Aus dem Institut für Public Health
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Herausforderungen der Wahrscheinlichkeitsschätzung in Health Data Science: methodische Überlegungen und Anwendungen

The probability estimation problem in the Health Data Science framework: methodological considerations and applications

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Marco Piccininni

aus Bari, Italien

Datum der Promotion: 17.09.2021

Table of Contents

1 List of abbreviations

APOE ε4	Apolipoprotein E allele ε4
BIS	Berlin Initiative Study
COVID-19	Coronavirus disease 2019
C-index	Concordance index
CSF	Cerebrospinal fluid
CSFtau	Tau protein in cerebrospinal fluid
DAG	Directed acyclic graph
HDL	High-density lipoprotein
HbA1c	Hemoglobin A1c
ICI	Integrated calibration index
ISTAT	Italian National Institute of Statistics
Lasso	Least absolute shrinkage and selection operator
MB	Markov blanket
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SCORE	Systematic coronary risk evaluation
SCORE H	Systematic coronary risk evaluation for populations at high risk
SCORE L	Systematic coronary risk evaluation for populations at low risk
SCORE OP	Systematic coronary risk evaluation older persons
SCORE OP H	Systematic coronary risk evaluation older persons for populations at high risk
SCORE OP L	Systematic coronary risk evaluation older persons for populations at low risk
SCORE OP H5	Systematic coronary risk evaluation older persons for populations at high risk and 5-year time frame
SCORE OP L5	Systematic coronary risk evaluation older persons for populations at low risk and 5-year time frame
UK	United Kingdom

2 Abstract

Health Data Science is a health discipline engaged in three tasks: Description, Prediction, and Counterfactual Prediction. In this dissertation, I explored similarities and differences in the statistical methods for probability estimation within the framework of these tasks, relying on clinical and public health relevant applications.

Description focuses on studying the occurrence of health events in populations using surveillance systems. The probability of a new event, or incidence proportion, cannot be directly computed using information from surveillance systems. However, it is possible to estimate the incidence rate, which can be used to approximate the incidence proportion of a theoretical cohort. We estimated the all-cause mortality of the Italian city Nembro, which was severely affected by the COVID-19 pandemic. We used data from national and local registries to estimate the monthly all-cause mortality rates from 2012 to April 2020. We found that the all-cause mortality rate in Nembro increased dramatically in March 2020: it was 154.4 per 1,000 person-years, which corresponded approximately to a 1.3% probability of dying within one month.

The use of prediction models to estimate the probability of an event is widespread in medicine. The main challenge in Prediction is making sure that a model performs well for individuals outside of the development setting. Therefore, it is crucial to assess the transportability of a model. We conducted an external validation of the SCORE OP, a risk score recently developed to predict the risk of fatal cardiovascular events in European older persons. We assessed discrimination and calibration (using projections for the 10-year versions) of the SCORE OP using data from the Berlin Initiative Study. We found that the SCORE OP overestimated the true risk for older persons within Berlin.

Counterfactual Prediction aims at answering “what if” questions, estimating the probability of an outcome in different worlds in which different interventions are applied. This task is rooted in counterfactual thinking, and relies on prior causal knowledge summarized in causal graphs. In clinical examples and simulations, we examined the role of these elements (focusing on the principle of independent mechanisms and the Markov Blanket) in informing modeling strategies for probability estimation in the factual world.

Each Health Data Science task deals with the probability estimation problem differently, according to its challenges and objectives. Recently, the exchange of tools, statistical techniques, and theoretical concepts between Prediction and Counterfactual Prediction

has made important scientific advancements possible and opened several research tracks for future exploration.

3 Zusammenfassung

Health Data Science ist eine Gesundheitsdisziplin, die sich mit drei Aufgaben befasst: Beschreibung, Vorhersage und kontrafaktische Vorhersage. In meiner Dissertation untersuchte ich Ähnlichkeiten und Unterschiede in den Methoden dieser drei Aufgaben, wobei ich mich auf klinische und gesundheitsrelevante Anwendungen stützte.

Die Beschreibung konzentriert sich auf die Untersuchung des Auftretens von Gesundheitsereignissen in Populationen unter Verwendung von Surveillance. Die Wahrscheinlichkeit eines neuen Ereignisses kann jedoch nicht direkt anhand von Surveillance Daten berechnet werden. Es ist es jedoch möglich, die Inzidenzrate, die zur Annäherung an den Inzidenzanteil einer theoretischen geschlossenen Kohorte verwendet werden kann, abzuschätzen.

Als Anwendung schätzten wir die Gesamtmortalität von Nembro, einer italienischen Kleinstadt, die von der COVID-19-Pandemie schwer getroffen wurde. Wir verwendeten Daten aus nationalen und lokalen Registern, um die monatlichen Gesamtmortalitätsraten der Einwohner von Nembro von 2012 bis April 2020 zu schätzen. Wir stellten fest, dass die Gesamtmortalitätsrate in Nembro im März 2020 dramatisch anstieg. Sie betrug 154,4 pro 1.000 Personenjahre, was einer Sterbewahrscheinlichkeit von 1,3% entsprach.

Die Verwendung von Risikovorhersagemodellen zur Schätzung der Wahrscheinlichkeit eines Ereignisses ist in der Medizin sehr verbreitet. Die größte Herausforderung bei der Vorhersage besteht darin, sicherzustellen, dass das Modell auch bei Personen funktioniert, die nicht zur Entwicklung des Models herangezogen wurden. Daher ist es wichtig, externe Validierungen durchzuführen, um die Nutzung des Modells in anderen Settings sicherzusellen.

Wir haben eine externe Validierung des SCORE OP, eines Risikoscores zur Vorhersage des Risikos tödlicher kardiovaskulärer Ereignisse bei älteren Menschen in Europa, durchgeführt. Wir bewerteten die Diskrimination und Kalibrierung des SCORE OP unter Verwendung von Daten aus der Berliner Initiative Studie. Wir stellten fest, dass die SCORE OP-Gleichungen das tatsächliche Risiko bei älteren Menschen in Berlin erheblich überschätzt und nicht angewandt werden sollte.

Die kontrafaktische Vorhersage zielt darauf ab, "Was-wäre-wenn"-Fragen zu beantworten. Die Vorhersagen stützen sich auf kausales Vorwissen und die Verwendung von Kausalgraphen. Wir untersuchten in klinischen Beispielen und Simulationen die Rolle dieser Elemente, insbesondere des Prinzips der unabhängigen Wirkmechanismen und der Markov-Blanket-Konzepte.

