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XV	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XVI	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XVII	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XVIII	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XIX	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XX	DATASUS	2012
Hospital morbidity rate according to IDC 10 causes, Chapter XXI	DATASUS	2012

[1]IBGE:Brazilian Institute of Statistics and Demographics

[2]DATASUS: Department of Informatics of the Health System (DATASUS)

[3]TSE: Electoral Supreme Court

[a] Rates in the article were estimated by the author with data collected from the sources list.

ES 2 Table with the balance of covariates, before and after propensity score matching with means, standardized mean difference (SMD) in % and p values for T-test.

.

Covariate		Treated (mean)	Control (mean)	SMD[1] (%)	P value of T test[2]
Number of residents	**Before**	0.256	-0.265	43.055	0.000
	After	-0.024	-0.007	-2.287	0.814
Life expectancy	**Before**	-0.121	0.125	24.529	0.014
	After	-0.055	-0.090	3.493	0.768
Older individuals (%)	**Before**	-0.115	0.119	22.695	0.020
	After	0.025	0.126	9.752	0.355
Dependendants (%)	**Before**	0.112	-0.116	21.619	0.023
	After	0.069	0.053	1.711	0.880
Women (%)	**Before**	0.068	-0.071	19.480	0.171
	After	0.077	0.167	12.195	0.245
Livebirths per capita 2010	**Before**	0.102	-0.106	19.741	0.038
	After	0.021	-0.122	14.295	0.187
Married people (%)	**Before**	-0.089	0.092	17.498	0.072
	After	-0.006	0.084	8.856	0.430
Evangelic (%)	**Before**	0.002	-0.002	0.387	0.972
	After	-0.024	-0.047	2.498	0.823
Disabled (%)	**Before**	-0.097	0.101	20.450	0.049
	After	0.014	0.054	4.052	0.723
Populational density	**Before**	-0.050	0.052	12.132	0.316
	After	0.004	-0.005	0.980	0.939
Urbanization rate (%)	**Before**	-0.106	0.110	22.587	0.031
	After	-0.011	0.051	6.637	0.586
Households with refrigerator (%)	**Before**	-0.161	0.167	28.516	0.001
	After	-0.060	-0.065	0.490	0.964
Households with computer (%)	**Before**	-0.048	0.050	9.375	0.333
	After	-0.039	-0.005	3.252	0.753
Households with car (%)	**Before**	-0.062	-0.062	12.526	0.212
	After	-0.048	-0.001	1.080	0.922
Household density (%)	**Before**	0.123	-0.128	22.158	0.012
	After	0.019	-0.055	7.563	0.496
Household with electricity (%)	**Before**	-0.132	0.137	22.995	0.007

	After	-0.022	0.040	6.514	0.565
Arborization	**Before**	-0.111	0.116	23.308	0.024
	After	-0.002	0.024	2.753	0.827
Paved area	**Before**	-0.189	0.196	41.244	0.000
	After	-0.064	-0.073	0.996	0.933
White individuals (%)	**Before**	-0.048	0.050	9.654	0.332
	After	-0.011	-0.140	13.086	0.240
Literacy rate (%)	**Before**	-0.052	0.054	10.255	0.290
	After	-0.058	-0.079	2.025	0.847
Higher education completed (%)	**Before**	-0.050	0.052	9.450	0.309
	After	0.001	0.126	11.206	0.248
Secondary school completed (%)	**Before**	-0.091	0.095	19.032	0.065
	After	-0.087	-0.139	5.047	0.659
Individuals living in the municipality	**Before**	-0.032	0.033	6.486	0.522
for less than 10 years (%)	**After**	0.008	-0.075	8.618	0.412
Average income	**Before**	-0.062	0.064	12.162	0.209
	After	-0.055	-0.015	3.806	0.704
Unemployment rate	**Before**	-0.005	0.005	1.064	0.917
	After	-0.037	-0.010	2.914	0.815
Child labour (%)	**Before**	0.041	-0.042	8.571	0.412
	After	0.045	-0.028	7.721	0.522
Retirees (%)	**Before**	-0.087	0.090	17.574	0.078
	After	0.033	0.172	13.823	0.205
Individuals working + 45 hour per week (%)	**Before**	-0.074	0.076	14.768	0.137
	After	-0.006	0.216	21.198	0.057
Low income children (%)	**Before**	0.080	-0.083	16.324	0.107
	After	0.064	0.054	1.040	0.926
GDP_per_capita	**Before**	-0.007	0.008	1.228	0.882
	After	-0.043	-0.063	1.658	0.858
Gini Index	**Before**	0.147	-0.152	30.824	0.003
	After	0.046	0.018	3.057	0.799
Coverage of families with public health care (%)	**Before**	-0.091	0.094	17.042	0.065
	After	-0.069	0.112	16.236	0.122
Individuals working in a different municipality (%)	**Before**	-0.134	0.139	30.611	0.007
	After	-0.069	-0.101	3.394	0.770
Hospital beds per capita (2012)	**Before**	0.002	-0.002	0.528	0.963
	After	-0.038	0.006	5.255	0.659
Hospital admission rate by collection of material (puncture/biopsy) (2012)	**Before**	0.071	-0.073	13.957	0.153

	After	0.001	-0.153	16.551	0.098
Hospital admission rate by emergency appointments (2012)	**Before**	-0.067	0.070	19.511	0.177
	After	-0.037	-0.139	12.954	0.162
Hospital admission rate by treatment of infectious diseases and parasites (2012)	**Before**	0.064	-0.066	11.686	0.196
	After	-0.015	-0.047	3.561	0.758
Hospital admission rate by treatment of blood and other immunity system diseases (2012)	**Before**	-0.042	0.043	9.619	0.401
	After	0.013	-0.090	10.624	0.347
Hospital admission rate by treatment of metabolic and nutritional diseases (2012)	**Before**	0.015	-0.015	2.764	0.769
	After	-0.035	0.076	12.487	0.295
Hospital admission rate by treatment of nervous system related diseases (2012)	**Before**	0.019	-0.019	3.682	0.704
	After	-0.042	0.061	11.555	0.336
Hospital admission rate by treatment of cardiovascular diseases (2012)	**Before**	-0.047	0.049	10.415	0.340
	After	-0.058	-0.062	0.479	0.965
Hospital admission rate by treatment of digestive system related diseases (2012)	**Before**	-0.013	0.014	2.680	0.792
	After	0.012	0.409	39.781	0.006
Hospital admission rate by treatment of skin related diseases (2012)	**Before**	0.045	-0.046	8.007	0.362
	After	0.042	-0.008	4.369	0.654
Hospital admission rate by treatment of musculoskeletal system related diseases (2012)	**Before**	0.011	-0.012	2.082	0.821
	After	-0.039	0.063	10.552	0.372
Hospital admission rate by treatment of pregnancy, childbirth and puerperium related diseases (2012)	**Before**	-0.039	0.041	8.301	0.430
	After	0.024	-0.051	7.210	0.486
Hospital admission rate by treatment of patients under long term care (2012)	**Before**	-0.030	0.031	5.457	0.541
	After	0.016	0.053	2.908	0.768
Hospital admission rate by treatment of ear and respiratory tract diseases (2012)	**Before**	0.062	-0.064	11.411	0.211
	After	-0.029	-0.020	0.916	0.926
Hospital admission rate by treatment of the genitourinary system diseases (2012)	**Before**	-0.058	0.060	14.421	0.241
	After	-0.065	-0.069	0.474	0.965

Hospital admission rate by treatment of affections occurred during neonatal period (2012)	**Before**	-0.017	0.018	3.391	0.727
	After	-0.062	-0.116	5.108	0.635
Hospital admission rate by treatment of mental diseases (2012)	**Before**	-0.038	0.039	9.009	0.451
	After	-0.056	-0.149	11.838	0.336
Hospital admission rate by procedure, chemotherapy (2012)	**Before**	-0.021	0.022	5.313	0.670
	After	-0.036	0.030	8.514	0.549
Hospital admission rate by oncology (general) (2012)	**Before**	-0.031	0.032	6.639	0.536
	After	0.014	0.079	6.772	0.562
Hospital admission rate by nephrology (general) (2012)	**Before**	-0.063	0.065	13.237	0.203
	After	-0.022	-0.002	1.873	0.842
Hospital admission rate by trauma and injuries (2012)	**Before**	0.045	-0.047	8.541	0.361
	After	0.011	-0.098	9.844	0.314
Hospital admission rate by intoxications and poisoning (2012)	**Before**	0.039	-0.040	7.224	0.434
	After	-0.030	0.004	3.256	0.701
Hospital admission rate by complications due to health procedures (2012)	**Before**	-0.076	0.079	22.506	0.126
	After	-0.057	-0.053	0.429	0.972
Hospital admission rate by labor and childbirth (2012)	**Before**	0.028	-0.029	5.704	0.568
	After	-0.025	0.058	8.449	0.484
Hospital admission rate by type of admissions and procedures - Total (2012)	**Before**	-0.052	0.054	10.592	0.295
	After	-0.075	-0.038	4.004	0.702
Outpatient care admissions rate by type of attendance: appointment (2012)	**Before**	-0.104	0.108	25.405	0.036
	After	-0.080	-0.001	9.046	0.433
Outpatient care admissions rate by type of attendance: urgency (2012)	**Before**	0.000	0.000	0.048	0.995
	After	0.068	-0.058	9.246	0.280
Outpatient care admissions rate by type of attendance: not informed (2012)	**Before**	0.101	-0.104	17.452	0.040
	After	-0.025	0.004	3.245	0.777
Outpatient care admissions rate by procedure, diagnosis (2012)	**Before**	-0.142	0.147	31.519	0.004
	After	-0.096	0.017	11.676	0.325
Outpatient care admissions rate by procedure, clinical (2012)	**Before**	-0.086	0.089	20.439	0.082

	After	-0.036	-0.096	6.571	0.512
Outpatient care admissions rate by procedure, surgery (2012)	**Before**	-0.067	0.070	14.659	0.176
	After	-0.017	0.138	15.262	0.237
Outpatient care admissions rate by procedure, organ transplantation (2012)	**Before**	0.028	-0.029	4.325	0.573
	After	-0.035	0.073	14.755	0.185
Outpatient care admissions rate by procedure, for medication dispensary (2012)	**Before**	-0.078	0.081	17.819	0.117
	After	-0.066	0.010	8.639	0.463
Outpatient care admissions rate by procedure, orthoses, prostheses and related materials (2012)	**Before**	0.039	-0.041	8.039	0.426
	After	0.033	0.142	10.716	0.426
Outpatient care admissions rate by procedure, primary care complementary actions (2012)	**Before**	-0.002	0.002	0.364	0.971
	After	-0.042	-0.081	3.558	0.693
Hospital admissions rate by procedure, clinical (2012)	**Before**	-0.060	0.062	14.433	0.230
	After	-0.109	-0.280	21.386	0.033
Hospital admissions rate by procedure, surgery (2012)	**Before**	-0.007	0.007	1.350	0.895
	After	-0.059	-0.026	3.686	0.726
Hospital morbidity rate according to ICD 10, Chapter I (2012)	**Before**	0.065	-0.068	11.872	0.184
	After	0.005	-0.028	3.477	0.756
Hospital morbidity rate according to ICD 10, Chapter II (2012)	**Before**	-0.110	0.114	25.448	0.026
	After	-0.069	0.029	11.656	0.336
Hospital morbidity rate according to ICD 10, Chapter III (2012)	**Before**	-0.036	0.037	8.204	0.469
	After	0.007	-0.091	10.159	0.372
Hospital morbidity rate according to ICD 10, Chapter IV (2012)	**Before**	0.002	-0.002	0.393	0.968
	After	-0.043	0.033	8.802	0.452
Hospital morbidity rate according to ICD 10, Chapter V (2012)	**Before**	-0.040	0.042	9.601	0.419
	After	-0.058	-0.146	11.218	0.364
Hospital morbidity rate according to ICD 10, Chapter VI (2012)	**Before**	-0.077	0.080	15.786	0.120
	After	-0.058	0.120	16.882	0.188
Hospital morbidity rate according to ICD 10, Chapter VII (2012)	**Before**	-0.065	0.068	28.983	0.193
	After	-0.071	-0.082	2.365	0.848
Hospital morbidity rate according to ICD 10, Chapter VIII (2012)	**Before**	-0.015	0.016	4.189	0.759

	After	-0.002	0.007	1.205	0.910
Hospital morbidity rate according to ICD 10, Chapter IX (2012)	**Before**	-0.067	0.070	14.823	0.175
	After	-0.067	-0.073	0.773	0.945
Hospital morbidity rate according to ICD 10, Chapter X (2012)	**Before**	0.051	-0.053	9.572	0.299
	After	-0.034	-0.045	1.120	0.909
Hospital morbidity rate according to ICD 10, Chapter XI (2012)	**Before**	-0.083	0.086	17.507	0.094
	After	-0.064	0.196	27.406	0.029
Hospital morbidity rate according to ICD 10, Chapter XII (2012)	**Before**	0.013	-0.014	2.287	0.789
	After	0.014	-0.142	13.328	0.137
Hospital morbidity rate according to ICD 10, Chapter XIII (2012)	**Before**	-0.078	0.081	17.678	0.115
	After	-0.091	-0.098	0.735	0.948
Hospital morbidity rate according to ICD 10, Chapter XIV (2012)	**Before**	-0.098	0.101	21.689	0.049
	After	-0.044	-0.009	3.604	0.720
Hospital morbidity rate according to ICD 10, Chapter XV (2012)	**Before**	0.007	-0.008	1.525	0.880
	After	-0.005	-0.026	2.073	0.857
Hospital morbidity rate according to ICD 10, Chapter XVI (2012)	**Before**	-0.011	0.011	2.114	0.828
	After	-0.040	-0.079	3.597	0.738
Hospital morbidity rate according to ICD 10, Chapter XVII (2012)	**Before**	-0.011	0.011	2.086	0.829
	After	-0.027	0.060	8.330	0.411
Hospital morbidity rate according to ICD 10, Chapter XVIII (2012)	**Before**	-0.090	0.093	50.366	0.073
	After	-0.070	-0.106	9.144	0.385
Hospital morbidity rate according to ICD 10, Chapter XIX (2012)	**Before**	-0.023	0.024	4.675	0.640
	After	-0.006	0.022	2.796	0.807
Hospital morbidity rate according to ICD 10, Chapter XXI (2012)	**Before**	-0.029	0.030	4.897	0.559
	After	0.033	0.074	3.019	0.748
Immunization dosages per capita (2012)	**Before**	-0.032	0.033	6.632	0.525
	After	-0.074	0.033	11.453	0.310
Proportional mortality by undetermined causes of death (2012)	**Before**	0.106	-0.110	21.558	0.032
	After	0.042	0.082	3.968	0.720
Political party of the mayor belonging to the federal government coalition in 2010	**Before**	0.502	0.495	1.525	0.880
	After	0.497	0.464	6.466	0.594
Region Northeast	**Before**	0.284	0.253	6.862	0.488

	After	0.289	0.280	17.715	0.120
Region North	**Before**	0.109	0.082	8.620	0.364
	After	0.094	0.032	21.014	0.034
Region Southeast	**Before**	0.184	0.330	37.532	0.001
	After	0.242	0.378	31.773	0.015
Region South	**Before**	0.363	0.263	20.804	0.032
	After	0.309	0.271	8.205	0.474

[1]The standardized means difference (SMD) in % is the mean difference as a percentage of average standard deviation: $100(\bar{x}_p - \bar{x}_t)/ sd_p$, where for each covariate $\bar{x}_p$ and $\bar{x}_t$ are the means before and after matching and sd_p corresponds to the standard deviation of treatment observations. [2]Before matching, the two sample t-test is used, after matching, the paired t-test is used.

ARTIGO 2

[Publicado na *Health Policy and Technology*]

TITLE: Machine learning and national health data to improve evidence: finding segmentation in individuals without private health insurance.

Joana Raquel Raposo dos Santos [1] [2] ¶ *,Carlos Matias Dias[2] [3] ¶, Alexandre Chiavegatto Filho[1] ¶

[1] Department of Epidemiology, Faculty of Public Health, University of São Paulo, São Paulo, Brazil

[2] Department of Epidemiology, National Health Institute, Lisbon, Portugal

[3] Center for Research in Public Health, National School of Public Health, NOVA University Lisbon, Lisbon, Portugal

[1]Current Address: Department of Epidemiology, National Health Institute, Lisbon, Portugal

[2] Current Address: Department of Epidemiology, Faculty of Public Health, University of São Paulo, São Paulo, Brazil

3 Current Address: Center for Research in Public Health, National School of Public Health, NOVA University of Lisbon, Lisbon, Portugal

*Corresponding author

joanasantos@usp.br (JS)

¶ These authors contributed equally to this work.

Abstract

Objective: Individuals without private health insurance have less access to healthcare, therefore are more prone to experience poor health when compared to those who have. Segmentation is an approach to find homogenous groups of people with the purpose of tailoring services and products. In public policy, segmentation might be used to identify characteristics and needs of specific

groups and deliver targeted programs and spare costs. We aim to identify and describe segments within the uninsured population to aid targeted policy actions and improve health.

Methods: We used secondary data collected from a representative, nationwide health survey (n=18,204). For the purpose of our analysis, we included data from individuals who answered “no” to the question: “Do you have private health insurance?” (n=12,134). Variables pertaining information on socio-demographic, health status, access and care were used. A multiple correspondence analysis was performed to find principal components followed by a hierarchical cluster.

Results: We found three clusters. The first (54.12% of our sample) composed by a group of young, middle aged and professionally active individuals without health problems. The second (36.70%), a cluster of aging individuals composed especially by elderly women, either retired or fulfilling domestic tasks, with a long-term health problem. The last (9.17%) composed by elder people, with long-term health problem and scoring low in mental health related questions.

Conclusion: Our study found three clusters (profiles of individuals) among the uninsured. Ultimately, our findings aim to support policy makers to deliver customized actions to improve health and provide cost-effective policies.

Introduction

Private health insurance is an important factor in terms of access to health care for most countries, which in turn have been shown to contribute to better health (VIACAVA; SOUZA-JÚNIOR; SZWARCWALD, 2005) (SCHNEIDER, 2004) (WILPER et al., 2009). Even in countries with a universal coverage public health system, owing a health insurance facilitates the access to healthcare, especially when the public health system is overwhelmed and is unable to respond on time to the health needs of the citizen. (EIRA, 2010)

However, owing an insurance requires either financial capacity or that the employer support that expense (THE HENRY J KAISER FAMILY FOUNDATION, 2016). Evidence shows it is very related to socioeconomic factors like the level of education, number of household assets and position within the work market (VIACAVA; SOUZA-JÚNIOR; SZWARCWALD, 2005), thus uninsured people are in a more fragile position regarding healthcare.

Nationwide information about this population can be found in databases with demographic, socio-economic and health related variables. In particular, national health surveys (NHS) provide not only representative samples but also large datasets of variables that are useful to characterize individuals in terms of their demographic, social, health perception and health status, access to health care, health care knowledge practices and health care use. Further, policies that specifically target groups can have an important role to public sector management, especially as universal approaches can sometimes be less effective and more expensive. (MCLAREN; PETIT, 2018)

Data mining techniques are extensively used by organizations to custom services and products (RYGIELSKI; WANG; YEN, 2002) (SHAW et al., 2001) (ARIMOND; ELFESSI, 2001) with the aim of improving customer relationship management. Data mining frequently uses machine-learning algorithms to discover patterns in data that can highlight information to improve knowledge about customers and support better decision processes. In the health sector, similar techniques have been used to classify patients with lung cancer, based on clinically measurable disease-specific variables (C.M.; V.H.; H.B., 2017). Another use of clustering in healthcare was to identify interactions of specific anti bodies and asthma (FONTANELLA et al., 2018) or to analyze patterns of behavioral risk factors (KHAN; UDDIN; ISLAM, 2019). These methods are successfully taking place across different areas, therefore it is interesting to apply them to health surveys´ databases to gain insight regarding populations´ health.

This study aimed to identify and describe homogeneous groups of individuals who do not have health private insurance in a National Health Survey to support policy-makers to tailor health policies.

Methods

The Portuguese National Health Survey (NHS) is a cross sectional study, conducted by the National Health Institute of Portugal and Statistics Portugal. The NHS is framed by the European Health Interview Survey (EHIS), which is a legal obligation for European Union (EU) member states (EUROPEAN COMMISSION, 2013). It aims to collect data for policy planning and monitoring, as well as providing data that allows comparisons between countries (EUROPEAN PARLIAMENT AND COUNCIL OF THE EU., 2009). The EHIS takes place every 5 years. So far, 2 waves have occurred: 2006-2009 and 2013-2015. The second wave of EHIS, corresponds to the NHS carried out in 2014 and is the survey under analysis. Survey participants (people aged 15 years and over) were selected by a probability multistage sampling stratified by region (NUTS II). This corresponded to a sample of 18, 204 individuals. The response rate was 80.8%. Further information about the questionnaire, sampling and procedures can be found in the methodological report (INE, 2018). The implementation of NHS is a legally binding commitment for EU member states, under the authority and coordination of Eurostat, regulated by the Commission regulation 141/2013, thus no Ethics Committee approval is needed. Also, under this legal framework, participants´ consent is not required. The data that support the findings of this study are available from the National Health Institute but restrictions apply to the availability of these data, which were used exclusively for the current study, and so are not publicly available. Data are available from National Health Institute for researchers who meet the criteria for access to confidential data.

Secondary data collected in the frame of the NHS was used to identify clusters of individuals without a private health insurance. This was possible because we only included individuals who answered “no” to the question “Do you have health

insurance?" (n=12,134). We removed individuals with missing answers because it overestimates the percentage of explained inertia in the process of multiple correspondence analysis (HUSSON, 2016) and this is recommended for cluster analysis (KASSAMBARA, 2018). All questions, variables and descriptions that compose the NHS questionnaire are in the S1 File. Out of 213 total variables, we removed those that did not add new information or presented duplicated information. We also removed variables with filters that contained "not applicable" answers, because these either relate to individuals with specific characteristics, such as people with severe limitation issues, or relate exclusively to women. We estimated the body mass index, grouped this variable into four categories and removed weight and height variables as well as variables that had only one level. We ended with 97 variables, corresponding to 356 categories. Variables with low frequencies, considered as under 2% of the sum for each variable were recoded. The fluxogram for the variables can be found in S2 File. Multiple correspondence analysis (MCA) was used to obtain a set of continuous and uncorrelated variables that retain as much information as possible (JOSSE, 2010). Then, agglomerative hierarchical clustering on the principal components was performed because combining MCA and hierarchical clustering techniques delivers clusters that are more stable (16). Further information regarding multiple correspondence analysis and hierarchical clustering can be found elsewhere (HUSSON, F. JOSSE, J, PAGÉS, 2008). The R packages *ca* and *factoMineR* was used for the analysis (HUSSON et al., 2018).

Multiple Correspondence Analysis

MCA is an exploratory multivariate technique that aims to analyze associations between multiple, categorical variables, but it is often used as a dimensionality reduction technique (FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS, 2005) (JOSSE; HUSSON; LÊ, 2008). This technique is similar to Principal Component Analysis (PCA), as it projects data points in a low Euclidean dimensional space to provide the best representation of the data. The dimensions aim to explain the highest amount of variability (inertia) in the data. More

information about this technique can be found elsewhere (FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS, 2005).

MCA requires preprocessing the data to achieve robust results. In particular, categories with low frequencies need to be regrouped to avoid having too much importance in the construction of dimensions. This is due to the fact that the contribution of a category to an axis is product of its relative frequency and its distance to the barycenter, or center of the axis (CRC; MURRELL,). Thus, variables with less than 2% of answers in a category were recoded (FRANCO, 2016).

To perform MCA, it is suggested that we choose supplementary variables (HJELLBREKKE, 2007). These do not contribute to the construction of dimensions. However, they are important because their position in the plot support the understanding of these variables in the space and their association with remaining variables. For example, categories and variables that are closer to each other are more similar and those that are further from the barycenter are more different. Then, we included the following variables as supplementary: type of household, number of residents and their ages within the household (8 variables corresponding to 29 categories). Remaining variables were set as active (89 variables and 327 active categories). After MCA, we used the scree plot (NGUYEN; HOLMES, 2019) to find the optimal number of dimensions.

Because MCA uses categorical variables, each dimension explains a low proportion of variance. As a result, a high number of dimensions must be chosen. However, a solution is to look at adjusted inertia instead, which gives improved percentages and rescales the multiple correspondence analysis results (NENADIC; GREENACRE, 2005). MCA is performed in a binary matrix coded with mutually exclusive 0 and 1. Thus, the matrix is always left with several columns filled with zeros. The inertia of the table becomes artificially inflated and the true percentage of explained variance underestimated. Greenacre (ABDI; VALENTIN, 2007) proposed to look at the inertia of the off diagonal blocks of a Burt Matrix, which has eigenvalues equal to the squared eigenvalues of the

analysis of the indicator matrix. Therefore, we looked at adjusted inertia when choosing the number of dimensions.

Hierarchical clustering on principal components

Clustering is an unsupervised method that aims to identify individuals with similar characteristics. It works by maximizing the inter-class and minimizing intra-class variances (FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS, 2005). Hence, the homogeneity of a cluster is characterized by the within inertia. Hierarchical clustering has the advantage that one does not need to define a number of clusters *a priori* (KASSAMBARA, 2018) when comparing to partioning clustering such as K-means. It works by gradually joining the most similar individuals and is represented by a dendrogram. We used the Euclidean distance to find similarity between units and Ward for partitioning (Análise Multivariada de Dados, 2005). Ward´s criterion minimizes the total within-cluster variance. To this end, at each step, it finds the pair of clusters that leads to minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centers. At the initial step, all individuals are clusters containing a single point.

When clustering on the axes from a MCA, it is possible to establish a hierarchy of clusters (HJELLBREKKE, 2007) because the principal component coordinates are used as quantitative variables (JOSSE, 2010). Groups are formed according to their homogeneity (Análise Multivariada de Dados, 2005) and the number of clusters chosen is based on the number of groups which best retains the inertia. In particular, we aim at a ratio between two successive within-cluster inertias that is as low as possible. The dendrogram is also a good indicative of how many clusters to keep. The height of the vertical lines represent the distance between points or clusters. The higher these lines, are the more different the clusters are.

Clusters are described in terms of variables and categories. To describe our clusters, we used the percentage of individuals in a certain category that are in a cluster (in the tables 2-4 referred as % cluster and the percentage of individuals in the cluster within that category (hereby defined as % sample)) (JOSSE, 2010). More detailed information regarding these methods is found elsewhere (FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS, 2005).

Results

Descriptive statistics

The descriptive analysis of the entire set of variables (97) is shown in S3 File. For illustrative purposes, Table 1 provides a brief characterization of our sample.

The sample is comprised by 42.5% men and 57.5% women. Regarding age groups, around 45% is over 60 years old and 11.5% is between 15 and 29 years old. Our sample has participants from all regions, yet, around 20% lives in Lisbon and Tagus Valley. In addition, almost half of our sample (46.11%) is composed by individuals who finished primary school (corresponding to 4 years of schooling) and 62% has 9 years of formal education. More than 1/3 of respondents has a formal job and 1/3 is retired. Almost 50% is between the first and second quartile of income and nearly 72% perceive their health as good or fair. The majority – 62% - does not have any long-standing health problem.

Table 1. Characterization of our sample (n=12, 348)

VARIABLE	DESCRIPTION	COUNT	%
REGION	North	1835	15.12
	Center	1540	12.69

	Lisbon and Tagus Valley	2520	20.77
	Alentejo	1053	8.68
	Algarve	2073	17.08
	Madeira	1489	12.27
	Azores	1624	13.38
SEX	Male	5157	42.50
	Female	6977	57.50
AGE GROUP	15-19	497	4.10
	20-24	470	3.87
	25-29	439	3.62
	30-34	612	5.04
	35-39	828	6.82
	40-44	875	7.21
	45-49	875	7.21
	50-54	986	8.13
	55-59	1003	8.27
	60-64	1096	9.03
	65-69	1185	9.77
	70-74	1031	8.50
	75-79	1020	8.41
	80-84	772	6.36
	85+	445	3.67
EDUCATIONAL LEVEL	None	1872	15.43
	Primary education	5595	46.11
	Lower secondary education	1937	15.96
	Upper secondary education	1526	12.58
	Tertiary education; bachelor level or equivalent	1204	9.92

SELF-DECLARED LABOUR STATUS	Carries out a job or profession, including unpaid work for a family business or holding, an apprenticeship or paid traineeship	4371	36.02
	Unemployed	1471	12.12
	Pupil, student, further training, unpaid work experience	651	5.37
	In retirement or early retirement or has given up business	4450	36.67
	Permanently disabled and Other inactive person	347	2.86
	Fulfilling domestic tasks	844	6.96
NET MONTHLY EQUIVALISED INCOME OF THE HOUSEHOLD	Below first quintile	3056	25.19
	Between 1st quintile and 2nd quintile	2795	23.03
	Between 3rd quintile and 4th quintile	2494	20.55
	Between 4th quintile and 5th quintile	2170	17.88
	Between 4th quintile and 5th quintile	1619	13.34
SELF PERCEIVED GENERAL HEALTH	Very good	1076	8.87
	Good	3688	30.39
	Fair	5096	42.00
	Bad	1721	14.18
	Very bad	553	4.56

LONG-STANDING HEALTH PROBLEM:	Yes	7919	65.26
SUFFER FROM ANY ILLNESS OR HEALTH PROBLEM OF A DURATION OF AT LEAST SIX MONTHS	No	4215	34.74

Multiple Correspondence Analysis

After performing the MCA, we used the screeplot on the adjusted inertia to find the optimal number of dimensions. This criteria allowed us to identify 5 dimensions, corresponding to an explained inertia percentage of 12.9% in the indicator matrix without adjustment. However, looking at adjusted inertia, this corresponds to 81.3% of explained data variability. (NGUYEN; HOLMES, 2019)

Supplementary variables

We plotted the map with the supplementary categories (S4 file), to identify the structure of our dataset. We observed a coherent pattern that groups households with children (HHTYPE=21, 23 and 24) and is located at the left top of the plot, meaning these are similar variables. At the bottom right, we found households without children (HHTYPE=22) in proximity to households of single individuals, indicating there are similarities. Further explanation on this analysis is in S4 file.

Active variable-categories

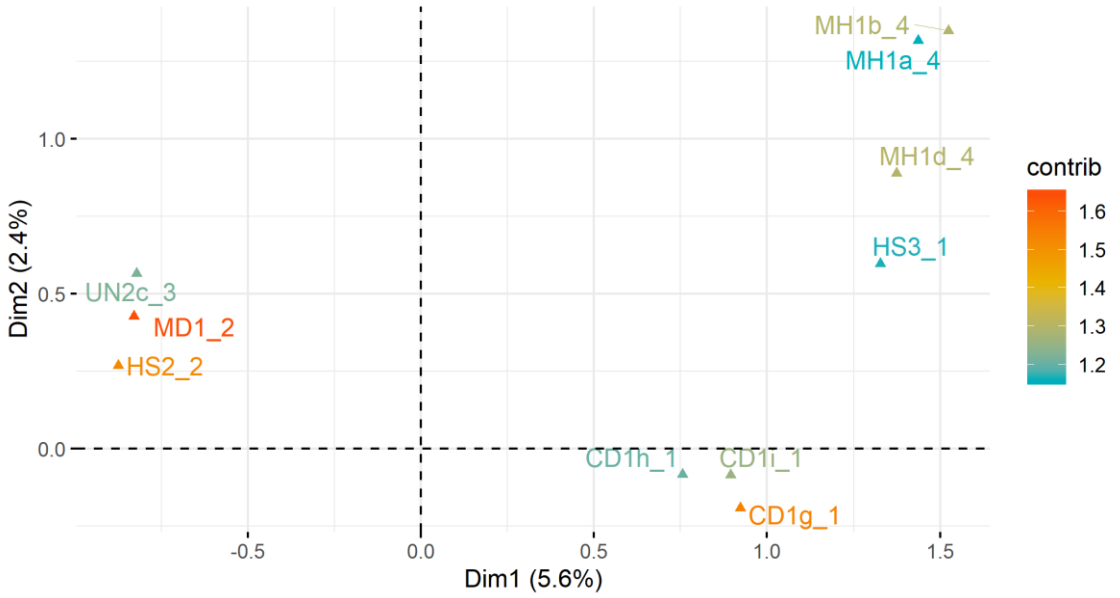
We present a table in supplementary material (S5 File) with contributions of all categories to each one of five dimensions. S5 File also shows the relative contribution of variables, through categories´ cumulative percentage. Cumulative contribution for each dimension varies from 0% to 12.5%. Variables that contribute the most to almost all dimensions are age group, education level and labour status.

For illustrative purposes we also plot categories with the highest contributions for the construction of the map (Fig. 1 - 3), but due to cluttering we present the top

10 categories (HJELLBREKKE, 2007). Dimension 1 to 5 explained 12.9% of total inertia, corresponding to 81.3% of adjusted inertia. Without adjustment it might seem a low percentage to explain inertia, but it is expected in MCA due to the fact that we have a high number of categories in our dataframe (ABDI; VALENTIN, 2007). Thus when selecting the number of dimensions it is more appropriate to consider adjusted inertia.

[Fig1 to Fig 3]

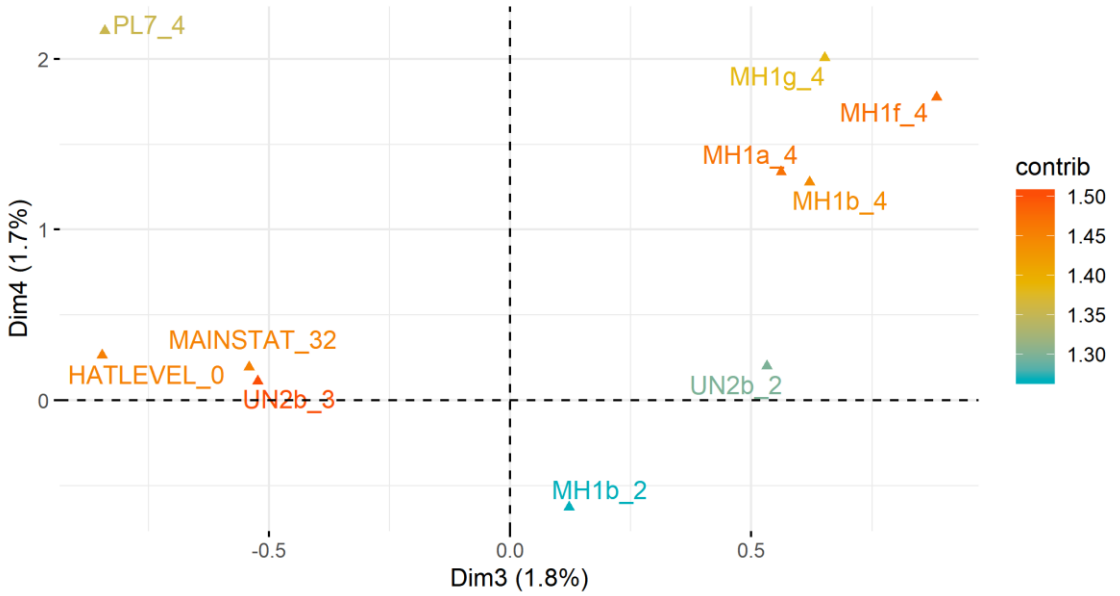
Figure 1. Map of the top highest contributed categories to dimensions 1 and 2. Contribution means the contribution of a point to the inertia of an axis.



Legend: MD1_2: No use of medicines prescribed by a doctor during the past two weeks;

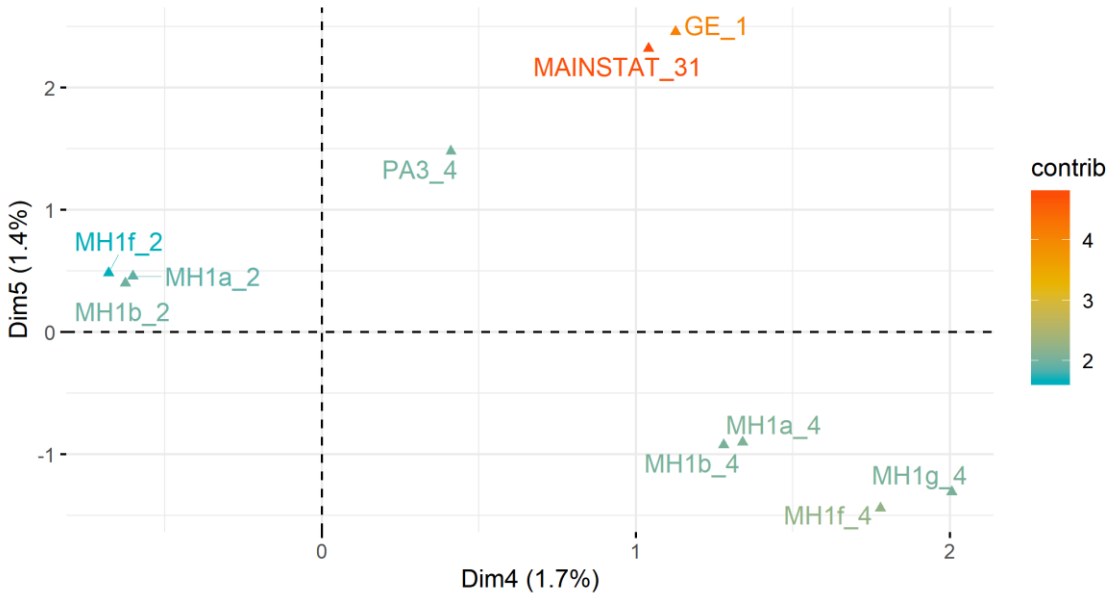
HS2_2: No long-standing health problem; CD1g_1: People suffering from arthrosis (arthritis excluded) in the past 12 months; MH1a_4: People feeling little pleasure in doing things over the last 2 weeks nearly every day; UN2C_3: People without the need to afford prescribed medicines in the past 12 months; HS3_1: People severely limited in daily activities for at the least the past 6 months; CD1h_1 People suffering from a low back disorder or other chronic back defect in the past 12 months.

Figure 2. Map of the top highest contributed categories to dimensions 3 and 4. Contribution means the contribution of a point to the inertia of an axis.



Legend: UN2b_3: People who could not afford dental examination or treatment in the past 12 months; MH1f_4 People feeling bad about themselves, feeling being a failure over the last 2 weeks nearly every day; MH1a_4: People feeling little pleasure in doing things over the last 2 weeks nearly every day; MH1b_4: People feeling down, depressed or hopeless over the last 2 weeks nearly every day; MH1g_4: People having trouble concentrating on things, such as reading the newspaper or watching television, over the last 2 weeks nearly every day HATLEVEL_0: People without schooling MAINSTAT_32: People in retirement and MH1b_2: People feeling down, depressed or hopeless over the last 2 weeks several days.

Figure 3. Map of the top highest contributed categories to dimensions 4 and 5. Contribution means the contribution of a point to the inertia of an axis.



Legend: MAINSTAT_31: People who are pupils, students, training or unpaid working experience. GE_1: Age group 15-19. MH1a_2: People who feel little interest or pressure doing things in the past 2 weeks during several days. MH1b_2: People who felt down in the past 2 weeks during several days MH1f_4: People who felt bad about themselves in the past 2 weeks during several days MH1a_2: People who feel little interest or pressure doing things in the past 2 weeks everyday. MH1b_2: People who felt down in the past 2 weeks everyday MH1g_2: People who had trouble concentrating on things, such as reading the newspaper or watching television, over the last 2 weeks everyday PA3_4: People whose cholesterol was never measured by a health professional.

Fig. 1 corresponds to categories with highest contribution to dimension 1 (along the x axis) and dimension 2 (along y axis). The position of categories indicate on the right side, at the top and further from the center of the axis, individuals who reported feeling depressed (MH1a_4) and with severe limitations in daily activities (HS3_1), meaning these were considered similar. At the bottom and close to each other, we find people who feel back pain (CD1h_1) and have arthrosis as chronic illness (CD1g_1). On the other hand, as different individuals, situated further and on opposite sides, we find individuals who do not report any long term condition (HS2_2) and did not consume any medication (MD1_2). Regarding dimensions 3 and 4 (fig 2), there is a cluster of mental health variables with high contributions for these two dimensions. At the bottom, people who reported that could not afford dental examination or treatment (UN2b_2). On the opposite side people who did not need dental examination (UN2b_3), without primary school (HATLEVEL_0) and retired (MAINSTAT_32). Finally, Fig. 3 shows that the youngest age group

and the working class category have high contributions and are kept together meaning these were considered similar categories.

Overall, we observe a cluster of categories that reflects a homogeneous group of healthy individuals showed in dimension 1 and 2, one cluster of individuals scoring low in mental health questions (dimension 3) and one of older people (dimension 4). As expected, dimension 5 shows a group of mental health categories with better score in the opposite side of those in dimension 3.

Cluster characterization

The dendrogram (Fig 4) and the inter-cluster inertia gain bar plot (Fig 5) suggested three clusters. The inertia-gain bar plot shows that the increase in inertia going from 3 to 4 clusters decreases more than going from 2 to 3. In addition, there is a slight increase in the within-cluster inertia ratio when we break into 4 clusters instead of 3 (this ratio increases from 0.831 to 0.876). As we want the ratio between two successive clusters to be as low as possible, we selected three groups. Cluster sizes and profiles are described in Tables 2-4 for clusters 1, 2 and 3 respectively. Yet, for purposes of simplicity, in these tables we show overrepresented categories that are more informative like sex, age, region, and health condition. Remaining variables - with abbreviated codes to allow an easier reading - are in supplementary material (S6 File).

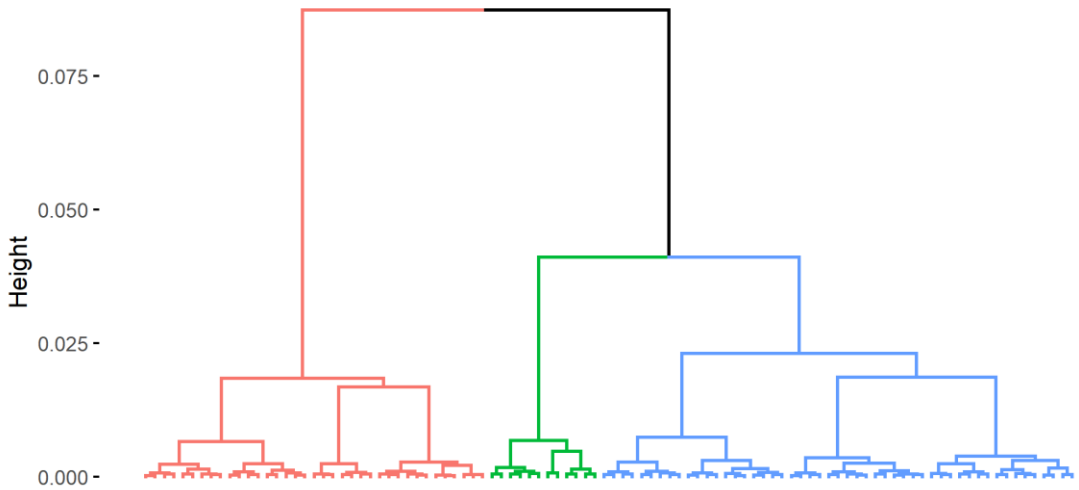


Figure 4. Dendrogram. The dendogram suggests partitioning into three clusters, indicated by the different colors and the height of the arms of the trees.

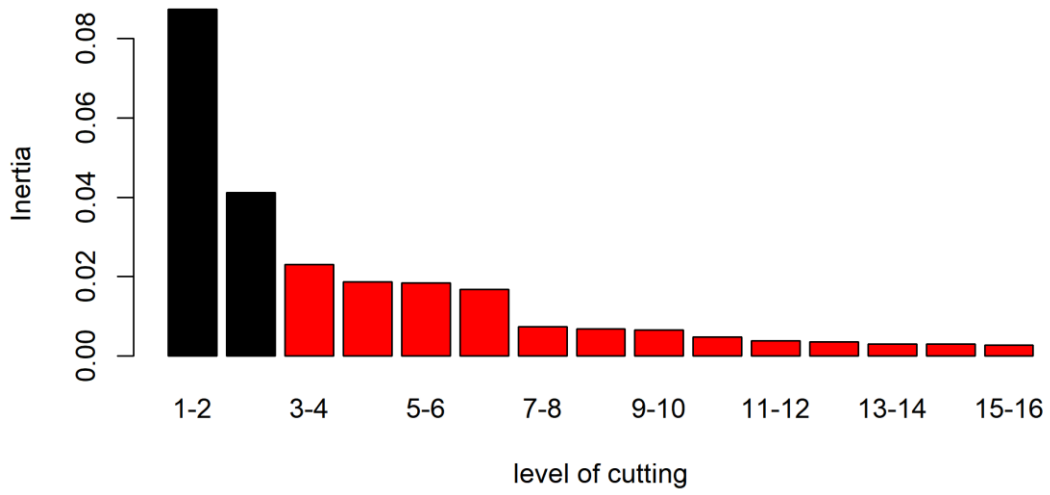


Figure 5. Within-cluster inertia bar plot.

Cluster 1: Healthy (self-perceived) individuals

The first cluster is composed of 6,567 individuals, which corresponds to 52.4% of our sample. This cluster is mainly composed of young and middle aged, healthy individuals. More than half of participants who are male (62.9%) are in this cluster. This cluster has almost all of the individuals in the sample within the age range from 15 to 39 years old. This a cluster where 58.8% of individuals have at least the secondary school (over 9 years): 23.15% has 9 years, 20.12% 12 years and 15.52% has a bachelors degree or equivalent. This cluster also has a high percentage of individuals in the upper quartiles of income: 67% and 61% of individuals in the 4^{th} and 5^{th} quartiles of income belong to cluster 1, which means that 36% of individuals in the cluster are in this range of income. Sixty-two and 65% of people in cluster 1 live in the Azores and Madeira islands, respectively. This correspondents to almost a third of people in cluster 1 (16% and 14%). Also, 84% of this cluster is composed by individuals who are professionally active – these are either employed, students or unemployed. As far as self-health is concerned for this cluster, it is composed by 94.4% of people who have very good

health, and declare not having any long-standing health problem. Further, 86% of individuals who declared they have not needed prescribed medicines fell into this cluster.

Table 2: Description of cluster 1, according to socio-demographic and health related most represented categories (percentage in the cluster and in the sample).

CLUSTER 1	**CATEGORIES**	**% in the cluster**	**% in the sample**
6567 (54.12%)	Age group = 15-19	98.19	7.43
	Age group = 20-24	95.32	6.82
	Age group = 25-29	94.53	6.32
	Self-perceived general health = Very Good	94.42	15.47
	Age group = 30-34	93.63	8.73
	Age group = 35-39	92.27	11.63
	Long-standing health problem: suffer from any illness or health problem of a duration of at least six months = No	90.13	57.85
	Age group = 40-44	88.00	11.73

	Upper secondary education	86.57	20.12
	Self-perceived general health = Good	85.76	48.17
	Educational level = Bachelor or equivalent	84.63	15.52
	Age group = 45-49	84.23	11.22
	Last appointment with family doctor = Over 12 months	80.97	36.93
	Lower secondary education	78.47	23.15
	General activity limitation: limitation in activities people usually do because of health problems for at least the past six months = Not at all	75.84	84.12
	Age group = 50-54	70.99	10.66
	Income = 5th quartile (richest)	67.14	16.55
	Region = Madeira	65.02	16.08
	Sex = Male	62.9	49.4
	Region = Azores	62.73	14.22
	Income = 4th quartile	61.8	20.42
	Age group = 55-59	61.62	9.41
	Pupil, student, further training, unpaid work experience	98.46	9.76

	Carries out a job or profession, including unpaid work for a family business or holding, an apprenticeship or paid traineeship	85.11	56.65
	Occupation = Unemployed	78.59	17.6
	Could not afford prescribed medicines in the past 12 months no need	86.23	37.29
	Unmet need for health care in the past 12 months due to long waiting list(s) = No need	86.17	24.94

Cluster 2: Healthy (aging) individuals

This cluster is composed by 4,454 individuals (35.63% of the sample). Forty percent of female participants fell into this cluster. Seventy percent of the cluster are individuals with 65 to 84 years old, meaning this is a cluster composed by older adults. Eighty four percent of the cluster has less than 4 years of schooling, including no schooling at all. Around 44.15% of participants in the second quartile of income belong to this cluster as well as 19.74% who come from Algarve and 23.6% from Lisbon. Retired individuals and people fulfilling domestic tasks represent the majority of the cluster. Self-health perception is rated as bad (28%) or fair (59%). More than half (51%) of individuals with some long term health condition are in cluster 2, as well as 40.2% of people who are severely limited. Not surprisingly, 90% of participants of the cluster visited the family doctor less than a year ago. Highly reported diseases in this cluster are high blood pressure (56.6%), diabetes (23.73%) as well as and arthrosis (58.24%).

Table 3: Description of cluster 2, according to socio-demographic and health related most represented categories (percentage in the cluster and in the sample).

CLUSTER 2	CATEGORIES	% in the cluster	% in the sample
4454 (36.70%)	Age group = 80-84	73.32	12.71
	Age group = 70-74	72.16	16.7
	Age group = 85+	69.89	6.98
	Age group = 75-79	69.31	15.87
	Age group = 65-69	66.75	17.76
	Self-perceived general health = Bad	64.26	24.83
	General activity limitation: limitation in activities people usually do because of health problems for at least the past six months = Limited	63.7	51.21

	Educational level= No schooling	62.39	26.22
	Self-perceived general health =Fair	51.35	58.76
	Long-standing health problem: suffer from any illness or health problem of a duration of at least six months = Yes	51.31	91.22
	Educational level= Primry school	46.47	58.37
	Age group = 60-64	44.16	10.87
	Income = Second quartile	44.15	27.71
	Last appointment with family doctor = Up to 12 months	43.97	90.21
	Region = Algarve	42.4	19.74
	Region = Lisbon and Tagus Valley	41.71	23.6
	General activity limitation: limitation in activities people usually do because of health problems for at least the past six months = Severly limited	40.19	11.45

	Sex = Female	40.03	62.71
	Occupation = In retirement or early retirement or has given up business	68.56	68.5
	Occupation = Fulfilling domestic tasks	53.91	10.22
	Suffering from diabetes in the past 12 months = Yes	65.86	23.73
	Suffering from high blood pressure in the past 12 months = Yes	64.79	58.24
	Suffering from high blood pressure in the past 12 months = Yes	62.05	56.65
	Suffering from urinary incontinence, problems in controlling the bladder in the past 12 months = Yes	60.19	15.72

Cluster 3: Psychologically vulnerable individuals

This is the smallest cluster with 1,113 individuals (8.9% of the total sample). Sixty percent of participants in this cluster is over 70 years old in total and 77% are women. Almost one fourth of people who have no schooling fell into this cluster and around 10% of people from Lisbon and Algarve. A major characteristic of this

cluster is that it is overrepresented by individuals who reported having often or daily: symptoms or feelings of depression, problems to focus on doing daily routine and restless. Not surprisingly, 95% of people in the cluster rate their self-health as bad or very bad, 13% have some long health problem, and 48.9% of people in the cluster are severely limited in daily activities.

Table 4: Description of cluster 3, according to socio-demographic and health related most represented categories (percentage in the cluster and in the sample).

CLUSTER 3	**CATEGORIES**	**% in the cluster**	**% in the sample**
1113 (9.17%)	Self perceived general health = Very bad	67.09	33.33
	General activity limitation: limitation in activities people usually do because of health problems for at least the past six months = Severely limited	48.86	55.71
	Self perceived general health = Bad	28.24	43.67
	Age group = 85+	26.29	10.51
	Educational level = No schooling	24.25	40.79
	Age group = 75-79	22.16	20.31

	Age group = 80-84	21.37	14.82
	Age group = 70-74	15.23	14.11
	Long-standing health problem: suffer from any illness or health problem of a duration of at least six months = Yes	13.74	97.75
	Sex = Female	12.34	77.36
	General activity limitation: limitation in activities people usually do because of health problems for at least the past six months = Limited	11.06	35.58
	Last appointment with family doctor = Up to 12 months	10.71	87.96
	Lisbon and Tagus Valley	10.48	23.72
	Region = Alentejo	10.32	19.23
	De facto marital status = Person not living in a consensual union	9.52	97.66
	Extent of feeling bad about yourself, feeling being a failure over the last 2 weeks = Almost every day	84.17	28.66

	Extent of having trouble concentrating on things, such as reading the newspaper or watching television, over the last 2 weeks = Almost every day	80.95	22.91
	Extent of moving or speaking so slowly that other people could have noticed or being so fidgety or restless, over the last 2 weeks = Almost every day	75.63	29.83
	Extent of feeling down, depressed or hopeless over the last 2 weeks = Almost every day	72.51	47.17
	Extent of having little interest or pleasure in doing things over the last 2 weeks = Almost every day	70.11	44.47

Discussion

In this paper we identified and analyzed clusters of individuals who do not own private health insurance, using a national health survey database to support cluster-driven health policy actions. We found 3 subgroups within people without health insurance, which may provide insights to specific policy actions.

The first cluster showed a group of individuals who perceived themselves as having good health, and did not report any limitation in general activities or long-standing health problems. Supplementary variables showed that this cluster was composed of single individuals and couples with children aged less than 25, indicating this is a cluster of working age and young and middle aged adults. Further, men at the top levels of income range who reported having the last appointment with the family doctor over 12 months characterize this cluster. This

suggests there is a low utilization of health services even by men with middle and high income. This is in line with other studies that have shown men are less prone to go to the doctor even when controlled for multiple variables like education or income. (LEVORATO et al., 2014) (TRAVASSOS et al., 2002) In fact, considering their last appointment with family doctor was over 12 months, this may also suggest they do not have regular medical appointments as a means of prevention. This lack of preventive attitude towards self-health by men has already been previously documented (BAKER et al., 2014) (CHOR et al., 2008) The second cluster was composed mainly of individuals who perceive their health as fair, but present some long-standing health issues, which were frequently chronic diseases like diabetes, arthrosis and high blood pressure. This is an aging cluster composed especially by women. In terms of access and care, these individuals visited the physician less than a month ago which reveals some extension of health care use. This is in line with one study that demonstrated that for the Portuguese population, healthcare utilization rates are higher for older women. (QUINTAL; LOURENÇO; FERREIRA, 2012) Yet, the same study did not find a statistical association for diseases and health care utilization as one would expect from the composition of this cluster, given the high prevalence of chronic diseases within this cluster. Finally, the third cluster presents older individuals who show depressive symptoms as well as very bad or bad self-reported health and severe limitations in activity. Such a cluster of health characteristics are in line with a previous study that demonstrated a correlation between depressive symptoms and current health status. (FISKE; GATZ; PEDERSEN, 2003). In addition, the last and non-health-related life events explain most of the relationship between age and depressive symptoms. (FISKE; GATZ; PEDERSEN, 2003)

Following from the segmentation described, results show it is of interest to build tailored services and custom health care packages and campaigns. For example, the first cluster is composed of younger, middle-aged and healthy individuals, who report very little use of health services and no health problems, thus calling for more actions that promote preventive behaviors towards this group. As they

are in working age, one option might be working-place health and well-being actions. A second package can be tailored to older individuals, retired and with specific health problems, practices and needs. In particular, transversal actions that target ageing issues, promoted and implemented in the public health services and with a particular focus on older, retired women and actions that target aging related diseases like high blood pressure, diabetes and arthrosis can be worth. Finally, a third package focused on the needs of individuals with depression and the psychologically vulnerable. This package could include increased availability of psychiatrists or psychological appointments and programs that can treat these individuals taking into account these are mainly women at older ages.

It is worth to mention that similar methodologies are carried out in data mining to do customer segmentation in marketing and other business activities (BERRY; LINOFF, 1997), as it allows the targeting of different groups more efficiently. Also, in the public health sector, MCA and/or cluster analysis has been used to assess the difficulties in access to tuberculosis services by users (RUFFI et al., 2009), and to evaluate hospital pharmacy services (NASCIMENTO, 2011) but to our knowledge it has not been tested for uninsured individuals.

Regarding the limitations, we highlight that the structure of our dataset did not allow us to reduce the number of variables. This finding was expected due to the large number of categories (FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS, 2005) and has been previously reported. Future research on each cluster will be useful to understand in depth each cluster of individuals.

In conclusion, we believe this study adds value to public health as it identifies and characterizes profiles of individuals who are uninsured. We described three clusters, with specific characteristics that can be targeted by policy makers to improve their health. In addition, the study demonstrates that national health surveys, with a large set of variables, are very rich in information about population and clustering can be usefull by bringing insights, inform and support public policies.

References

ABDI, H.; VALENTIN, D. **Multiple Correspondence Analysis.** Disponível em: </www.utd.edu/~herve>. Acesso em: 29 jun. 2020.

AIRAKSINEN, J. et al. Prediction of long-term absence due to sickness in employees: Development and validation of a multifactorial risk score in two cohort studies. **Scandinavian Journal of Work, Environment and Health**, v. 44, n. 3, p. 274–282, 2018.

ALBOUKADEL KASSAMBARA. **Cluster validation essentials**.

AMARAL, F. **Introdução à Ciencia de Dados mineração de dados e big data**. Rio Janeiro: Alta Books, 2015.

Análise Multivariada de Dados. [s.l: s.n.]

ARIMOND, G.; ELFESSI, A. A clustering method for categorical data in tourism market segmentation research. **Journal of Travel Research**, v. 39, n. 4, p. 391–397, 2001.

ATHEY, S. Machine Learning and Causal Inference for Policy Evaluation. p. 5–6, 2015.

AUSTIN, P. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. **Statist. Med.**, n. December 2006, p. 3385–3397, 2009.

AUSTIN, P. C. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. **Pharmaceutical Statistics**, v. 10, n. 2, p. 150–161, 2011.

BAKER, P. et al. The men's health gap: Men must be included in the global health equity agenda. **Bulletin of the World Health Organization**, v. 92, n. 8, p. 618–620, 2014.

BERRY, M. J.; LINOFF, G. **Data Mining Techniques: For Marketing, Sales,**

and Customer Support. [s.l: s.n.]

BHATLA, N.; JYOTI, K. An analysis of heart disease prediction using different data mining techniques. **International Journal of Engineering Research & Tecnology**, v. 1, n. 8, p. 1–4, 2012.

BOOT, C. R. L. et al. Prediction of long-term and frequent sickness absence using company data. **Occupational Medicine**, v. 67, n. 3, p. 176–181, 2017.

BREIMAN, L. Statistical Modeling: The Two Cultures. **Statistical Science**, v. 16, n. 3, p. 199–231, 2001a.

BREIMAN, L. Random Forests. **Machine Learning**, n. 45, p. 5–32, 2001b.

BREIMAN, L. et al. Classification and regression trees. **Classification and Regression Trees**, p. 1–358, 2017.

C.M., L.; V.H., V. B.; H.B., F. Application of unsupervised analysis techniques to lung cancer patient data. **PLoS ONE**, v. 12, n. 9, p. 1–18, 2017. Disponível em: <http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L618273302%0Ahttp://dx.doi.org/10.1371/journal.pone.0184370>.

CHENG LI. A Gentle Introduction to Gradient Boosting. **Seamless R and C++ Integration with Rcpp**, p. 3–18, 2013. Disponível em: <http://link.springer.com/10.1007/978-1-4614-6868-4_1>.

CHOR, J. S. Y. et al. Are men seeking medical advice too late? Contacts to general practitioners and hospital admissions in Denmark 2005. **Journal of Public Health**, v. 30, n. 1, p. 110–111, 2008.

CONASS. **Programa Mais Médicos-Nota Tecnica 23/2013.**Brasilia, 2013. . Disponível em: <http://maismedicos.gov.br/>.

CONCOLATO, C. E.; CHEN, L. M. Data Science: A New Paradigm in the Age of Big-Data Science and Analytics. **New Mathematics and Natural Computation**, v. 13, n. 2, p. 119–143, 2017.

COVENEY, P. V; DOUGHERTY, E. R.; HIGHFIELD, R. R. Big data need big

theory too Subject Areas : Author for correspondence : **Philosophical Transactions of the Royal Society A**, 2016.

CRC, H.; MURRELL, P. **Exploratory Multivariate Analysis by Example Using R Introduction to Data Technologies**. [s.l: s.n.]

DATASUS. **DATASUS**. Disponível em: <http://datasus.saude.gov.br/>. Acesso em: 7 jul. 2017a.

DATASUS. **Programa Mais Médicos. Governo Federal.** Disponível em: <http://maismedicos.gov.br/conheca-programa>. Acesso em: 23 ago. 2018b.

DHAR, V. Data Science and Prediction Vasant Dhar Professor, Stern School of Business Director, Center for Digital Economy Research. **Communications of the ACM**, n. May, p. 64–73, 2012. Disponível em: <http://archive.nyu.edu/bitstream/2451/31553/2/Dhar-DataScience.pdf>.

DHL. Big Data in Logistics. **DHL Customer Solutions & Innovation**, n. December, p. 1–30, 2013.

DIAS, C. M. 25 anos de Inquérito Nacional de Saúde em Portugal. **Revista Portuguesa de Saúde Pública**, p. 51–60, 2009.

DONOHO, D. 50 Years of Data Science. **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745–766, 2017. Disponível em: <https://doi.org/10.1080/10618600.2017.1384734>.

EIRA, A. **A Saúde em Portugal: A procura de cuidados de saúde privados**. 2010. Faculdade Economia Porto, 2010. Disponível em: <https://repositorio-aberto.up.pt/bitstream/10216/26931/2/A saude em Portugal A procura privada de cuidados de saude Ana Eira.pdf>.

ELIZABETH A. STUART, BRIAN K. LEE, AND JUSTIN LESSLER, S. Improving propensity score weighting using machine learning. **Statist. Med.**, v. 29, n. 3, p. 337/346, 2010.

EUROPEAN COMMISSION. COMMISSION REGULATION (EU) No 141/2013 of 19 February 2013 implementing Regulation (EC) No 1338/2008 of the

European Parliament and of the Council on Community statistics on public health and health and safety at work, as regards statistics based on the E. v. 2013, n. 141, 2013. Disponível em: <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32013R0141>.

EUROPEAN PARLIAMENT AND COUNCIL OF THE EU. Regulation (EC) N.223/2009. . 2009, p. 164–173.

EUROSTAT. **EUROPEAN HEALTH INTERVIEW SURVEY (EHIS).** Disponível em: <https://ec.europa.eu/eurostat/web/microdata/european-health-interview-survey>. Acesso em: 10 out. 2019.

FERREIRA, K. J. Analytics for an Online Retailer: Demand Forecasting and Price Optimization. v. 39, n. 5, p. 293–296, 2008.

FERRI, C.; HERNÁNDEZ-ORALLO, J.; MODROIU, R. **An experimental comparison of performance measures for classificationPattern Recognition Letters**, 2009. .

FISKE, A.; GATZ, M.; PEDERSEN, N. L. Depressive Symptoms and Aging: The Effects of Illness and Non-Health-Related Events. **Journals of Gerontology - Series B Psychological Sciences and Social Sciences**, v. 58, n. 6, p. 320–328, 2003.

FONTANELLA, S. et al. Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: A cross-sectional analysis within a population-based birth cohort. **PLoS Medicine**, v. 15, n. 11, p. 1–22, 2018.

FRANCO, G. Di. Multiple correspondence analysis : one only or several techniques ? **Quality & Quantity**, p. 1299–1315, 2016. Disponível em: <http://dx.doi.org/10.1007/s11135-015-0206-0>.

FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS. **Exploratory Multivariate Analysis using R**. [s.l.] CRC Press, 2005.

GRILLON, C. et al. Scale-up of HIV self-testing. **The Lancet HIV**, v. 5, n. 7, p.

e324–e326, 2018. Disponível em: <http://dx.doi.org/10.1016/S2352-3018(18)30095-X>.

HAIR, JR, RE ANDERSON, B. W. **Analise Multivariada de Dados**. 5ª ed. Porto Alegre: Bookman, 2005.

HASTIE, T. et al. **An Introduction to Statistical Learning, Springer Texts**. CA, USA: Springer, 2013. v. 102

HINDMAN, M. Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. **Annals of the American Academy of Political and Social Science**, v. 659, n. 1, p. 48–62, 2015. Disponível em: <https://doi.org/10.1177/0002716215570279>.

HJELLBREKKE, J. **Multiple Correspondence Analysis for Social Sciences**. New York: Routledge. Taylor and Francis Group., 2007.

HUSSON, F. JOSSE, J, PAGÉS, J. **FactoMineR. Hierarchical Clustering on Principal Components.** Disponível em: <http://factominer.free.fr>.

HUSSON, A. F. et al. Package ‘ FactoMineR ’. 2018.

HUSSON, F. **Handling missing values in MCA**, 2016. .

ILMARINEN, J. The Work Ability Index (WAI). **Occupational Medicine**, v. 57, n. 2, p. 160–160, 2006.

INE, INSTITUTO NACIONAL DE SAÚDE DOUTOR RICARDO JORGE, I. **Inquérito Nacional Saúde**Lisboa, 2016. .

INE. **DOCUMENTO METODOLÓGICO.National Health Survey**. [s.l: s.n.].

IZBICKI, R. Machine Learning sob a ótica estatística. 2016.

JOSSE, J. Technical Report – Agrocampus. n. September, 2010.

JOSSE, J.; HUSSON, F.; LÊ, S. **Multiple Correspondence Analysis**, 2008. .

JOSSELIN, J.-M.; LE MAUX, B. Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education. In:

SPRINGER (Ed.). **Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education.** 1st 2017 ed. [s.l.] Springer International Publishing, 2017. p. 489–531.

KASSAMBARA, A. **Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis Book 1)**. [s.l: s.n.]

KHAN, A.; UDDIN, R.; ISLAM, S. M. S. **Clustering patterns of behavioural risk factors for cardiovascular diseases in Bangladeshi adolescents: A population-based studyHealth Policy and Technology**, 2019. .

KIVIMÄKI, M. et al. Sickness absence as a risk marker of future disability pension: The 10-town study. **Journal of Epidemiology and Community Health**, v. 58, n. 8, p. 710–711, 2004.

KOCAKULAH, M. C. et al. Absenteeism Problems And Costs: Causes, Effects And Cures. **International Business & Economics Research Journal (IBER)**, v. 15, n. 3, p. 89–96, 2016.

KOHAVI, R. Cross validation number.pdf. v. 5, p. 7, 1995. Disponível em: <http://robotics.stanford.edu/~ronnyk>.

KOLANOVIC, M.; KRISHNAMACHARI, R. T. Big Data and AI Strategies. n. May, p. 280, 2017.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Supervised Machine Learning: A Review of Classification Techniques. **Artificial Intelligence Review**, v. 26, p. 159–190, 2006.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling [Hardcover]**. [s.l: s.n.]

KUMAR, M.; KUMAR, M. International Journal of Computer Science and Mobile Computing Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. **International Journal of Computer Science and Mobile Computing**, v. 5, n. 2, p. 24–33, 2016. Disponível em: <www.ijcsmc.com>.

LANDIS, G. K. The measurement of observer agreement for categorical data.

Biometrics, 1977.

LEG/UFPR. **Métodos baseados em árvores.** Disponível em: <http://cursos.leg.ufpr.br/>. Acesso em: 17 jun. 2020.

LEITE, W. **Practical propensity score methods using R**. USA: SAGE Publications, 2017.

LEVORATO, C. D. et al. Fatores associados à procura por serviços de saúde numa perspectiva relacional de gênero. **Ciencia e Saude Coletiva**, v. 19, n. 4, p. 1263–1274, 2014.

LINDEN, A.; YARNOLD, P. R. Combining machine learning and matching techniques to improve causal inference in program evaluation. **Journal of Evaluation in Clinical Practice**, v. 22, n. 6, p. 864–870, 2016.

LUND, T. et al. Using administrative sickness absence data as a marker of future disability pension: The prospective DREAM study of Danish private sector employees. **Occupational and Environmental Medicine**, v. 65, n. 1, p. 28–31, 2008.

MCLAREN, L.; PETIT, R. Universal and targeted policy to achieve health equity: a critical analysis of the example of community water fluoridation cessation in Calgary, Canada in 2011. **Critical Public Health**, v. 28, n. 2, p. 153–164, 2018. Disponível em: <http://doi.org/10.1080/09581596.2017.1361015>.

MICHIE, D. Methodologies from machine learning in data analysis and software. **Computer Journal**, v. 34, n. 6, p. 559–565, 1991.

MICHIE, S.; WILLIAMS, S. Reducing work related psychological ill health and sickness absence: A systematic literature review. **Occupational and Environmental Medicine**, v. 60, n. 1, p. 3–9, 2003.

MIGUEL A. CARREIRA-PERPINAN. **Dimensionality Reduction**. [s.l: s.n.]

MIKE LOUKIDES. **What is Data Science?** 1st editio ed. [s.l.] O'Reilly Media, 2016.

MINISTERIO SAUDE. **TC 80: Cooperação Técnica entre o Ministério da Saúde e a Organização Pan-Americana da Saúde - Projeto Acesso da população Brasileira à Atenção Básica em Saúde.**Brasilia, 2013. .

MOLNAR, C. **Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.** Disponível em: <https://christophm.github.io/>. Acesso em: 27 jun. 2020.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [s.l: s.n.]

NASCIMENTO, A. do. **Avaliação de farmácias hospitalares Brasileiras utilizando Análise de Correspondência Múltipla.** 2011. UFRJ - Universidade Federal do Rio de Janeiro., 2011.

NENADIC, O.; GREENACRE, M. Computation of Multiple Correspondence Analysis, with Code in R. **Ssrn**, p. 25–27, 2005.

NGUYEN, L. H.; HOLMES, S. Ten quick tips for effective dimensionality reduction. **PLoS Computational Biology**, v. 15, n. 6, p. 1–19, 2019.

PAHO/WHO. **Ministério da Saúde reforça cooperação com a OPAS para continuidade do Mais Médicos**. Disponível em: <http://portalms.saude.gov.br/noticias/agencia-saude/42756-ministerio-da-saude-reforca-cooperacao-com-a-opas-para-continuidade-do-mais-medicos>.

PEDROSO, R. T.; JUHÁSOVÁ, M. B.; HAMANN, E. M. **A ciência baseada em evidências nas políticas públicas para reinvenção da prevenção ao uso de álcool e outras drogasInterface - Comunicação, Saúde, Educação**, 2019. .

POHAR, M.; BLAS, M.; TURK, S. Comparison of Logistic Regression and Linear Discriminant Analysis : A Simulation Study. **Metodološki zvezki**, v. 1, n. 1, p. 143–161, 2004.

QUINTAL, C.; LOURENÇO, Ó.; FERREIRA, P. **Utilização de cuidados de saúde pela população idosa portuguesa: Uma análise por género e classes latentesRevista Portuguesa de Saude Publica**, 2012. .

RAMOS, M. C.; SILVA, E. N. da. Como usar a abordagem da Política

Informada por Evidência na saúde pública? **Saúde em Debate**, v. 42, n. 116, p. 296–306, 2018. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-11042018000100296&lng=pt&tlng=pt>.

RUBIN, D. B. Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. **Journal of Educational Psychology**, v. 66, n. 5, p. 688–701, 1974.

RUFFI, A. et al. Difi culdades de acesso a serviços de saúde para diagnóstico de tuberculose em municípios do Brasil Diffi culties in the accessibility to health services for tuberculosis. v. 43, n. 3, p. 389–397, 2009.

RUSSO, L. X. et al. Primary care physicians and infant mortality: Evidence from Brazil. **PLoS ONE**, v. 14, n. 5, p. 1–16, 2019.

RYGIELSKI, C.; WANG, J.; YEN, D. C. Data mining techniques for customer relationship management. v. 24, p. 483–502, 2002.

SAMUEL, A. L. Some studies in Machine Learning using the Game of Checkers. **IBM J. RES. DEVELOP**, v. 44, n. 1, 2000.

SANTOS, H. G. dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. p. 207, 2018.

SCHEFFER, M. et al. **Demografia médica no Brasil 2015**. São Paulo: Departamento de Medicina Preventiva, Faculdade de Medicina da USP. Conselho Regional de Medicina do Estado de São Paulo. Conselho Federal de Medicina., 2015.

SCHNEIDER, P. Why should the poor insure? Theories of decision-making in the context of health insurance. **Health Policy and Planning**, v. 19, n. 6, p. 349–355, 2004.

SHADISH, W. R. Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings. **Psychological Methods**, v. 15, n. 1, p. 3–17, 2010.

SHAW, M. J. et al. Knowledge management and data mining for marketing. p. 127–137, 2001.

SHI, L. et al. Primary care, infant mortality, and low birth weight in the states of the USA. **Journal of Epidemiology and Community Health**, v. 58, n. 5, p. 374–380, 2004.

SILVER, N. **Statistics Views**. Disponível em: <https://www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html>. Acesso em: 7 ago. 2019.

THE HENRY J KAISER FAMILY FOUNDATION. **Key Facts about the Uninsured PopulationThe Henry J Kaiser Family Foundation**, 2016. . Disponível em: <https://kaiserfamilyfoundation.files.wordpress.com/2013/09/8488-key-facts-about-the-uninsured-population.pdf>.

TRAVASSOS, C. et al. Utilização dos serviços de saúde no Brasil: Gênero, características familiares e condição social. **Revista Panamericana de Salud Publica/Pan American Journal of Public Health**, v. 11, n. 5–6, p. 365–373, 2002.

UNIVERSITY OF CHICAGO. **Data Science for Social Good.**

VAN DEN BERG, T. I. J.; ELDERS, L. A. M.; BURDORF, A. The effects of work-related and individual factors on work ability: A systematic review. **Promotion of Work Ability Towards Productive Aging - Selected Papers of the 3rd International Symposium on Work Ability**, p. 15–18, 2009.

VIACAVA, F.; SOUZA-JÚNIOR, P. R. B. de; SZWARCWALD, C. L. Cobertura da população brasileira com 18 anos ou mais por plano de saúde privado: uma análise dos dados da Pesquisa Mundial de Saúde. **Cad Saude Publica**, v. 21, n. supl.1, p. S119–S128, 2005.

WENG, S. F. et al. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? **PLoS ONE**, v. 12, n. 4, p. 1–14, 2017.

WILPER, A. P. et al. Health insurance and mortality in US adults. **American Journal of Public Health**, v. 99, n. 12, p. 2289–2295, 2009.

SUPPLEMENTARY MATERIAL

S1 File. National Health Survey Questionnaire.

S2 File. Fluxogram regarding variable selection for clustering.