Jede Aufgabe der Gesundheitsdatenwissenschaft lehnt das Wahrscheinlichkeitsschätzproblem entsprechend ihren eigenen Herausforderungen, Methoden und Zielen ab. In den letzten Jahren ermöglichte der Austausch von Werkzeugen, statistischen Techniken und theoretischen Konzepten zwischen Vorhersage und kontrafaktischer Vorhersage wichtige wissenschaftliche Fortschritte und eröffnete mehrere neue Wege für zukünftige biomedizinische Forschung.

4 Introduction

Health Data Science is an umbrella term for the application of modern data analysis techniques to address problems in health and medicine. Health Data Science emerged over time from the natural convergence of Epidemiology, Biostatistics, and the more recently formalized field of Data Science[1]. The later scientific field generates knowledge from data, using modern analytical techniques and critical thinking to solve “complex real-world health problems”[2].

Hernàn et al. believe that defining Data Science by its tasks, rather than by its activities and tools, is more rigorous and presents a unique opportunity to bridge the historical gaps of traditional statistics[3]. According to this perspective, Data Science should be defined as the discipline engaged in three tasks: Description, Prediction and Counterfactual Prediction[3].

Despite these three tasks being characterized by different languages, assumptions, aims, and set of procedures, they often rely on the same statistical concepts. While this dependence on the same concepts can be disorientating in health research[3,4], the consideration of the same statistical problems in Description, Prediction and Counterfactual Prediction makes it possible to study the differences between their approaches. A statistical problem that is crucial in all three Health Data Science tasks is the estimation of probability.

Descriptive epidemiology predominantly aims its attention at the examination of the occurrence of diseases or other health-related events in populations[5]. A common goal in descriptive epidemiology is to estimate the frequency of new health-related events in a specified time frame in a given population[5]. This quantity is called “incidence” and can be quantified according to two different metrics: the “incidence proportion” and the “incidence rate”[6]. Estimating the incidence of health-related events for specific geographic regions over time is crucial to better allocate healthcare resources and to detect temporal or spatial differences that may indicate a difference in the underlying distribution of risk factors[7,8].

When a surveillance system (e.g. a population-based registry) records all the new occurrences of the event of interest in a specific geographic region, the incidence rate can be estimated in a highly cost-effective way[7]. The incidence rate is the ratio between the number of new cases recorded by the registry during the study period and the overall number of person-time spent at risk from the individuals living in the geographic region during the study period[7]. The denominator's quantity can be approximated, under reasonable assumptions, using information on the number of individuals living in the region[7,9].

The incidence rate is less interpretable than the incidence proportion, which is a probability[6,7]. The incidence proportion, however, cannot be directly computed in geographically defined populations[7]. Probabilities can only be computed in “closed” populations[6], while geographically defined populations observed over a calendar time period are “dynamic” or “open” populations[7,9]. Fortunately, in a closed population, a known mathematical formula describes the relationship between the incidence rate and the incidence proportion of a non-recurrent inevitable event (e.g., all-cause death)[6]. Under a set of (typically reasonable) assumptions[6,7], the incidence rate estimated from an open population can be used to approximate the incidence proportion of a corresponding theoretical closed population of interest[6,7,9]. In this way it can be interpreted as a probability.

The estimation of probability is also crucial in Prediction, the second Health Data Science task. Risk prediction models for binary outcomes are widely used in the medical field to inform or support decision making[10,11]. According to Moons et al., “risk prediction models use predictors (covariates) to estimate the absolute probability or risk that a certain outcome is present (diagnostic prediction model) or will occur within a specific time period (prognostic prediction model) in an individual with a particular predictor profile”[11]. A clinical risk prediction model is usually developed by applying statistical techniques (e.g. regression models, machine learning algorithms) to a “development” dataset which contains information about the predictors and the outcome for a sample of individuals[10,11]. However, the actual aim of a risk prediction model is to predict the probability of the outcome for individuals outside the development dataset, for whom the outcome is still unknown, and outside the underlying population from which the development sample was drawn[12]. The trade-off between a model’s good fit on the development dataset and its ability to provide valid predictions outside the development dataset is crucial in risk prediction[10,11,13]. Good performances in the development setting do not ensure similarly accurate performances in other populations[10,12]. The practice of assessing a risk prediction model's performance by applying it to a sample of individuals drawn from an underlying population different from the development one is called "external validation"[10,12]. The external validation assesses the "transportability" (or "generalizability") of a risk prediction model to a different and specific setting[10,12]. This assessment is a necessary step before a model can be used to estimate probabilities in clinical practice[10,12].

Prediction relies on the association between the covariates and the outcome to predict the value of the outcome variable for newly observed individuals[4,12,14]. Counterfactual Prediction, on the other hand, aims at predicting the value of the outcome variable in a hypothetical "counterfactual" world characterized by different actions (e.g. treatment or intervention)[3,4]. Therefore, Counterfactual Prediction is rooted in causal reasoning rather than mere probabilistic reasoning[15]. Counterfactual Prediction answers "what if" questions and comprises the foundation of causal effects estimation and causal inference, which are essential for decision making[3]. Even this task, if the outcome of interest is binary, ultimately consists of estimating a probability[4]. Unlike Description or Prediction, however, Counterfactual Prediction heavily relies on subject matter knowledge[3]. Prior knowledge about the causal structure is crucial for Counterfactual Prediction[3] and is often summarized using Directed Acyclic Graphs (DAGs)[3,4,6,15–17]. DAGs are graphical tools designed to map all a priori knowledge in the form of assumptions about the causal mechanism underlying the data generation process. These tools are generally used in epidemiology to define analytic strategies to deal with selection, confounding, and measurement bias when estimating causal effects[4,6,16,17]. More generally, a DAG describes a causal structure and, under some assumptions, provides qualitative information about the conditional independence between variables[4,16,17]. Because of these characteristics, DAGs are valuable resources in probability estimation problems when causal mechanisms are known.

5 Objectives

The aim of my doctoral project is to explore, in clinical and public health relevant applications, the statistical methods for probability estimation in the framework of the three Health Data Science tasks.

Specifically, the research objectives are:

1. Study I (Publication 1): Estimate, in a descriptive study based on registry data, the incidence of all-cause death (all-cause mortality) over time for an Italian city hardly hit by the COVID-19 pandemic[18].
2. Study II (Publication 2): Assess the performance, in an external validation study, of the SCORE OP clinical risk prediction model in Berlin older persons population[19].
3. Study III (Publication 3): Investigate how the use of DAGs and causal thinking, typical of Counterfactual Prediction, can improve modeling strategies for probability estimation problems[20].

6 Methods

6.1 Description task: all-cause mortality rate estimation using registry data

6.1.1 Population

Nembro is a small city with 11,505 inhabitants[21], situated in the province of Bergamo, Lombardy, Italy[18].

In 2018 the province of Bergamo was characterized by a high life expectancy at birth (83.28 years), similar to the corresponding Lombardy regional value (83.44 years)[22].

Between 2009 and 2016, the two leading causes of death in the province were cardiovascular disease and cancer (specifically bronchial and lung cancer for men and breast cancer for women), representing 66.9% of the total deaths[23]. Within this time interval, Bergamo province had a higher mortality rate for all-type cancers compared to the region, but similar mortality rates for cardiovascular disease, bronchial and lung cancers, and breast cancer[23].

Towards the end of February 2020, the province of Bergamo became one of the major hot spots of the COVID-19 outbreak in Italy[24]. Hospitals in the area rapidly became overwhelmed by the surging number of infections. On March 8th the Italian government imposed a regional lockdown in the attempt to halt the spread of the virus[24]. Nembro was one of the first Italian cities to register SARS-CoV-2 infections during the outbreak and was severely affected by the COVID-19 pandemic[25].

6.1.2 Data

To compute an all-cause mortality rate (incidence rate of death events) two quantities are needed: the numerator and the denominator of the rate.

We sought to compute the Nembro all-cause mortality rates for each month from January 2012 to April 2020 to describe variation over time[18]. The numerator of each rate was the number of all-cause deaths that occurred among Nembro residents during a given month[18]. The denominator was the number of person-years the population of Nembro residents spent at risk of dying during the month[18].

The number of monthly all-cause deaths that occurred in Nembro from January 2012 to November 2019 was retrieved from the public dataset of the Italian National Institute of Statistics (ISTAT)[22]. The number of monthly all-cause deaths for the remaining months (December 2019 to April 2020) was obtained from the Nembro official registration office, thanks to the special authorization from the mayor of the city[18]. Despite not belonging to official ISTAT data sources, this death registry, managed by the local authority, promptly records the number of resident death, and is a valuable and high-quality data source during a pandemic emergency.

Since all-cause death is a non-recurrent event and prevalence is not of concern, the number of person-years spent at risk every month was estimated using official information about the population size. The number of residents alive at the beginning of each month, from January 2012 to December 2019, was obtained from ISTAT official data[22]. The number of residents alive at the beginning of January 2020 was obtained from the Nembro official registration office yearly report[21].

Due to data limitations, the mortality rate for April was estimated using only information based on the first eleven days[18].

6.1.3 Statistical analysis

The computation of the overall amount of person-time spent at risk can be conceptualized as a geometrical problem. Let's consider a plot in which the time from the beginning of the study is on the x-axis and the population of individuals at risk is on the y-axis. The overall amount of person-time spent at risk during the study is the area under the curve describing the size of the population at risk over time[9]. The computation of the overall amount of person-time spent at risk consists of finding the solution of a definite integral. While this conceptualization is often ignored in practice, it comprises the theoretical rationale behind the classical approaches for the estimation of the rate's denominator[9]. Classical formulas for the estimation of this quantity are based on the assumption of a linear change of the population at risk over time. Under this assumption, the area under the curve of the population size function is simply the area of a trapezoid[9].

To estimate the amount of person-years spent at risk by the population of Nembro, we used a flexible and easy to implement approximation. We approximated the area under the curve of the population size function as the sum of the area of smaller rectangles corresponding to each day of the study period. Each small rectangle has the base equal to one (one day) and the height equal to the number of residents alive on that day. This approach follows the same rationale of the Reimann sum for integral approximations.

Hence, we estimated the amount of person-years spent at risk in two steps.

First, the number of residents alive each day from January 1st, 2012 to January 1st, 2020, was estimated by interpolating data of the population size on the first day of the month, using a spline regression model[18]. Specifically, we used linear splines, assuming a linear change of the population size between the first days of two consecutive months[18].

Meanwhile the number of alive residents between January 1st and April 11th in 2020 was a projection based on two different models. The projections were obtained as the weighted average of the predictions from the last segment of the spline regression, and the predictions of a third-degree polynomial regression fitted on the whole time period[18]. The weights were chosen to avoid function discontinuity and to give more importance to long term trends[18].

The second step consists of summing up the number of Nembro residents alive every day within each month[18]. This operation corresponds to the abovementioned "sum of smaller rectangles", where the result of the sum approximates the overall number of person-days spent at risk in the month. Finally, person-days were converted into person-years through dividing by 365.25.

In a sensitivity analysis, we estimated the monthly all-cause mortality rates using the same approach while assuming a drop in the population size after February 20th, 2020, proportionate to a dramatic increase in deaths or emigration[18].

All analyses were conducted using R-3.6.0 and RStudio v1.1.456[18].

6.2 Prediction task: external validation of the SCORE OP risk prediction model

6.2.1 Risk prediction models

The Systematic COronary Risk Evaluation (SCORE)[26] is a prognostic risk prediction model developed in 2003 using European populations data. The SCORE predicts the probability of an individual dying from cardiovascular diseases within 10 years based on age, sex, total cholesterol, systolic blood pressure, and smoking status. It has two versions: the SCORE for populations at high cardiovascular risk and the SCORE for populations at low cardiovascular risk[26]. We will refer to them as SCORE H and SCORE L, respectively. The use of SCORE is currently recommended by the European Society of Cardiology guidelines for individuals between 40 and 65 years old[27].

On several occasions, risk prediction models developed using data from middle-aged individuals demonstrated poor performance when applied in older person populations[28]. This was attributed to a different association between classical risk factors and cardiovascular endpoints in older people compared to middle-aged adults[28]. For this reason, in 2015, a new version of the SCORE, called SCORE Older Persons (SCORE OP) was developed using data from older European individuals only[28].

The SCORE OP was designed to predict the probability of fatal cardiovascular events among Europeans aged 65 or older, using the same predictors included in the SCORE with the addition of diabetes status and HDL cholesterol[28]. Overall, there are four SCORE OP equations[28]: two which produce 5-year predictions in high and low cardiovascular risk populations (SCORE OP H5 and SCORE OP L5), and two producing 10-year predictions in high and low cardiovascular risk populations (SCORE OP H and SCORE OP L).