S3 File. Characterization of our sample (n=12148) for all 97 variables.

S4 File. List of supplementary variables used in MCA and analysis.

S5 File. Relative contribution of active categories and cumulative percentage of variables for dimensions 1-5.

S6 File. Description of clusters 1-3, according to overrepresented categories (percentage in the cluster and in the sample).

S1 File. National health survey questionnaire.

Nr.	Abbreviation	Description	Answer categories and codes
1	PID	Identification number of respondent.	10 - digit number – 1 missing
2	HHID	Identification number of household.	10-digit number – 1 missing
3	PRIMSTRAT	Primary strata as used in the selection of the sample	0001 – 9999 – 2 not applicable (no stratification)
4	PSU	Primary sampling units as used in the selection of the sample	0001 – 9999 – 2 not applicable (no multistage sampling)
5	WGT	Final individual weight	Number 5.3 – 1 missing
6	PROXY	Was the selected person interviewed or someone else (proxy interview)	1 Person himself/herself 2 Other member of the household 3 Someone else outside the household – 1 missing
7	REFYEAR	Reference year of the interview	4 digits
8	REFMONTH	Reference month of the interview	1 – 12 – 1 missing
9	INTMETHOD	Data collection method used	10 Postal, non-electronic version 11 Postal, electronic version (e-mail) 20 Face-to-face, non-electronic version 21 Face-to-face, electronic version

			30 Telephone, non-electronic version 31 Telephone, electronic version 40 Use of internet 50 Mixed mode collection (e.g.: both postal and interview to collect data)
10	INTLANG	Language used for interview	3-digit codes (Standard Code List Eurostat) – 1 missing
11	COUNTRY		
12	NUTS		
13	REGION	Region of residence	Coding according to NUTS at 2-digit level – 1 missing
14	NUTS2	Residence according to NUTS2	NUTS2
15	DEG_URB	Degree of urbanisation	1 Densely-populated area 2 Intermediate-populated area 3 Thinly-populated area – 1 missing
16	SEX	Sex of the person	1 Male 2 Female
17	AGE	Age of the person	
18	BIRTHPLACE		-1 missing 10 Portugal 21 Country in EU 22 Country outside the EU
19	CITIZEN		-1 missing 10 Portugal 21 Country in EU 22 Country outside the EU

20	MARSTALEGAL	Legal marital status	1 Never married and never been in a registered partnership 2 Married or in a registered partnership 3 Widowed or in registered partnership that ended with death of partner (not remarried or in new registered partnership) 4 Divorced or in registered partnership that was legally dissolved (not remarried or in new registered partnership) – 1 missing (don't know, refusal)
21	MARSTADE-FACTO	De facto marital status	1 Person living in a consensual union 2 Person not living in a consensual union – 1 missing (don't know, refusal)
22	IN1	Always lived in Portugal	1 Yes 2 No -1 Missing (don´t know, refusal)
23	IN2	Number of years living in Portugal	-1 Missing (don´t know, refusal) -2 not applicable
24	IN3	Number of years since came back to Portugal	-1 Missing (don´t know, refusal) -2 not applicable
25	IN4	Number of years of residence in Portugal	-1 Missing (don´t know, refusal) -2 not applicable
26	HATLEVEL	Highest level of education completed (Educational attainment)	-1 Missing 0 -None 1 — Primary education

			2 — Lower secondary education 3 — Upper secondary education 4 — Post-secondary but non-tertiary education 5 — Tertiary education; short-cycle 6 — Tertiary education; bachelor level or equivalent
27	MAINSTAT	Self-declared labour status	10 Carries out a job or profession, including unpaid work for a family business or holding, an apprenticeship or paid trai- neeship, etc. 20 Unemployed 31 Pupil, student, further training, unpaid work experience 32 In retirement or early retirement or has given up business 33 Permanently disabled 34 In compulsory military or community service 35 Fulfilling domestic tasks 36 Other inactive person – 1 missing (don't know, refusal)
28	FT_PT	Full or part-time work	1 Full-time 2 Part-time

			– 1 missing (don't know, refusal) – 2 not applicable
29	JOBSTAT	Status in employment	10 Self-employed 21 Employee with a permanent job/work contract of unlimited duration 22 Employee with a temporary job/work contract of limited duration – 1 missing (don't know, refusal) – 2 not applicable
30	JOBISCO	Occupation in employment	ISCO-08 coded at 2-digit level – 1 missing (don't know, refusal) – 2 not applicable
31	LOCNACE	Economic sector in employment	NACE Rev. 2 — sections – 1 missing (don't know, refusal) – 2 not applicable
32	fT_PT	Number of persons living in household, including the respondent	0 – 98 – 1 missing (don't know, refusal)
33	HHNBPERS_0_4	Number of persons aged 4 or younger	0 – 98 – 1 missing (don't know, refusal)
34	HHNBPERS_5_13	Number of persons aged from 5 to 13	0 – 98 – 1 missing (don't know, refusal)
35	HHNBPERS_14_15	Number of persons aged from 14 to 15	0 – 98 – 1 missing (don't know, refusal)

36	HHNBPERS_16_24	Number of persons aged from 16 to 24	0 – 98 – 1 missing (don't know, refusal)
37	HHNBPERS_25_64	Number of persons aged from 25 to 64	0 – 98 – 1 missing (don't know, refusal)
38	HHNBPERS_65plus	Number of persons aged 65 and over	0 – 98 – 1 missing (don't know, refusal)
39	HHTYPE	Type of household	10 One-person household 21 Lone parent with child(ren) aged less than 25 22 Couple without child(ren) aged less than 25 23 Couple with child(ren) aged less than 25 24 Couple or lone parent with child(ren) aged less than 25 and other persons living in household 25 Other type of household – 1 missing (don't know, refusal)
40	HH_ACT	Number of persons aged 16 – 64 in the household who are employed	0 – 98 – 1 missing (don't know, refusal)
41	HH_INACT	Number of persons aged 16 – 64 in the household who are unemployed or are economically inactive	0 – 98 – 1 missing (don't know, refusal)
42	HHINCOME	Net monthly equivalised income of the household	1 Below 1st quintile 2 Between 1st quintile and 2nd quintile 3 Between 2nd quintile and 3rd quintile 4 Between 3rd quintile and 4th quintile 5 Between 4th quintile and 5th quintile

			– 1 missing (don’t know, refusal)
43	IN5	Does own a Health subsistem	-1 Missing (don´t know, refusal) -2 not applicable 1 ADSE 2ADM 3SAD/PSP 4SAD/GNR 5SAMS 6OTHER
44	IN6	Does own a private health plan?	-1 Missing 1yes 2 No
45	IN7	Coverage of the health plan	-2 Not applicable -1 Missing 1 Only hospital admission 2 Admission, appointments and diagnosis and therapeutic complementary services 3 Admission, appointments and diagnosis and therapeutic complementary services and medicines 4 other combination of packages 5 all
46	HS1	Self perceived general health	-1 missing 1 Very good 2 Good 3 fair

			4 Bad 5 very bad
47	HS2	Long-standing health problem: Suffer from any illness or health problem of a duration of at least six months	-1 missing 1Yes 2 No
48	HS3	General activity limitation: Limitation in activities people usually do because of health problems for at least the past six months	-1 missing 1 Severely 2 Limited 3 Not at all
49	CD1a	Suffering from ashtma in the past 12 months	1. Yes 2.No - 1 missing (don´t know, refusal)
50	CD1b	Suffering from chronic bronchitis, chronic obstructive pulmonary disease or emphysema in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
51	CD1c	Suffering from a myocardial infarction (heart attack) in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
52	CD1d	Suffering from a coronary heart disease or angina pectoris in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
53	CD1e	Suffering from high blood pressure in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
54	CD1f	Suffering from a stroke (cerebral haemorrhage, cerebral thrombosis) in the past 12 months	1 Yes 2 No

			– 1 missing (don't know, refusal)
55	CD1g	Suffering from arthrosis (arthritis excluded) in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
56	CD1h	Suffering from a low back disorder or other chronic back defect in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
57	CD1i	Suffering from a neck disorder or other chronic neck defect in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
58	CD1j	Suffering from diabetes in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
59	CD1k	Suffering from an allergy, such as rhinitis, eye inflammation, dermatitis, food allergy or other (allergic asthma excluded) in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
60	CD1l	Suffering from cirrhosis of the liver in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
61	CD1m	Suffering from urinary incontinence, problems in controlling the bladder in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
62	CD1n	Suffering from kidney problems in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)

63	CD1o	Suffering from depression in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
64	AC1a	Occurrence of a road traffic accident in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
65	AC1b	Occurrence of an accident at home in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
66	AC1c	Occurrence of a leisure accident in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
67	AC2	Most serious medical care intervention for the most serious accident in the past 12 months	1 Admission to a hospital or any other health facility 2 A doctor or nurse 3 No intervention was needed – 1 missing (don't know, refusal) – 2 not applicable
68	AW1	Absent from work due to personal health problems in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
69	AW2	Number of days of absence from work due to personal health problems in the past 12 months	Number of days – 1 missing (don't know, refusal) – 2 not applicable
70	PL1	Wearing glasses or contact lenses	1 Yes 2 No

			3 Blind or can not see at all – 1 missing (don't know, refusal)
71	PL2	Difficulty in seeing, even when wearing glasses or contact lenses	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable
72	PL3	Use of a hearing aid	1 Yes 2 No 3 Profoundly deaf – 1 missing (don't know, refusal)
73	PL4	Difficulty in hearing what is said in a conversation with one other person in a quiet room even when using a hearing aid	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable
74	PL5	Difficulty in hearing what is said in a conversation with one other person in a noisier room even when using a hearing aid	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable

75	in8	Difficulty speaking	– 1 missing (don't know, refusal) – 2 not applicable 1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do
76	PL6	Difficulty in walking half a km on level ground without the use of any aid	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal)
77	in9	Difficulty in walking 200m on level ground without the use of any aid	– 1 missing (don't know, refusal) – 2 not applicable 1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do
78	PL7	Difficulty in walking up or down 12 steps	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal)
79	PC1a	Difficulty in feeding yourself	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do

			– 1 missing (don't know, refusal) – 2 not applicable
80	PC1b	Difficulty in getting in and out of a bed or chair	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable
81	PC1c	Difficulty in dressing and undressing	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable
82	in15	Extent to which is difficult wash hands and face	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable

83	PC1d	Difficulty in using toilets	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable
84	PC1e	Difficulty in bathing or showering	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do – 1 missing (don't know, refusal) – 2 not applicable
85	PC2	Usually receiving help with one or more self-care activities: feeding yourself, getting in and out of a bed or chair, dressing and undressing, using toilets, bathing or showering	1 Yes, with at least one activity 2 No – 1 missing (don't know, refusal) – 2 not applicable
86	PC3	Need to receive help or more help with one or more self-care activities: feeding yourself, getting in and out of a bed or chair, dressing and undressing, using toilets, bathing or showering	1 Yes, with at least one activity 2 No – 1 missing (don't know, refusal) – 2 not applicable
87	HA1a	Difficulty in preparing meals	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do 5 Not applicable (never tried it or do not need to do it)

			– 1 missing (don't know, refusal) – 2 not applicable
88	HA1b	Difficulty in using the telephone	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do 5 Not applicable (never tried it or do not need to do it) – 1 missing (don't know, refusal) – 2 not applicable
89	HA1c	Difficulty to do shopping	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do 5 Not applicable (never tried it or do not need to do it) – 1 missing (don't know, refusal) – 2 not applicable
90	HA1d	Difficulty in managing medication	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do 5 Not applicable (never tried it or do not need to do it) – 1 missing (don't know, refusal) – 2 not applicable

91	HA1e	Difficulty in doing light housework	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do 5 Not applicable (never tried it or do not need to do it)
92	HA1f	Extent to which is difficult to do ocasional domestic and heavy tasks	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do 5 Not applicable (never tried it or do not need to do it)
93	HA1g	Extent to which takes care of financial issues	1 No difficulty 2 Some difficulty 3 A lot of difficulty 4 Cannot do at all/Unable to do 5 Not applicable (never tried it or do not need to do it)
94	HA2	Has help in domestic tasks	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
95	HA3	Need help in domestic tasks	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
96	PN1	Intensity of bodily pain during the past 4 weeks	1 None 2 Very mild 3 Mild 4 Moderate

			5 Severe 6 Very severe – 1 missing (don't know, refusal)
97	PN2	Extent that pain interfered with normal work during the past 4 weeks (including both work outside the home and housework)	1 Not at all 2 A little bit 3 Moderately 4 Quite a bit 5 Extremely – 1 missing (don't know, refusal)
98	MH1a	Extent of having little interest or pleasure in doing things over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day – 1 missing (don't know, refusal)
99	MH1b	Extent of feeling down, depressed or hopeless over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day – 1 missing (don't know, refusal)
100	MH1c	Extent of having trouble falling or staying asleep, or sleeping too much over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day – 1 missing (don't know, refusal)
101	MH1d	Extent of feeling tired or having little energy over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day

			– 1 missing (don't know, refusal)
102	MH1e	Extent of having poor appetite or overeating over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day – 1 missing (don't know, refusal)
103	MH1f	Extent of feeling bad about yourself, feeling being a failure over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day – 1 missing (don't know, refusal)
104	MH1g	Extent of having trouble concentrating on things, such as reading the newspaper or watching television, over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day – 1 missing (don't know, refusal)
105	MH1h	Extent of moving or speaking so slowly that other people could have noticed or being so fidgety or restless, over the last 2 weeks	1 Not at all 2 Several days 3 More than half the days 4 Nearly every day – 1 missing (don't know, refusal)
106	HO1	Admission as an inpatient in a hospital in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
107	HO2	Number of nights spent as a patient in a hospital in the past 12 months	Number 1 – 99 – 1 missing (don't know, refusal)

			– 2 not applicable
108	HO3	Admission as a day patient in a hospital in the past 12 months	1 Yes 2 No – 1 missing (don't know, refusal)
109	HO4	Number of times admitted as a day patient in a hospital in the past 12 months	Number 1 – 300 – 1 missing (don't know, refusal) – 2 not applicable
110	AM1	Last appointment with dentist	-1 missing 1 Less than 6 months 2 6 to 12 months 3 Over 12 months 4 Never
111	IN16	Reason for the last appointment with the dentist	-2 Not applicable -1 Missing 1 Pain 2 tooth extraction 3 to do dental prothesis 4 To know the current dental situation 5 Yearly check up 6 Oral hygiene 7 restoration of teeth 8 other
112	IN17	Reason for never having visited a dentist	-2 Not applicable -1 Missing 1 Never needed 2 There is no dentist nearby 3 Difficult to set an appointment 4 Expensive 5 Other

113	IN18	Frequency brushing teeth	-2 Not applicable -1 Missing 1 Everyday in the morning, after lunch and before bed 2 Everyday in the morning and before bed 3 Everyday in the morning 4 Everyday before bed 4 A few times per week 5 Less then once week 6 Never
114	AM2	Last appointment with family doctor	-1 missing 1 less than 12 months 2 12 or more 3 Never
115	AM3	Last appointment with family doctor (number)	-2 Not applicable -1 Missing
116	AM4	Last appointment with specialist doctor	-1 missing 1 less than 12 months 2 12 or more 3 Never
117	AM5	Last appointment with specialist doctor (number)	-2 Not applicable -1 Missing
118	AM6a	Consultation with phisioterapist in the past 12 months	1 Yes 2 No -1 Missing (don´t know, refusal)
119	AM6b	Consultation of a psychologist or psychotherapist in the past 12 months	1 Yes 2 No -1 missing (don´t know, refusal)
120	AM7	Use of any home care services for personal needs during the past 12 months	1 Yes 2 No

			-1 missing (don´t know, refusal)
121	MD1	Use of any medicines prescribed by a doctor during the past two weeks (excluding contraception)	1 Yes 2 No -1 missing (don´t know, refusal)
122	MD2	Use of any medicines, herbal medicines or vitamins not prescribed by a doctor during the past two weeks	1 Yes 2 No -1 missings (don´t know, refusal)
123	PA1	Last time of vaccination against flu.	Month/year (MMYYYY) Never or too long ago (0) – 1 missing (don't know, refusal)
124	IN19	Last time of vaccination against tetanus.	
125	PA2	Last time of blood pressure measurement by a health professional	1 Within the past 12 months 2 1 to less than 3 years 3 3 to less than 5 years 4 More than 5 years 5 Never – 1 missing (don't know, refusal)
126	PA3	Last time of blood cholesterol measurement by a health professional	1 Within the past 12 months 2 1 to less than 3 years 3 3 to less than 5 years 4 More than 5 years 5 Never – 1 missing (don't know, refusal)
127	PA4	Last time of blood sugar measurement by a health professional	1 Within the past 12 months 2 1 to less than 3 years 3 3 to less than 5 years 4 More than 5 years 5 Never

			– 1 missing (don’t know, refusal)
128	PA5	Last time of a faecal occult blood test	1 Within the past 12 months 2 1 to less than 2 years 3 2 to less than 3 years 4 More than 3 years 5 Never – 1 missing (don’t know, refusal)
129	PA6	Last time of a colonoscopy	1 Within the past 12 months 2 1 to less than 5 years 3 5 to less than 10 years 4 More than 10 years 5 Never – 1 missing (don’t know, refusal)
130	PA7	Last time of a mammography (breast X-ray)	1 Within the past 12 months 2 1 to less than 2 years 3 2 to less than 3 years 4 More than 3 years 5 Never
131	PA8	Last time of a cervical smear test	1 Within the past 12 months 2 1 to less than 2 years 3 2 to less than 3 years 4 More than 3 years 5 Never – 1 missing (don’t know, refusal) – 2 not applicable
132	NT1	Current pregnancy	1 Yes 2 No

			– 1 missing (don't know, refusal) – 2 not applicable
133	NT2	Pregnancy in the last 12 months	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
134	IN20	Contraception pill in the last 12 months	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
135	IN21	Main contraception method used	1 Pill 2 Male condom 3 IUD 4 Diafragm 5 Hormonal injection 6 Implant 7 Vasectomy 8 Body temperatur control 9 Abstinence -1 missing -2 not applicable
136	IN22	Reason to not use contraception	1 Breastfeeding 2 Planning pregnancy 3 No sexual activity 4 Health reasons 5 Menopause 6 Other -1 missing -2 not applicable

137	IN23	Had a last pregnancy	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
138	IN24	Year of the last pregnancy	
139	nt3	Breastfeeding	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
140	IN25	Breastfeeding time (weeks)	
141	IN26	Which week of pregnancy was the first penatal appointment	33 33plus – 1 missing (don't know, refusal) – 2 not applicable
142	IN27	Tobacco consumed during pregnancy	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
143	IN28	Use of the next day contraception pill	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable

144	UN1a	Unmet need for health care in the past 12 months due to long waiting list(s)	1 Yes 2 No 3 No need – 1 missing (don't know, refusal) – 2 not applicable
145	UN1b	Unmet need for health care in the past 12 months due to distance or transportation problems	1 Yes 2 No 3 No need – 1 missing (don't know, refusal) – 2 not applicable
146	UN2a	Could not afford medical examination or treatment in the past 12 months	1 Yes 2 No 3 No need – 1 missing (don't know, refusal)
147	UN2b	Could not afford dental examination or treatment in the past 12 months	1 Yes 2 No 3 No need – 1 missing (don't know, refusal)
148	UN2c	Could not afford prescribed medicines in the past 12 months	1 Yes 2 No 3 No need – 1 missing (don't know, refusal)
149	UN2d	Could not afford mental health care (by a psychologist or a psychiatrist for example) in the past 12 months	1 Yes 2 No 3 No need – 1 missing (don't know, refusal)
150	BM1	Height without shoes	Number in cm

			– 1 missing (don't know, refusal)
151	BM2	Weight without clothes and shoes	Number in kg – 1 missing (don't know, refusal)
152	PE1	Physical effort of working tasks (both paid and unpaid work activities included)	1 Mostly sitting or standing 2 Mostly walking or tasks of moderate physical effort 3 Mostly heavy labour or physically demanding work 4 Not performing any working tasks – 1 missing (don't know, refusal)
153	PE2	Number of days in a typical week walking to get to and from places at least 10 minutes continuously	Number of days 1 – 7 0 I never carry out such physical activities – 1 missing (don't know, refusal)
154	PE3	Time spent on walking to get to and from places on a typical day	1 10 – 29 minutes per day 2 30 – 59 minutes per day 3 1 hour to less than 2 hours per day 4 2 hours to less than 3 hours per day 5 3 hours or more per day – 1 missing (don't know, refusal) – 2 not applicable
155	PE4	Number of days in a typical week bicycling to get to and from places at least 10 minutes continuously	Number of days 1 – 7 0 I never carry out such physical activities – 1 missing (don't know, refusal)

156	PE5	Time spent on bicycling to get to and from places on a typical day	1 10 – 29 minutes per day 2 30 – 59 minutes per day 3 1 hour to less than 2 hours per day 4 2 hours to less than 3 hours per day 5 3 hours or more per day – 1 missing (don't know, refusal) – 2 not applicable
157	PE6	Number of days in a typical week doing sports, fitness or recreational (leisure) physical activities that cause at least a small increase in breathing or heart rate for at least 10 minutes continuously	Number of days 1 – 7 0 I never carry out such physical activities – 1 missing (don't know, refusal)
158	PE7	Time spent on doing sports, fitness or recreational (leisure) physical activities in a typical week	HHMM (hours/minutes) – 1 missing (don't know, refusal) – 2 not applicable
159	PE8	Number of days in a typical week doing muscle-strengthening activities	Number of days 1 – 7 0 I never carry out such physical activities – 1 missing (don't know, refusal)
160	IN29	Number of meals a day	– 1 missing (don't know, refusal)
161	IN30	Milk derived products consumed the day before	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
162	IN31	Soup consumed the day before	2 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable

163	IN32	Bread consumed the day before	3 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
164	IN33	Meat consumed the day before	4 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
165	IN34	Fish consumed the day before	5 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
166	IN35	Potatoes rice or pasta consumed the day before	6 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
167	IN36	Beans consumed the day before	7 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
168	IN37	Chocolates or cake consumed the day before	8 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
169	IN38	Soda consumed the day before	9 Yes 2 No – 1 missing (don't know, refusal)

			– 2 not applicable
170	IN39	Natural juice consumed the day before	10 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
171	IN40	Other	11 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
172	IN41	Fast food consumed the day before	12 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
173	IN42	Frozen package food consumed the day before	13 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
174	FV1	Frequency of eating fruit, excluding juice	1 Once or more a day 2 4 to 6 times a week 3 1 to 3 times a week 4 Less than once a week
175	FV2	Number of portions of fruit a day, excluding juice	Number 1 – 99 – 1 missing (don't know, refusal) – 2 not applicable
176	FV3	Frequency of eating vegetables or salad, excluding juice and potatoes	1 Once or more a day 2 4 to 6 times a week 3 1 to 3 times a week

			4 Less than once a week 5 Never – 1 missing (don't know, refusal)
177	FV4	Number of portions of vegetables or salad, excluding juice and potatoes a day	Number 1 – 99 – 1 missing (don't know, refusal) – 2 not applicable
178	IN43	Tobbacco status	1 Everyday 2 Occasionally 3 Was a smoker once 4 Not a smoker -1 missing
179	in44	type of tobacco	2 Packaged cigarettes 2 Handmade cigarettes 3 Charutos 4 Cigarilhas 5 Pipe 7 Other -1 missing -2 not applicable
180	SK3	Average number of cigarettes a day	Number 1 – 99 – 1 missing (don't know, refusal) – 2 not applicable
181	IN45	Age quit smoking	Number 1 – 99
182	IN46	Type of help to quit	– 1 missing (don't know, refusal) – 2 not applicable
183	IN47	Age started smoking in a daily basis	Number 1 – 99
184	IN48	Use of ecigarette	– 1 missing (don't know, refusal) – 2 not applicable

185	SK4	Frequency of exposure to tobacco smoke indoors	1 Never or almost never 2 Less than 1 hour per day 3 1 hour or more a day – 1 missing (don't know, refusal)
186	IN49	Place of exposure to smokehand exposure	
187	AL1	Frequency of consumption of an alcoholic drink of any kind (beer, wine, cider, spirits, cocktails, premixes, liqueurs, homemade alcohol…) in the past 12 months	1 Every day or almost 2 5 – 6 days a week 3 3 – 4 days a week 4 1 – 2 days a week 5 2 – 3 days in a month 6 Once a month 7 Less than once a month 8 Not in the past 12 months, as I no longer drink alcohol 9 Never, or only a few sips or tries, in my whole life – 1 missing (don't know, refusal)
188	AL2	Frequency of consumption of an alcoholic drink for Monday — Thursday	1 On all 4 days 2 On 3 of the 4 days 3 On 2 of the 4 days 4 On 1 of the 4 days 5 On none of the 4 days – 1 missing (don't know, refusal) – 2 not applicable
189	AL3	Number of alcoholic (standard) drinks on average on one of the days (Monday to Thursday)	1 16 or more drinks a day 2 10 – 15 drinks a day 3 6 – 9 drinks a day 4 4 – 5 drinks a day 5 3 drinks a day 6 2 drinks a day

			7 1 drink a day 8 0 drink a day – 1 missing (don’t know, refusal) – 2 not applicable
190	AL4	Frequency of consumption of an alcoholic drink for Friday — Sunday	1 On all 3 days 2 On 2 of the 3 days 3 On 1 of the 3 days 4 On none of the 3 days – 1 missing (don’t know, refusal) – 2 not applicable
191	AL5	Number of alcoholic (standard) drinks on average on one of the days (Friday — Sunday)	1 16 or more drinks a day 2 10 – 15 drinks a day 3 6 – 9 drinks a day 4 4 – 5 drinks a day 5 3 drinks a day 6 2 drinks a day 7 1 drink a day 8 0 drink a day – 1 missing (don’t know, refusal) – 2 not applicable
192	AL6	Frequency of risky single-occasion drinking (equivalent of 60 g of pure ethanol or more) during the past 12 months	1 Every day or almost 2 5 – 6 days a week 3 3 – 4 days a week 4 1 – 2 days a week 5 2 – 3 days in a month 6 Once a month 7 Less than once a month 8 Not in the past 12 months 9 Never in my whole life

			– 1 missing (don't know, refusal) – 2 not applicable
193	IN50	In most ways my life is close to my ideal.	7 - Strongly agree 6 - Agree 5 - Slightly agree 4 - Neither agree nor disagree 3 - Slightly disagree 2 - Disagree 1 - Strongly disagree -1 Missing
194	IN51	The conditions of my life are excellent.	7 - Strongly agree 6 - Agree 5 - Slightly agree 4 - Neither agree nor disagree 3 - Slightly disagree 2 - Disagree 1 - Strongly disagree -1 Missing
195	IN52	I am satisfied with my life.	7 - Strongly agree 6 - Agree 5 - Slightly agree 4 - Neither agree nor disagree 3 - Slightly disagree 2 - Disagree 1 - Strongly disagree -1 Missing

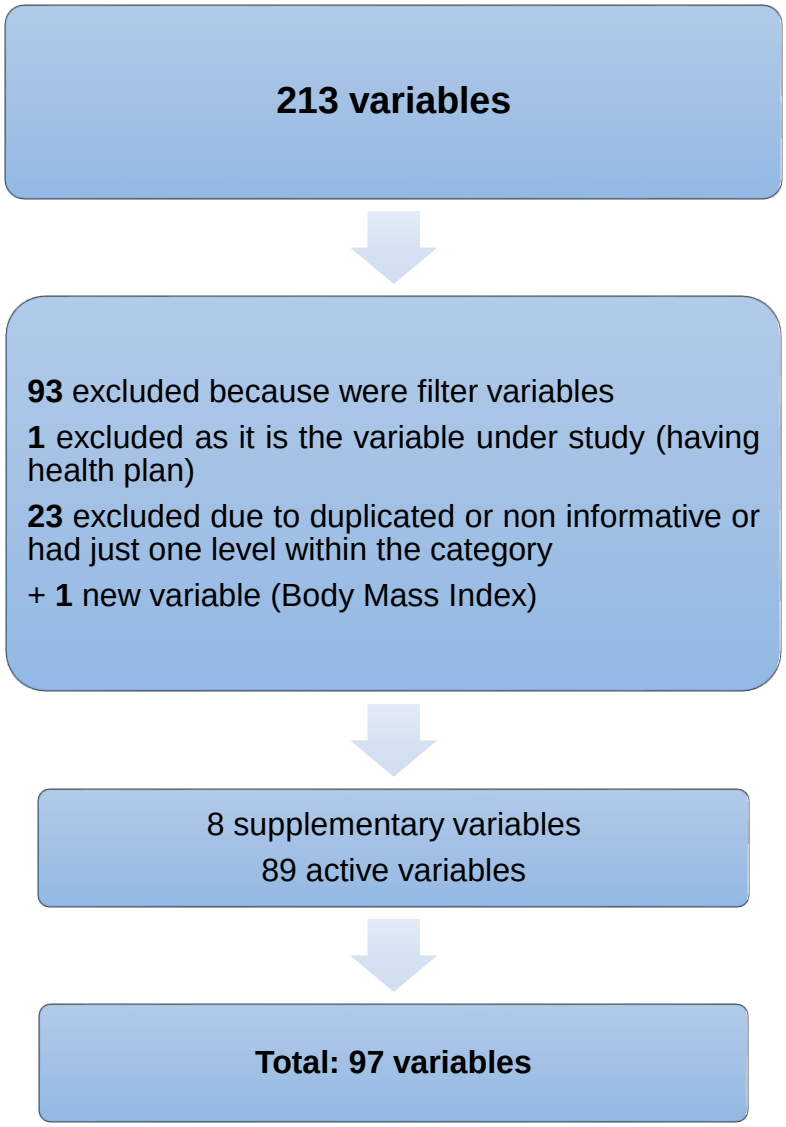
196	IN53	So far I have gotten the important things I want in life.	7 - Strongly agree 6 - Agree 5 - Slightly agree 4 - Neither agree nor disagree 3 - Slightly disagree 2 - Disagree 1 - Strongly disagree -1 Missing
197	IN54	If I could live my life over, I would change almost nothing.	7 - Strongly agree 6 - Agree 5 - Slightly agree 4 - Neither agree nor disagree 3 - Slightly disagree 2 - Disagree 1 - Strongly disagree -1 Missing
198	SS1	Number of close people to count on in case of serious personal problems	1 None 2 1 or 2 3 3 to 5 4 6 or more – 1 missing (don't know, refusal)
199	SS2	Degree of concern shown by other people in what the person is doing	1 A lot of concern and interest 2 Some concern and interest 3 Uncertain 4 Little concern and interest 5 No concern and interest – 1 missing (don't know, refusal)

200	SS3	How easy is it to get practical help from neighbors in case of need	1 Very easy 2 Easy 3 Possible 4 Difficult 5 Very difficult – 1 missing (don't know, refusal)
201	IC1	Providing care or assistance to one or more persons suffering from some age problem, chronic health condition or infirmity, at least once a week (professional activities excluded)	1 Yes 2 No – 1 missing (don't know, refusal)
202	IC2	Prevailing relationship of the person(s) suffering from any chronic condition or infirmity or due to old age being provided with care or assistance at least once a week from the respondent	1 Member(s) of respondent's family 2 Non-member(s) of respondent's family – 1 missing (don't know, refusal) – 2 not applicable
203	IC3	Number of hours per week the respondent provides care or assistance to the person(s) suffering from any chronic condition or infirmity or due to old age	1 Less than 10 hours per week 2 At least 10 but less than 20 hours per week 3 20 hours per week or more – 1 missing (don't know, refusal) – 2 not applicable
204	IN10	Are you limited to your house?	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable

205	IN11	Limited being sit on a chair	1 Yes 2 No 3. wheelchair – 1 missing (don't know, refusal) – 2 not applicable
206	IN12	Are you bedridden	1 Yes 2 No – 1 missing (don't know, refusal) – 2 not applicable
207	IN13	Extent to which the bedridden is dependent on other	1 Alone,without difficulties 2 Alone with difficulties 3. Only with help from others – 1 missing (don't know, refusal) – 2 not applicable
208	IN14	Cause for the long term incapacity	1 Road accident, excluding occupational accidents 2 Road accident during occupational time 3. Road accident from work/house 4. Occupational accident, excluding road accident 5.domestic accident 6 Accident during leisure 7Accident in school 8Ilness, exclude occupational 9 Occupational 10 Other – 1 missing (don't know, refusal) – 2 not applicable

209	IN55	Expenses with medical appointments in the last 2 weeks of the household income	– 1 missing (don't know, refusal) – 2 not applicable
210	IN56	Expenses with diagnostics in the last 2 weeks of the household income	– 1 missing (don't know, refusal) – 2 not applicable
211	IN57	Expenses with medicines in the last 2 weeks of the household income	– 1 missing (don't know, refusal) – 2 not applicable
212	IN58	Expenses with surgeries and treatments in the last 2 weeks of the household income	– 1 missing (don't know, refusal) – 2 not applicable
213	IN59	Expenses with other treatments in the last 2 weeks of the household income	– 1 missing (don't know, refusal) – 2 not applicable

S2 File. Fluxogram regarding variable selection for the analysis.



.

S3 File. Descriptive analysis of the 97 variables.