To compute SCORE and SCORE OP predictions, we relied on published coefficients, parameters, and formulas[26,28]. The published SCORE formula is based on a “mathematical inconsistent model” that allows the computed probabilities to be higher than one[29]. We applied a simple mathematical correction proposed by Støvring et al. as the first step of their competing risk approach to prevent this error from occurring[29].

6.2.2 Data

We externally validated the SCORE OP using data from the Berlin Initiative Study (BIS)[19]. The BIS is a population-based cohort study of 2,069 Berlin older persons (aged 70 or older) recruited from the database of a large health insurance company. Further details of the study design and recruitment strategy can be found in the original paper[30].
Information on the following SCORE and SCORE OP predictors were obtained during in-person interviews at the baseline: age, sex, systolic blood pressure (average of two measurements), HDL cholesterol level, total cholesterol level, self-reported smoking status (current status), diabetes mellitus presence (self-reported antidiabetic treatment use and/or HbA1c level higher than 6.5%)[19].

Information about the date of death and causes of death were obtained from multiple sources: general practitioners, direct contact with relatives, insurance records, Berlin death certificate archive, and, for in-hospital deaths, medical discharge letters[19]. When no information on the causes was available, in the primary analyses, deaths were assumed to be due to non-cardiovascular reasons[19].

The BIS recruitment process started in November 2009, while the end of the follow-up for this external validation study was set to September 30th, 2015[19]. Therefore, each individual was followed from recruitment until death, last available study visit (loss to follow-up), or the last day of September 2015 (administrative censorship)[19].

We only included participants with follow-up information and who self-reported no history of myocardial infarction at baseline to maintain consistency with SCORE and SCORE OP development study criteria[19]. We further excluded individuals without information on all SCORE or SCORE OP predictors[19]. The BIS was approved by the local ethics committee of the Charité – Universitätsmedizin Berlin (Ref. EA2/009/08).

6.2.3 Statistical analysis

We predicted the 5-year risk of fatal cardiovascular events for all individuals in the study according to SCORE OP H5 and SCORE OP L5 risk prediction models[19].

Performance of the 5-year version of the SCORE OP was then assessed, comparing the predicted risks with the actual outcomes[19]. Two crucial metrics were considered in the performance evaluation of each risk prediction model: calibration and discrimination.

To assess calibration, individuals were grouped based on deciles of the predicted risk. For each group, the average predicted probability and the estimated actual probability obtained from a Kaplan-Meier estimator were computed[19]. The overall number of predicted and actual events across all decile groups were contrasted to assess calibration-in-the-large[19]. Average predictions and actual probabilities for each decile group were also reported to ensure a more detailed calibration assessment. Discrimination was, instead, measured using the C-index on all the available follow-up time[19].

In a secondary analysis, we assessed the calibration of the 10-year versions of the SCORE and SCORE OP risk prediction models[19]. Since the observed follow-up was shorter than 10 years, we used as actual probabilities the 10-year projections obtained from a Weibull regression survival model (instead of the Kaplan-Meier estimates) for each risk score in the secondary analysis[19].

In a sensitivity analysis, we assessed risk prediction model performance under the assumption that deaths ascribed to unknown reasons were all due to cardiovascular diseases[19]. Additionally, to understand the magnitude of the results at the population level, all analyses were rerun on five resampled datasets representative of the size and age-sex structure of the 2010 Berlin older person population[19].

Finally, we compared the SCORE and SCORE OP 10-year risk predictions graphically[19]. The comparison was conducted on the age range between 60 and 100 years for three hypothetical risk profiles (high, medium, and low risk) for men and women separately[19].

All analyses were conducted using R-3.4.3 and RStudio v1.0.153[19].

6.3 Counterfactual Prediction task: DAGs and causal thinking in probability estimation

Counterfactual Prediction is rooted in counterfactual thinking and relies heavily on prior causal knowledge and the use of causal graphs. We examined the role of these elements, typical of Counterfactual Prediction, in informing modeling strategies for probability estimation problems in the “factual” world. Specifically, we focused on two concepts pertaining to the causal inference and causal discovery worlds: the principle of independent mechanisms and the Markov Blanket.

6.3.1 Principle of independent mechanisms

Using the words of Peters et al., the principle of independent mechanisms assumes that “the causal generative process of a system’s variables is composed of autonomous modules that do not inform or influence each other”[15]. According to this principle, a causal process can be seen as a chain of independent physical mechanisms, meaning that each physical mechanism is independent of the input it receives and, therefore, from the previous mechanisms on the chain[15].

The consequences of this principle are clear, even when considering the simple scenario with only two variables: one Cause and its Effect.

The joint probability distribution of the variables Cause and Effect P(Cause,Effect) can be written in two different ways[15,31]:

$$P(\text{Cause},\text{Effect}) = P(\text{Cause}|\text{Effect})P(\text{Effect}) = P(\text{Effect}|\text{Cause})P(\text{Cause})$$

P(Effect|Cause) maps probabilistically the values of the variable Cause to the values of the variable Effect, representing the causal mechanism. P(Cause), on the other hand, is the distribution of the variable Cause, representing the input of the causal mechanism.

According to the principle of the independent mechanisms, P(Cause) and P(Effect|Cause) are independent[15,31]. As a consequence, the conditional distribution of the Effect given the Cause P(Effect|Cause) and the marginal distribution of the Cause P(Cause) change independently across different joint distributions[15,20,31].

This independence in the second factorization induces a dependency between P(Cause|Effect) and P(Effect) in the first factorization[15,31]. Therefore, P(Cause|Effect) (which is only a mathematical relationship not representing a causal mechanism) and P(Effect) tend to change together across different joint distributions[15,20,31].

This implies that the probability of a disease estimated using its effect as predictor (anticausal direction) could be less transportable to other settings characterized by a different joint distribution[15,20,31]. This principle, which has a long history and was recently well described by Peters et al.[15], has important implications for the transportability of diagnostic clinical risk prediction models. We will demonstrate this using two hypothetical clinical examples in the field of neurodegenerative diseases.

6.3.2 Markov Blanket

Predictor selection is a step of paramount importance when estimating the probability of an outcome through a risk prediction model. Predictor selection can be described as the challenge of choosing the lowest possible number of variables that contain sufficient information to predict the outcome.

Thus an intuitive idea might be to exclude, from all the available variables, those that do not provide additional information about the outcome.

When we quantify the “additional information” in terms of “conditional dependence”[32], predictor selection is reduced to a Markov Blanket identification problem. In order to define the Markov Blanket, we need to introduce some notation: let Y be the random outcome variable that we are interested in predicting; let **X** be the set of all available variables (without Y), which we will assume includes all variables relevant in the causal processes involving Y. Then, the Markov Blanket of the outcome variable MB(Y) is the smallest subset of **X** that satisfies the following relationship[20,33,34]:

$$\forall\ K \in \mathbf{X} - MB(Y): \Pr(Y|MB(Y),K) = \Pr(Y|MB(Y))$$

All other variables of **X** are conditionally independent from Y (and therefore do not provide additional information about the outcome) once we condition on the Markov Blanket set of Y[20,33,34]. If we have a preference for a lower number of predictors and we measure the performance in terms of calibration, in an ideal regression setting, the variables in the Markov Blanket of the outcome are the only variables needed for optimal prediction[33].

The concept of Markov Blanket was first formalized in 1988 by Pearl[35] and was later used for predictor selection[32]. Tsamardinos proved that a variable is strongly relevant[36,37] to predict the outcome *if and only if* it is included in the Markov Blanket of the outcome[37].

Another positive characteristic of the Markov Blanket concept that highlights its applicability is the possibility of identifying the Markov Blanket using DAGs.

Let us imagine that we have a DAG representing all the causal processes involving Y and describing the causal relationships among Y and all the variables in **X**. If the joint distribution of the variables is Markovian and faithful with respect to the described DAG (two common assumptions in causal inference and causal discovery), then the Markov Blanket of Y can be easily identified as all the parents of the Y node, all the children of the Y node, and all the parents of Y node's children[15,20,33–35].

This implies that if all variables involved in the causal processes of the outcome are known, and we are able to accurately describe the causal relationships among all the variables in a DAG, then we can use this graphical rule to confidently identify the relevant predictors to be included in the clinical risk prediction model.

To show this, we ran a simulation of 100,000 different scenarios, each based on a random DAG, according to which a dataset with 10,000 observations, 24 candidate predictors, and one binary outcome was generated. In each scenario, the calibration of eight different prediction tools was compared using the 10 fold cross-validation Integrated Calibration Index (ICI)[38]. Specifically, the eight prediction tools were: a logistic regression model with Markov Blanket set variables as independent variables, a a logistic regression model with all 24 candidate predictors as independent variables, a logistic regression model with all variables with a path to the outcome node as independent variables, a logistic regression model with parents of the outcome node as independent variables, a logistic lasso regression model with all 24 candidate predictors as input, a logistic ridge regression model with all 24 candidate predictors as input, a logistic elastic net regression model with alpha equal to 0.5 and all 24 candidate predictors as input, and a random forest algorithm with all 24 candidate predictors as input. Details of the simulation and the simulation code are available in the original paper and supplement material[20]. All analyses were conducted using R-3.6.3[20].

7 Results

7.1 Description task: all-cause mortality rate estimation using registry data

In total, 1,116 Nembro residents died between January 1st, 2012 and April 11th, 2020[18]. Specifically, 112 Nembro residents died in 2012, 112 died in 2013, 95 died in 2014, 119 died in 2015, 126 died in 2016, 109 died in 2017, 128 died in 2018, 121 died in 2019, and 194 died in the first 102 days of 2020[18]. More Nembro residents died during March 2020 alone (151 deaths) than during the entire year 2019[18].

The monthly all-cause mortality rate from January 2012 to February 2020 ranged from 1.0 per 1,000 person-years to 21.5 per 1,000 person-years (see Figure 1)[18]. Strikingly, the all-cause mortality in March 2020 was 154.4 per 1,000 person-years, almost eleven times the all-cause mortality recorded in March 2019 (14.3 per 1,000 person-years) and by far the highest rate observed in the last 8 years (Figure 1)[18]. In April 2020 (11 days considered), the all-cause mortality rate decreased to 23.0 per 1,000 person-years (Figure 1)[18].

Since in the sensitivity analysis scenario a loss in person-years is assumed, all-cause mortality rates estimated in the sensitivity analysis are slightly higher: 15.3 per 1,000 person-years in February 2020, 155.7 per 1,000 person-years in March 2020, and 23.5 per 1,000 person-years in April 2020[18].

Information about deceased Nembro residents' demographics and official COVID-19 deaths, obtained from other data sources, are reported in the original paper[18].