VARIABLE	CATEGORIES	DESCRIPTION	COUNT	%
REGION	NUTS.1	North	1835	15.12
	NUTS.2	Center	1540	12.69
	NUTS.3	Lisbon and Tagus Valley	2520	20.77
	NUTS.4	Alentejo	1053	8.68
	NUTS.5	Algarve	2073	17.08
	NUTS.6	Madeira	1489	12.27
	NUTS.7	Azores	1624	13.38
DEGREE OF URBANIZATION	DEG_URB.1	Densely-populated area	3265	26.91
	DEG_URB.2	Intermediate-populated area	3934	32.42
	DEG_URB.3	Thinly populated area	4935	40.67
SEX	SEX.1	Male	5157	42.50
	SEX.2	Female	6977	57.50
AGE GROUP	GE.1	15-19	497	4.10
	GE.2	20-24	470	3.87
	GE.3	25-29	439	3.62
	GE.4	30-34	612	5.04
	GE.5	35-39	828	6.82
	GE.6	40-44	875	7.21
	GE.7	45-49	875	7.21
	GE.8	50-54	986	8.13
	GE.9	55-59	1003	8.27
	GE.10	60-64	1096	9.03
	GE.11	65-69	1185	9.77
	GE.12	70-74	1031	8.50
	GE.13	75-79	1020	8.41
	GE.14	80-84	772	6.36
	GE.15	85+	445	3.67

LEGAL MARITAL STATUS	MARSTALEGAL.1	Never married and never been in a registered partnership	2812	23.17
	MARSTALEGAL.2	Married or in a registered partnership	6273	51.70
	MARSTALEGAL.3	Widowed or in registered partnership that ended with death of partner (not remarried or in new registered partnership)	2002	16.50
	MARSTALEGAL.4	Divorced or in registered partnership that was legally dissolved (not remarried or in new registered partnership)	1047	8.63
DE FACTO MARITAL STATUS	MARSTADEFACTO.1	Person living in a consensual union	718	5.92
	MARSTADEFACTO.2	Person not living in a consensual union	11416	94.08
ALWAYS LIVED IN PORTUGAL	IN1.1	Yes	10127	83.46
	IN1.2	No	2007	16.54
EDUCATIONAL LEVEL	HATLEVEL.0	None	1872	15.43
	HATLEVEL.1	Primary education	5595	46.11
	HATLEVEL.2	Lower secondary education	1937	15.96
	HATLEVEL.3	Upper secondary education	1526	12.58

	HATLEVEL.4	Tertiary education; bachelor level or equivalent	1204	9.92
SELF-DECLARED LABOUR STATUS	MAINSTAT.10	Carries out a job or profession, including unpaid work for a family business or holding, an apprenticeship or paid traineeship	4371	36.02
	MAINSTAT.20	Unemployed	1471	12.12
	MAINSTAT.31	Pupil, student, further training, unpaid work experience	651	5.37
	MAINSTAT.32	In retirement or early retirement or has given up business	4450	36.67
	MAINSTAT.33	Permanently disabled and Other inactive person	347	2.86
	MAINSTAT.35	Fulfilling domestic tasks	844	6.96
NUMBER OF PERSONS LIVING IN HOUSEHOLD, INCLUDING THE RESPONDENT	HHNBPERS.1	1	2865	23.61
	HHNBPERS.2	2	4352	35.87
	HHNBPERS.3	3	2532	20.87
	HHNBPERS.4	4	1678	13.83
	HHNBPERS.5	5	467	3.85
	HHNBPERS.6	6+	240	1.98
NUMBER OF PERSONS AGED 4 OR YOUNGER	HHNBPERS_0_4.0	0	11249	92.71
	HHNBPERS_0_4.1	1+	885	7.29
NUMBER OF PERSONS AGED 5-13	HHNBPERS_5_13.0	0	10149	83.64
	HHNBPERS_5_13.1	1	1573	12.96
	HHNBPERS_5_13.2	2+	412	3.40
	HHNBPERS_14_15.0	0	11499	94.77

NUMBER OF PERSONS AGED 14-15	HHNBPERS_14_15.1	1+	635	5.23
NUMBER OF PERSONS AGED 16-24	HHNBPERS_16_24.0	0	9908	81.65
	HHNBPERS_16_24.1	1	1765	14.55
	HHNBPERS_16_24.2	2+	461	3.80
NUMBER OF PERSONS AGED 25-64	HHNBPERS_25_64.0	0	3620	29.83
	HHNBPERS_25_64.1	1	2873	23.68
	HHNBPERS_25_64.2	2	4655	38.36
	HHNBPERS_25_64.3	3+	986	8.13
NUMBER OF PERSONS AGED 65+	HHNBPERS_65PLUS.0	0	6768	55.78
	HHNBPERS_65PLUS.1	1	3188	26.27
	HHNBPERS_65PLUS.2	2+	2178	17.95
TYPE OF HOUSEHOLD	HHTYPE.10	One-person household	2865	23.61
	HHTYPE.21	Lone parent with child(ren) aged less than 25	547	4.51
	HHTYPE.22	Couple without child (ren) aged less than 25	3366	27.74
	HHTYPE.23	Couple with child (ren) aged less than 25	2758	22.73
	HHTYPE.24	Couple or lone parent with child(ren) aged less than 25 and other persons living in household	346	2.85
	HHTYPE.25	Other type of household	2252	18.56
NUMBER OF PERSONS AGED 16 – 64 IN THE HOUSEHOLD WHO ARE EMPLOYED	HH_ACT.0	0	5740	47.31
	HH_ACT.1	1	3169	26.12
	HH_ACT.2	2	2729	22.49
	HH_ACT.3	3+	526	4.09
NUMBER OF PERSONS AGED 16 – 64 IN THE	HH_INACT.0	0	6706	55.27
	HH_INACT.1	1	3499	28.84

HOUSEHOLD WHO ARE UNEMPLOYED OR ARE ECONOMICALLY INACTIVE	HH_INACT.2	2	1486	12.25
	HH_INACT.3	3+	443	3.65
NET MONTHLY EQUIVALISED INCOME OF THE HOUSEHOLD	HHINCOME.1	Below first quintile	3056	25.19
	HHINCOME.2	Between 1st quintile and 2nd quintile	2795	23.03
	HHINCOME.3	Between 3rd quintile and 4th quintile	2494	20.55
	HHINCOME.4	Between 4th quintile and 5th quintile	2170	17.88
	HHINCOME.5	Between 4th quintile and 5th quintile	1619	13.34
SELF PERCEIVED GENERAL HEALTH	HS1.1	Very good	1076	8.87
	HS1.2	Good	3688	30.39
	HS1.3	Fair	5096	42.00
	HS1.4	Bad	1721	14.18
	HS1.5	Vey bad	553	4.56
LONG-STANDING HEALTH PROBLEM:	HS2.1	Yes	7919	65.26
SUFFER FROM ANY ILLNESS OR HEALTH PROBLEM OF A DURATION OF AT LEAST SIX MONTHS	HS2.2	No	4215	34.74
GENERAL ACTIVITY LIMITATION:	HS3.1	Severely limited	1269	10.46
LIMITATION IN ACTIVITIES PEOPLE USUALLY DO BECAUSE OF HEALTH PROBLEMS FOR AT LEAST THE PAST SIX MONTHS	HS3.2	Limted	3581	29.51
	HS3.3	Nit at all	7284	60.03
	CD1A.1	Yes	677	5.58

				94.42
SUFFERING FROM ASHTMA IN THE PAST 12 MONTHS	CD1A.2	No	11457	
SUFFERING FROM CHRONIC BRONCHITIS, CHRONIC OBSTRUCTIVE PULMONARY DISEASE OR EMPHYSEMA IN THE PAST 12 MONTHS	CD1B.1	Yes	884	7.29
	CD1B.2	No	11250	92.71
SUFFERING FROM A MYOCARDIAL INFARCTION (HEART ATTACK) IN THE PAST 12 MONTHS	CD1C.1	Yes	306	2.52
	CD1C.2	No	11828	97.48
SUFFERING FROM A CORONARY HEART DISEASE OR ANGINA PECTORIS IN THE PAST 12 MONTHS	CD1D.1	Yes	783	6.45
	CD1D.2	No	11351	93.55
SUFFERING FROM HIGH BLOOD PRESSURE IN THE PAST 12 MONTHS	CD1E.1	Yes	4066	33.51
	CD1E.2	No	8068	66.49
SUFFERING FROM A STROKE (CEREBRAL HAEMORRHAGE, CEREBRAL THROMBOSIS) IN THE PAST 12 MONTHS	CD1F.1	Yes	347	2.86
	CD1F.2	No	11787	97.14
SUFFERING FROM ARTHROSIS (ARTHRITIS EXCLUDED) IN THE PAST 12 MONTHS	CD1G.1	Yes	4004	33.00
	CD1G.2	No	8130	67.00
SUFFERING FROM A LOW BACK DISORDER OR OTHER CHRONIC BACK DEFECT IN THE PAST 12 MONTHS	CD1H.1	Yes	4807	39.62
	CD1H.2	No	7327	60.38

SUFFERING FROM A NECK DISORDER OR OTHER CHRONIC NECK DEFECT IN THE PAST 12 MONTHS	CD1I.1	Yes	3610	29.75
	CD1I.2	No	8524	70.25
SUFFERING FROM DIABETES IN THE PAST 12 MONTHS	CD1J.1	Yes	1605	13.23
	CD1J.2	No	10529	86.77
SUFFERING FROM AN ALLERGY, SUCH AS RHINITIS, EYE INFLAMMATION, DERMATITIS, FOOD ALLERGY OR OTHER (ALLERGIC ASTHMA EXCLUDED) IN THE PAST 12 MONTHS	CD1K.1	Yes	2215	18.25
	CD1K.2	No	9919	81.75
SUFFERING FROM URINARY INCONTINENCE, PROBLEMS IN CONTROLLING THE BLADDER IN THE PAST 12 MONTHS	CD1M.1	Yes	1163	9.58
	CD1M.2	No	10971	90.42
SUFFERING FROM KIDNEY PROBLEMS IN THE PAST 12 MONTHS	CD1N.1	Yes	834	6.87
	CD1N.2	No	11300	93.13
SUFFERING FROM DEPRESSION IN THE PAST 12 MONTHS	CD1O.1	Yes	1775	14.63
	CD1O.2	No	10359	85.37
OCCURRENCE OF AN ACCIDENT AT HOME IN THE PAST 12 MONTHS	AC1B.1	Yes	461	3.80
	AC1B.2	No	11673	96.20
WEARING GLASSES OR CONTACT LENSES	PL1.1		7525	62.02
	PL1.2		4609	37.98
USE OF A HEARING AID	PL3.1		332	2.74
	PL3.2		11802	97.26
DIFFICULTY SPEAKING	IN8.1		11888	97.97
	IN8.2		246	2.03
DIFFICULTY IN WALKING HALF A KM ON LEVEL	PL6.1	No difficulty	10139	83.56
	PL6.2	Some difficulty	981	8.08

GROUND WITHOUT THE USE OF ANY AID	PL6.3	A lot of difficulty	717	5.91
	PL6.4	Cannot at all	297	2.45
DIFFICULTY IN WALKING UP OR DOWN 12 STEPS	PL7.1	No difficulty	9805	80.81
	PL7.2	Some difficulty	1281	10.56
	PL7.3	A lot of difficulty	794	6.54
	PL7.4	Cannot at all	254	2.09
INTENSITY OF BODILY PAIN DURING THE PAST 4 WEEKS	PN1.1	None	5094	41.98
	PN1.2	Very mild	1151	9.49
	PN1.3	Mild	1793	14.78
	PN1.4	Moderate	1948	16.05
	PN1.5	Severe	1616	13.32
	PN1.6	Very severe	532	4.38
EXTENT THAT PAIN INTERFERED WITH NORMAL WORK DURING THE PAST 4 WEEKS (INCLUDING BOTH WORK OUTSIDE THE HOME AND HOUSEWORK)	PN2.1	Not at all	6934	57.15
	PN2.2	A little bit	2281	18.80
	PN2.3	Moderately	1410	11.62
	PN2.4	Quite a bit	1206	9.94
	PN2.5	Extremely	303	2.50
EXTENT OF HAVING LITTLE INTEREST OR PLEASURE IN DOING THINGS OVER THE LAST 2 WEEKS	MH1A.1	Not at all	7958	65.58
	MH1A.2	Several days	2932	24.16
	MH1A.3	More than half of the days	538	4.43
	MH1A.4	Almost every day	706	5.82
EXTENT OF FEELING DOWN, DEPRESSED OR HOPELESS OVER THE LAST 2 WEEKS	MH1B.1	Not at all	7684	63.33
	MH1B.2	Several days	3155	26.00
	MH1B.3	More than half of the days	571	4.71
	MH1B.4	Almost every day	724	5.97
EXTENT OF HAVING TROUBLE FALLING OR STAYING ASLEEP, OR SLEEPING TOO	MH1C.1	Not at all	6803	56.07
	MH1C.2	Several days	3125	25.75
	MH1C.3	More than half of the days	780	6.43

MUCH OVER THE LAST 2 WEEKS	MH1C.4	Almost every day	1426	11.75
EXTENT OF FEELING TIRED OR HAVING LITTLE ENERGY OVER THE LAST 2 WEEKS	MH1D.1	Not at all	5994	49.40
	MH1D.2	Several days	4210	34.70
	MH1D.3	More than half of the days	810	6.68
	MH1D.4	Almost every day	1120	9.23
EXTENT OF HAVING POOR APPETITE OR OVEREATING OVER THE LAST 2 WEEKS	MH1E.1	Not at all	9841	81.10
	MH1E.2	Several days	1652	13.61
	MH1E.3	More than half of the days	299	2.46
	MH1E.4	Almost every day	342	2.82
EXTENT OF FEELING BAD ABOUT YOURSELF, FEELING BEING A FAILURE OVER THE LAST 2 WEEKS	MH1F.1	Not at all	9221	75.99
	MH1F.2	Several days	2132	17.57
	MH1F.3	More than half of the days	402	3.31
	MH1F.4	Almost every day	379	3.12
EXTENT OF HAVING TROUBLE CONCENTRATING ON THINGS, SUCH AS READING THE NEWSPAPER OR WATCHING TELEVISION, OVER THE LAST 2 WEEKS	MH1G.1	Not at all	9731	80.20
	MH1G.2	Several days	1759	14.50
	MH1G.3	More than half of the days	329	2.71
	MH1G.4	Almost every day	315	2.60
EXTENT OF MOVING OR SPEAKING SO SLOWLY THAT OTHER PEOPLE COULD HAVE NOTICED OR BEING SO FIDGETY OR RESTLESS, OVER THE LAST 2 WEEKS	MH1H.1	Not at all	10373	85.49
	MH1H.2	Several days	1322	10.90
	MH1H.3	More than half of the days	439	3.62
ADMISSION AS AN INPATIENT IN A HOSPITAL IN THE PAST 12 MONTHS	HO1.1	Yes	1179	9.72
	HO1.2	No	10955	90.28
	HO3.1	Yes	5121	42.20

ADMISSION AS A DAY PATIENT IN A HOSPITAL IN THE PAST 12 MONTHS	HO3.2	No	7013	57.80
LAST APPOINTMENT WITH DENTIST	AM1.1	Less than 6 months	3026	24.94
	AM1.2	6-12 mo	1795	14.79
	AM1.3	+12	6902	56.88
	AM1.4	Never	411	3.39
FREQUENCY BRUSHING TEETH	IN18.1	Everyday in the morning, after lunch and before bed	2526	20.82
	IN18.2	Everyday in the morning and before bed	5009	41.28
	IN18.3	Everyday in the morning	1400	11.54
	IN18.4	Everyday before bed	1473	12.14
	IN18.5	A few times per week/Less than 1*	983	8.10
	IN18.6	Never	743	6.12
LAST APPOINTMENT WITH FAMILY DOCTOR	AM2.1	Up to 12 months	9139	75.32
	AM2.2	+12/Never	2995	24.68
LAST APPOINTMENT WITH SPECIALIST DOCTOR	AM4.1	Less than 12	5430	44.75
	AM4.2	12+	5633	46.42
	AM4.3	Never	1071	8.83
CONSULTATION WITH PHISIOTERAPIST IN THE PAST 12 MONTHS	AM6A.1	Yes	1148	9.46
	AM6A.2	No	10986	90.54
CONSULTATION OF A PSYCHOLOGIST OR PSYCHOTHERAPIST IN THE PAST 12 MONTHS	AM6B.1	Yes	690	5.69
	AM6B.2	No	11444	94.31
	AM7.1	Yes	378	3.12

USE OF ANY HOME CARE SERVICES FOR PERSONAL NEEDS DURING THE PAST 12 MONTHS	AM7.2	No	11756	96.88
USE OF ANY MEDICINES PRESCRIBED BY A DOCTOR DURING THE PAST TWO WEEKS (EXCLUDING CONTRACEPTION)	MD1.1	Yes	7758	63.94
	MD1.2	No	4376	36.06
USE OF ANY MEDICINES, HERBAL MEDICINES OR VITAMINS NOT PRESCRIBED BY A DOCTOR DURING THE PAST TWO WEEKS	MD1.1.1 (MD2.1)	Yes	2744	22.61
	MD1.1.2 (MD2.2)	No	9390	77.39
LAST TIME OF VACCINATION AGAINST TETANUS.	IN19.1	Less than 10 years	9180	75.66
	IN19.2	More than 10 years	2420	19.94
	IN19.3	Never	534	4.40
LAST TIME OF BLOOD PRESSURE MEASUREMENT BY A HEALTH PROFESSIONAL	PA2.1	Up to 12 months	9477	78.10
	PA2.2	1 to 3 y	1900	15.66
	PA2.3	+3 y	400	3.30
	PA2.4	Never	357	2.94
LAST TIME OF BLOOD CHOLESTEROL MEASUREMENT BY A HEALTH PROFESSIONAL	PA3.1	Up to 12 months	8489	69.96
	PA3.2	1 to 3 y	2391	19.70
	PA3.3	+3 y	502	4.14
	PA3.4	Never	752	6.20
LAST TIME OF BLOOD SUGAR MEASUREMENT BY A HEALTH PROFESSIONAL	PA4.1	Up to 12 months	8395	69.19
	PA4.2	1 to 3 y	2374	19.56
	PA4.3	+3 y	523	4.31
	PA4.4	Never	842	6.94
	PA5.1	Up to 1 y	1391	11.46

LAST TIME OF A FAECAL OCCULT BLOOD TEST	PA5.2	1-2y	817	6.73
	PA5.3	2-3y	395	3.26
	PA5.4	+3y	926	7.63
	PA5.5	Never	8605	70.92
LAST TIME OF A COLONOSCOPY	PA6.1	Up to 1 y	547	4.51
	PA6.2	1-2y	1444	11.90
	PA6.3	2-3y	560	4.62
	PA6.4	+3y	265	2.18
	PA6.5	Never	9318	76.79
UNMET NEED FOR HEALTH CARE IN THE PAST 12 MONTHS DUE TO LONG WAITING LIST(S)	UN1A.1	Yes	2656	21.89
	UN1A.2	No	7577	62.44
	UN1A.3	No need	1901	15.67
UNMET NEED FOR HEALTH CARE IN THE PAST 12 MONTHS DUE TO DISTANCE OR TRANSPORTATION PROBLEMS	UN1B.1	Yes	373	3.07
	UN1B.2	No	9649	79.52
	UN1B.3	No need	2112	17.41
COULD NOT AFFORD MEDICAL EXAMINATION OR TREATMENT IN THE PAST 12 MONTHS	UN2A.1	Yes	1344	11.08
	UN2A.2	No	7582	62.49
	UN2A.3	No need	3208	26.44
COULD NOT AFFORD DENTAL EXAMINATION OR TREATMENT IN THE PAST 12 MONTHS	UN2B.1	Yes	2765	22.79
	UN2B.2	No	4070	33.54
	UN2B.3	No need	5299	43.67
COULD NOT AFFORD PRESCRIBED MEDICINES IN THE PAST 12 MONTHS	UN2C.1	Yes	1117	9.21
	UN2C.2	No	8177	67.39
	UN2C.3	No need	2840	23.41
COULD NOT AFFORD MENTAL HEALTH CARE (BY A PSYCHOLOGIST OR A PSYCHIATRIST FOR EXAMPLE) IN THE PAST 12 MONTHS	UN2D.1	Yes	347	2.86
	UN2D.2	No	725	5.97
	UN2D.3	No need	11062	91.17

PHYSICAL EFFORT OF WORKING TASKS (BOTH PAID AND UNPAID WORK ACTIVITIES INCLUDED)	PE1.1	Mostly sitting or standing	5700	46.98
	PE1.2	Mostly walking or tasks of moderate physical effort	4699	38.73
	PE1.3	Mostly heavy labour or physically demanding work	1088	8.97
	PE1.4	Not performing any working tasks	647	5.33
NUMBER OF DAYS IN A TYPICAL WEEK WALKING TO GET TO AND FROM PLACES AT LEAST 10 MINUTES CONTINUOUSLY	PE2.0	0 (Never)	4955	40.84
	PE2.1	1	441	3.63
	PE2.2	2	704	5.80
	PE2.3	3	672	5.54
	PE2.4	4	333	2.74
	PE2.5	5	1345	11.08
	PE2.6	6	335	2.76
	PE2.7	7	3349	27.60
NUMBER OF DAYS IN A TYPICAL WEEK BICYCLING TO GET TO AND FROM PLACES AT LEAST 10 MINUTES CONTINUOUSLY	PE4.0	0	11418	94.10
	PE4.1	1-3	427	3.74
	PE4.2	4+	289	2.53
NUMBER OF DAYS IN A TYPICAL WEEK DOING SPORTS, FITNESS OR RECREATIONAL (LEISURE) PHYSICAL ACTIVITIES THAT CAUSE AT LEAST A SMALL INCREASE IN BREATHING OR HEART RATE FOR AT LEAST 10 MINUTES CONTINUOUSLY	PE6.0	0	8673	71.48
	PE6.1	1	486	4.01
	PE6.2	2	876	7.22
	PE6.3	3	625	5.15
	PE6.4	4	1474	12.15
	PE8.0	0	11144	91.84

NUMBER OF DAYS IN A TYPICAL WEEK DOING MUSCLE-STRENGTHENING ACTIVITIES	PE8.1	1+	990	8.16
NUMBER OF MEALS A DAY	IN29.0	Up to 3	1445	11.91
	IN29.1	3+	10689	88.09
FREQUENCY OF EATING FRUIT, EXCLUDING JUICE	FV1.1	Once or more a day	8704	71.73
	FV1.2	4-6 week	1369	11.28
	FV1.3	1-3 week	1408	11.60
	FV1.4	Less than 1 week	653	5.38
FREQUENCY OF EATING VEGETABLES OR SALAD, EXCLUDING JUICE AND POTATOES	FV3.1	Once or more a day	6099	50.26
	FV3.2	4-6 week	2799	23.07
	FV3.3	1-3 week	2522	20.78
	FV3.4	Less than 1 week	714	5.88
TOBBACCO STATUS	IN43.1	Everyday	1878	15.48
	IN43.2	Occassionally	292	2.41
	IN43.3	Smoker once	2526	20.82
	IN43.4	Not smoker	7438	61.30
FREQUENCY OF EXPOSURE TO TOBACCO SMOKE INDOORS	SK4.1	Never	11321	93.30
	SK4.2	Lessthan 1h day	316	2.60
	SK4.3	1+ day	497	4.10
FREQUENCY OF CONSUMPTION OF AN ALCOHOLIC DRINK OF ANY KIND (BEER, WINE, CIDER, SPIRITS, COCKTAILS, PREMIXES, LIQUEURS, HOMEMADE ALCOHOL…) IN THE PAST 12 MONTHS	AL1.1	Every day or almost	3012	24.82
	AL1.2	5-6 days w	399	3.29
	AL1.3	3-4 days w	1242	10.24
	AL1.4	1-2 days w	671	5.53
	AL1.5	2 3 in month	757	6.24
	AL1.6	Once a month	1311	10.80
	AL1.7	Less than once	1382	11.39
	AL1.8	Not in the past 12 months/never	3360	27.69
	IN50.1	Stronlgy agree	814	6.71

IN MOST WAYS MY LIFE IS CLOSE TO MY IDEAL.	IN50.2	agree	1449	11.94
	IN50.3	Slightly Agree	852	7.02
	IN50.4	Neither agree or Disagree	1309	10.79
	IN50.5	Slightly disagree	3426	28.23
	IN50.6	Disagree	3762	31.00
	IN50.7	Strongly disagree	522	4.30
THE CONDITIONS OF MY LIFE ARE EXCELLENT.	IN51.1	Stronlgy agree	670	5.52
	IN51.2	agree	2072	17.08
	IN51.3	Slightly Agree	1234	10.17
	IN51.4	Neither agree or Disagree	1219	10.05
	IN51.5	Slightly disagree	3810	31.40
	IN51.6	Disagree	2705	22.29
	IN51.7	Strongly disagree	424	3.49
I AM SATISFIED WITH MY LIFE.	IN52.1	Stronlgy agree	415	3.42
	IN52.2	agree	1338	11.03
	IN52.3	Slightly Agree	876	7.22
	IN52.4	Neither agree or Disagree	936	7.71
	IN52.5	Slightly disagree	2993	24.67
	IN52.6	Disagree	4830	39.81
	IN52.7	Strongly disagree	746	6.15
SO FAR I HAVE GOTTEN THE IMPORTANT THINGS I WANT IN LIFE.	IN53.1	Stronlgy agree	267	2.20
	IN53.2	agree	1178	9.71
	IN53.3	Slightly Agree	808	6.66
	IN53.4	Neither agree or Disagree	935	7.71
	IN53.5	Slightly disagree	3139	25.87
	IN53.6	Disagree	4964	40.91
	IN53.7	Strongly disagree	843	6.95

IF I COULD LIVE MY LIFE OVER, I WOULD CHANGE ALMOST NOTHING.	IN54.1	Stronlgy agree	933	7.69
	IN54.2	agree	2555	21.06
	IN54.3	Slightly Agree	1399	11.53
	IN54.4	Neither agree or Disagree	862	7.10
	IN54.5	Slightly disagree	2091	17.23
	IN54.6	Disagree	3302	27.21
	IN54.7	Strongly disagree	992	8.18
NUMBER OF CLOSE PEOPLE TO COUNT ON IN CASE OF SERIOUS PERSONAL PROBLEMS	SS1.1	None	390	3.21
	SS1.2	1 or 2	5031	41.46
	SS1.3	3 to 5	4654	38.36
	SS1.4	6+	2059	16.97
DEGREE OF CONCERN SHOWN BY OTHER PEOPLE IN WHAT THE PERSON IS DOING	SS2.1	None	7135	58.80
	SS2.2	1 or 2	3747	30.88
	SS2.3	3 to 5	889	7.33
	SS2.4	6+	363	2.99
HOW EASY IS IT TO GET PRACTICAL HELP FROM NEIGHBORS IN CASE OF NEED	SS3.1	Very easy	1745	14.38
	SS3.2	Easy	4394	36.21
	SS3.3	Possible	3404	28.05
	SS3.4	Difficult	1771	14.60
	SS3.5	Very difficult	820	6.76
PROVIDING CARE OR ASSISTANCE TO ONE OR MORE PERSONS SUFFERING FROM SOME AGE PROBLEM, CHRONIC HEALTH CONDITION OR INFIRMITY, AT LEAST ONCE A WEEK (PROFESSIONAL ACTIVITIES EXCLUDED)	IC1.1	Yes	1530	12.61
	IC1.2	No	10604	87.39
BODY MASS INDEX	IMC_C.1	Up to 24.9	5026	41.38
	IMC_C.2	25-29.9	4777	39.37

IMC_C.3	+30	2331	19.21

S4 File. Map with supplementary categories and detailed analysis.

Supplementary variables

Number of persons living in household, including the respondent

Number of persons aged 4 or younger

Number of persons aged from 5 to 13

Number of persons aged from 13 to 15

Number of persons aged from 16 to 24

Number of persons aged from 25 to 64

Number of persons aged 65 plus

Type of Household

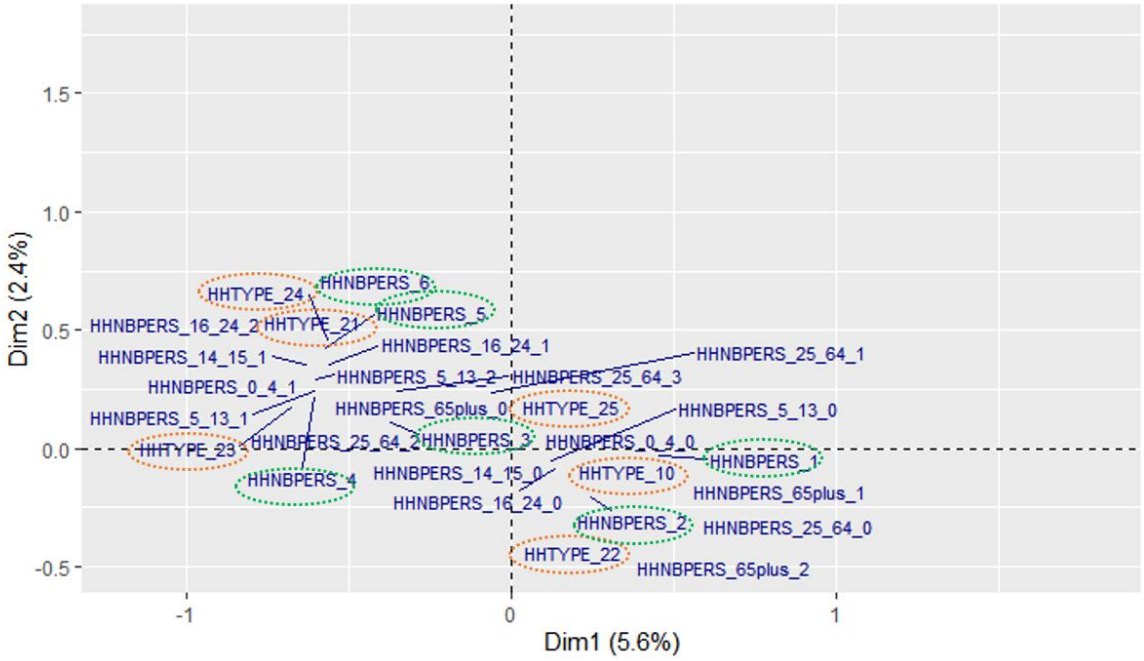


Fig 1. Map with supplementary categories and their labels on the first plane.

HHTYPE_10: One-person household; HHTYPE_21: Lone parent with child(ren) aged less than 25; HHTYPE_22: Couple without child(ren) aged less than 25; HHTYPE_23: Couple with child(ren) aged less than 25; HHTYPE_24: Couple or lone parent with child(ren) aged less than 25 and other persons living in household; HHTYPE_25: Other type of household.

For the following variables the last number in the acronym relates to the number of people: HHNBPERS: Number of persons living in household, including the respondent; HHNBPERS_0_4: Number of persons aged 4 or younger; HHNBPERS_5_13: Number of persons aged from 5 to 13; HHNBPERS_14_15: Number of persons aged from 13 to 15; HHNBPERS_16_24: Number of persons aged from 16 to 24; HHNBPERS_25_64: Number of persons aged from 25 to 64; HHNBPERS_65plus: Number of persons aged over 65.

In this plot we observe the location of supplementary variables. We find that with the variable HHTYPE (type of household), households with children are placed together, such as single parent with children less than 25 years old (HHTYPE=21), with a couple with children less than 25 years old (HHTYPE=23), couple or single parent with child(ren) aged less than 25 as well as other persons living in household (HHTYPE=24) were all in the same quadrant of the map. The closest variables – type of household 21 and 24 - mean these categories have similar distributions. We also observed that households with a couple without children aged less than 25 (HHTYPE=22) are positioned in an antagonist quadrant, meaning this category is inversely associated to the ones described above. Also, categories closer to the origin, such as one-person households (HHTYPE=10), means that this category is less different. Regarding the variable number of persons living in the household (HHNBPERS), we observe households with 3 or more people are in the same quadrant and close to the types 21, 24 e 23. which are inversely associated with households with a couple without children aged less than 25 (HHTYPE=22) and one-person (HHTYPE=10). This means that households with dependents are inversely associated with households without minors or any dependents. A similar analysis can be done with remaining variables and support the observation of a profile in these households. For

example, at the bottom left, close to households without child(ren) aged less than 25 is located the category HHNBPERS_2 (households with 2 people), HHNBPERS_16_24_0 (households without people 16 to 24 years old). At the top, in the left square, households with children are joined together (HHTYPE 21, 23 and 24), without residents over 64, but with a large variation of categories pointing out residents over 5 years old. Thus, it seems there are two large and distinct groups of households: younger, with children and older.

S5 FILE. Relative contribution of categories and cumulative percentage of variables to dimensions 1-5 construction.

			CONTRIBUTION DIMENSION 1		CONTRIBUTION DIMENSION 2		CONTRIBUTION DIMENSION 3		CONTRIBUTION DIMENSION 4		CONTRIBUTION DIMENSION 5	
	Variables	**Categories**	**Contribution (in permille)**	**Contribution (%)**	**Contribution (in permille)**	**Contribution (%)**	**Contribution (in permille)**	**Contribution (%)**	**Contribution (in permille)**	**Contribution (%)**	**Contribution (in permille)**	**Contribution (%)**
1		NUTS.1	0	0	1	0.1	2	0.2	0	0	0	0
2		NUTS.2	0	0	0	0	0	0	0	0	1	0.1
3		NUTS.3	0	0	2	0.2	0	0	1	0.1	0	0
4	REGION	NUTS.4	0	0	0	0	2	0.2	0	0	0	0
5		NUTS.5	0	0	0	0	5	0.5	1	0.1	0	0
6		NUTS.6	0	0	2	0.2	0	0	0	0	2	0.2
7		NUTS.7	0	0	0	0	0	0	0	0	0	0
			0	**0**	**5**	**0.5**	**9**	**0.9**	**2**	**0.2**	**3**	**0.3**
8		DEG_URB.1	0	0	0	0	6	0.6	0	0	0	0
9	DEG_URB	DEG_URB.2	0	0	0	0	1	0.1	0	0	0	0
10		DEG_URB.3	1	0.1	0	0	7	0.7	0	0	0	0
			0	**0.1**	**7**	**0.7**	**28**	**2.8**	**3**	**0.3**	**5**	**0.5**
11	SEX	SEX.1	3	0.3	0	0	6	0.6	0	0	15	1.5
12		SEX.2	2	0.2	0	0	4	0.4	0	0	11	1.1

			0	**0.5**	**0**	**0**	**10**	**1**	**0**	**0**	**26**	**2.6**
13		GE.1	4	0.4	4	0.4	2	0.2	13	1.3	75	7.5
14		GE.2	1	0.1	0	0	0	0	4	0.4	2	0.2
15		GE.3	1	0.1	4	0.4	1	0.1	0	0	1	0.1
16		GE.4	3	0.3	3	0.3	5	0.5	0	0	0	0
17		GE.5	4	0.4	2	0.2	10	1	3	0.3	1	0.1
18		GE.6	4	0.4	1	0.1	12	1.2	3	0.3	3	0.3
19		GE.7	3	0.3	0	0	12	1.2	5	0.5	3	0.3
20	AGE GROUP	GE.8	3	0.3	3	0.3	3	0.3	2	0.2	17	1.7
21		GE.9	2	0.2	1	0.1	3	0.3	0	0	1	0.1
22		GE.10	2	0.2	1	0.1	3	0.3	0	0	0	0
23		GE.11	3	0.3	1	0.1	3	0.3	0	0	1	0.1
24		GE.12	2	0.2	1	0.1	3	0.3	1	0.1	3	0.3
25		GE.13	1	0.1	1	0.1	2	0.2	2	0.2	6	0.6
26		GE.14	0	0	0	0	3	0.3	4	0.4	6	0.6
27		GE.15	0	0	0	0	1	0.1	4	0.4	6	0.6
			33	**3.3**	**22**	**2.2**	**63**	**6.3**	**41**	**4.1**	**125**	**12.5**
28		MARSTALEGAL.1	7	0.7	12	1.2	6	0.6	2	0.2	29	2.9
29	LEGAL MARITAL STATUS	MARSTALEGAL.2	0	0	5	0.5	0	0	1	0.1	17	1.7
30		MARSTALEGAL.3	9	0.9	1	0.1	15	1.5	1	0.1	8	0.8

31		MARSTALEGAL.4	0	0	1	0.1	3	0.3	5	0.5	7	0.7
			9	**0.9**	**7**	**0.7**	**18**	**1.8**	**7**	**0.7**	**32**	**3.2**
32	DE FACTO MARITAL STATUS	MARSTADEFACTO.1	1	0.1	2	0.2	3	0.3	1	0.1	1	0.1
33		MARSTADEFACTO.2	0	0	0	0	0	0	0	0	0	0
				1	**9**	**0.9**	**21**	**2.1**	**8**	**0.8**	**33**	**3.3**
34	ALWAYS LIVED IN PORTUGAL	IN1.1	0	0	0	0	0	0	0	0	1	0.1
35		IN1.2	1	1	2	0.2	2	0.2	0	0	5	0.5
			0.1	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
36	EDUCATIONAL LEVEL	HATLEVEL.0	9	0.9	0	0	25	2.5	3	0.3	7	0.7
37		HATLEVEL.1	1	0.1	0	0	3	0.3	4	0.4	9	0.9
38		HATLEVEL.2	3	0.3	2	0.2	5	0.5	0	0	3	0.3
39		HATLEVEL.3	4	0.4	0	0	10	1	0	0	1	0.1
40		HATLEVEL.4	3	0.3	2	0.2	15	1.5	2	0.2	0	0
			20	**2**	**4**	**0.4**	**58**	**5.8**	**9**	**0.9**	**20**	**2**
41	SELF-DECLARED LABOUR STATUS	MAINSTAT.10	8	0.8	0	0	14	1.4	2	0.2	16	1.6
42		MAINSTAT.20	1	0.1	13	1.3	1	0.1	8	0.8	1	0.1
43		MAINSTAT.31	5	0.5	4	0.4	4	0.4	14	1.4	87	8.7
44		MAINSTAT.32	12	1.2	10	1	25	2.5	3	0.3	0	0
45		MAINSTAT.33	2	0.2	3	0.3	0	0	2	0.2	0	0
46		MAINSTAT.35	1	0.1	0	0	0	0	3	0.3	2	0.2

			29	2.9	30	3	44	4.4	32	3.2	106	10.6
47	NUMBER OF PERSONS LIVING IN HOUSEHOLD, INCLUDING THE RESPONDENT	HHNBPERS.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
48		HHNBPERS.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
49		HHNBPERS.3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
50		HHNBPERS.4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
51		HHNBPERS.5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
52		HHNBPERS.6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
53	NUMBER OF PERSONS AGED 4 OR YOUNGER	HHNBPERS_0_4.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
54		HHNBPERS_0_4.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
55	NUMBER OF PERSONS AGED 5-13	HHNBPERS_5_13.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
56		HHNBPERS_5_13.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
57		HHNBPERS_5_13.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
58	NUMBER OF PERSONS AGED 14-15	HHNBPERS_14_15.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
59		HHNBPERS_14_15.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
60	NUMBER OF PERSONS AGED 16-24	HHNBPERS_16_24.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
61		HHNBPERS_16_24.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

62		HHNBPERS_16_24.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
63	NUMBER OF PERSONS AGED 25-64	HHNBPERS_25_64.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
64		HHNBPERS_25_64.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
65		HHNBPERS_25_64.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
66		HHNBPERS_25_64.3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
67	NUMBER OF PERSONS AGED 65+	HHNBPERS_65PLUS.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
68		HHNBPERS_65PLUS.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
69		HHNBPERS_65PLUS.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
70	TYPE OF HOUSEHOLD	HHTYPE.10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
71		HHTYPE.21	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
72		HHTYPE.22	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
73		HHTYPE.23	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
74		HHTYPE.24	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
75		HHTYPE.25	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
76	NUMBER OF PERSONS AGED 16 – 64 IN THE HOUSEHOLD	HH_ACT.0	9	0.9	1	0.1	18	1.8	0	0	2	0.2
77		HH_ACT.1	2	0.2	3	0.3	6	0.6	3	0.3	1	0.1
78		HH_ACT.2	7	0.7	0	0	10	1	1	0.1	0	0

79	WHO ARE EMPLOYED	HH_ACT.3	1	0.1	0	0	1	0.1	0	0	0	0
			19	**1.9**	**4**	**0.4**	**35**	**3.5**	**4**	**0.4**	**3**	**0.3**
80	NUMBER OF PERSONS AGED 16 – 64 IN THE HOUSEHOLD WHO ARE UNEMPLOYED OR ARE ECONOMICALLY INACTIVE	HH_INACT.0	1	0.1	6	0.6	4	0.4	2	0.2	0	0
81		HH_INACT.1	1	0.1	3	0.3	4	0.4	2	0.2	1	0.1
82		HH_INACT.2	0	0	3	0.3	2	0.2	1	0.1	1	0.1
83		HH_INACT.3	1	0.1	3	0.3	0	0	0	0	7	0.7
			3	**0.3**	**15**	**1.5**	**10**	**1**	**5**	**0.5**	**9**	**0.9**
84	NET MONTHLY EQUIVALISED INCOME OF THE HOUSEHOLD	HHINCOME.1	0	0	9	0.9	2	0.2	2	0.2	3	0.3
85		HHINCOME.2	1	0.1	0	0	3	0.3	0	0	0	0
86		HHINCOME.3	0	0	0	0	0	0	0	0	0	0
87		HHINCOME.4	1	0.1	2	0.2	2	0.2	0	0	2	0.2
88		HHINCOME.5	2	0.2	6	0.6	7	0.7	4	0.4	1	0.1
			4	**0.4**	**17**	**1.7**	**14**	**1.4**	**6**	**0.6**	**6**	**0.6**
89	SELF PERCEIVED GENERAL HEALTH	HS1.1	8	0.8	1	0.1	2	0.2	9	0.9	10	1
90		HS1.2	12	1.2	0	0	1	0.1	1	0.1	1	0.1
91		HS1.3	1	0.1	5	0.5	1	0.1	14	1.4	2	0.2
92		HS1.4	15	1.5	0	0	1	0.1	0	0	3	0.3
93		HS1.5	10	1	6	0.6	0	0	20	2	1	0.1
			26	**2.6**	**11**	**1.1**	**2**	**0.2**	**34**	**3.4**	**6**	**0.6**