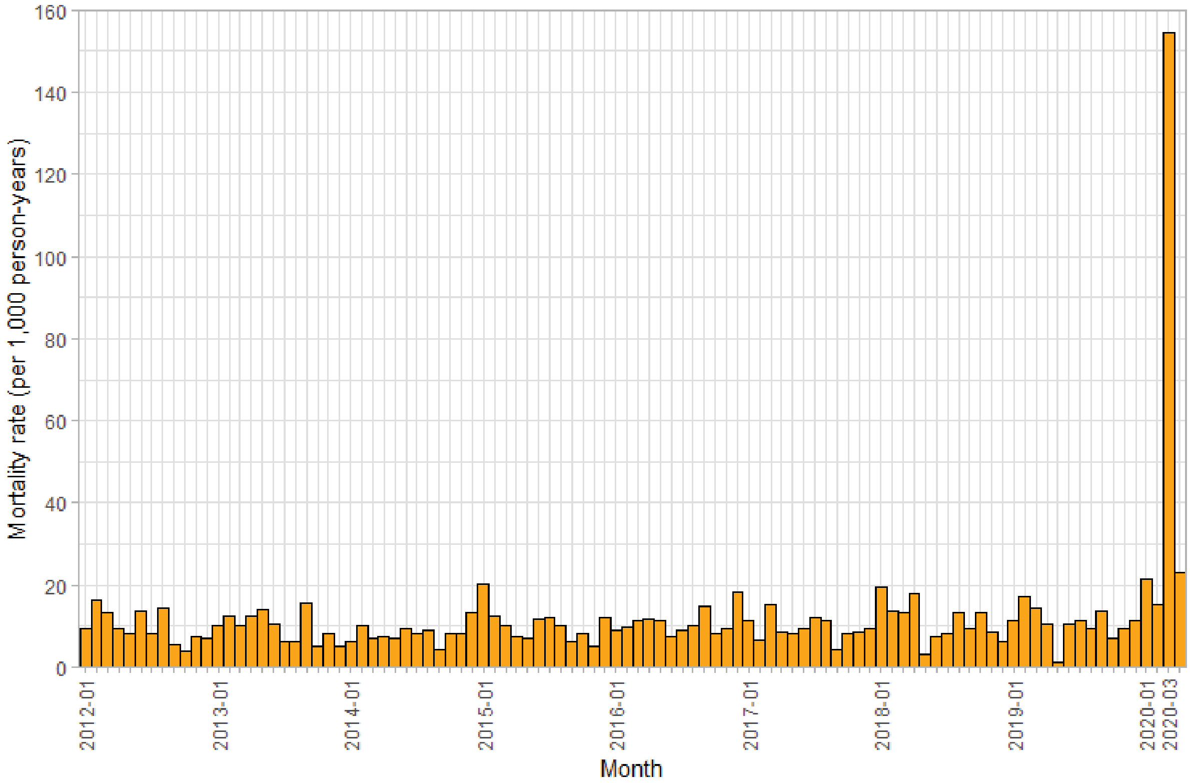


Figure 1. Nembro residents monthly all-cause mortality rates (per 1,000 person-years) from January 2012 to April 2020[18]. April 2020 mortality rate is estimated only based on the first eleven days information. The first month of each year and March 2020 are indicated on the x-axis.

.

7.2 Prediction task: external validation of the SCORE OP risk prediction model

Out of 2,069 BIS participants, 1,657 were included in this validation study[19]. Overall, 412 participants were excluded: 102 because no follow-up information was available, 23 because no information on myocardial infarction history was available, 273 because they self-reported at the baseline to have experienced a myocardial infarction, and 14 because information on some of the SCORE OP predictors was missing[19]. The median follow-up time was 4.8 years, and the overall amount of person-time spent at risk during the study period was 7,370.3 person-years[19]. During the study follow-up, 118 deaths due to cardiovascular diseases were observed[19].

The SCORE OP H5 and the SCORE OP L5 predicted 302 and 215 fatal cardiovascular events in 5 years, respectively[19]. The number of fatal cardiovascular events estimated with the Kaplan-Meier estimator for the same time interval was only 142[19]. Therefore, relying on the calibration-in-the-large assessment, both the SCORE OP H5 (predicted/actual ratio=2.13) and the SCORE OP L5 (predicted/actual ratio=1.51) overestimated the true risk[19]. The SCORE OP H5 showed a systematic overestimation across all decile groups (Table 1)[19]. Similarly, the SCORE OP L5 average predicted risk was higher than the probability of the event as estimated by the Kaplan-Meier estimator in 9 out of 10 decile groups (Table 1)[19].

These same results were confirmed in the secondary analysis on the 10-year time span. The SCORE OP H predicted 677 events, despite only 399 actual events being projected (predicted/actual ratio=1.70) and overestimated the true risk in all decile groups (Table 1)[19]. Similarly, the SCORE OP L overestimated the true risk in eight decile groups (Table 1)[19].

We estimated a C-index of 0.79 (0.75 to 0.83) and 0.80 (0.75 to 0.83) for the SCORE OP H and the SCORE OP L respectively[19]. The estimated C-index for the SCORE H and the SCORE L were 0.72 (0.67 to 0.76) and 0.72 (0.67 to 0.77) respectively[19]. Despite the fact that the SCORE OP equations showed slightly better discrimination ability compared to the SCORE equations on the observed follow-up, they showed less calibration compared to the SCORE equations on the 10-year interval[19]. The SCORE L demonstrated an underestimation of the true risk in the calibration-in-the-large assessment (predicted/actual ratio=0.67). In contrast, the SCORE H showed relatively good calibration both in terms of calibration-in-the-large (predicted/actual ratio=0.97) and decile group comparison (Table 1)[19].

	SCORE OP H5		SCORE OP L5		SCORE OP H		SCORE OP L		SCORE H		SCORE L	
Decile group	Average prediction	Actual probability	Average prediction	Actual probability	Average prediction	Actual probability*	Average prediction	Actual probability*	Average prediction	Actual probability*	Average prediction	Actual probability*
1	0.02662	0.02081	0.01498	0.02079	0.08292	0.06445	0.05281	0.06479	0.05997	0.06082	0.04096	0.08115
2	0.04264	0.0218	0.02582	0.01295	0.13083	0.06522	0.08496	0.04326	0.08921	0.06321	0.06059	0.06365
3	0.05918	0.01278	0.03809	0.02141	0.17836	0.04259	0.11725	0.06496	0.11454	0.16344	0.07623	0.10567
4	0.08099	0.01979	0.05345	0.02633	0.23806	0.0665	0.15957	0.06588	0.1429	0.06488	0.09345	0.16534
5	0.10729	0.02685	0.07157	0.02772	0.30611	0.08725	0.20944	0.06663	0.17111	0.12481	0.11346	0.06463
6	0.13931	0.04849	0.09449	0.0401	0.38324	0.15489	0.26769	0.17472	0.20475	0.27377	0.13555	0.20213
7	0.18859	0.07226	0.12709	0.06508	0.48985	0.21994	0.35471	0.22253	0.24413	0.2341	0.16518	0.29727
8	0.25377	0.12521	0.17874	0.1155	0.61016	0.3841	0.46414	0.35076	0.29291	0.2507	0.2005	0.25683
9	0.34997	0.23465	0.25076	0.22386	0.74797	0.6237	0.60052	0.60843	0.37195	0.45189	0.259	0.44614
10	0.57292	0.27283	0.44448	0.30469	0.9157	0.69801	0.81788	0.73361	0.55269	0.61689	0.41168	0.63402

Table 1. Average predictions and actual probabilities for each decile group for all examined risk score equations[19].
*10-year actual probabilities are estimated using Weibull regression models to project beyond the observed follow-up.

Overall, the results obtained in the above-described analyses were consistent with the ones obtained from the sensitivity analyses[19].
When we compared the predicted risks of SCORE and SCORE OP across age for three different hypothetical risk profiles (low, medium, and high-risk profile), we found that the SCORE OP predicted higher risks compared to the SCORE in all risk profiles for female persons aged over 75 years old and for male persons aged over 78 years old[19].