94	LONG-STANDING HEALTH PROBLEM: SUFFER FROM ANY ILLNESS OR HEALTH PROBLEM OF A DURATION OF AT LEAST SIX MONTHS	HS2.1	11	1.1	2	0.2	0	0	1	0.1	0	0
95		HS2.2	20	2	4	0.4	0	0	1	0.1	0	0
			31	**3.1**	**6**	**0.6**	**0**	**0**	**2**	**0.2**	**0**	**0**
96	GENERAL ACTIVITY LIMITATION: LIMITATION IN ACTIVITIES PEOPLE USUALLY DO BECAUSE OF HEALTH PROBLEMS FOR AT LEAST THE PAST SIX MONTHS	HS3.1	14	1.4	7	0.7	0	0	18	1.8	1	0.1
97		HS3.2	9	0.9	2	0.2	0	0	9	0.9	4	0.4
98		HS3.3	13	1.3	0	0	0	0	0	0	1	0.1
			36	**3.6**	**9**	**0.9**	**0**	**0**	**27**	**2.7**	**6**	**0.6**
99	SUFFERING FROM ASHTMA IN THE PAST 12 MONTHS	CD1A.1	2	0.2	0	0	0	0	0	0	1	0.1
100		CD1A.2	0	0	0	0	0	0	0	0	0	0
			2	**0.2**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0.1**

101	SUFFERING FROM CHRONIC BRONCHITIS, CHRONIC OBSTRUCTIVE PULMONARY DISEASE OR EMPHYSEMA IN THE PAST 12 MONTHS	CD1B.1	4	0.4	0	0	0	0	0	0	0	0
102		CD1B.2	0	0	0	0	0	0	0	0	0	0
			4	**0.4**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
103	SUFFERING FROM A MYOCARDIAL INFARCTION (HEART ATTACK) IN THE PAST 12 MONTHS	CD1C.1	3	0.3	0	0	1	0.1	3	0.3	0	0
104		CD1C.2	0	0	0	0	0	0	0	0	0	0
			3	**0.3**	**0**	**0**	**1**	**0.1**	**3**	**0.3**	**0**	**0**
105	SUFFERING FROM A CORONARY HEART DISEASE OR ANGINA PECTORIS IN THE PAST 12 MONTHS	CD1D.1	7	0.7	0	0	2	0.2	3	0.3	0	0
106		CD1D.2	0	0	0	0	0	0	0	0	0	0
			7	**0.7**	**0**	**0**	**2**	**0.2**	**3**	**0.3**	**0**	**0**
107		CD1E.1	14	1.4	4	0.4	4	0.4	0	0	0	0

108	SUFFERING FROM HIGH BLOOD PRESSURE IN THE PAST 12 MONTHS	CD1E.2	7	0.7	2	0.2	2	0.2	0	0	0	0
			21	**2.1**	**6**	**0.6**	**6**	**0.6**	**0**	**0**	**0**	**0**
109	SUFFERING FROM A STROKE (CEREBRAL HAEMORRHAGE, CEREBRAL THROMBOSIS) IN THE PAST 12 MONTHS	CD1F.1	3	0.3	0	0	1	0.1	4	0.4	0	0
110		CD1F.2	0	0	0	0	0	0	0	0	0	0
			3	**0.3**	**0**	**0**	**1**	**0.1**	**4**	**0.4**	**0**	**0**
111	SUFFERING FROM ARTHROSIS (ARTHRITIS EXCLUDED) IN THE PAST 12 MONTHS	CD1G.1	21	2.1	2	0.2	2	0.2	0	0	1	0.1
112		CD1G.2	10	1	1	0.1	1	0.1	0	0	1	0.1
			31	**3.1**	**3**	**0.3**	**3**	**0.3**	**0**	**0**	**2**	**0.2**
113	SUFFERING FROM A LOW BACK DISORDER OR OTHER CHRONIC BACK DEFECT	CD1H.1	17	1.7	0	0	0	0	3	0.3	0	0
114		CD1H.2	11	1.1	0	0	0	0	2	0.2	0	0

	IN THE PAST 12 MONTHS											
			28	**2.8**	**0**	**0**	**0**	**0**	**5**	**0.5**	**0**	**0**
115	SUFFERING FROM A NECK DISORDER OR OTHER CHRONIC NECK DEFECT IN THE PAST 12 MONTHS	CD1I.1	18	1.8	0	0	0	0	2	0.2	1	0.1
116		CD1I.2	8	0.8	0	0	0	0	1	0.1	0	0
			26	**2.6**	**0**	**0**	**0**	**0**	**3**	**0.3**	**1**	**0.1**
117	SUFFERING FROM DIABETES IN THE PAST 12 MONTHS	CD1J.1	7	0.7	2	0.2	3	0.3	0	0	0	0
118		CD1J.2	1	0.1	0	0	0	0	0	0	0	0
			8	**0.8**	**2**	**0.2**	**3**	**0.3**	**0**	**0**	**0**	**0**
119	SUFFERING FROM AN ALLERGY, SUCH AS RHINITIS, EYE INFLAMMATION, DERMATITIS, FOOD ALLERGY OR OTHER (ALLERGIC ASTHMA EXCLUDED) IN	CD1K.1	2	0.2	0	0	8	0.8	0	0	2	0.2
120		CD1K.2	0	0	0	0	2	0.2	0	0	0	0

	THE PAST 12 MONTHS											
			2	**0.2**	**0**	**0**	**10**	**1**	**0**	**0**	**2**	**0.2**
121	SUFFERING FROM URINARY INCONTINENCE, PROBLEMS IN CONTROLLING THE BLADDER IN THE PAST 12 MONTHS	CD1M.1	11	1.1	0	0	2	0.2	3	0.3	0	0
122		CD1M.2	1	0.1	0	0	0	0	0	0	0	0
			12	**1.2**	**0**	**0**	**2**	**0.2**	**3**	**0.3**	**0**	**0**
123	SUFFERING FROM KIDNEY PROBLEMS IN THE PAST 12 MONTHS	CD1N.1	7	0.7	0	0	0	0	1	0.1	0	0
124		CD1N.2	1	0.1	0	0	0	0	0	0	0	0
			20	**2**	**0**	**0**	**2**	**0.2**	**4**	**0.4**	**0**	**0**
125	SUFFERING FROM DEPRESSION IN THE PAST 12 MONTHS	CD1O.1	13	1.3	5	0.5	12	1.2	0	0	0	0
126		CD1O.2	2	0.2	1	0.1	2	0.2	0	0	0	0
			15	**1.5**	**6**	**0.6**	**14**	**1.4**	**0**	**0**	**0**	**0**
127	OCCURRENCE OF AN ACCIDENT AT HOME IN THE	AC1B.1	2	0.2	0	0	0	0	2	0.2	0	0
128		AC1B.2	0	0	0	0	0	0	0	0	0	0

	PAST 12 MONTHS											
			2	**0.2**	**0**	**0**	**0**	**0**	**2**	**0.2**	**0**	**0**
129	WEARING GLASSES OR CONTACT LENSES	PL1.1	3	0.3	5	0.5	0	0	0	0	0	0
130		PL1.2	5	0.5	8	0.8	0	0	0	0	0	0
			8	**0.8**	**13**	**1.3**	**0**	**0**	**0**	**0**	**0**	**0**
131	USE OF A HEARING AID	PL3.1	1	0.1	1	0.1	2	0.2	0	0	0	0
132		PL3.2	0	0	0	0	0	0	0	0	0	0
			1	**0.1**	**1**	**0.1**	**2**	**0.2**	**0**	**0**	**0**	**0**
133	DIFFICULTY SPEAKING	IN8.1	0	0	0	0	0	0	0	0	0	0
134		IN8.2	2	0.2	4	0.4	0	0	7	0.7	1	0.1
			2	**0.2**	**4**	**0.4**	**0**	**0**	**7**	**0.7**	**1**	**0.1**
135	DIFFICULTY IN WALKING HALF A KM ON LEVEL GROUND WITHOUT THE USE OF ANY AID	PL6.1	4	0.4	0	0	2	0.2	1	0.1	1	0.1
136		PL6.2	7	0.7	0	0	2	0.2	1	0.1	4	0.4
137		PL6.3	10	1	1	0.1	4	0.4	4	0.4	3	0.3
138		PL6.4	5	0.5	3	0.3	3	0.3	23	2.3	0	0
			26	**2.6**	**4**	**0.4**	**11**	**1.1**	**29**	**2.9**	**8**	**0.8**
139	DIFFICULTY IN WALKING UP OR DOWN 12 STEPS	PL7.1	5	0.5	0	0	2	0.2	1	0.1	1	0.1
140		PL7.2	8	0.8	0	0	2	0.2	1	0.1	4	0.4
141		PL7.3	11	1.1	1	0.1	3	0.3	4	0.4	3	0.3

142		PL7.4	5	0.5	4	0.4	3	0.3	25	2.5	1	0.1
			29	**2.9**	**5**	**0.5**	**10**	**1**	**31**	**3.1**	**9**	**0.9**
143		PN1.1	13	1.3	0	0	0	0	3	0.3	2	0.2
144		PN1.2	0	0	0	0	0	0	0	0	0	0
145	INTENSITY OF BODILY PAIN	PN1.3	0	0	1	0.1	0	0	4	0.4	0	0
146	DURING THE PAST 4 WEEKS	PN1.4	3	0.3	1	0.1	0	0	5	0.5	1	0.1
147		PN1.5	10	1	0	0	1	0.1	0	0	1	0.1
148		PN1.6	6	0.6	4	0.4	1	0.1	12	1.2	1	0.1
			32	**3.2**	**6**	**0.6**	**2**	**0.2**	**24**	**2.4**	**5**	**0.5**
149	EXTENT THAT	PN2.1	13	1.3	0	0	0	0	2	0.2	2	0.2
150	PAIN INTERFERED	PN2.2	1	0.1	2	0.2	0	0	9	0.9	1	0.1
151	WITH NORMAL WORK DURING	PN2.3	5	0.5	0	0	0	0	5	0.5	3	0.3
152	THE PAST 4 WEEKS	PN2.4	12	1.2	2	0.2	0	0	1	0.1	1	0.1
153	(INCLUDING BOTH WORK OUTSIDE THE HOME AND HOUSEWORK)	PN2.5	5	0.5	6	0.6	0	0	25	2.5	4	0.4
			36	**3.6**	**10**	**1**	**0**	**0**	**42**	**4.2**	**11**	**1.1**
154	EXTENT OF	MH1A.1	7	0.7	4	0.4	2	0.2	2	0.2	2	0.2
155	HAVING LITTLE	MH1A.2	4	0.4	0	0	1	0.1	22	2.2	15	1.5
156	INTEREST OR	MH1A.3	4	0.4	3	0.3	1	0.1	0	0	2	0.2
157	PLEASURE IN DOING THINGS	MH1A.4	9	0.9	18	1.8	4	0.4	26	2.6	14	1.4

	OVER THE LAST 2 WEEKS											
			24	**2.4**	**25**	**2.5**	**8**	**0.8**	**50**	**5**	**33**	**3.3**
158	EXTENT OF FEELING	MH1B.1	9	0.9	5	0.5	2	0.2	4	0.4	2	0.2
159	DOWN, DEPRESSED	MH1B.2	5	0.5	0	0	1	0.1	25	2.5	13	1.3
160	OR HOPELESS OVER	MH1B.3	5	0.5	4	0.4	1	0.1	0	0	2	0.2
161	THE LAST 2 WEEKS	MH1B.4	10	1	19	1.9	5	0.5	24	2.4	15	1.5
			29	**2.9**	**28**	**2.8**	**9**	**0.9**	**53**	**5.3**	**32**	**3.2**
162	EXTENT OF HAVING	MH1C.1	7	0.7	1	0.1	2	0.2	3	0.3	3	0.3
163	TROUBLE FALLING	MH1C.2	1	0.1	0	0	0	0	11	1.1	8	0.8
164	OR STAYING ASLEEP,	MH1C.3	3	0.3	1	0.1	1	0.1	2	0.2	3	0.3
165	OR SLEEPING TOO MUCH OVER THE LAST 2 WEEKS	MH1C.4	9	0.9	5	0.5	5	0.5	5	0.5	4	0.4
			20	**2**	**7**	**0.7**	**8**	**0.8**	**21**	**2.1**	**18**	**1.8**
166	EXTENT OF FEELING	MH1D.1	10	1	2	0.2	3	0.3	4	0.4	6	0.6
167	TIRED OR HAVING	MH1D.2	1	0.1	1	0.1	1	0.1	19	1.9	13	1.3
168	LITTLE ENERGY	MH1D.3	5	0.5	2	0.2	1	0.1	1	0.1	5	0.5
169	OVER THE LAST 2 WEEKS	MH1D.4	13	1.3	13	1.3	4	0.4	22	2.2	9	0.9
			29	**2.9**	**18**	**1.8**	**9**	**0.9**	**46**	**4.6**	**33**	**3.3**

170	EXTENT OF HAVING POOR APPETITE OR OVEREATING OVER THE LAST 2 WEEKS	MH1E.1	2	0.2	2	0.2	1	0.1	0	0	1	0.1
171		MH1E.2	4	0.4	3	0.3	2	0.2	7	0.7	11	1.1
172		MH1E.3	3	0.3	3	0.3	1	0.1	0	0	0	0
173		MH1E.4	4	0.4	8	0.8	4	0.4	16	1.6	6	0.6
			13	**1.3**	**16**	**1.6**	**8**	**0.8**	**23**	**2.3**	**18**	**1.8**
174	EXTENT OF FEELING BAD ABOUT YOURSELF, FEELING BEING A FAILURE OVER THE LAST 2 WEEKS	MH1F.1	5	0.5	5	0.5	2	0.2	1	0.1	1	0.1
175		MH1F.2	7	0.7	3	0.3	2	0.2	20	2	13	1.3
176		MH1F.3	5	0.5	6	0.6	2	0.2	1	0.1	0	0
177		MH1F.4	7	0.7	17	1.7	6	0.6	25	2.5	20	2
			24	**2.4**	**31**	**3.1**	**12**	**1.2**	**47**	**4.7**	**34**	**3.4**
178	EXTENT OF HAVING TROUBLE CONCENTRATING ON THINGS, SUCH AS READING THE NEWSPAPER OR WATCHING TELEVISION, OVER THE LAST 2 WEEKS	MH1G.1	3	0.3	3	0.3	1	0.1	0	0	1	0.1
179		MH1G.2	6	0.6	2	0.2	2	0.2	11	1.1	14	1.4
180		MH1G.3	4	0.4	4	0.4	1	0.1	0	0	0	0
181		MH1G.4	5	0.5	13	1.3	3	0.3	26	2.6	13	1.3
			18	**1.8**	**22**	**2.2**	**7**	**0.7**	**37**	**3.7**	**28**	**2.8**
182		MH1H.1	2	0.2	2	0.2	1	0.1	0	0	0	0

183	EXTENT OF MOVING OR SPEAKING SO SLOWLY THAT OTHER PEOPLE COULD HAVE NOTICED OR BEING SO FIDGETY OR RESTLESS, OVER THE LAST 2 WEEKS	MH1H.2	5	0.5	4	0.4	3	0.3	8	0.8	8	0.8
184		MH1H.3	7	0.7	15	1.5	3	0.3	20	2	7	0.7
			14	**1.4**	**21**	**2.1**	**7**	**0.7**	**28**	**2.8**	**15**	**1.5**
185	ADMISSION AS AN INPATIENT IN A HOSPITAL IN THE PAST 12 MONTHS	HO1.1	5	0.5	0	0	1	0.1	4	0.4	0	0
186		HO1.2	0	0	0	0	0	0	0	0	0	0
			5	**0.5**	**0**	**0**	**1**	**0.1**	**4**	**0.4**	**0**	**0**
187	ADMISSION AS A DAY PATIENT IN A HOSPITAL IN THE PAST 12 MONTHS	HO3.1	4	0.4	2	0.2	7	0.7	1	0.1	0	0
188		HO3.2	3	0.3	1	0.1	5	0.5	1	0.1	0	0
			7	**0.7**	**3**	**0.3**	**12**	**1.2**	**2**	**0.2**	**0**	**0**
189	LAST APPOINTMENT WITH DENTIST	AM1.1	1	0.1	2	0.2	21	2.1	1	0.1	1	0.1
190		AM1.2	1	0.1	1	0.1	7	0.7	0	0	0	0
191		AM1.3	1	0.1	1	0.1	13	1.3	1	0.1	1	0.1

192		AM1.4	0	0	3	0.3	9	0.9	0	0	0	0
			3	**0.3**	**7**	**0.7**	**50**	**5**	**2**	**0.2**	**2**	**0.2**
193		IN18.1	0	0	1	0.1	3	0.3	1	0.1	1	0.1
194		IN18.2	1	0.1	0	0	7	0.7	0	0	0	0
195	FREQUENCY BRUSHING	IN18.3	0	0	0	0	2	0.2	0	0	0	0
196	TEETH	IN18.4	0	0	0	0	1	0.1	0	0	1	0.1
197		IN18.5	1	0.1	0	0	6	0.6	0	0	1	0.1
198		IN18.6	2	0.2	3	0.3	17	1.7	1	0.1	0	0
			4	**0.4**	**4**	**0.4**	**36**	**3.6**	**2**	**0.2**	**3**	**0.3**
199	LAST APPOINTMENT	AM2.1	3	0.3	6	0.6	4	0.4	0	0	0	0
200	WITH FAMILY DOCTOR	AM2.2	8	0.8	19	1.9	11	1.1	0	0	0	0
			11	**1.1**	**25**	**2.5**	**15**	**1.5**	**0**	**0**	**0**	**0**
201	LAST	AM4.1	4	0.4	4	0.4	14	1.4	1	0.1	0	0
202	APPOINTMENT WITH	AM4.2	2	0.2	1	0.1	8	0.8	2	0.2	0	0
203	SPECIALIST DOCTOR	AM4.3	2	0.2	5	0.5	4	0.4	0	0	1	0.1
			8	**0.8**	**10**	**1**	**26**	**2.6**	**3**	**0.3**	**1**	**0.1**
204	CONSULTATIO N WITH PHISIOTERAPI	AM6A.1	2	0.2	1	0.1	6	0.6	1	0.1	0	0
205	ST IN THE PAST 12 MONTHS	AM6A.2	0	0	0	0	1	0.1	0	0	0	0

			2	**0.2**	**1**	**0.1**	**7**	**0.7**	**1**	**0.1**	**0**	**0**
206	CONSULTATION OF A PSYCHOLOGIST OR PSYCHOTHERAPIST IN THE PAST 12 MONTHS	AM6B.1	2	0.2	3	0.3	22	2.2	0	0	0	0
207		AM6B.2	0	0	0	0	1	0.1	0	0	0	0
			2	**0.2**	**3**	**0.3**	**23**	**2.3**	**0**	**0**	**0**	**0**
208	USE OF ANY HOME CARE SERVICES FOR PERSONAL NEEDS DURING THE PAST 12 MONTHS	AM7.1	4	0.4	1	0.1	4	0.4	12	1.2	0	0
209		AM7.2	0	0	0	0	0	0	0	0	0	0
			4	**0.4**	**1**	**0.1**	**4**	**0.4**	**12**	**1.2**	**0**	**0**
210	USE OF ANY MEDICINES PRESCRIBED BY A DOCTOR DURING THE PAST TWO WEEKS	MD1.1	10	1	7	0.7	0	0	0	0	0	0
211		MD1.2	19	1.9	12	1.2	0	0	0	0	0	0
			29	**2.9**	**19**	**1.9**	**0**	**0**	**0**	**0**	**0**	**0**
212	USE OF ANY MEDICINES, HERBAL MEDICINES OR	MD2.1	0	0	0	0	6	0.6	1	0.1	1	0.1
213		MD2.2	0	0	0	0	2	0.2	0	0	0	0

	VITAMINS NOT PRESCRIBED BY A DOCTOR DURING THE PAST TWO WEEKS											
			0	**0**	**0**	**0**	**8**	**0.8**	**1**	**0.1**	**1**	**0.1**
214	LAST TIME OF VACCINATION AGAINST TETANUS.	IN19.1	0	0	1	0.1	5	0.5	0	0	0	0
215		IN19.2	0	0	1	0.1	8	0.8	0	0	0	0
216		IN19.3	1	0.1	2	0.2	11	1.1	0	0	2	0.2
			1	**0.1**	**4**	**0.4**	**24**	**2.4**	**0**	**0**	**2**	**0.2**
217	LAST TIME OF BLOOD PRESSURE MEASUREMENT BY A HEALTH PROFESSIONAL	PA2.1	3	0.3	7	0.7	5	0.5	0	0	0	0
218		PA2.2	6	0.6	10	1	9	0.9	1	0.1	0	0
219		PA2.3	2	0.2	8	0.8	7	0.7	0	0	0	0
220		PA2.4	2	0.2	8	0.8	2	0.2	1	0.1	20	2
			13	**1.3**	**33**	**3.3**	**23**	**2.3**	**2**	**0.2**	**20**	**2**
221	LAST TIME OF BLOOD CHOLESTEROL MEASUREMENT BY A HEALTH PROFESSIONAL	PA3.1	4	0.4	11	1.1	7	0.7	0	0	1	0.1
222		PA3.2	4	0.4	8	0.8	8	0.8	1	0.1	2	0.2
223		PA3.3	2	0.2	9	0.9	8	0.8	1	0.1	0	0
224		PA3.4	4	0.4	14	1.4	2	0.2	3	0.3	41	4.1
			14	**1.4**	**42**	**4.2**	**25**	**2.5**	**5**	**0.5**	**44**	**4.4**

225	LAST TIME OF BLOOD SUGAR MEASUREMENT BY A HEALTH PROFESSIONAL	PA4.1	4	0.4	11	1.1	7	0.7	0	0	1	0.1
226		PA4.2	4	0.4	7	0.7	7	0.7	1	0.1	2	0.2
227		PA4.3	2	0.2	10	1	8	0.8	1	0.1	0	0
228		PA4.4	3	0.3	13	1.3	3	0.3	1	0.1	31	3.1
			13	**1.3**	**41**	**4.1**	**25**	**2.5**	**3**	**0.3**	**34**	**3.4**
229	LAST TIME OF A FAECAL OCCULT BLOOD TEST	PA5.1	2	0.2	4	0.4	2	0.2	0	0	2	0.2
230		PA5.2	1	0.1	2	0.2	0	0	0	0	0	0
231		PA5.3	0	0	0	0	0	0	0	0	0	0
232		PA5.4	1	0.1	0	0	0	0	0	0	0	0
233		PA5.5	1	0.1	2	0.2	0	0	0	0	1	0.1
			3	**0.3**	**4**	**0.4**	**0**	**0**	**0**	**0**	**1**	**0.1**
234	LAST TIME OF A COLONOSCOPY	PA6.1	1	0.1	1	0.1	1	0.1	0	0	1	0.1
235		PA6.2	2	0.2	3	0.3	0	0	0	0	2	0.2
236		PA6.3	1	0.1	1	0.1	0	0	0	0	1	0.1
237		PA6.4	0	0	0	0	0	0	0	0	0	0
238		PA6.5	1	0.1	2	0.2	0	0	0	0	1	0.1
			5	**0.5**	**7**	**0.7**	**1**	**0.1**	**0**	**0**	**5**	**0.5**
239	UNMET NEED FOR HEALTH CARE IN THE PAST 12 MONTHS DUE TO LONG	UN1A.1	4	0.4	0	0	4	0.4	2	0.2	0	0
240		UN1A.2	0	0	5	0.5	1	0.1	1	0.1	0	0
241		UN1A.3	9	0.9	18	1.8	21	2.1	0	0	0	0

	WAITING LIST(S)											
			13	**1.3**	**23**	**2.3**	**26**	**2.6**	**3**	**0.3**	**0**	**0**
242	UNMET NEED FOR HEALTH CARE IN THE PAST 12 MONTHS DUE TO DISTANCE OR TRANSPORTATION PROBLEMS	UN1B.1	3	0.3	3	0.3	0	0	0	0	0	0
243		UN1B.2	1	0.1	5	0.5	4	0.4	0	0	0	0
244		UN1B.3	8	0.8	16	1.6	20	2	0	0	0	0
			12	**1.2**	**24**	**2.4**	**24**	**2.4**	**0**	**0**	**0**	**0**
245	COULD NOT AFFORD MEDICAL EXAMINATION OR TREATMENT IN THE PAST 12 MONTHS	UN2A.1	3	0.3	10	1	5	0.5	8	0.8	1	0.1
246		UN2A.2	1	0.1	15	1.5	4	0.4	2	0.2	0	0
247		UN2A.3	9	0.9	16	1.6	23	2.3	0	0	0	0
			13	**1.3**	**41**	**4.1**	**32**	**3.2**	**10**	**1**	**1**	**0.1**
248	COULD NOT AFFORD DENTAL EXAMINATION OR TREATMENT IN THE PAST 12 MONTHS	UN2B.1	1	0.1	9	0.9	3	0.3	15	1.5	2	0.2
249		UN2B.2	1	0.1	6	0.6	22	2.2	3	0.3	1	0.1
250		UN2B.3	0	0	0	0	28	2.8	1	0.1	0	0
			2	**0.2**	**15**	**1.5**	**53**	**5.3**	**19**	**1.9**	**3**	**0.3**

251	COULD NOT AFFORD PRESCRIBED MEDICINES IN THE PAST 12 MONTHS	UN2C.1	4	0.4	8	0.8	3	0.3	7	0.7	1	0.1
252		UN2C.2	2	0.2	10	1	1	0.1	1	0.1	0	0
253		UN2C.3	12	1.2	13	1.3	8	0.8	0	0	0	0
			18	**1.8**	**31**	**3.1**	**12**	**1.2**	**8**	**0.8**	**1**	**0.1**
254	COULD NOT AFFORD MENTAL HEALTH CARE (BY A PSYCHOLOGIST OR A PSYCHIATRIST FOR EXAMPLE) IN THE PAST 12 MONTHS	UN2D.1	3	0.3	7	0.7	8	0.8	2	0.2	1	0.1
255		UN2D.2	2	0.2	1	0.1	17	1.7	0	0	0	0
256		UN2D.3	0	0	1	0.1	2	0.2	0	0	0	0
			5	**0.5**	**9**	**0.9**	**27**	**2.7**	**2**	**0.2**	**1**	**0.1**
257	PHYSICAL EFFORT OF WORKING TASKS (BOTH PAID AND UNPAID WORK ACTIVITIES INCLUDED)	PE1.1	1	0.1	0	0	0	0	2	0.2	8	0.8
258		PE1.2	1	0.1	0	0	0	0	4	0.4	2	0.2
259		PE1.3	1	0.1	1	0.1	0	0	3	0.3	8	0.8
260		PE1.4	4	0.4	4	0.4	5	0.5	18	1.8	0	0
			7	**0.7**	**5**	**0.5**	**5**	**0.5**	**27**	**2.7**	**18**	**1.8**
261	NUMBER OF DAYS IN A TYPICAL	PE2.0	1	0.1	0	0	1	0.1	2	0.2	1	0.1
262		PE2.1	0	0	0	0	0	0	0	0	0	0

263	WEEK	PE2.2	0	0	0	0	0	0	0	0	0	0
264	WALKING TO GET TO AND	PE2.3	0	0	0	0	0	0	0	0	0	0
265	FROM PLACES AT LEAST 10	PE2.4	0	0	0	0	0	0	0	0	0	0
266	MINUTES CONTINUOUSL	PE2.5	1	0.1	0	0	1	0.1	0	0	3	0.3
267	Y	PE2.6	0	0	0	0	0	0	0	0	0	0
268		PE2.7	0	0	0	0	0	0	1	0.1	0	0
			2	**0.2**	**0**	**0**	**2**	**0.2**	**3**	**0.3**	**4**	**0.4**
269	NUMBER OF	PE4.0	0	0	0	0	0	0	0	0	0	0
270	DAYS IN A TYPICAL WEEK	PE4.1	1	0.1	0	0	0	0	0	0	0	0
271	BICYCLING TO GET TO AND FROM PLACES AT LEAST 10 MINUTES CONTINUOUSL Y	PE4.2	0	0	0	0	1	0.1	0	0	0	0
			1	**0.1**	**0**	**0**	**1**	**0.1**	**0**	**0**	**0**	**0**
272	NUMBER OF	PE6.0	1	0.1	0	0	4	0.4	1	0.1	2	0.2
273	DAYS IN A TYPICAL	PE6.1	1	0.1	0	0	1	0.1	0	0	1	0.1
274	WEEK DOING SPORTS,	PE6.2	1	0.1	0	0	4	0.4	1	0.1	4	0.4
275	FITNESS OR RECREATIONA L (LEISURE)	PE6.3	1	0.1	0	0	2	0.2	1	0.1	1	0.1
276	PHYSICAL ACTIVITIES THAT CAUSE AT LEAST A	PE6.4	1	0.1	1	0.1	2	0.2	1	0.1	1	0.1

	SMALL INCREASE IN BREATHING OR HEART RATE FOR AT LEAST 10 MINUTES CONTINUOUSLY											
			5	**0.5**	**1**	**0.1**	**13**	**1.3**	**4**	**0.4**	**9**	**0.9**
277	NUMBER OF DAYS IN A TYPICAL WEEK DOING MUSCLE-STRENGTHENING ACTIVITIES	PE8.0	0	0	0	0	1	0.1	0	0	0	0
278		PE8.1	3	0.3	0	0	9	0.9	4	0.4	5	0.5
			3	**0.3**	**0**	**0**	**10**	**1**	**4**	**0.4**	**5**	**0.5**
279	NUMBER OF MEALS A DAY	IN29.0	0	0	9	0.9	0	0	2	0.2	1	0.1
280		IN29.1	0	0	1	0.1	0	0	0	0	0	0
			0	**0**	**10**	**1**	**0**	**0**	**2**	**0.2**	**1**	**0.1**
281	FREQUENCY OF EATING FRUIT, EXCLUDING JUICE	FV1.1	0	0	5	0.5	0	0	0	0	1	0.1
282		FV1.2	0	0	2	0.2	0	0	0	0	1	0.1
283		FV1.3	0	0	6	0.6	0	0	1	0.1	1	0.1
284		FV1.4	0	0	8	0.8	0	0	0	0	0	0
			0	**0**	**21**	**2.1**	**0**	**0**	**1**	**0.1**	**3**	**0.3**
285		FV3.1	0	0	6	0.6	1	0.1	1	0.1	1	0.1

286	FREQUENCY	FV3.2	0	0	0	0	0	0	0	0	0	0
287	OF EATING VEGETABLES	FV3.3	0	0	4	0.4	1	0.1	1	0.1	0	0
288	OR SALAD, EXCLUDING JUICE AND POTATOES	FV3.4	0	0	7	0.7	1	0.1	0	0	2	0.2
			0	**0**	**17**	**1.7**	**3**	**0.3**	**2**	**0.2**	**3**	**0.3**
289		IN43.1	3	0.3	12	1.2	1	0.1	3	0.3	6	0.6
290	TOBBACCO	IN43.2	1	0.1	1	0.1	1	0.1	0	0	0	0
291	STATUS	IN43.3	0	0	2	0.2	1	0.1	0	0	13	1.3
292		IN43.4	2	0.2	1	0.1	0	0	1	0.1	11	1.1
			6	**0.6**	**33**	**3.3**	**6**	**0.6**	**6**	**0.6**	**33**	**3.3**
293	FREQUENCY	SK4.1	0	0	0	0	0	0	0	0	0	0
294	OF EXPOSURE TO TOBACCO	SK4.2	0	0	1	0.1	1	0.1	0	0	0	0
295	SMOKE INDOORS	SK4.3	1	0.1	4	0.4	0	0	0	0	0	0
			1	**0.1**	**5**	**0.5**	**1**	**0.1**	**0**	**0**	**0**	**0**
296	FREQUENCY	AL1.1	0	0	2	0.2	10	1	1	0.1	26	2.6
297	OF CONSUMPTIO	AL1.2	0	0	0	0	0	0	0	0	1	0.1
298	N OF AN ALCOHOLIC	AL1.3	2	0.2	0	0	2	0.2	0	0	0	0
299	DRINK OF ANY	AL1.4	1	0.1	0	0	3	0.3	0	0	1	0.1
300	KIND (BEER, WINE, CIDER,	AL1.5	1	0.1	0	0	4	0.4	0	0	2	0.2
301	SPIRITS, COCKTAILS,	AL1.6	0	0	0	0	3	0.3	0	0	2	0.2
302	PREMIXES, LIQUEURS,	AL1.7	2	0.2	0	0	1	0.1	0	0	0	0

303	HOMEMADE ALCOHOL…) IN THE PAST 12 MONTHS	AL1.8	3	0.3	0	0	0	0	0	0	16	1.6
			9	**0.9**	**2**	**0.2**	**23**	**2.3**	**1**	**0.1**	**48**	**4.8**
304		IN50.1	1	0.1	9	0.9	1	0.1	1	0.1	5	0.5
305		IN50.2	3	0.3	9	0.9	0	0	5	0.5	0	0
306	IN MOST	IN50.3	1	0.1	1	0.1	0	0	4	0.4	0	0
307	WAYS MY LIFE IS CLOSE TO	IN50.4	1	0.1	0	0	2	0.2	2	0.2	0	0
308	MY IDEAL.	IN50.5	0	0	2	0.2	0	0	3	0.3	0	0
309		IN50.6	2	0.2	6	0.6	0	0	10	1	0	0
310		IN50.7	1	0.1	1	0.1	2	0.2	10	1	3	0.3
			9	**0.9**	**28**	**2.8**	**5**	**0.5**	**35**	**3.5**	**8**	**0.8**
311		IN51.1	3	0.3	19	1.9	4	0.4	0	0	9	0.9
312		IN51.2	3	0.3	8	0.8	0	0	10	1	1	0.1
313	THE CONDITIONS	IN51.3	0	0	1	0.1	0	0	4	0.4	0	0
314	OF MY LIFE	IN51.4	0	0	0	0	1	0.1	2	0.2	0	0
315	ARE EXCELLENT.	IN51.5	1	0.1	6	0.6	0	0	0	0	0	0
316		IN51.6	2	0.2	5	0.5	0	0	14	1.4	3	0.3
317		IN51.7	1	0.1	0	0	1	0.1	13	1.3	11	1.1
			10	**1**	**39**	**3.9**	**6**	**0.6**	**43**	**4.3**	**24**	**2.4**
318		IN52.1	4	0.4	18	1.8	4	0.4	3	0.3	11	1.1
319		IN52.2	5	0.5	11	1.1	1	0.1	5	0.5	0	0

320		IN52.3	1	0.1	2	0.2	0	0	4	0.4	0	0
321		IN52.4	0	0	0	0	0	0	5	0.5	0	0
322	I AM SATISFIED WITH MY LIFE.	IN52.5	0	0	1	0.1	0	0	6	0.6	0	0
323		IN52.6	3	0.3	8	0.8	1	0.1	8	0.8	0	0
324		IN52.7	1	0.1	2	0.2	1	0.1	16	1.6	7	0.7
			14	**1.4**	**42**	**4.2**	**7**	**0.7**	**47**	**4.7**	**18**	**1.8**
325		IN53.1	1	0.1	10	1	2	0.2	0	0	5	0.5
326		IN53.2	2	0.2	12	1.2	0	0	5	0.5	0	0
327	SO FAR I HAVE GOTTEN THE	IN53.3	0	0	3	0.3	0	0	4	0.4	0	0
328	IMPORTANT	IN53.4	0	0	1	0.1	1	0.1	3	0.3	1	0.1
329	THINGS I WANT IN LIFE.	IN53.5	0	0	0	0	0	0	2	0.2	0	0
330		IN53.6	1	0.1	7	0.7	0	0	5	0.5	0	0
331		IN53.7	1	0.1	2	0.2	2	0.2	12	1.2	3	0.3
			5	**0.5**	**35**	**3.5**	**5**	**0.5**	**31**	**3.1**	**9**	**0.9**
332		IN54.1	2	0.2	11	1.1	5	0.5	1	0.1	4	0.4
333		IN54.2	1	0.1	3	0.3	1	0.1	7	0.7	0	0
334	IF I COULD LIVE MY LIFE OVER, I	IN54.3	0	0	0	0	0	0	3	0.3	0	0
335	WOULD	IN54.4	0	0	0	0	1	0.1	0	0	0	0
336	CHANGE ALMOST	IN54.5	0	0	1	0.1	0	0	0	0	0	0
337	NOTHING.	IN54.6	1	0.1	4	0.4	1	0.1	7	0.7	1	0.1
338		IN54.7	0	0	2	0.2	0	0	11	1.1	3	0.3
			4	**0.4**	**21**	**2.1**	**8**	**0.8**	**29**	**2.9**	**8**	**0.8**

339	NUMBER OF CLOSE PEOPLE TO COUNT ON IN CASE OF SERIOUS PERSONAL PROBLEMS	SS1.1	0	0	3	0.3	0	0	2	0.2	2	0.2
340		SS1.2	1	0.1	1	0.1	3	0.3	3	0.3	0	0
341		SS1.3	0	0	1	0.1	0	0	1	0.1	0	0
342		SS1.4	0	0	1	0.1	2	0.2	5	0.5	0	0
			1	**0.1**	**6**	**0.6**	**5**	**0.5**	**11**	**1.1**	**2**	**0.2**
343	DEGREE OF CONCERN SHOWN BY OTHER PEOPLE IN WHAT THE PERSON IS DOING	SS2.1	0	0	1	0.1	0	0	5	0.5	2	0.2
344		SS2.2	0	0	0	0	0	0	5	0.5	1	0.1
345		SS2.3	0	0	1	0.1	0	0	1	0.1	2	0.2
346		SS2.4	0	0	3	0.3	0	0	1	0.1	1	0.1
			0	**0**	**5**	**0.5**	**0**	**0**	**12**	**1.2**	**6**	**0.6**
347	HOW EASY IS IT TO GET PRACTICAL HELP FROM NEIGHBORS IN CASE OF NEED	SS3.1	0	0	2	0.2	1	0.1	4	0.4	1	0.1
348		SS3.2	0	0	1	0.1	0	0	0	0	0	0
349		SS3.3	0	0	0	0	1	0.1	1	0.1	0	0
350		SS3.4	0	0	1	0.1	0	0	1	0.1	0	0
351		SS3.5	0	0	3	0.3	2	0.2	0	0	0	0
			0	**0**	**7**	**0.7**	**4**	**0.4**	**6**	**0.6**	**1**	**0.1**
352	PROVIDING CARE OR ASSISTANCE TO ONE OR MORE PERSONS	IC1.1	0	0	0	0	2	0.2	3	0.3	0	0
353		IC1.2	0	0	0	0	0	0	0	0	0	0

	SUFFERING FROM SOME AGE PROBLEM, CHRONIC HEALTH CONDITION OR INFIRMITY, AT LEAST ONCE A WEEK											
			0	**0**	**0**	**0**	**2**	**0.2**	**3**	**0.3**	**0**	**0**
354		IMC_C.1	2	0.2	2	0.2	2	0.2	1	0.1	8	0.8
355	BODY MASS INDEX	IMC_C.2	0	0	2	0.2	1	0.1	0	0	5	0.5
356		IMC_C.3	2	0.2	0	0	0	0	1	0.1	1	0.1
			4	**0.4**	**4**	**0.4**	**3**	**0.3**	**2**	**0.2**	**14**	**1.4**

NA: Supplementary variables thus these do not account to explaining variability in dimensions.

S6 File. Description of cluster 1 to 3, according to the most represented categories (percentage in the cluster and in the sample).

TABLE 1. Description of cluster 1, according to socio-demographic and health related most represented categories (percentage in the cluster and in the sample).

VARIABLE	CATEGORY	Percentage_in_cluster	Percentage_in_sample
MAINSTAT	MAINSTAT_31	98.46	9.76
GE	GE_1	98.19	7.43
GE	GE_2	95.32	6.82
GE	GE_3	94.53	6.32
HS1	HS1_1	94.42	15.47
GE	GE_4	93.63	8.73
GE	GE_5	92.27	11.63
MD1	MD1_2	90.74	60.47
HS2	HS2_2	90.13	57.85
HHTYPE	HHTYPE_23	89.34	37.52
HHNBPERS_0_4	HHNBPERS_0_4_1	88.93	11.98
PA2	PA2_3	88.75	5.41
GE	GE_6	88	11.73
PA2	PA2_4	87.96	4.78
HHNBPERS_16_24	HHNBPERS_16_24_2	87.2	6.12
HH_ACT	HH_ACT_2	87.14	36.21

HHNBPERS_5_13	HHNBPERS_5_13_2	86.89	5.45
HHNBPERS_5_13	HHNBPERS_5_13_1	86.65	20.76
HATLEVEL	HATLEVEL_3	86.57	20.12
IN43	IN43_2	86.3	3.84
UN2c	UN2c_3	86.23	37.29
UN1a	UN1a_3	86.17	24.94
HS1	HS1_2	85.76	48.17
PA4	PA4_3	85.47	6.81
HHTYPE	HHTYPE_24	85.26	4.49
MAINSTAT	MAINSTAT_10	85.11	56.65
PA3	PA3_4	85.11	9.75
PA3	PA3_3	84.86	6.49
HHNBPERS	HHNBPERS_4	84.74	21.65
HATLEVEL	HATLEVEL_4	84.63	15.52
HHNBPERS_14_15	HHNBPERS_14_15_1	84.57	8.18
HH_ACT	HH_ACT_3	84.48	6.38
GE	GE_7	84.23	11.22
HHNBPERS_16_24	HHNBPERS_16_24_1	83.4	22.42
UN1b	UN1b_3	83.24	26.77
PA2	PA2_2	83.11	24.04

MARSTADEFACTO	MARSTADEFACTO_1	82.87	9.06
MARSTALEGAL	MARSTALEGAL_1	82.43	35.3
HHNBPERS	HHNBPERS_5	82.23	5.85
PE8	PE8_1	82.12	12.38
HHTYPE	HHTYPE_21	81.17	6.76
SK4	SK4_2	81.01	3.9
AM2	AM2_2	80.97	36.93
PE4	PE4_1	80.8	5.25
PA4	PA4_4	80.64	10.34
IN43	IN43_1	80.56	23.04
HHNBPERS_65plus	HHNBPERS_65plus_0	79.45	81.88
HHNBPERS_25_64	HHNBPERS_25_64_2	79.27	56.19
AL1	AL1_3	79.23	14.98
MAINSTAT	MAINSTAT_20	78.59	17.6
UN2a	UN2a_3	78.49	38.34
HATLEVEL	HATLEVEL_2	78.47	23.15
PE1	PE1_3	77.48	12.84
PN1	PN1_1	77.09	59.8
HH_INACT	HH_INACT_3	76.98	5.19
PE6	PE6_1	76.75	5.68

HHNBPERS	HHNBPERS_6	76.67	2.8
SK4	SK4_3	76.46	5.79
HHNBPERS_25_64	HHNBPERS_25_64_3	76.17	11.44
HS3	HS3_3	75.84	84.12
AL1	AL1_4	75.71	7.74
PA4	PA4_2	75.65	27.35
PA3	PA3_2	75.53	27.5
AM4	AM4_3	75.26	12.27
HHNBPERS	HHNBPERS_3	75.2	28.99
PN2	PN2_1	74.1	78.24
PE6	PE6_3	74.08	7.05
CD1g	CD1g_2	73.58	91.09
HH_ACT	HH_ACT_1	72.93	35.19
CD1h	CD1h_2	72.74	81.16
PE6	PE6_2	72.6	9.68
PL1	PL1_2	71.97	50.51
AL1	AL1_5	71.07	8.19
GE	GE_8	70.99	10.66
CD1e	CD1e_2	70.98	87.21
MH1d	MH1d_1	70.72	64.55

IN50	IN50_7	68.77	5.47
MARSTALEGAL	MARSTALEGAL_4	68.77	10.96
CD1i	CD1i_2	68.59	89.04
MH1c	MH1c_1	67.65	70.08
MH1b	MH1b_1	67.61	79.11
HH_INACT	HH_INACT_2	67.16	15.2
HHINCOME	HHINCOME_5	67.14	16.55
HH_INACT	HH_INACT_1	67.02	35.71
AL1	AL1_2	66.92	4.07
AM1	AM1_1	66.79	30.78
AM1	AM1_2	66.63	18.21
MH1a	MH1a_1	66.21	80.23
PL7	PL7_1	65.81	98.26
PE2	PE2_5	65.72	13.46
AL1	AL1_6	65.37	13.05
IMC_c	IMC_c_1	65.34	50.01
FV1	FV1_4	65.24	6.49
PE4	PE4_2	65.05	2.86
NUTS	NUTS_7	65.02	16.08
FV1	FV1_3	64.35	13.8

PL6	PL6_1	63.97	98.77
MH1f	MH1f_1	63.69	89.43
IN51	IN51_7	63.68	4.11
HO3	HO3_2	63.44	67.75
UN2b	UN2b_2	63.39	39.29
SEX	SEX_1	62.9	49.4
IN18	IN18_2	62.81	47.91
PE1	PE1_2	62.78	44.92
NUTS	NUTS_6	62.73	14.22
SS3	SS3_3	62.07	32.18
HHINCOME	HHINCOME_4	61.8	20.42
GE	GE_9	61.62	9.41
IN52	IN52_6	61.47	45.21
IN52	IN52_7	61.39	6.97
PE6	PE6_4	61.33	13.77
MH1g	MH1g_1	61.3	90.83
PA6	PA6_5	61.09	86.68
AM4	AM4_2	60.91	52.25
PA5	PA5_5	60.74	79.59
IC	IC_1	60.72	14.15

SS2	SS2_2	60.42	34.48
CD1o	CD1o_2	60.34	95.19
IN50	IN50_6	60.1	34.43
CD1j	CD1j_2	60.09	96.35
IN51	IN51_6	60.07	24.74
SS1	SS1_4	59.93	18.79
IN53	IN53_7	59.91	7.69
HHNBPERS_25_64	HHNBPERS_25_64_1	59.69	26.12
IN50	IN50_5	59.63	31.11
MH1e	MH1e_1	59.52	89.19
IN52	IN52_5	59.51	27.12
IN18	IN18_1	59.42	22.86
DEG_URB	DEG_URB_2	59.25	35.5
MH1h	MH1h_1	59.21	93.53
CD1m	CD1m_2	59.05	98.64
IN29	IN29_0	58.96	12.97
IN51	IN51_5	58.74	34.08
MD2	MD2_1	58.67	24.52
IN53	IN53_6	58.36	44.11
PN1	PN1_2	58.21	10.2

IN54	IN54_5	58.11	18.5
IN19	IN19_1	57.69	80.65
FV1	FV1_2	57.63	12.01
DEG_URB	DEG_URB_1	57.61	28.64
SS1	SS1_3	57.56	40.79
IN54	IN54_6	57.54	28.93
CD1d	CD1d_2	57.29	99.03
CD1n	CD1n_2	57.11	98.26
IN1	IN1_1	57.06	87.99
HO1	HO1_2	57.02	95.11
PE2	PE2_7	56.7	28.92
CD1b	CD1b_2	56.64	97.03
HHTYPE	HHTYPE_25	56.62	19.42
UN2d	UN2d_3	56.35	94.91
CD1k	CD1k_2	56.2	84.88
AM6a	AM6a_2	55.65	93.1
AM7	AM7_2	55.63	99.59
CD1f	CD1f_2	55.44	99.51
CD1a	CD1a_2	55.38	96.62
CD1c	CD1c_2	55.33	99.66

AC1b	AC1b_2	55.23	98.17
PL3	PL3_2	55.16	99.13
AM6b	AM6b_2	55.13	96.07
IN8	IN8_1	54.9	99.39

Table 2: Description of cluster 2, according to the most represented categories (percentage in the cluster and in the sample).

VARIABLE	categories	Percentage_in_cluster	Percentage_in_sample
PL7	PL7_2	73.69	21.19
GE	GE_14	73.32	12.71
GE	GE_12	72.16	16.7
PL6	PL6_2	71.46	15.74
HHNBPERS_25_64	HHNBPERS_25_64_0	70.88	57.61
GE	GE_15	69.89	6.98
HHNBPERS_65plus	HHNBPERS_65plus_2	69.38	33.92
GE	GE_13	69.31	15.87
PL3	PL3_1	68.67	5.12
MAINSTAT	MAINSTAT_32	68.56	68.5
GE	GE_11	66.75	17.76

MARSTALEGAL	MARSTALEGAL_3	66.73	30
CD1j	CD1j_1	65.86	23.73
CD1g	CD1g_1	64.79	58.24
HS1	HS1_4	64.26	24.83
HS3	HS3_2	63.7	51.21
HATLEVEL	HATLEVEL_0	62.39	26.22
CD1e	CD1e_1	62.05	56.65
PN2	PN2_3	61.49	19.47
CD1m	CD1m_1	60.19	15.72
HH_ACT	HH_ACT_0	59.41	76.56
CD1d	CD1d_1	59.13	10.4
CD1i	CD1i_1	59.06	47.87
HHNBPERS_65plus	HHNBPERS_65plus_1	57.65	41.27
PN1	PN1_4	56.57	24.74
PA6	PA6_4	56.23	3.35
CD1c	CD1c_1	56.21	3.86
CD1h	CD1h_1	55.88	60.31
PN2	PN2_2	55.41	28.38
PL7	PL7_3	55.04	9.81
PA6	PA6_1	55.03	6.76

HHTYPE	HHTYPE_22	54.99	41.56
PL6	PL6_3	54.39	8.76
PA6	PA6_2	54.09	17.53
MAINSTAT	MAINSTAT_35	53.91	10.22
MH1c	MH1c_3	53.21	9.32
PN1	PN1_5	53.16	19.29
MD1	MD1_1	53	92.32
CD1b	CD1b_1	52.94	10.51
CD1n	CD1n_1	52.88	9.9
MH1f	MH1f_2	52.81	25.28
PA5	PA5_1	52.41	16.37
HHNBPERS	HHNBPERS_1	51.76	33.3
HHTYPE	HHTYPE_10	51.76	33.3
MH1b	MH1b_2	51.6	36.55
HS1	HS1_3	51.35	58.76
HS2	HS2_1	51.31	91.22
MH1a	MH1a_2	51.09	33.63
PA5	PA5_2	51.04	9.36
CD1f	CD1f_1	51.01	3.97
IN1	IN1_2	50.97	22.97

PA6	PA6_3	50.89	6.4
MH1g	MH1g_2	50.65	20
IN18	IN18_6	50.61	8.44
CD1o	CD1o_1	49.97	19.91
HHNBPERS	HHNBPERS_2	49.4	48.27
MH1b	MH1b_3	49.04	6.29
AL1	AL1_7	48.99	15.2
PN2	PN2_4	48.84	13.22
PA5	PA5_3	48.35	4.29
MH1d	MH1d_3	48.27	8.78
MH1h	MH1h_2	47.88	14.21
MH1a	MH1a_3	47.77	5.77
IN19	IN19_3	47.75	5.73
SS3	SS3_1	47.56	18.63
MH1d	MH1d_2	47.17	44.59
HATLEVEL	HATLEVEL_1	46.47	58.37
IMC_c	IMC_c_3	46.46	24.32
IN18	IN18_5	46.39	10.24
HO1	HO1_1	46.23	12.24
HH_INACT	HH_INACT_0	46.03	69.31

AC1b	AC1b_1	45.99	4.76
PL1	PL1_1	45.77	77.32
MH1e	MH1e_2	45.76	16.97
PA4	PA4_1	45.4	85.56
PA3	PA3_1	45.36	86.46
MH1c	MH1c_2	45.31	31.79
UN2d	UN2d_1	45.24	3.52
HO3	HO3_1	44.93	51.66
PN1	PN1_3	44.79	18.03
CD1a	CD1a_1	44.61	6.78
PA5	PA5_4	44.49	9.25
UN2c	UN2c_1	44.49	11.16
UN2c	UN2c_2	44.39	81.5
UN2a	UN2a_2	44.32	75.44
GE	GE_10	44.16	10.87
HHINCOME	HHINCOME_2	44.15	27.71
AL1	AL1_1	44.12	29.84
AM6a	AM6a_1	44.08	11.36
UN1a	UN1a_1	44.01	26.25
AM2	AM2_1	43.97	90.21

AM4	AM4_1	43.68	53.26
PA2	PA2_1	43.57	92.7
IN18	IN18_4	43.52	14.39
UN1b	UN1b_1	43.16	3.61
AM1	AM1_3	43	66.64
AM7	AM7_1	42.86	3.64
IN43	IN43_4	42.73	71.35
IN19	IN19_2	42.64	23.17
IN54	IN54_7	42.64	9.5
AL1	AL1_8	42.56	32.11
CD1k	CD1k_1	42.48	21.13
NUTS	NUTS_5	42.4	19.74
UN2b	UN2b_3	42.31	50.34
PE1	PE1_1	42.23	54.04
HHNBPERS_16_24	HHNBPERS_16_24_0	42.1	93.65
IN50	IN50_4	42.09	12.37
MH1c	MH1c_4	42.01	13.45
IN18	IN18_3	41.86	13.16
HHNBPERS_5_13	HHNBPERS_5_13_0	41.8	95.24
IMC_c	IMC_c_2	41.76	44.79

NUTS	NUTS_3	41.71	23.6
DEG_URB	DEG_URB_3	41.54	46.03
UN1b	UN1b_2	41.52	89.94
IN52	IN52_2	41.48	12.46
IN54	IN54_4	41.07	7.95
UN2d	UN2d_2	40.55	6.6
UN1a	UN1a_2	40.52	68.93
PE6	PE6_0	40.32	78.51
FV1	FV1_1	40.22	78.6
HS3	HS3_1	40.19	11.45
IN52	IN52_3	40.18	7.9
SEX	SEX_2	40.03	62.71
SS1	SS1_2	39.65	44.79
MARSTALEGAL	MARSTALEGAL_2	39.5	55.64
UN2a	UN2a_1	39.43	11.9
IN51	IN51_2	39.29	18.28
SS2	SS2_1	39.03	62.53
HHNBPERS_0_4	HHNBPERS_0_4_0	38.95	98.36
IN43	IN43_3	38.68	21.94
PE8	PE8_0	38.64	96.68

SS3	SS3_2	38.6	38.08
PE2	PE2_7	38.19	28.72
FV3	FV3_1	38.17	52.27
MARSTADEFACTO	MARSTADEFACTO_2	38.17	97.82
MD2	MD2_2	38.13	80.38
SK4	SK4_1	38.13	96.92
HHNBPERS_14_15	HHNBPERS_14_15_0	37.99	98.07
IN29	IN29_1	37.87	90.88
PE4	PE4_0	37.5	96.14
IC	IC_2	37.26	88.71

Table 3: Description of cluster 3, according to the most represented categories (percentage in the cluster and in the sample).

Variable	Category	Percentage_in_cluster	Percentage_in_sample
MH1f	MH1f_4	84.17	28.66
MH1g	MH1g_4	80.95	22.91
MH1h	MH1h_3	75.63	29.83
PL7	PL7_4	74.02	16.89
MH1b	MH1b_4	72.51	47.17
PN2	PN2_5	71.62	19.5

MH1a	MH1a_4	70.11	44.47
PL6	PL6_4	67.34	17.97
HS1	HS1_5	67.09	33.33
MH1e	MH1e_4	66.08	20.31
MH1d	MH1d_4	57.41	57.77
PN1	PN1_6	54.51	26.06
MH1f	MH1f_3	54.48	19.68
IN52	IN52_1	54.22	20.22
IN8	IN8_2	51.22	11.32
AM7	AM7_1	50	16.98
MH1g	MH1g_3	48.94	14.47
HS3	HS3_1	48.86	55.71
MH1e	MH1e_3	48.49	13.03
PL6	PL6_3	44.21	28.48
PL7	PL7_3	42.44	30.28
PE1	PE1_4	41.73	24.26
CD1f	CD1f_1	39.77	12.4
IN51	IN51_1	38.21	23
PN2	PN2_4	38.06	41.24
MAINSTAT	MAINSTAT_33	38.04	11.86

UN1b	UN1b_1	37.27	12.49
MH1c	MH1c_4	37.17	47.62
IN53	IN53_1	37.08	8.89
UN2d	UN2d_1	36.89	11.5
CD1c	CD1c_1	36.6	10.06
MH1b	MH1b_3	36.08	18.51
MH1a	MH1a_3	34.76	16.8
CD1n	CD1n_1	33.45	25.07
CD1d	CD1d_1	32.69	23
CD1o	CD1o_1	32.23	51.39
CD1m	CD1m_1	32.16	33.6
MH1d	MH1d_3	30	21.83
HS1	HS1_4	28.24	43.67
AC1b	AC1b_1	27.98	11.59
IN52	IN52_2	26.61	31.99
HO1	HO1_1	26.55	28.12
IN50	IN50_1	26.54	19.41
GE	GE_15	26.29	10.51
IN18	IN18_6	25.84	17.25
PN1	PN1_5	25.8	37.47

CD1b	CD1b_1	25	19.86
HATLEVEL	HATLEVEL_0	24.25	40.79
AM6b	AM6b_1	24.2	15
IN54	IN54_1	23.47	19.68
CD1a	CD1a_1	22.6	13.75
MH1h	MH1h_2	22.47	26.68
IN50	IN50_2	22.22	28.93
GE	GE_13	22.16	20.31
UN2c	UN2c_1	22.11	22.19
UN2a	UN2a_1	21.95	26.5
UN2d	UN2d_2	21.93	14.29
PL6	PL6_2	21.81	19.23
GE	GE_14	21.37	14.82
MARSTALEGAL	MARSTALEGAL_3	21.08	37.92
IN53	IN53_2	21.05	22.28
CD1i	CD1i_1	21	68.1
MH1c	MH1c_3	20.77	14.56
CD1g	CD1g_1	20.6	74.12
SS2	SS2_4	20.11	6.56
IN19	IN19_3	20.04	9.61

SS1	SS1_1	20	7.01
CD1j	CD1j_1	19.19	27.67
PL7	PL7_2	19.13	22.01
MH1g	MH1g_2	19.1	30.19
PA6	PA6_1	18.65	9.16
FV3	FV3_4	18.49	11.86
CD1h	CD1h_1	18.39	79.42
HHNBPERS_25_64	HHNBPERS_25_64_0	17.76	57.77
IN51	IN51_2	17.42	32.43
MH1f	MH1f_2	17.35	33.24
CD1e	CD1e_1	17.29	63.16
MH1e	MH1e_2	16.95	25.16
AL1	AL1_7	16.79	20.84
AM1	AM1_4	16.79	6.2
SS3	SS3_5	16.59	12.22
AM6a	AM6a_1	16.46	16.98
MAINSTAT	MAINSTAT_32	16.34	65.32
IN52	IN52_3	15.98	12.58
HHNBPERS_65plus	HHNBPERS_65plus_1	15.78	45.19
UN1a	UN1a_1	15.55	37.11

HHNBPERS	HHNBPERS_1	15.46	39.8
HHTYPE	HHTYPE_10	15.46	39.8
AL1	AL1_8	15.39	46.45
IN53	IN53_3	15.35	11.14
GE	GE_12	15.23	14.11
PA6	PA6_3	15.18	7.64
HH_ACT	HH_ACT_0	15.17	78.26
HHNBPERS_65plus	HHNBPERS_65plus_2	14.88	29.11
PL3	PL3_1	14.16	4.22
PA6	PA6_2	14.13	18.33
IMC_c	IMC_c_3	14.03	29.38
PE2	PE2_0	13.91	61.9
IN18	IN18_5	13.84	12.22
HS2	HS2_1	13.74	97.75
HO3	HO3_1	13.71	63.07
PA5	PA5_1	13.59	16.98
MD1	MD1_1	13.53	94.34
AM4	AM4_1	13.41	65.41
UN2b	UN2b_1	13.38	33.24
PN2	PN2_3	13.33	16.89

IN50	IN50_3	13.26	10.15
IN53	IN53_4	13.16	11.05
PA5	PA5_4	12.96	10.78
IN29	IN29_0	12.94	16.8
PA6	PA6_4	12.83	3.05
CD1k	CD1k_1	12.69	25.25
IN54	IN54_2	12.52	28.75
PE2	PE2_1	12.47	4.94
PA5	PA5_3	12.41	4.4
HHINCOME	HHINCOME_1	12.4	34.05
SEX	SEX_2	12.34	77.36
IN50	IN50_4	12.3	14.47
IN19	IN19_2	11.94	25.97
HHINCOME	HHINCOME_2	11.56	29.02
IN54	IN54_4	11.48	8.89
IN43	IN43_4	11.43	76.37
PE6	PE6_0	11.36	88.5
IN51	IN51_3	11.35	12.58
PE1	PE1_1	11.3	57.86
PA4	PA4_1	11.19	84.37

PA3	PA3_1	11.11	84.73
FV1	FV1_2	11.1	13.66
AM1	AM1_3	11.08	68.73
HS3	HS3_2	11.06	35.58
PL1	PL1_1	11.04	74.66
HH_INACT	HH_INACT_0	10.98	66.13
SS1	SS1_2	10.95	49.51
IN18	IN18_4	10.93	14.47
PA2	PA2_1	10.86	92.45
HHTYPE	HHTYPE_22	10.84	32.79
HHNBPERS	HHNBPERS_2	10.75	42.05
DEG_URB	DEG_URB_3	10.74	47.62
AM2	AM2_1	10.71	87.96
PN1	PN1_4	10.63	18.6
HHNBPERS_16_24	HHNBPERS_16_24_0	10.54	93.8
NUTS	NUTS_3	10.48	23.72
HHNBPERS_5_13	HHNBPERS_5_13_0	10.45	95.33
MH1a	MH1a_2	10.44	27.49
NUTS	NUTS_5	10.32	19.23
FV3	FV3_3	10.19	23.09

SS2	SS2_1	9.85	63.16
UN2c	UN2c_2	9.81	72.06
PE8	PE8_0	9.73	97.39
HHNBPERS_0_4	HHNBPERS_0_4_0	9.67	97.75
PE4	PE4_0	9.65	99.01
HHNBPERS_14_15	HHNBPERS_14_15_0	9.57	98.92
IC	IC_2	9.57	91.19
MARSTADEFACTO	MARSTADEFACTO_2	9.52	97.66
SK4	SK4_1	9.48	96.41

ARTIGO 3

Artigo em avaliação em revista científica internacional.

Title: A machine learning approach for classifying work absenteeism.

Santos, J[1 2], Silva, A[1] , Matias-Dias, C [2] , Chiavegatto Filho, ADP [1]

[1] School of Public Health - University of São Paulo, São Paulo, Brazil

[2] Department of Epidemiology, Institute of Public Health Dr Ricardo Jorge, Lisbon, Portugal

ABSTRACT

Absence from work attributed to diseases and accidents is increasing in most countries. Several studies show that the economic burden can be high and that it is important to identify and support these workers with tailored programs so that they are targeted earlier and job absenteeism is prevented. **OBJECTIVE:** This study aims to develop a predictive risk model for labor sickness absence using a national health dataset. **METHODS:** Data from the Portuguese national health interview survey was used (n=18,104). We analyzed data from all participants who were employed (n=6, 249). The dataset was split into two subsets: the first to train the model and find the best hyperparameters for each algorithm, and the second to evaluate the performance of the best model in new unseen data (70% and 30%, respectively). We tested three machine learning algorithms for classification tasks: logistic regression, random forest and gradient boosted trees. All models were compared in terms of sensitivity, specificity, precision and AUC. **RESULTS:** Random forests obtained the best AUC with 0.666, followed by logistic regression with 0.662 and gradient boosting trees with 0.660 in the test set. It also performed the best in terms of precision (67%) and specificity (99%). On the other hand, logistic regression obtained the best F1 (45%) followed by extreme Gradient Boosting with 44%. The former achieved 58% sensitivity and the latter 52%, but both showed poor precision. For the random forests algorithm the top predictors were: having some long standing health problem in the past 6 months, suffering from neck or back pain, arthrosis, depression, self-rated health and overall

satisfaction with life and status in employment. **DISCUSSION:** We developed three models that predict job absenteeism that achieved only fair performance, with an AUC of around 65-66%. Contrary to random forest, logistic regression and gradient boosting trees obtained good sensitivity at cost of poor precision. Variable importance analysis showed a set of characteristics that contributed the most to prediction scores. In conclusion, the variables from the national survey were not enough to achieve a high predictive performance of work absenteeism.

BACKGROUND

Absence from work attributed to disease or accident is an important concern as it represents a large economic and health burden (KOCAKULAH et al., 2016). As life expectancy increases along with the prevalence of chronic diseases, higher prevalences of disability are expected in the short and long run (KOCAKULAH et al., 2016). As public reforms that push retirement into later ages are increasing worldwide, labor absenteeism also tends to rise and this is especially challenging in countries with high levels of social protection, since it is associated with sickness and disability related pensions (GRILLON et al., 2018) (KIVIMÄKI et al., 2004).

Previous studies have predicted employees in risk of sickness absence (BOOT et al., 2017) (AIRAKSINEN et al., 2018). However, to our knowledge, none has analyzed data collected from a national and representative health survey. Predicting those who are more prone to be absent from work can be valuable because identifying in advance these workers can help decision makers from social services and companies to plan and support these individuals with preventive care (MICHIE; WILLIAMS, 2003).

Machine learning algorithms have shown to have good predictive performance for many health outcomes (WENG et al., 2017) (BHATLA; JYOTI, 2012) (KUMAR; KUMAR, 2016). There is also good evidence documenting how well they perform in large datasets, how good they handle correlated data and the fact that they do not have assumptions regarding the distribution of variables (KOLANOVIC; KRISHNAMACHARI, 2017). Machine learning algorithms model a function inductively from the information contained in the data used for training. The objective is to predict

the value of the variable of interest as accurately as possible, in addition to providing the model with a good generalization capacity, i.e. having a good performance in the face of data that were not used in training. Throughout the learning process, parameters of the predictor function are adjusted to the data in order to minimize the prediction error, and increase the generalization capacity. The area under the receiver operating characteristic (ROC) curve (AUC) is the most common performance measure for machine learning algorithms in classification tasks. This metric is the probability that a model has of selecting the positive example from two chosen at random. The closer the AUC is to 1, the better the classification algorithm.

OBJECTIVES

This study aimed to develop a predictive risk model for labor sickness absence from a national health dataset.

METHODS

Dataset

This is an observational and cross-sectional study based on data from the 2014 Portuguese National Health Survey (INS 2014), with a probability multistage sampling procedure stratified by NUTS II region (2002) to population 15 and over living in private households (n=18.104). Data collection was carried out between 10 September and 15 December 2014 based on computer-assisted personal interviews (CAPI) or computer-assisted web interviews (CAWI), with a 80.8% response rate nationwide. Further details on the questionnaire, sampling selection and fieldwork procedures are available from the methodological document (INE, 2018) as well as from the INS 2014 survey report (INE, INSTITUTO NACIONAL DE SAÚDE DOUTOR RICARDO JORGE, 2016). The questionnaire comprises of 263 questions regarding sociodemographic information, health status, disease and chronic conditions, accidents and injuries, absence from work due to health problems, functional limitations, personal care and

household activities mental health and satisfaction with life, pain, health care and access, unmet healthcare needs, medicine health care use, health determinants, social support and provision of informal care. The INS is part of the European Health Interview Survey (EHIS), coordinated by EUROSTAT and runs every five years in all member states to inform, monitor and compare health between countries. This is a polítical commitment regulated according to the EC n.1338/2008. As such, Committee of Ethics and informed consent of participants do not apply. Confidentiality and anonymization rules are strictly regulated and access to microdata is available upon request (EUROPEAN PARLIAMENT AND COUNCIL OF THE EU., 2009).

For the purpose of this study we analyzed participants who answered being employed (n=7,790). After removing missing values we ended with a sample of 6,249 participants. Out of these, 1.622 (26%) had been absent from work due to health problems in the past 12 months and 4,627 had not (74%).

Data pre processing

We initially selected sociodemographic variables, as well as variables related to health status, determinants and self-reported diseases, in a total of 60 (Annex 1). These variables contain information about age, sex, region, education, income, occupation, household characteristics, owing health insurance, self-health perception and long standing health problem, whether the participant has a chronic disease, BMI and any intensity pain experience felt lately. Other questions include mental wellbeing and satisfaction with life, as well as practices regarding physical activity, tobacco, alcohol, eating habits and perception regarding social support. As our database does not allow to be certain about when each event happened, we removed variables that could have occurred after being absent from work. In particular, we excluded variables that asked about the health status of participants in the last 4 weeks, such as intensity of bodily pain in the last 4 weeks, as well as those regarding mental health, to minimize the chances that they are the consequence of work absence. All features were categorical.

Data Split

The sample was randomly split into 70% and 30%, stratified according to the response variable distribution, to balance the class distributions within the splits. (KUHN; JOHNSON, 2013) Training set was composed of 4,375 respondents, with a class distribution of 1,136 and 3,239, corresponding to *absent* and *not absent* respectively. The test set was composed of 1,874 respondents, with a class distribution of 486 and 1,388, corresponding to *absent* and *non absent* respectively.

Model Evaluation

The performance of three popular machine learning algorithms (logistic regression, random forest and gradient boosted trees) was analyzed on the test set. With the exception of logistic regression that does not have hyperparameters, all algorithms had their hyperparameters tuned with 10-fold cross-validation with Bayesian optimization. Predictive performance was assessed using the area under the ROC curve (AUC). Other performance metrics calculated for each algorithm include F1-score, precision, sensitivity, and specificity. For tree-based algorithms we used the shapley additive explanations (SHAP) values as an approach to explain the output of our models (MOLNAR, 2020). SHAP is a popular technique to interpret machine learning decisions (MOLNAR, 2020).

RESULTS

Baseline characteristics of our sample.

Table 1 shows some of the characteristics of our sample. Given the large number of variables in our models we present a table with the socio-demographic variables and some health variables. Remaining variables description are in Annex 2.

	Total		Absent		Non absent	
Sex	n	%	n	%	n	%
male	3057	48.92	712	43.90	2345	50.68
female	3192	51.08	910	56.10	2282	49.32
Age categories						
15-29	615	9.84	153	9.43	462	9.98
30-34	718	11.49	178	10.97	540	11.67
35-39	983	15.73	256	15.78	727	15.71
40-44	978	15.65	253	15.60	725	15.67
45-49	886	14.18	216	13.32	670	14.48
50-54	833	13.33	225	13.87	608	13.14
55-59	673	10.77	193	11.90	480	10.37
60-64	404	6.47	112	6.91	292	6.31
65+	159	2.54	36	2.22	123	2.66
Region						
North	933	14.93	251	15.47	682	14.74
Center	1042	16.67	278	17.14	764	16.51
Lisbon and Tagus Valley	1048	16.77	278	17.14	770	16.64
Alentejo	758	12.13	189	11.65	569	12.30
Algarve	804	12.87	196	12.08	608	13.14
Madeira	808	12.93	212	13.07	596	12.88
Azores	856	13.70	218	13.44	638	13.79
Educational level						
None	139	2.22	39	2.40	100	2.16
Primary education	2024	32.39	548	33.79	1476	31.90
Lower secondary education	1263	20.21	319	19.67	944	20.40
Upper secondary education	1397	22.36	345	21.27	1052	22.74
Tertiary education	1426	22.82	371	22.87	1055	22.80
Full time or part time employed						
Yes	5844	93.52	1524	93.96	4320	93.37
No	405	6.48	98	6.04	307	6.63
Status in employment						
Self-employed	1031	16.50	209	12.89	822	17.77
Employee with a permanent job/work contract of unlimited duration	4388	70.22	1212	74.72	3176	68.64
Employee with a temporary job/work contract of limited duration	830	13.28	201	12.39	629	13.59

Net monthly equalized income of the household						
Below 1st quintile	699	11.19	186	11.47	513	11.09
Between 1st quintile and 2nd quintile	978	15.65	272	16.77	706	15.26
Between 2nd quintile and 3rd quintile	1362	21.80	348	21.45	1014	21.91
Between 3rd quintile and 4th quintile	1615	25.84	401	24.72	1214	26.24
Between 4th quintile and 5th quintile	1595	25.52	415	25.59	1180	25.50
Self perceived general health						
Very good	901	14.42	140	8.63	761	16.45
Good	2940	47.05	625	38.53	2315	50.03
Fair	2109	33.75	672	41.43	1437	31.06
Bad/Very bad	299	4.78	185	11.41	114	2.46
Long-standing health problem: Suffer from any illness or health problem of a duration of at least six months						
Yes	2941	47.06	1001	61.71	1940	41.93
No	3308	52.94	621	38.29	2687	58.07
General activity limitation						
Severely limited	225	3.60	164	10.11	61	1.32
Limted	1067	17.07	516	31.81	551	11.91
Not at all	4957	79.32	942	58.08	4015	86.77

Table 1. Baseline characteristics of the sample, according to sociodemographic and important health variables.

Our sample is composed of 6,249 individuals. It is well balanced with 49% men and 51% women. Around 42% are between 35 and 49 years old, which corresponds to a significant part of the professionally active population. Around 60% has at least the lower secondary education and 70% has a permanent job. Around 34% of these individuals live in the center and in the Lisbon and Tagus valley.

Model results

Table 2 presents the predictive performance on the test data for the three machine learning algorithms (logistic regression, random forest, and gradient boosted trees). Risk scores distribution for each algorithm is presented in Figure 1.

	Gradient Boosting	**Random Forest**	**Logistic Regression**
Precision	0.3847	0.667	0.370
Sensitivity/Recall	0.5175	0.021	0.583
Specificity	0.7097	0.996	0.651
F1 Score	0.4413	0.040	0.453
AUC	0.6543	0.666	0.662

Table 2. Performance metrics for the three models in the test set.