7.3 Counterfactual Prediction task: DAGs and causal thinking in probability estimation

7.3.1 Principle of independent mechanisms

Imagine that we are interested in developing a diagnostic clinical risk prediction model to predict the presence of Alzheimer's disease (Y=1, while Y=0 denotes absence of Alzheimer's disease)[20]. In this example, the only predictor we will use is the APOE ε4 allele status: X = 1 denotes allele presence, while X = 0 denotes allele absence[20].

As APOE ε4 allele is a known risk factor for Alzheimer's disease[39], we could represent the relationship between the two variables with the DAG in Figure 2a.

We assume that the allele APOE ε4 is a direct cause of the disease status and that the two variables have no common causes. Under this scenario, we could collect cross-sectional data about Alzheimer's disease presence and APOE ε4 allele status in a defined population A, and then use a logistic regression to predict the presence of Alzheimer's disease[20].

The logistic regression equation would then encode the four conditional probabilities, Pr(Y = 1|X = 1), Pr(Y = 1|X = 0), Pr(Y = 0|X = 1), and Pr(Y = 0|X = 0), which define the conditional distribution $\mathbb{P}$(Y|X)[20]. The marginal distribution $\mathbb{P}$(X) is fully defined from the prevalence of the APOE ε4 allele (Pr(X = 1)) because we are dealing with a binary variable[20].

Let us imagine that we would like to apply this recently developed diagnostic clinical risk prediction model in another population, labeled as population B, in which we know the prevalence of the APOE ε4 allele to be different[20].

In this case, APOE ε4 allele status (X) is the cause and $\mathbb{P}$(Y|X) represents the causal mechanism. Hence, according to the principle of independent mechanisms, a change in the cause distribution $\mathbb{P}$(X) does not give any information on $\mathbb{P}$(Y|X) in the population B[15,20,31].

The conclusion is that if the causal mechanism $\mathbb{P}$(Y|X) is the same in the two populations (A and B) the diagnostic clinical risk prediction model developed in population A will lead to valid predictions in population B[20]. However, if the underlying causal mechanism changed from population A to population B, no information on the change could be obtained knowing only the difference in the predictor distribution[15,20,31].

Consequently, the diagnostic clinical risk prediction model developed in population A is still our best guess[15,20,31].

In this example, our knowledge of the causal structure suggests that using the diagnostic clinical risk prediction model developed from population A, in the new population B is the right choice[15,20,31].

Imagine now that we want to predict the presence of Alzheimer’s disease using a binary variable that indicates if the concentration of tau protein measured in the cerebrospinal fluid is high (CSFtau), instead of the APOE ε4 allele status[20]. Since the CSFtau levels are a product of Alzheimer's disease[40], assuming that no other variables are relevant, the DAG can be represented as described in Figure 2b (Y indicates Alzheimer’s disease status, while K indicates CSFtau levels).

As before, we can develop a diagnostic clinical risk prediction model using a logistic regression and cross-sectional data from a defined population C[20].

Similarly to the previous example, the logistic regression would describe the conditional distribution $\mathbb{P}(Y|K)$, while the marginal distribution $\mathbb{P}(K)$ would be described by the CSFtau prevalence ($Pr(K = 1)$)[20].

Imagine we want to apply this developed diagnostic clinical risk prediction model to a new population, labeled as population D, in which we know that the prevalence of high CSFtau is different[20]. In this case, $\mathbb{P}(Y|K)$ does not represent a causal mechanism, and we are in the anticausal scenario, where the predictor K is an effect of the disease status (Y). According to the principle of independent mechanisms, $\mathbb{P}(Y|K)$ is not independent of $\mathbb{P}(K)$. Consequently, the change in the prevalence of CSFtau is likely to occur together with a change in $\mathbb{P}(Y|K)$[15,20,31]. We could conclude that the diagnostic clinical risk prediction model developed in population C would lead to invalid predictions in the population D, because it describes a mathematical relationship that is unlikely to hold in the population D[20]. This reasoning is correct, even though the causal mechanism between the Alzheimer’s disease status and the CSFtau levels is the same for both populations[20].

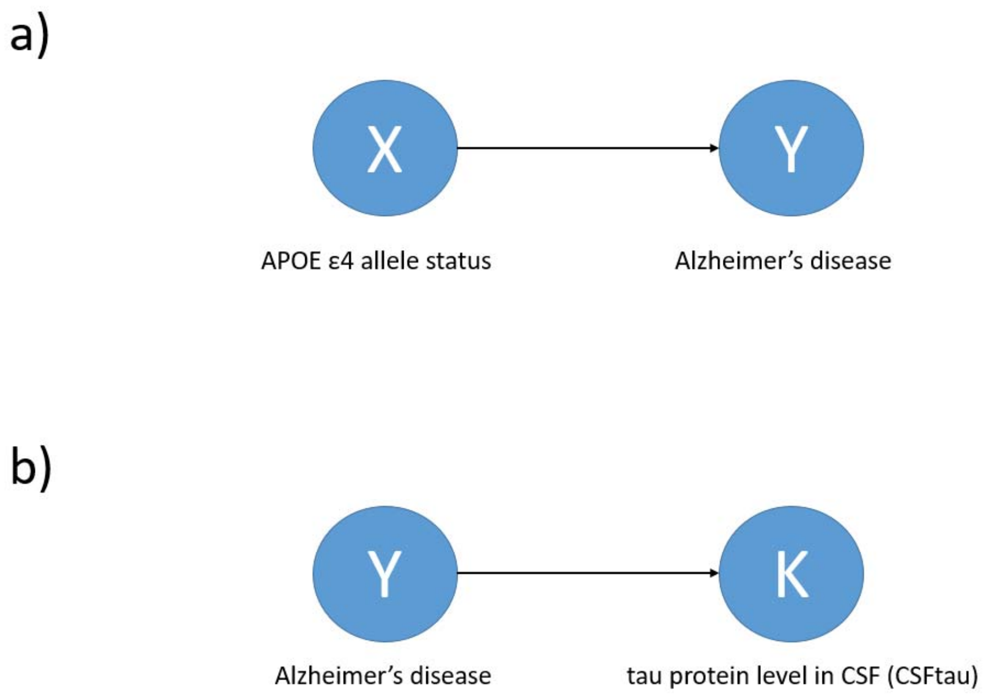


Figure 2. Directed acyclic graphs for the two simplified examples[20]. CSF=cerebrospinal fluid

7.3.2 Markov Blanket

To illustrate the use of the Markov Blanket concept in predictor selection, we generated 100,000 datasets each based on a random DAG[20].