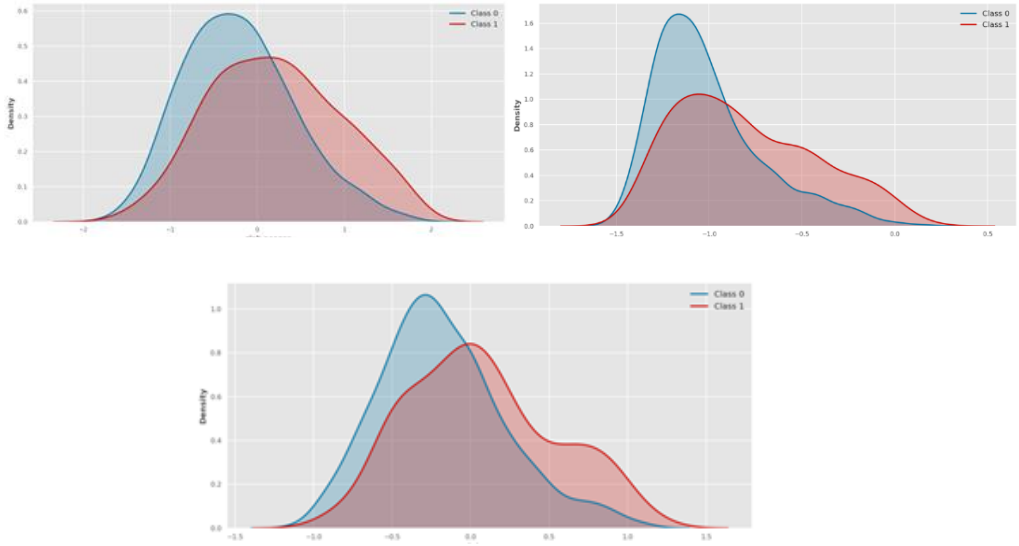


Fig 1. Distribution of the standard probabilities on the x axis for three models: at top left logistic regression, top right Random Forest and at the bottom Gradient Boosting Trees.

Legend: The figure shows that the standard probabilites for each class meaning these risks overlap and are not well separated, reflecting the challenge in distinguishing both classes with features in our dataset.

All models showed very similar performance in terms of AUC (Table 2). The Random Forest obtained the best AUC and specificity (0.666 and 0.996 respectively), but performed poorly for sensitivity with 0.021. Gradient Boosting Tree also obtained good AUC (0.6543) but the and F1-Score improved (0.518 and 0.441). In terms of precison (i.e. positive predictive value), Random Forests was the best algorithm (0.667).

The SHAP plot (Figure 2) ranks by decreasing order of importance the variables for the Random Forests algorithm. We observe that not having any long-standing health problem, suffering from a low back disorder or other chronic back defect in the past 12 months, suffering from a neck disorder or other chronic neck defect, depression, self-health perception, suffering from arthrosis, status in employment (self-employed or employed with a job of unlimited duration) and satisfaction with life were the most important predictive variables.

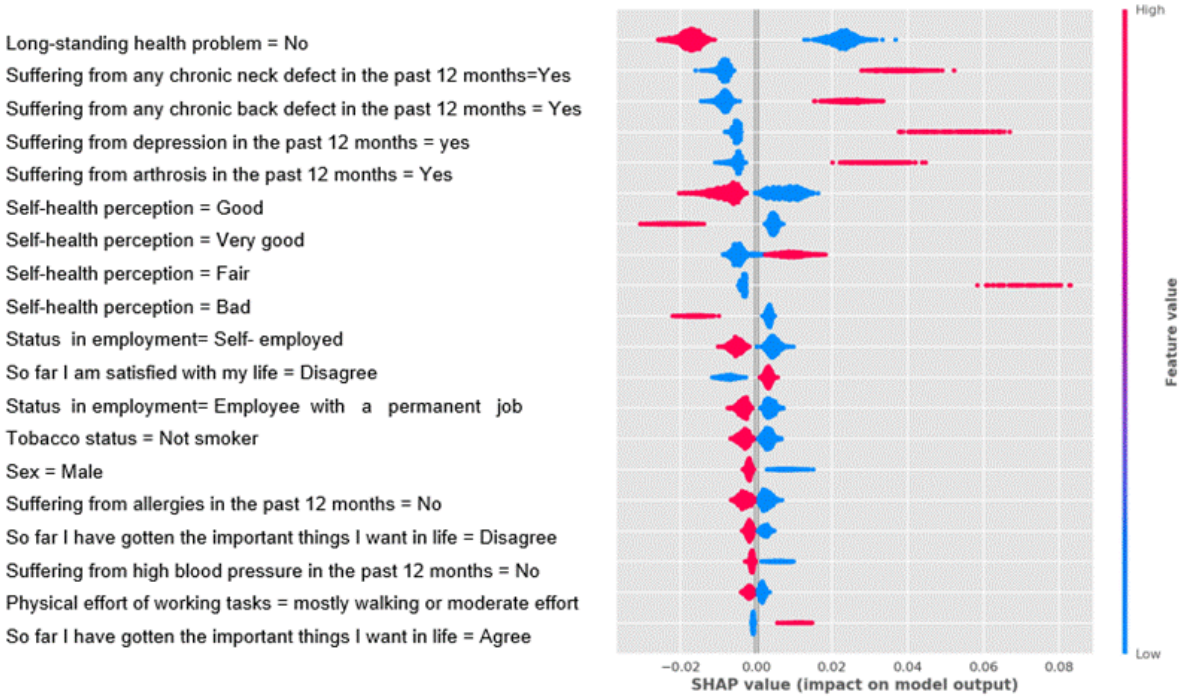


Fig 2. SHAP plot for Random Forest model.

Legend: In this summary plot, the order of the columns represents the amount of information the column is accountable for in the prediction, with a similar interpretation to feature importance in tree-based models. Each dot in the visualization represents one prediction. If the actual value in the dataset was high, the color is pink; blue indicates the actual value being low. However, these are categorical values, meaning binary values are high when is 1 and low if is 0. For categories with more than one class, like self-health perception pink is 1 if the class for that line/ variable is present and 0 otherwise. The x-axis represents the SHAP value, which is the impact on the model output. The greater the SHAP value, the higher the impact for the prediction.

We can observe that suffering from chronic neck or back defect has a high predictive power in absenteeism, as well as depression and arthrosis. Likewise, negative SHAP values push the probability to lower values of our dependent variable, meaning very good and good health and being self-employed increase the probability of not being absent. For gradient boosting trees, practicing physical activities for 3 days a week and being 50-54 years old decreased the probability of being absent.

DISCUSSION

Our study applied a predictive model to identify in advance individuals who are more at risk of being absent from work due to health problems. We used three popular machine learning models for binary classification problems, with a previous history of good predictive performance (BREIMAN et al., 2017) (BREIMAN, 2001b). We obtained an AUC of 0.666 for the random forest model when evaluating the model in unseen data. Other studies found better performance (but for the long term abstinence and using different data and methods) with an AUC of 0.720 (BOOT et al., 2017) and 0.730 (AIRAKSINEN et al., 2018). All three algorithms had a similar low AUC of around 65%-66%.

The most important metric for companies in this example is probably the positive predictive value, so that the percentage of false positives is kept low, decreasing the probability of useless costly spending in preventive measures. In this case, the random forests algorithm achieved an interesting result of 0.667. The density plots however show a small separation ability between classes, indicating that even adopting other approaches such as the cutoff point will not improve our model significantly. It also shows that increasing sensitivity will also decrease precision which will in turn increase the number of false positives. Yet, even though there is this constraint regarding the predictive ability of variables in our dataset, we demonstrate nation-wide data can bring insights regarding profile, health and well-being of patients.

For the random forests algorithm, we found a set of diseases to be highly predictive of absence, namely chronic back and neck pain. This is in line with a systematic review that concluded that poor musculoskeletal capacity is positively associated with a lower score in the Work Ability Index (WAI) (VAN DEN BERG; ELDERS; BURDORF, 2009), a questionnaire that measures the work ability of employees, by taking into account sick leaves (ILMARINEN, 2006), as well as poor working postures and poor ergonomic factors at the workplace. Although these variables are not in our study, it is consistent with our main predictors. Further, bad self-rated health was found to be a top predictor with a high SHAP value. It is also interesting to note that regarding the impact of these variables in the model, we observe good and very good health perception to have low

SHAP values, therefore reducing the predictive probability, while fair health is in the center of the plot meaning it had low impact in the prediction.

Our study has the limitation of only using variables available in a cross-sectional national survey, which could mean that some variables occur after the outcome. Although many variables such as sex, age and region are reasonable to consider that happened mostly before absenteeism, this is not true for every variable. The lack of strong predictors affected the performance of our models, which is clear from density plots that show overlapping classes.

CONCLUSION

Our study found that it was difficult to identify workers who are at high risk of absenteeism due to chronic and preventable circumstances using a national health survey. By including more predictive variables, possibly through a questionnaire designed with this purpose, a risk score could support the creation of a system that assists in the prediction of absenteeism by companies or other social services, providing companies with information to anticipate preventive measures.

REFERENCES

ABDI, H.; VALENTIN, D. **Multiple Correspondence Analysis.** Disponível em: </www.utd.edu/~herve>. Acesso em: 29 jun. 2020.

AIRAKSINEN, J. et al. Prediction of long-term absence due to sickness in employees: Development and validation of a multifactorial risk score in two cohort studies. **Scandinavian Journal of Work, Environment and Health**, v. 44, n. 3, p. 274–282, 2018.

ALBOUKADEL KASSAMBARA. **Cluster validation essentials**.

AMARAL, F. **Introdução à Ciencia de Dados mineração de dados e big data**. Rio Janeiro: Alta Books, 2015.

Análise Multivariada de Dados. [s.l: s.n.]

ARIMOND, G.; ELFESSI, A. A clustering method for categorical data in tourism market segmentation research. **Journal of Travel Research**, v. 39, n. 4, p. 391–397, 2001.

ATHEY, S. Machine Learning and Causal Inference for Policy Evaluation. p. 5–6, 2015.

AUSTIN, P. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. **Statist. Med.**, n. December 2006, p. 3385–3397, 2009.

AUSTIN, P. C. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. **Pharmaceutical Statistics**, v. 10, n. 2, p. 150–161, 2011.

BAKER, P. et al. The men's health gap: Men must be included in the global health equity agenda. **Bulletin of the World Health Organization**, v. 92, n. 8, p. 618–620, 2014.

BERRY, M. J.; LINOFF, G. **Data Mining Techniques: For Marketing, Sales, and Customer Support**. [s.l: s.n.]

BHATLA, N.; JYOTI, K. An analysis of heart disease prediction using different data mining techniques. **International Journal of Engineering Research & Tecnology**, v. 1, n. 8, p. 1–4, 2012.

BOOT, C. R. L. et al. Prediction of long-term and frequent sickness absence using company data. **Occupational Medicine**, v. 67, n. 3, p. 176–181, 2017.

BREIMAN, L. Statistical Modeling: The Two Cultures. **Statistical Science**, v. 16, n. 3, p. 199–231, 2001a.

BREIMAN, L. Random Forests. **Machine Learning**, n. 45, p. 5–32, 2001b.

BREIMAN, L. et al. Classification and regression trees. **Classification and Regression Trees**, p. 1–358, 2017.

C.M., L.; V.H., V. B.; H.B., F. Application of unsupervised analysis techniques to lung

cancer patient data. **PLoS ONE**, v. 12, n. 9, p. 1–18, 2017. Disponível em: <http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L618273302%0Ahttp://dx.doi.org/10.1371/journal.pone.0184370>.

CHENG LI. A Gentle Introduction to Gradient Boosting. **Seamless R and C++ Integration with Rcpp**, p. 3–18, 2013. Disponível em: <http://link.springer.com/10.1007/978-1-4614-6868-4_1>.

CHOR, J. S. Y. et al. Are men seeking medical advice too late? Contacts to general practitioners and hospital admissions in Denmark 2005. **Journal of Public Health**, v. 30, n. 1, p. 110–111, 2008.

CONASS. **Programa Mais Médicos-Nota Tecnica 23/2013.**Brasilia, 2013. . Disponível em: <http://maismedicos.gov.br/>.

CONCOLATO, C. E.; CHEN, L. M. Data Science: A New Paradigm in the Age of Big-Data Science and Analytics. **New Mathematics and Natural Computation**, v. 13, n. 2, p. 119–143, 2017.

COVENEY, P. V; DOUGHERTY, E. R.; HIGHFIELD, R. R. Big data need big theory too Subject Areas : Author for correspondence : **Philosophical Transactions of the Royal Society A**, 2016.

CRC, H.; MURRELL, P. **Exploratory Multivariate Analysis by Example Using R Introduction to Data Technologies**. [s.l: s.n.]

DATASUS. **DATASUS**. Disponível em: <http://datasus.saude.gov.br/>. Acesso em: 7 jul. 2017a.

DATASUS. **Programa Mais Médicos. Governo Federal.** Disponível em: <http://maismedicos.gov.br/conheca-programa>. Acesso em: 23 ago. 2018b.

DHAR, V. Data Science and Prediction Vasant Dhar Professor, Stern School of Business Director, Center for Digital Economy Research. **Communications of the ACM**, n. May, p. 64–73, 2012. Disponível em: <http://archive.nyu.edu/bitstream/2451/31553/2/Dhar-DataScience.pdf>.

DHL. Big Data in Logistics. **DHL Customer Solutions & Innovation**, n. December,

p. 1–30, 2013.

DIAS, C. M. 25 anos de Inquérito Nacional de Saúde em Portugal. **Revista Portuguesa de Saúde Pública**, p. 51–60, 2009.

DONOHO, D. 50 Years of Data Science. **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745–766, 2017. Disponível em: <https://doi.org/10.1080/10618600.2017.1384734>.

EIRA, A. **A Saúde em Portugal: A procura de cuidados de saúde privados**. 2010. Faculdade Economia Porto, 2010. Disponível em: <https://repositorio-aberto.up.pt/bitstream/10216/26931/2/A saude em Portugal A procura privada de cuidados de saude Ana Eira.pdf>.

ELIZABETH A. STUART, BRIAN K. LEE, AND JUSTIN LESSLER, S. Improving propensity score weighting using machine learning. **Statist. Med.**, v. 29, n. 3, p. 337/346, 2010.

EUROPEAN COMMISSION. COMMISSION REGULATION (EU) No 141/2013 of 19 February 2013 implementing Regulation (EC) No 1338/2008 of the European Parliament and of the Council on Community statistics on public health and health and safety at work, as regards statistics based on the E. v. 2013, n. 141, 2013. Disponível em: <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32013R0141>.

EUROPEAN PARLIAMENT AND COUNCIL OF THE EU. Regulation (EC) N.223/2009. . 2009, p. 164–173.

EUROSTAT. **EUROPEAN HEALTH INTERVIEW SURVEY (EHIS).** Disponível em: <https://ec.europa.eu/eurostat/web/microdata/european-health-interview-survey>. Acesso em: 10 out. 2019.

FERREIRA, K. J. Analytics for an Online Retailer: Demand Forecasting and Price Optimization. v. 39, n. 5, p. 293–296, 2008.

FERRI, C.; HERNÁNDEZ-ORALLO, J.; MODROIU, R. **An experimental comparison of performance measures for classificationPattern Recognition Letters**, 2009. .

FISKE, A.; GATZ, M.; PEDERSEN, N. L. Depressive Symptoms and Aging: The Effects of Illness and Non-Health-Related Events. **Journals of Gerontology - Series B Psychological Sciences and Social Sciences**, v. 58, n. 6, p. 320–328, 2003.

FONTANELLA, S. et al. Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: A cross-sectional analysis within a population-based birth cohort. **PLoS Medicine**, v. 15, n. 11, p. 1–22, 2018.

FRANCO, G. Di. Multiple correspondence analysis : one only or several techniques ? **Quality & Quantity**, p. 1299–1315, 2016. Disponível em: <http://dx.doi.org/10.1007/s11135-015-0206-0>.

FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS. **Exploratory Multivariate Analysis using R**. [s.l.] CRC Press, 2005.

GRILLON, C. et al. Scale-up of HIV self-testing. **The Lancet HIV**, v. 5, n. 7, p. e324–e326, 2018. Disponível em: <http://dx.doi.org/10.1016/S2352-3018(18)30095-X>.

HAIR, JR, RE ANDERSON, B. W. **Analise Multivariada de Dados**. 5ª ed. Porto Alegre: Bookman, 2005.

HASTIE, T. et al. **An Introduction to Statistical Learning, Springer Texts**. CA, USA: Springer, 2013. v. 102

HINDMAN, M. Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. **Annals of the American Academy of Political and Social Science**, v. 659, n. 1, p. 48–62, 2015. Disponível em: <https://doi.org/10.1177/0002716215570279>.

HJELLBREKKE, J. **Multiple Correspondence Analysis for Social Sciences**. New York: Routledge. Taylor and Francis Group., 2007.

HUSSON, F. JOSSE, J, PAGÉS, J. **FactoMineR. Hierarchical Clustering on Principal Components.** Disponível em: <http://factominer.free.fr>.

HUSSON, A. F. et al. Package ‘ FactoMineR ’. 2018.

HUSSON, F. **Handling missing values in MCA**, 2016. .

ILMARINEN, J. The Work Ability Index (WAI). **Occupational Medicine**, v. 57, n. 2, p.

160–160, 2006.

INE, INSTITUTO NACIONAL DE SAÚDE DOUTOR RICARDO JORGE, I. **Inquérito Nacional Saúde**Lisboa, 2016. .

INE. **DOCUMENTO METODOLÓGICO.National Health Survey**. [s.l: s.n.].

IZBICKI, R. Machine Learning sob a ótica estatística. 2016.

JOSSE, J. Technical Report – Agrocampus. n. September, 2010.

JOSSE, J.; HUSSON, F.; LÊ, S. **Multiple Correspondence Analysis**, 2008. .

JOSSELIN, J.-M.; LE MAUX, B. Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education. In: SPRINGER (Ed.). **Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education.** 1st 2017 ed. [s.l.] Springer International Publishing, 2017. p. 489–531.

KASSAMBARA, A. **Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis Book 1)**. [s.l: s.n.]

KHAN, A.; UDDIN, R.; ISLAM, S. M. S. **Clustering patterns of behavioural risk factors for cardiovascular diseases in Bangladeshi adolescents: A population-based studyHealth Policy and Technology**, 2019. .

KIVIMÄKI, M. et al. Sickness absence as a risk marker of future disability pension: The 10-town study. **Journal of Epidemiology and Community Health**, v. 58, n. 8, p. 710–711, 2004.

KOCAKULAH, M. C. et al. Absenteeism Problems And Costs: Causes, Effects And Cures. **International Business & Economics Research Journal (IBER)**, v. 15, n. 3, p. 89–96, 2016.

KOHAVI, R. Cross validation number.pdf. v. 5, p. 7, 1995. Disponível em: <http://robotics.stanford.edu/~ronnyk>.

KOLANOVIC, M.; KRISHNAMACHARI, R. T. Big Data and AI Strategies. n. May, p. 280, 2017.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Supervised Machine Learning: A Review of Classification Techniques. **Artificial Intelligence Review**, v. 26, p. 159–190, 2006.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling [Hardcover]**. [s.l: s.n.]

KUMAR, M.; KUMAR, M. International Journal of Computer Science and Mobile Computing Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. **International Journal of Computer Science and Mobile Computing**, v. 5, n. 2, p. 24–33, 2016. Disponível em: <www.ijcsmc.com>.

LANDIS, G. K. The measurement of observer agreement for categorical data. **Biometrics**, 1977.

LEG/UFPR. **Métodos baseados em árvores.** Disponível em: <http://cursos.leg.ufpr.br/>. Acesso em: 17 jun. 2020.

LEITE, W. **Practical propensity score methods using R**. USA: SAGE Publications, 2017.

LEVORATO, C. D. et al. Fatores associados à procura por serviços de saúde numa perspectiva relacional de gênero. **Ciencia e Saude Coletiva**, v. 19, n. 4, p. 1263–1274, 2014.

LINDEN, A.; YARNOLD, P. R. Combining machine learning and matching techniques to improve causal inference in program evaluation. **Journal of Evaluation in Clinical Practice**, v. 22, n. 6, p. 864–870, 2016.

LUND, T. et al. Using administrative sickness absence data as a marker of future disability pension: The prospective DREAM study of Danish private sector employees. **Occupational and Environmental Medicine**, v. 65, n. 1, p. 28–31, 2008.

MCLAREN, L.; PETIT, R. Universal and targeted policy to achieve health equity: a critical analysis of the example of community water fluoridation cessation in Calgary, Canada in 2011. **Critical Public Health**, v. 28, n. 2, p. 153–164, 2018. Disponível em: <http://doi.org/10.1080/09581596.2017.1361015>.

MICHIE, D. Methodologies from machine learning in data analysis and software. **Computer Journal**, v. 34, n. 6, p. 559–565, 1991.

MICHIE, S.; WILLIAMS, S. Reducing work related psychological ill health and sickness absence: A systematic literature review. **Occupational and Environmental Medicine**, v. 60, n. 1, p. 3–9, 2003.

MIGUEL A. CARREIRA-PERPINAN. **Dimensionality Reduction**. [s.l: s.n.]

MIKE LOUKIDES. **What is Data Science?** 1st editio ed. [s.l.] O'Reilly Media, 2016.

MINISTERIO SAUDE. **TC 80: Cooperação Técnica entre o Ministério da Saúde e a Organização Pan-Americana da Saúde - Projeto Acesso da população Brasileira à Atenção Básica em Saúde.**Brasilia, 2013. .

MOLNAR, C. **Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.** Disponível em: <https://christophm.github.io/>. Acesso em: 27 jun. 2020.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [s.l: s.n.]

NASCIMENTO, A. do. **Avaliação de farmácias hospitalares Brasileiras utilizando Análise de Correspondência Múltipla.** 2011. UFRJ - Universidade Federal do Rio de Janeiro., 2011.

NENADIC, O.; GREENACRE, M. Computation of Multiple Correspondence Analysis, with Code in R. **Ssrn**, p. 25–27, 2005.

NGUYEN, L. H.; HOLMES, S. Ten quick tips for effective dimensionality reduction. **PLoS Computational Biology**, v. 15, n. 6, p. 1–19, 2019.

PAHO/WHO. **Ministério da Saúde reforça cooperação com a OPAS para continuidade do Mais Médicos**. Disponível em: <http://portalms.saude.gov.br/noticias/agencia-saude/42756-ministerio-da-saude-reforca-cooperacao-com-a-opas-para-continuidade-do-mais-medicos>.

PEDROSO, R. T.; JUHÁSOVÁ, M. B.; HAMANN, E. M. **A ciência baseada em evidências nas políticas públicas para reinvenção da prevenção ao uso de álcool e outras drogasInterface - Comunicação, Saúde, Educação**, 2019. .

POHAR, M.; BLAS, M.; TURK, S. Comparison of Logistic Regression and Linear Discriminant Analysis : A Simulation Study. **Metodološki zvezki**, v. 1, n. 1, p. 143–161, 2004.

QUINTAL, C.; LOURENÇO, Ó.; FERREIRA, P. **Utilização de cuidados de saúde pela população idosa portuguesa: Uma análise por género e classes latentesRevista Portuguesa de Saude Publica**, 2012. .

RAMOS, M. C.; SILVA, E. N. da. Como usar a abordagem da Política Informada por Evidência na saúde pública? **Saúde em Debate**, v. 42, n. 116, p. 296–306, 2018. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-11042018000100296&lng=pt&tlng=pt>.

RUBIN, D. B. Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. **Journal of Educational Psychology**, v. 66, n. 5, p. 688–701, 1974.

RUFFI, A. et al. Difi culdades de acesso a serviços de saúde para diagnóstico de tuberculose em municípios do Brasil Diffi culties in the accessibility to health services for tuberculosis. v. 43, n. 3, p. 389–397, 2009.

RUSSO, L. X. et al. Primary care physicians and infant mortality: Evidence from Brazil. **PLoS ONE**, v. 14, n. 5, p. 1–16, 2019.

RYGIELSKI, C.; WANG, J.; YEN, D. C. Data mining techniques for customer relationship management. v. 24, p. 483–502, 2002.

SAMUEL, A. L. Some studies in Machine Learning using the Game of Checkers. **IBM J. RES. DEVELOP**, v. 44, n. 1, 2000.

SANTOS, H. G. dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. p. 207, 2018.

SCHEFFER, M. et al. **Demografia médica no Brasil 2015**. São Paulo: Departamento de Medicina Preventiva, Faculdade de Medicina da USP. Conselho Regional de Medicina do Estado de São Paulo. Conselho Federal de Medicina., 2015.

SCHNEIDER, P. Why should the poor insure? Theories of decision-making in the context of health insurance. **Health Policy and Planning**, v. 19, n. 6, p. 349–355, 2004.

SHADISH, W. R. Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings. **Psychological Methods**, v. 15, n. 1, p. 3–17, 2010.

SHAW, M. J. et al. Knowledge management and data mining for marketing. p. 127–137, 2001.

SHI, L. et al. Primary care, infant mortality, and low birth weight in the states of the USA. **Journal of Epidemiology and Community Health**, v. 58, n. 5, p. 374–380, 2004.

SILVER, N. **Statistics Views**. Disponível em: <https://www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html>. Acesso em: 7 ago. 2019.

THE HENRY J KAISER FAMILY FOUNDATION. **Key Facts about the Uninsured PopulationThe Henry J Kaiser Family Foundation**, 2016. . Disponível em: <https://kaiserfamilyfoundation.files.wordpress.com/2013/09/8488-key-facts-about-the-uninsured-population.pdf>.

TRAVASSOS, C. et al. Utilização dos serviços de saúde no Brasil: Gênero, características familiares e condição social. **Revista Panamericana de Salud Publica/Pan American Journal of Public Health**, v. 11, n. 5–6, p. 365–373, 2002.

UNIVERSITY OF CHICAGO. **Data Science for Social Good.**

VAN DEN BERG, T. I. J.; ELDERS, L. A. M.; BURDORF, A. The effects of work-related and individual factors on work ability: A systematic review. **Promotion of Work Ability Towards Productive Aging - Selected Papers of the 3rd International Symposium on Work Ability**, p. 15–18, 2009.

VIACAVA, F.; SOUZA-JÚNIOR, P. R. B. de; SZWARCWALD, C. L. Cobertura da população brasileira com 18 anos ou mais por plano de saúde privado: uma análise dos dados da Pesquisa Mundial de Saúde. **Cad Saude Publica**, v. 21, n. supl.1, p.

S119–S128, 2005.

WENG, S. F. et al. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? **PLoS ONE**, v. 12, n. 4, p. 1–14, 2017.

WILPER, A. P. et al. Health insurance and mortality in US adults. **American Journal of Public Health**, v. 99, n. 12, p. 2289–2295, 2009.

ANNEX 1

List of variables in our model:

[1] Region

[2] Degree of urbanization

[3] Sex

[4] Age group

[5] Legal marital status

[6] De facto marital status

[7] Has always lived in Portugal

[8] Educational level

[9] Full time or part time employed

[10] Status in employment

[12] Number of persons living in household, including the respondent

[13] Number of persons aged 4 or younger

[14] Number of persons aged from 5 to 13

[15] Number of persons aged from 14 to 15

[16] Number of persons aged from 16 to 24

[17] Number of persons aged from 25 to 64

[18] Number of persons aged 65 and over

[19] Type of household

[20] Number of persons aged 16 – 64 in the household who are employed

[21] Number of persons aged 16 – 64 in the household who are unemployed or are economically inactive

[22] Net monthly equalized income of the household

[23] Does own a private health plan?

[24] Self perceived general health

[25] Long-standing health problem: Suffer from any illness or health problem of a duration of at least six months

[26] Suffering from ashtma in the past 12 months

[27] Suffering from chronic bronchitis, chronic obstructive pulmonary disease or emphysema in the past 12 months

[28] Suffering from a myocardial infarction (heart attack) in the past 12 months

[29] Suffering from a coronary heart disease or angina pectoris in the past 12 months

[30] Suffering from high blood pressure in the past 12 months

[31] Suffering from a stroke (cerebral hemorrhage, cerebral thrombosis) in the past 12 months

[32] Suffering from arthrosis (arthritis excluded) in the past 12 months

[33] Suffering from a low back disorder or other chronic back defect in the past 12 months

[34] Suffering from a neck disorder or other chronic neck defect in the past 12 months

[35] Suffering from diabetes in the past 12 months

[36] Suffering from an allergy, such as rhinitis, eye inflammation, dermatitis, food allergy or other (allergic asthma excluded) in the past 12 months

[37] Suffering from cirrhosis of the liver in the past 12 months

[38] Suffering from urinary incontinence, problems in controlling the bladder in the past 12 months

[39] Suffering from kidney problems in the past 12 months

[40] Suffering from depression in the past 12 months

[41] Absent from work due to personal health problems in the past 12 months

[42] Physical effort of working tasks (both paid and unpaid work activities included)

[43] Number of days in a typical week doing sports, fitness or recreational (leisure) physical activities that cause at least a small increase in breathing or heart rate for at least 10 minutes continuously

[44] Number of days in a typical week doing muscle-strengthening activities

[45] Number of meals a day

[46] Frequency of eating fruit, excluding juice

[47] Frequency of eating vegetables or salad, excluding juice and potatoes

[48] Tobacco status

[49] Use of e-cigarette

[50] Frequency of exposure to tobacco smoke indoors

[51] Frequency of consumption of an alcoholic drink of any kind (beer, wine, cider, spirits, cocktails, premixes, liqueurs, homemade alcohol…) in the past 12 months

[52] Until what extent do you agree or disagree that, in most ways your life is close to my ideal

[53] Until what extent do you agree or disagree that you are satisfied with your life

[54] Until what extent do you agree or disagree that so far you have gotten the important things you want in life

[55] Until what extent do you agree or disagree that if you could live my life over, you would change almost nothing

[56] Until what extent do you agree or disagree that the conditions of my life are excellent

[57] Number of close people to count on in case of serious personal problems

[58] Degree of concern shown by other people in what the person is doing

[59] How easy is it to get practical help from neighbors in case of need

[60] Providing care or assistance to one or more persons suffering from some age problem, chronic health condition or infirmity, at least once a week

[61] Body Mass Index categories

ANNEX 2

A2. Description of our sample with total counts and percentages for total sample and according to those who absented from work and did not.

Sex	Total		Absent		Non absent	
	n	%	n	%	n	%
male	3057	48.92	712	43.90	2345	50.68
female	3192	51.08	910	56.10	2282	49.32
Age categories						
15-29	615	9.84	153	9.43	462	9.98
30-34	718	11.49	178	10.97	540	11.67
35-39	983	15.73	256	15.78	727	15.71
40-44	978	15.65	253	15.60	725	15.67
45-49	886	14.18	216	13.32	670	14.48
50-54	833	13.33	225	13.87	608	13.14
55-59	673	10.77	193	11.90	480	10.37
60-64	404	6.47	112	6.91	292	6.31
65+	159	2.54	36	2.22	123	2.66
Region						
North	933	14.93	251	15.47	682	14.74
Center	1042	16.67	278	17.14	764	16.51

Lisbon and Tagus Valley	1048	16.77	278	17.14	770	16.64
Alentejo	758	12.13	189	11.65	569	12.30
Algarve	804	12.87	196	12.08	608	13.14
Madeira	808	12.93	212	13.07	596	12.88
Azores	856	13.70	218	13.44	638	13.79
Degree of urbanization						
Densely-populated area	1919	30.71	518	31.94	1401	30.28
Intermediate-populated area	2129	34.07	525	32.37	1604	34.67
Thinly-populated area	2201	35.22	579	35.70	1622	35.06
Legal Marital status						
Single	1666	26.66	416	25.65	1250	27.02
Married	3608	57.74	920	56.72	2688	58.09
Widowed	210	3.36	66	4.07	144	3.11
Divorced	765	12.24	220	13.56	545	11.78
De facto marital status						
Person living in a consensual union	639	10.23	184	11.34	455	9.83
Person not living in a consensual union	5610	89.77	1438	88.66	4172	90.17
Always lived in Portugal						
Yes	5573	89.18	1446	89.15	4127	89.19
No	676	10.82	176	10.85	500	10.81
Educational level						

None	139	2.22	39	2.40	100	2.16
Primary education	2024	32.39	548	33.79	1476	31.90
Lower secondary education	1263	20.21	319	19.67	944	20.40
Upper secondary education	1397	22.36	345	21.27	1052	22.74
Tertiary education	1426	22.82	371	22.87	1055	22.80
Full time or part time employed						
Yes	5844	93.52	1524	93.96	4320	93.37
No	405	6.48	98	6.04	307	6.63
Status in employment						
Self-employed	1031	16.50	209	12.89	822	17.77
Employee with a permanent job/work contract of unlimited duration	4388	70.22	1212	74.72	3176	68.64
Employee with a temporary job/work contract of limited duration	830	13.28	201	12.39	629	13.59
Number of persons living in household, including the respondent						
1	920	14.72	239	14.73	681	14.72
2	1614	25.83	443	27.31	1171	25.31
3	1911	30.58	498	30.70	1413	30.54
4	1415	22.64	337	20.78	1078	23.30
5	288	4.61	74	4.56	214	4.63
6+	101	1.62	31	1.91	70	1.51
Number of persons aged 4 or younger						

Any	5345	85.53	1383	85.27	3962	85.63
1+	904	14.47	239	14.73	665	14.37
Number of persons aged from 5 to 13						
Any	4431	70.91	1174	72.38	3257	70.39
1	1428	22.85	346	21.33	1082	23.38
2+	390	6.24	102	6.29	288	6.22
Number of persons aged from 14 to 15						
Any	5780	92.49	1521	93.77	4259	92.05
1+	469	7.51	101	6.23	368	7.95
Number of persons aged from 16 to 24						
Any	4867	77.88	1285	79.22	3582	77.42
1	1121	17.94	281	17.32	840	18.15
2+	261	4.18	56	3.45	205	4.43
Number of persons aged from 25 to 64						
Any	124	1.98	24	1.48	100	2.16
1	1613	25.81	414	25.52	1199	25.91
2	3861	61.79	1000	61.65	2861	61.83
3+	651	10.42	184	11.34	467	10.09
Number of persons aged 65 and over						
Any	5593	89.50	1452	89.52	4141	89.50
1	492	7.87	131	8.08	361	7.80

2+	164	2.62	39	2.40	125	2.70
Type of household						
Other type of household	920	14.72	239	14.73	681	14.72
One person household	1307	20.92	1022	63.01	285	6.16
Lone parent with child(ren) aged less than 25	1059	16.95	296	18.25	763	16.49
Couple without child(ren) aged less than 25	2594	41.51	622	38.35	1972	42.62
Couple with child(ren) aged less than 25	207	3.31	61	3.76	146	3.16
Couple or lone parent with child(ren) aged less than 25 and other persons living in household	1082	17.31	302	18.62	780	16.86
Number of persons aged 16 – 64 in the household who are employed						
Any	35	0.56	26	1.60	9	0.19
1	2352	37.64	642	39.58	1710	36.96
2	3200	51.21	807	49.75	2393	51.72
3	645	10.32	165	10.17	480	10.37
Number of persons aged 16 – 64 in the household who are unemployed						
Any	4243	67.90	1086	66.95	3157	68.23
1	1558	24.93	412	25.40	1146	24.77
2 +	448	7.17	124	7.64	324	7.00
Net monthly equalized income of the household						
Below 1st quintile	699	11.19	186	11.47	513	11.09
Between 1st quintile and 2nd quintile	978	15.65	272	16.77	706	15.26
Between 2nd quintile and 3rd quintile	1362	21.80	348	21.45	1014	21.91

Between 3rd quintile and 4th quintile	1615	25.84	401	24.72	1214	26.24
Between 4th quintile and 5th quintile	1595	25.52	415	25.59	1180	25.50
Does own a private health plan?						
Yes	1599	25.59	404	24.91	1195	25.83
No	4650	74.41	1218	75.09	3432	74.17
Self perceived general health						
Very good	901	14.42	140	8.63	761	16.45
Good	2940	47.05	625	38.53	2315	50.03
Fair	2109	33.75	672	41.43	1437	31.06
Bad/V bad	299	4.78	185	11.41	114	2.46
Long-standing health problem: Suffer from any illness or health problem of a duration of at least six months						
Yes	2941	47.06	1001	61.71	1940	41.93
No	3308	52.94	621	38.29	2687	58.07
General activity limitation						
Severely limited	225	3.60	164	10.11	61	1.32
Limted	1067	17.07	516	31.81	551	11.91
Not at all	4957	79.32	942	58.08	4015	86.77
Suffering from ashtma in the past 12 months						
Yes	217	3.47	93	5.73	124	2.68
No	6032	96.53	1529	94.27	4503	97.32

Suffering from chronic bronchitis, chronic obstructive pulmonary disease or emphysema in the past 12 months						
Yes	201	3.22	75	4.62	126	2.72
No	6048	96.78	1547	95.38	4501	97.28
Suffering from a myocardial infarction (heart attack) in the past 12 months						
Yes	37	0.59	19	1.17	18	0.39
No	6212	99.41	1603	98.83	4609	99.61
Suffering from a coronary heart disease or angina pectoris in the past 12 months						
Yes	93	1.49	49	3.02	44	0.95
No	6156	98.51	1573	96.98	4583	99.05
Suffering from high blood pressure in the past 12 months						
Yes	1013	33.51	332	20.47	681	14.72
No	5236	33.51	1290	79.53	3946	85.28
Suffering from a stroke (cerebral hemorrhage, cerebral thrombosis) in the past 12 months						
Yes	33	0.53	14	0.86	19	0.41
No	11787	188.62	1608	99.14	4608	99.59
Suffering from arthrosis (arthritis excluded) in the past 12 months						
Yes	794	12.71	347	21.39	447	9.66
No	5455	87.29	1275	78.61	4180	90.34
Suffering from a low back disorder or other chronic back defect in the past 12 months						

Yes	1635	26.16	620	38.22	1015	21.94
No	4614	73.84	1002	61.78	3612	78.06
Suffering from a neck disorder or other chronic neck defect in the past 12 months						
Yes	1111	17.78	458	28.24	653	14.11
No	5138	82.22	1164	71.76	3974	85.89
Suffering from diabetes in the past 12 months						
Yes	312	4.99	92	5.67	220	4.75
No	5937	95.01	1530	94.33	4407	95.25
Suffering from an allergy, such as rhinitis, eye inflammation, dermatitis, food allergy or other (allergic asthma excluded) in the past 12 months						
Yes	1117	17.87	380	23.43	737	15.93
No	5132	82.13	1242	76.57	3890	84.07
Suffering from urinary incontinence, problems in controlling the bladder in the past 12 months						
Yes	113	1.81	58	3.58	55	1.19
No	6136	98.19	1564	96.42	4572	98.81
Suffering from kidney problems in the past 12 months						
Yes	160	2.56	73	4.50	87	1.88
No	6089	97.44	1549	95.50	4540	98.12
Suffering from depression in the past 12 months						
Yes	511	8.18	260	16.03	251	5.42

No	5738	91.82	1362	83.97	4376	94.58
Number of days in a typical week doing sports, fitness or recreational (leisure) physical activities that cause at least a small increase in breathing or heart rate for at least 10 minutes continuously						
I never carry out such physical activities 0	4069	65.11	1094	67.45	2975	64.30
1	376	6.02	85	5.24	291	6.29
2	643	10.29	155	9.56	488	10.55
3	442	7.07	104	6.41	338	7.30
4	191	3.06	50	3.08	141	3.05
5	223	3.57	44	2.71	179	3.87
6	72	1.15	21	1.29	51	1.10
7	233	3.73	69	4.25	164	3.54
Number of days in a typical week doing muscle-strengthening activities						
I never carry out such physical activities	5453	87.26	1449	89.33	4004	86.54
1+	990	15.84	173	10.67	623	13.46
Number of meals a day						
up to 2	729	11.67	195	12.02	534	11.54
3+	5520	88.33	1427	87.98	4093	88.46
Frequency of eating fruit, excluding juice						
1+ day	4548	72.78	1181	72.81	3367	72.77
4 to 6 per week	704	11.27	174	10.73	530	11.45
1 to 3 per week	703	11.25	186	11.47	517	11.17
less than once a week	294	4.70	81	4.99	213	4.60

Frequency of eating vegetables or salad, excluding juice and potatoes						
1+ day	3396	54.34	862	53.14	2534	54.77
4 to 6 per week	2799	44.79	380	23.43	1066	23.04
1 to 3 per week	2522	40.36	316	19.48	861	18.61
less than once a week	714	11.43	64	3.95	166	3.59
Tobacco status						
Everyday	1424	22.79	390	24.04	1034	22.35
Occasionally	240	3.84	60	3.70	180	3.89
Was a smoker once	1421	22.74	380	23.43	1041	22.50
Not a smoker	3164	50.63	792	48.83	2372	51.26
Use of e-cigarette						
Yes	54	0.86	9	0.55	45	0.97
No	6195	99.14	1613	99.45	4582	99.03
Frequency of exposure to tobacco smoke indoors						
Never or almost never	5707	91.33	1448	89.27	4259	92.05
Less than 1 hour per day	246	3.94	79	4.87	167	3.61
1 hour or more a day	296	4.74	95	5.86	201	4.34
Frequency of consumption of an alcoholic drink of any kind (beer, wine, cider, spirits, cocktails, premixes, liqueurs, homemade alcohol…) in the past 12 months						
Every day or almost	1454	23.27	358	22.07	1096	23.69
5 – 6 days a week	310	4.96	68	4.19	242	5.23
3 – 4 days a week	1067	17.07	266	16.40	801	17.31

1 – 2 days a week	533	8.53	145	8.94	388	8.39
2 – 3 days in a month	538	8.61	133	8.20	405	8.75
Once a month	773	12.37	208	12.82	565	12.21
Less than once a month	416	6.66	132	8.14	284	6.14
Not in the past 12 months, as I no longer drink alcohol/Never	1158	18.53	312	19.24	846	18.28
Until what extent do you agree or disagree that, in most ways your life is close to my ideal						
Strongly disagree	326	5.22	102	6.29	224	4.84
Disagree	506	8.10	165	10.17	341	7.37
Slightly disagree	347	5.55	119	7.34	228	4.93
Neither agree or disagree	547	8.75	159	9.80	388	8.39
Slightly agree	1941	31.06	481	29.65	1460	31.55
Agree	2220	35.53	522	32.18	1698	36.70
Strongly agree	362	5.79	74	4.56	288	6.22
Until what extent do you agree or disagree that you are satisfied with your life						
Strongly disagree	231	3.70	93	5.73	138	2.98
Disagree	864	13.83	274	16.89	590	12.75
Slightly disagree	610	9.76	167	10.30	443	9.57
Neither agree or disagree	609	9.75	156	9.62	453	9.79
Slightly agree	2255	36.09	574	35.39	1681	36.33
Agree	1466	23.46	326	20.10	1140	24.64
Strongly agree	214	3.42	32	1.97	182	3.93
Until what extent do you agree or disagree that so far you have gotten the important things you want in life						
Strongly disagree	104	1.66	47	2.90	57	1.23

Disagree	432	6.91	160	9.86	272	5.88
Slightly disagree	355	5.68	120	7.40	235	5.08
Neither agree or disagree	421	6.74	130	8.01	291	6.29
Slightly agree	1644	26.31	437	26.94	1207	26.09
Agree	2856	45.70	632	38.96	2224	48.07
Strongly agree	437	6.99	96	5.92	341	7.37
Until what extent do you agree or disagree that if you could live my life over, you would change almost nothing						
Strongly disagree	75	1.20	28	1.73	47	1.02
Disagree	405	6.48	162	9.99	243	5.25
Slightly disagree	359	5.74	97	5.98	262	5.66
Neither agree or disagree	362	5.79	103	6.35	259	5.60
Slightly agree	1612	25.80	410	25.28	1202	25.98
Agree	2872	45.96	683	42.11	2189	47.31
Strongly agree	564	9.03	139	8.57	425	9.19
Until what extent do you agree or disagree that the conditions of my life are excellent						
Strongly disagree	412	6.59	140	8.63	272	5.88
Disagree	1219	19.51	373	23.00	846	18.28
Slightly disagree	794	12.71	214	13.19	580	12.54
Neither agree or disagree	374	5.98	83	5.12	291	6.29
Slightly agree	1225	19.60	283	17.45	942	20.36
Agree	1729	27.67	410	25.28	1319	28.51
Strongly agree	496	7.94	119	7.34	377	8.15
Number of close people to count on in case of serious personal problems						

None	144	2.30	42	2.59	102	2.20
One or 2	2166	34.66	586	36.13	1580	34.15
3 to 5	2625	42.01	655	40.38	1970	42.58
6+	1314	21.03	339	20.90	975	21.07
Degree of concern shown by other people in what the person is doing						
A lot of concern and interest	3457	55.32	900	55.49	2557	55.26
Some concern and interest	2180	34.89	565	34.83	1615	34.90
Uncertain	463	7.41	114	7.03	349	7.54
Little or no concern and interest	149	2.38	43	2.65	106	2.29
How easy is it to get practical help from neighbors in case of need						
Very easy	696	11.14	198	12.21	498	10.76
Easy	2284	36.55	562	34.65	1722	37.22
Possible	2016	32.26	492	30.33	1524	32.94
Difficult	844	13.51	235	14.49	609	13.16
Very difficult	409	6.55	135	8.32	274	5.92
Providing care or assistance to one or more persons suffering from some age problem, chronic health condition or infirmity, at least once a week						
Yes	828	13.25	238	14.67	590	12.75
No	5421	86.75	1384	85.33	4037	87.25
Body Mass Index categories (BMI)						
Under	81	1.30	18	1.11	63	1.36
normal	2820	45.13	706	43.53	2114	45.69

overweight or obese	3348	53.58	898	55.36	2450	52.95

CONCLUSÃO

A ciência de dados é uma área de estudo com crescimento recente e caracterizada pelo volume e tipologia de dados, acesso e disponibilidade – ainda que se verifiquem preocupações e progressos quanto à gestão da confidencialidade dos mesmos e segurança dos cidadãos. Verifica-se atualmente que o volume de pesquisa e aplicabilidade acontece desproporcionalmente mais no setor privado em relação ao setor público, estimulando assim, importantes avanços nesse setor. O setor público, e, em particular, o setor de formulação de políticas parece continuar à margem desse desenvolvimento, apesar da integração desses conceitos poder ser essencial para políticas que se alicercem em uma evidência mais robusta. Face ao exposto, este trabalho procurou, de forma prática, explorar se a ciência de dados pode contribuir para otimizar a formulação de políticas públicas com exemplos concretos. Para isso foram utilizados bancos de dados coletados regularmente, e métodos causais e de machine learning foram aplixados para exemplificar como podem contribuir para apoiar decisões de natureza técnica e política. O intuito foi o de explorar, com exemplos reais, os benefícios e limitações dessa possibilidade, numa perspectiva de aprofundamento e aprimoramento das técnicas, contribuindo para a otimização de decisões e a melhor compreensão dos dados.

No artigo 1, foi realizada uma avaliação de um programa de saúde (Mais Médicos) que mostrou ter um impacto importante em alguns municípios brasileiros. Apesar de o objetivo inicial ter sido utilizar um algoritmo de predição como random forest para predizer uma probabilidade e esse valor ser usado no âmbito do propensity score matching para parear os municípios e, assim, avaliar o impacto, porém verificou-se que a estratégia não era adequada. Mesmo assim, esse artigo evidenciou a importância da continuidade do Mais Médicos para municípios com poucos médicos.

O artigo 2, por sua vez, pode ter uma contribuição permitindo a formulação de ações interventivas especificas junto aos sub-grupos identificados, de forma a melhorar a sua condição de saúde. Um ponto positivo é que foi possível conhecer o perfil demográfico, social, económico, de saúde e de comportamentos de acesso aos cuidados de saúde de grupos dentro dessa população mais vulnerável. Essas ações e programas podem ser mais eficientes no cumprimento dos objetivos e mais econômicas.

Por fim, o artigo 3 pode trazer uma contribuição para os serviços de saúde, nomeadamente para a saúde ocupacional, intervindo com antecedência em indivíduos da população profissionalmente ativa. Por exemplo, no caso em que uma empresa queira realizar medidas preventivas com trabalhadores com maior risco de se ausentarem por motivos de doença, tendo por base características demográficas, económicas entre outras. Com um modelo preditivo esses trabalhadores pordem ser identificados e acompanhados, com ações promovidas pela própria empresa.

De um modo geral, observou-se que a ciência de dados pode ter um papel na melhoria da evidência em políticas públicas, principalmente para ultrapassar dificuldades que por vezes abordagens mais tradicionais não solucionam, desde desafios mais macro (como o estabelecimento de contrafactuais em estudos quase experimentais) como mais micro (ao nível da gestão dos serviços com previsões acuradas para a alocação prioritária de recursos).

Relativamente aos desafios, um dos principais, e comum, a todos esses artigos foi articular bancos de dados com tantas variáveis. Em particular no artigo 2, essa dificuldade foi sentida porque um dos beneficios seria usar a maior parte das variáveis para a descrição dos grupos, a fim de se tornar uma descrição bem rica. Por outro lado, por ser um algoritmo não supervisionado, não foi possivel aplicar um método de seleção de variáveis. Nesse artigo, esse número também dificultou o reporte e visualização dos grupos. No artigo 1 essa preocupação também se notou, sobretudo na regressão logística, pois haveria muitas variáveis ruído nesse score. Porém, em PSM o importante não é o ajustamento do modelo, mas antes garantir que entravam todas as potenciais variáveis de confundimento. No artigo 3, o maior desafio foi garantir que haveria uma seleção

de variáveis que fizessem sentido na análise, sendo que uma das limitações foi ser um banco de dados transversal pelo que inviabilizou identificar exatamente quais variáveis pudessem realmente ser preditoras da variável de desfecho. Por causa dssa limitação decidiu-se usar apenas variáveis que não envolviam uma temporalidade, ou que certamente teriam ocorrido antes do desfecho como a idade, a educação, entre outras. Não só para esse artigo mas também para os restantes, foram adotados os ajustes necessários a contornar essas limitações, como descrito em cada um dos artigos.

REFERÊNCIAS

ABDI, H.; VALENTIN, D. **Multiple Correspondence Analysis.** Disponível em: </www.utd.edu/~herve>. Acesso em: 29 jun. 2020.

AIRAKSINEN, J. et al. Prediction of long-term absence due to sickness in employees: Development and validation of a multifactorial risk score in two cohort studies. **Scandinavian Journal of Work, Environment and Health**, v. 44, n. 3, p. 274–282, 2018.

ALBOUKADEL KASSAMBARA. **Cluster validation essentials**.

AMARAL, F. **Introdução à Ciencia de Dados mineração de dados e big data**. Rio Janeiro: Alta Books, 2015.

Análise Multivariada de Dados. [s.l: s.n.]

ARIMOND, G.; ELFESSI, A. A clustering method for categorical data in tourism market segmentation research. **Journal of Travel Research**, v. 39, n. 4, p. 391–397, 2001.

ATHEY, S. Machine Learning and Causal Inference for Policy Evaluation. p. 5–6, 2015.

AUSTIN, P. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. **Statist. Med.**, n. December 2006, p. 3385–3397, 2009.

AUSTIN, P. C. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. **Pharmaceutical Statistics**, v. 10, n. 2, p. 150–161, 2011.

BAKER, P. et al. The men's health gap: Men must be included in the global health equity agenda. **Bulletin of the World Health Organization**, v. 92, n. 8, p. 618–620, 2014.

BERRY, M. J.; LINOFF, G. **Data Mining Techniques: For Marketing, Sales,**

and Customer Support. [s.l: s.n.]

BHATLA, N.; JYOTI, K. An analysis of heart disease prediction using different data mining techniques. **International Journal of Engineering Research & Tecnology**, v. 1, n. 8, p. 1–4, 2012.

BOOT, C. R. L. et al. Prediction of long-term and frequent sickness absence using company data. **Occupational Medicine**, v. 67, n. 3, p. 176–181, 2017.

BREIMAN, L. Statistical Modeling: The Two Cultures. **Statistical Science**, v. 16, n. 3, p. 199–231, 2001a.

BREIMAN, L. Random Forests. **Machine Learning**, n. 45, p. 5–32, 2001b.

BREIMAN, L. et al. Classification and regression trees. **Classification and Regression Trees**, p. 1–358, 2017.

C.M., L.; V.H., V. B.; H.B., F. Application of unsupervised analysis techniques to lung cancer patient data. **PLoS ONE**, v. 12, n. 9, p. 1–18, 2017. Disponível em: <http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L618273302%0Ahttp://dx.doi.org/10.1371/journal.pone.0184370>.

CHENG LI. A Gentle Introduction to Gradient Boosting. **Seamless R and C++ Integration with Rcpp**, p. 3–18, 2013. Disponível em: <http://link.springer.com/10.1007/978-1-4614-6868-4_1>.

CHOR, J. S. Y. et al. Are men seeking medical advice too late? Contacts to general practitioners and hospital admissions in Denmark 2005. **Journal of Public Health**, v. 30, n. 1, p. 110–111, 2008.

CONASS. **Programa Mais Médicos-Nota Tecnica 23/2013.**Brasilia, 2013. . Disponível em: <http://maismedicos.gov.br/>.

CONCOLATO, C. E.; CHEN, L. M. Data Science: A New Paradigm in the Age of Big-Data Science and Analytics. **New Mathematics and Natural Computation**, v. 13, n. 2, p. 119–143, 2017.

COVENEY, P. V; DOUGHERTY, E. R.; HIGHFIELD, R. R. Big data need big theory too Subject Areas : Author for correspondence : **Philosophical Transactions of the Royal Society A**, 2016.

CRC, H.; MURRELL, P. **Exploratory Multivariate Analysis by Example Using R Introduction to Data Technologies**. [s.l: s.n.]

DATASUS. **DATASUS**. Disponível em: <http://datasus.saude.gov.br/>. Acesso em: 7 jul. 2017a.

DATASUS. **Programa Mais Médicos. Governo Federal.** Disponível em: <http://maismedicos.gov.br/conheca-programa>. Acesso em: 23 ago. 2018b.

DHAR, V. Data Science and Prediction Vasant Dhar Professor, Stern School of Business Director, Center for Digital Economy Research. **Communications of the ACM**, n. May, p. 64–73, 2012. Disponível em: <http://archive.nyu.edu/bitstream/2451/31553/2/Dhar-DataScience.pdf>.

DHL. Big Data in Logistics. **DHL Customer Solutions & Innovation**, n. December, p. 1–30, 2013.

DIAS, C. M. 25 anos de Inquérito Nacional de Saúde em Portugal. **Revista Portuguesa de Saúde Pública**, p. 51–60, 2009.

DONOHO, D. 50 Years of Data Science. **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745–766, 2017. Disponível em: <https://doi.org/10.1080/10618600.2017.1384734>.

EIRA, A. **A Saúde em Portugal: A procura de cuidados de saúde privados**. 2010. Faculdade Economia Porto, 2010. Disponível em: <https://repositorio-aberto.up.pt/bitstream/10216/26931/2/A saude em Portugal A procura privada de cuidados de saude Ana Eira.pdf>.

ELIZABETH A. STUART, BRIAN K. LEE, AND JUSTIN LESSLER, S. Improving propensity score weighting using machine learning. **Statist. Med.**, v. 29, n. 3, p. 337/346, 2010.

EUROPEAN COMMISSION. COMMISSION REGULATION (EU) No 141/2013 of 19 February 2013 implementing Regulation (EC) No 1338/2008 of the European Parliament and of the Council on Community statistics on public health and health and safety at work, as regards statistics based on the E. v. 2013, n. 141, 2013. Disponível em: <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32013R0141>.

EUROPEAN PARLIAMENT AND COUNCIL OF THE EU. Regulation (EC) N.223/2009. . 2009, p. 164–173.

EUROSTAT. **EUROPEAN HEALTH INTERVIEW SURVEY (EHIS).** Disponível em: <https://ec.europa.eu/eurostat/web/microdata/european-health-interview-survey>. Acesso em: 10 out. 2019.

FERREIRA, K. J. Analytics for an Online Retailer: Demand Forecasting and Price Optimization. v. 39, n. 5, p. 293–296, 2008.

FERRI, C.; HERNÁNDEZ-ORALLO, J.; MODROIU, R. **An experimental comparison of performance measures for classificationPattern Recognition Letters**, 2009. .

FISKE, A.; GATZ, M.; PEDERSEN, N. L. Depressive Symptoms and Aging: The Effects of Illness and Non-Health-Related Events. **Journals of Gerontology - Series B Psychological Sciences and Social Sciences**, v. 58, n. 6, p. 320–328, 2003.

FONTANELLA, S. et al. Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: A cross-sectional analysis within a population-based birth cohort. **PLoS Medicine**, v. 15, n. 11, p. 1–22, 2018.

FRANCO, G. Di. Multiple correspondence analysis : one only or several techniques ? **Quality & Quantity**, p. 1299–1315, 2016. Disponível em: <http://dx.doi.org/10.1007/s11135-015-0206-0>.

FRANÇOIS HUSSON; SEBASTIEN LÊ;JÉROME PAGÈS. **Exploratory Multivariate Analysis using R**. [s.l.] CRC Press, 2005.

GRILLON, C. et al. Scale-up of HIV self-testing. **The Lancet HIV**, v. 5, n. 7, p. e324–e326, 2018. Disponível em: <http://dx.doi.org/10.1016/S2352-3018(18)30095-X>.

HAIR, JR, RE ANDERSON, B. W. **Analise Multivariada de Dados**. 5ª ed. Porto Alegre: Bookman, 2005.

HASTIE, T. et al. **An Introduction to Statistical Learning, Springer Texts**. CA, USA: Springer, 2013. v. 102

HINDMAN, M. Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. **Annals of the American Academy of Political and Social Science**, v. 659, n. 1, p. 48–62, 2015. Disponível em: <https://doi.org/10.1177/0002716215570279>.

HJELLBREKKE, J. **Multiple Correspondence Analysis for Social Sciences**. New York: Routledge. Taylor and Francis Group., 2007.

HUSSON, F. JOSSE, J, PAGÉS, J. **FactoMineR. Hierarchical Clustering on Principal Components.** Disponível em: <http://factominer.free.fr>.

HUSSON, A. F. et al. Package ‘ FactoMineR ’. 2018.

HUSSON, F. **Handling missing values in MCA**, 2016. .

ILMARINEN, J. The Work Ability Index (WAI). **Occupational Medicine**, v. 57, n. 2, p. 160–160, 2006.

INE, INSTITUTO NACIONAL DE SAÚDE DOUTOR RICARDO JORGE, I. **Inquérito Nacional Saúde**Lisboa, 2016. .

INE. **DOCUMENTO METODOLÓGICO.National Health Survey**. [s.l: s.n.].

IZBICKI, R. Machine Learning sob a ótica estatística. 2016.

JOSSE, J. Technical Report – Agrocampus. n. September, 2010.

JOSSE, J.; HUSSON, F.; LÊ, S. **Multiple Correspondence Analysis**, 2008. .

JOSSELIN, J.-M.; LE MAUX, B. Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education. In: SPRINGER (Ed.). **Statistical Tools for Program Evaluation-Methods and Applications to Economic Policy, Public Health, and Education.** 1st 2017 ed. [s.l.] Springer International Publishing, 2017. p. 489–531.

KASSAMBARA, A. **Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis Book 1)**. [s.l: s.n.]

KHAN, A.; UDDIN, R.; ISLAM, S. M. S. **Clustering patterns of behavioural risk factors for cardiovascular diseases in Bangladeshi adolescents: A population-based studyHealth Policy and Technology**, 2019. .

KIVIMÄKI, M. et al. Sickness absence as a risk marker of future disability pension: The 10-town study. **Journal of Epidemiology and Community Health**, v. 58, n. 8, p. 710–711, 2004.

KOCAKULAH, M. C. et al. Absenteeism Problems And Costs: Causes, Effects And Cures. **International Business & Economics Research Journal (IBER)**, v. 15, n. 3, p. 89–96, 2016.

KOHAVI, R. Cross validation number.pdf. v. 5, p. 7, 1995. Disponível em: <http://robotics.stanford.edu/~ronnyk>.

KOLANOVIC, M.; KRISHNAMACHARI, R. T. Big Data and AI Strategies. n. May, p. 280, 2017.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Supervised Machine Learning: A Review of Classification Techniques. **Artificial Intelligence Review**, v. 26, p. 159–190, 2006.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling [Hardcover]**. [s.l: s.n.]

KUMAR, M.; KUMAR, M. International Journal of Computer Science and Mobile Computing Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. **International Journal of Computer Science and Mobile Computing**, v. 5, n. 2, p. 24–33, 2016. Disponível em: <www.ijcsmc.com>.

LANDIS, G. K. The measurement of observer agreement for categorical data. **Biometrics**, 1977.

LEG/UFPR. **Métodos baseados em árvores.** Disponível em: <http://cursos.leg.ufpr.br/>. Acesso em: 17 jun. 2020.

LEITE, W. **Practical propensity score methods using R**. USA: SAGE Publications, 2017.

LEVORATO, C. D. et al. Fatores associados à procura por serviços de saúde numa perspectiva relacional de gênero. **Ciencia e Saude Coletiva**, v. 19, n. 4, p. 1263–1274, 2014.

LINDEN, A.; YARNOLD, P. R. Combining machine learning and matching

techniques to improve causal inference in program evaluation. **Journal of Evaluation in Clinical Practice**, v. 22, n. 6, p. 864–870, 2016.

LUND, T. et al. Using administrative sickness absence data as a marker of future disability pension: The prospective DREAM study of Danish private sector employees. **Occupational and Environmental Medicine**, v. 65, n. 1, p. 28–31, 2008.

MCLAREN, L.; PETIT, R. Universal and targeted policy to achieve health equity: a critical analysis of the example of community water fluoridation cessation in Calgary, Canada in 2011. **Critical Public Health**, v. 28, n. 2, p. 153–164, 2018. Disponível em: <http://doi.org/10.1080/09581596.2017.1361015>.

MICHIE, D. Methodologies from machine learning in data analysis and software. **Computer Journal**, v. 34, n. 6, p. 559–565, 1991.

MICHIE, S.; WILLIAMS, S. Reducing work related psychological ill health and sickness absence: A systematic literature review. **Occupational and Environmental Medicine**, v. 60, n. 1, p. 3–9, 2003.

MIGUEL A. CARREIRA-PERPINAN. **Dimensionality Reduction**. [s.l: s.n.]

MIKE LOUKIDES. **What is Data Science?** 1st editio ed. [s.l.] O'Reilly Media, 2016.

MINISTERIO SAUDE. **TC 80: Cooperação Técnica entre o Ministério da Saúde e a Organização Pan-Americana da Saúde - Projeto Acesso da população Brasileira à Atenção Básica em Saúde.**Brasilia, 2013. .

MOLNAR, C. **Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.** Disponível em: <https://christophm.github.io/>. Acesso em: 27 jun. 2020.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [s.l: s.n.]

NASCIMENTO, A. do. **Avaliação de farmácias hospitalares Brasileiras utilizando Análise de Correspondência Múltipla.** 2011. UFRJ - Universidade Federal do Rio de Janeiro., 2011.

NENADIC, O.; GREENACRE, M. Computation of Multiple Correspondence Analysis, with Code in R. **Ssrn**, p. 25–27, 2005.

NGUYEN, L. H.; HOLMES, S. Ten quick tips for effective dimensionality reduction. **PLoS Computational Biology**, v. 15, n. 6, p. 1–19, 2019.

PAHO/WHO. **Ministério da Saúde reforça cooperação com a OPAS para continuidade do Mais Médicos**. Disponível em: <http://portalms.saude.gov.br/noticias/agencia-saude/42756-ministerio-da-saude-reforca-cooperacao-com-a-opas-para-continuidade-do-mais-medicos>.

PEDROSO, R. T.; JUHÁSOVÁ, M. B.; HAMANN, E. M. **A ciência baseada em evidências nas políticas públicas para reinvenção da prevenção ao uso de álcool e outras drogasInterface - Comunicação, Saúde, Educação**, 2019. .

POHAR, M.; BLAS, M.; TURK, S. Comparison of Logistic Regression and Linear Discriminant Analysis : A Simulation Study. **Metodološki zvezki**, v. 1, n. 1, p. 143–161, 2004.

QUINTAL, C.; LOURENÇO, Ó.; FERREIRA, P. **Utilização de cuidados de saúde pela população idosa portuguesa: Uma análise por género e classes latentesRevista Portuguesa de Saude Publica**, 2012. .

RAMOS, M. C.; SILVA, E. N. da. Como usar a abordagem da Política Informada por Evidência na saúde pública? **Saúde em Debate**, v. 42, n. 116, p. 296–306, 2018. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-11042018000100296&lng=pt&tlng=pt>.

RUBIN, D. B. Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. **Journal of Educational Psychology**, v. 66, n. 5, p. 688–701, 1974.

RUFFI, A. et al. Difi culdades de acesso a serviços de saúde para diagnóstico de tuberculose em municípios do Brasil Diffi culties in the accessibility to health services for tuberculosis. v. 43, n. 3, p. 389–397, 2009.

RUSSO, L. X. et al. Primary care physicians and infant mortality: Evidence from Brazil. **PLoS ONE**, v. 14, n. 5, p. 1–16, 2019.

RYGIELSKI, C.; WANG, J.; YEN, D. C. Data mining techniques for customer relationship management. v. 24, p. 483–502, 2002.

SAMUEL, A. L. Some studies in Machine Learning using the Game of Checkers. **IBM J. RES. DEVELOP**, v. 44, n. 1, 2000.

SANTOS, H. G. dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. p. 207, 2018.

SCHEFFER, M. et al. **Demografia médica no Brasil 2015**. São Paulo: Departamento de Medicina Preventiva, Faculdade de Medicina da USP. Conselho Regional de Medicina do Estado de São Paulo. Conselho Federal de Medicina., 2015.

SCHNEIDER, P. Why should the poor insure? Theories of decision-making in the context of health insurance. **Health Policy and Planning**, v. 19, n. 6, p. 349–355, 2004.

SHADISH, W. R. Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings. **Psychological Methods**, v. 15, n. 1, p. 3–17, 2010.

SHAW, M. J. et al. Knowledge management and data mining for marketing. p. 127–137, 2001.

SHI, L. et al. Primary care, infant mortality, and low birth weight in the states of the USA. **Journal of Epidemiology and Community Health**, v. 58, n. 5, p. 374–380, 2004.

SILVER, N. **Statistics Views**. Disponível em: <https://www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html>. Acesso em: 7 ago. 2019.

THE HENRY J KAISER FAMILY FOUNDATION. **Key Facts about the Uninsured PopulationThe Henry J Kaiser Family Foundation**, 2016. . Disponível em: <https://kaiserfamilyfoundation.files.wordpress.com/2013/09/8488-key-facts-about-the-uninsured-population.pdf>.

TRAVASSOS, C. et al. Utilização dos serviços de saúde no Brasil: Gênero,

características familiares e condição social. **Revista Panamericana de Salud Publica/Pan American Journal of Public Health**, v. 11, n. 5–6, p. 365–373, 2002.

UNIVERSITY OF CHICAGO. **Data Science for Social Good.**

VAN DEN BERG, T. I. J.; ELDERS, L. A. M.; BURDORF, A. The effects of work-related and individual factors on work ability: A systematic review. **Promotion of Work Ability Towards Productive Aging - Selected Papers of the 3rd International Symposium on Work Ability**, p. 15–18, 2009.

VIACAVA, F.; SOUZA-JÚNIOR, P. R. B. de; SZWARCWALD, C. L. Cobertura da população brasileira com 18 anos ou mais por plano de saúde privado: uma análise dos dados da Pesquisa Mundial de Saúde. **Cad Saude Publica**, v. 21, n. supl.1, p. S119–S128, 2005.

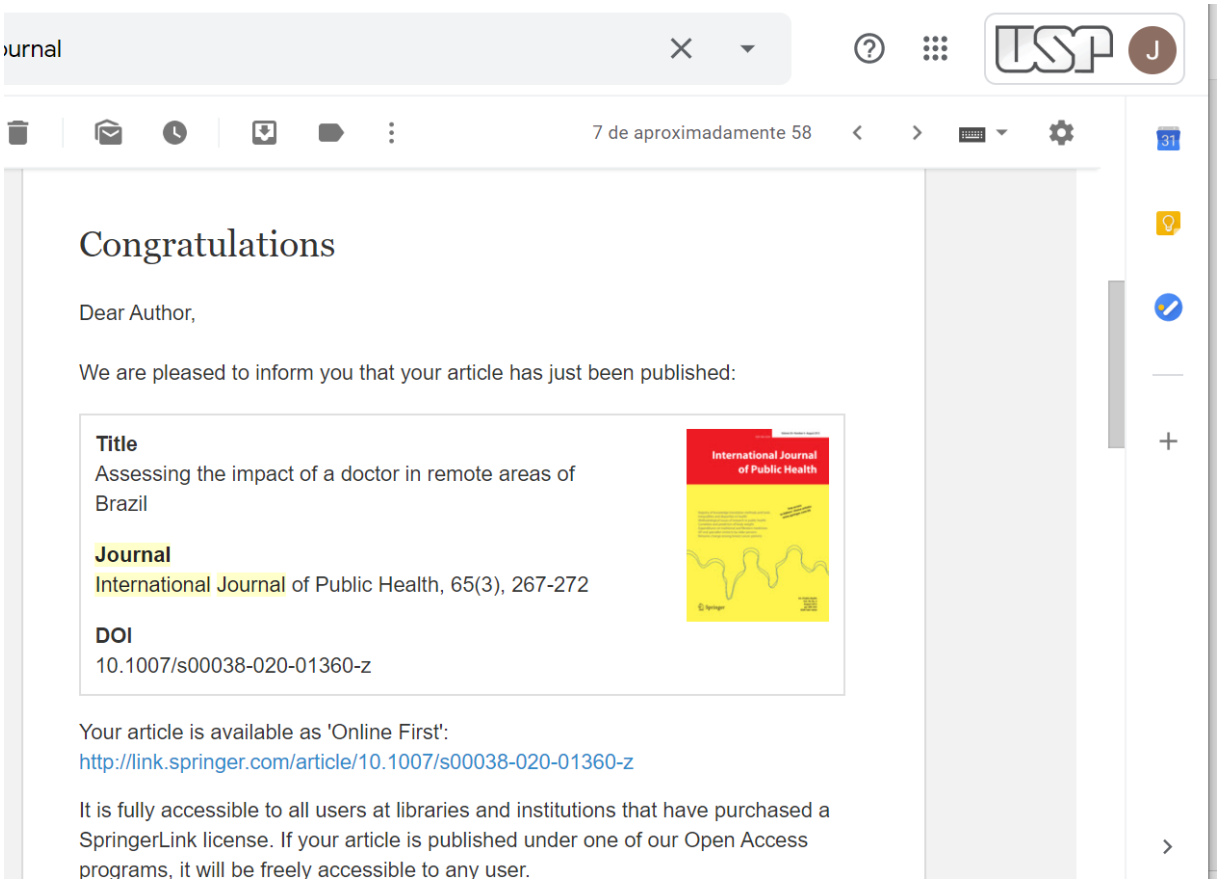
WENG, S. F. et al. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? **PLoS ONE**, v. 12, n. 4, p. 1–14, 2017.

WILPER, A. P. et al. Health insurance and mortality in US adults. **American Journal of Public Health**, v. 99, n. 12, p. 2289–2295, 2009.

ANEXOS

Anexo A

Comprovante de submissão e publicação do artigo *Assessing the impact of a doctor in remote areas of Brazil.*



urnal

7 de aproximadamente 58

Congratulations

Dear Author,

We are pleased to inform you that your article has just been published:

Title
Assessing the impact of a doctor in remote areas of Brazil

Journal
International Journal of Public Health, 65(3), 267-272

DOI
10.1007/s00038-020-01360-z

Your article is available as 'Online First':
http://link.springer.com/article/10.1007/s00038-020-01360-z

It is fully accessible to all users at libraries and institutions that have purchased a SpringerLink license. If your article is published under one of our Open Access programs, it will be freely accessible to any user.

Anexo B

Comprovante de submissão do artigo *Machine learning and national health data to improve evidence: finding segmentation in individuals without private insurance.*

CURRICULUM LATTES

CURRICULUM LATTES

CNPq

Currículo Lattes

Dados gerais | Formação | Atuação | Projetos | Produções | Eventos | Orientações | Bancas | +

Alexandre Dias Porto Chiavegatto Filho

Bolsista de Produtividade em Pesquisa do CNPq - Nível 2

Endereço para acessar este CV: http://lattes.cnpq.br/5517850224634709

ID Lattes: **5517850224634709**

Última atualização do currículo em 01/05/2020

Possui graduação em Economia pela FEA/USP, doutorado direto em Saúde Pública pela FSP/USP e pós-doutorado na Universidade de Harvard. É Professor Livre Docente do Departamento de Epidemiologia da FSP/USP, orientador dos programas de pós-graduação de Saúde Pública e de Epidemiologia da USP e coordenador de cursos sobre o uso do R para a análise de dados e sobre machine learning em saúde. Atuou como professor convidado (2016) e pesquisador visitante (2017 e 2019) na Universidade de Harvard. Atualmente é o Pesquisador Principal de pesquisas financiadas pela FAPESP, CNPq, PRP/USP e Fundação Lemann. Em 2015-2016 foi responsável pelo curso online Big Data em Saúde no Brasil da parceria USP-Coursera, que teve mais de 8.500 alunos matriculados e representantes de todos os Estados brasileiros. É o diretor do Laboratório de Big Data e Análise Preditiva em Saúde (Labdaps) da FSP/USP. Tem experiência em pesquisas na área de saúde pública, com ênfase em estatísticas de saúde e machine learning. **(Texto informado pelo autor)**

Identificação

Nome Alexandre Dias Porto Chiavegatto Filho

Nome em citações bibliográficas Chiavegatto Filho, A.D.P.;CHIAVEGATTO FILHO, A. D. P.;Chiavegatto Filho, Alexandre Dias Porto;Filho, Alexandre Dias Porto Chiavegatto;PORTO CHIAVEGATTO FILHO, ALEXANDRE DIAS;Chiavegatto Filho, Alexandre DP;Chiavegatto, Alexandre Dias Porto;Chiavegatto Filho, Alexandre;Filho, Alexandre Chiavegatto;CHIAVEGATTO FILHO, ALEXANDRE D P;CHIAVEGATTO FILHO, A D P;CHIAVEGATTO FILHO, ALEXANDRE D. P.;DIAS PORTO CHIAVEGATTO FILHO, ALEXANDRE