Out of the 100,000 generated datasets, 37,272 presented an exogenous outcome variable, 8,032 of which presented an outcome variable without any children nodes[20]. Therefore, the performance of the logistic regression with parents as predictors and the logistic regression with the Markov Blanket set or with variables with a path to the outcome as predictors could not be assessed for all datasets[20]. In general, the Markov Blanket-based logistic model showed an average ICI of 0.01882, the lowest across all prediction tools[20].

The Markov Blanket-based logistic model had the lowest average ICI (0.01956) even when only the 62,159 datasets in which all prediction tools had a computable ICI were considered[20]. In these datasets, the Markov Blanket-based logistic model included on average only 4.1 predictors[20]. Still, its ICI was, in direct comparison, lower or equal to the ICI obtained from the other prediction tools most of the times (from 56.36% to 97.37%, depending on the tool being compared)[20]. The distribution of the ICI for the eight prediction tools over the 62,159 datasets is represented in Figure 3.

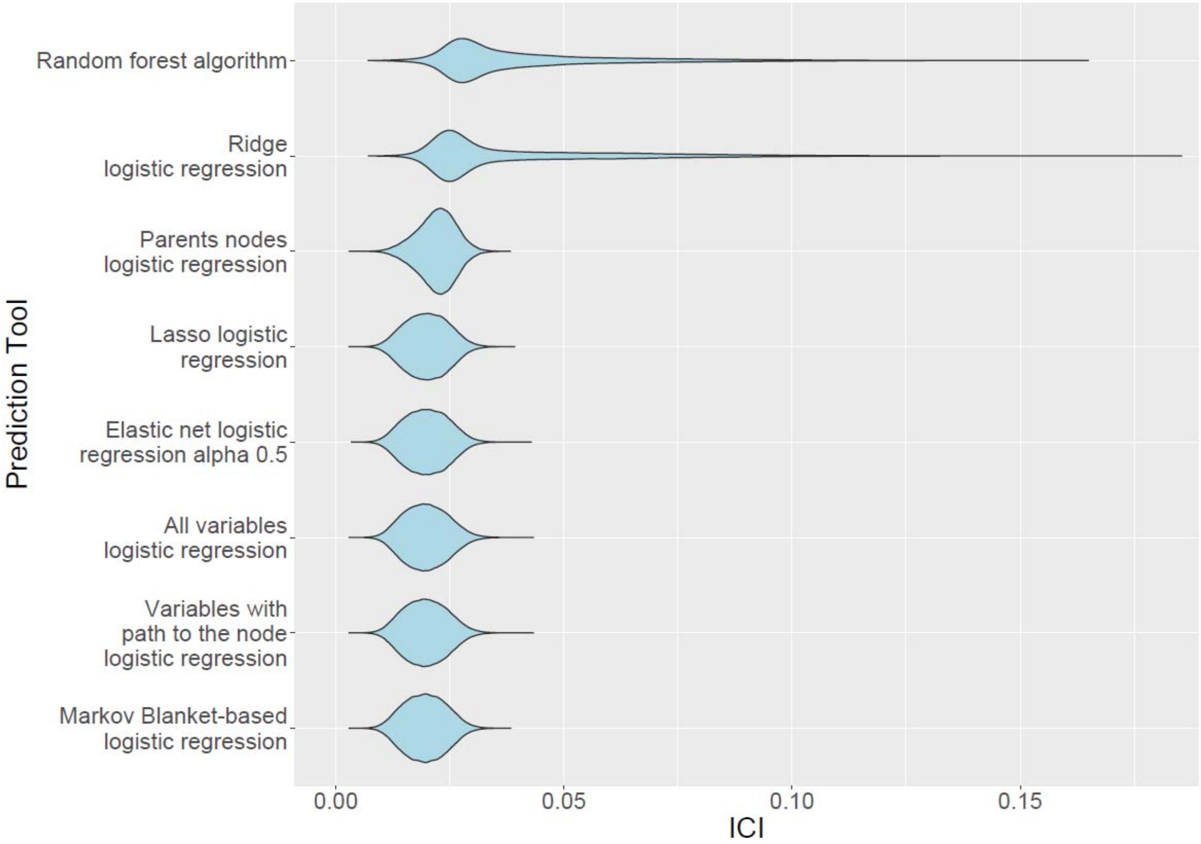


Figure 3. Violin plot of the distribution of the 10-fold cross-validated Integrated Calibration Index (ICI) for each prediction tool[20]. Only the 62,159 datasets in which the ICI was computable for all eight prediction tools were considered.

8 Discussion

8.1 Description task: all-cause mortality rate estimation using registry data

In the three projects, Description, Prediction, and Counterfactual prediction were coverd. Description consists of providing a compact quantitative representation of phenomena as they appear[3,14], and represents the basis for new hypotheses and research questions[41]. This task is particularly important when the research community is faced with a new phenomenon whose characteristics and behavior are unknown[41]. The newly emerged COVID-19 pandemic is a perfect opportunity to illustrate the value of descriptive studies. Description is particularly useful during such an emergency situation because of its potential to provide readily available results. Indeed, descriptive studies often rely on already available data, representing an efficient way to analyse trends, plan healthcare resources, and generate causal hypotheses[7,41]. In our application, we dealt with the statistical problem of probability estimation in the context of surveillance of incident cases (passive surveillance in particular), a specific type of descriptive study[41]. The probability of the event (also referred to as incidence proportion) cannot be directly computed by using only information from a registry surveilling an open population[7]. However, this quantity can be approximated using incidence rates, which can be estimated using information from the registry and administrative data about the population size at a specific time point[7,9].

In our study, we found that the all-cause mortality rate in Nembro increased dramatically in March 2020[18]. In this month, the all-cause mortality was 154.4 per 1,000 person-years[18], which corresponds approximately to a probability of 1.3% of dying in one month (using the exponential formula and relying on some known assumptions[6,7]). The mortality rate in March 2020 was almost 11 times the mortality rate recorded in March 2019, and the number of recorded deaths in March 2020 was higher than the number of deaths recorded during the entire previous year[18].

The increase in all-cause deaths observed during the first months of the COVID-19 pandemic in Nembro is not completely explained by official COVID-19 statistics. At the end of March 2020, the mayor of Nembro reported that only about 200 COVID-19 cases were confirmed in the city[25]. According to data obtained from Ondata[42], only 85 deaths in Nembro were officially attributed to COVID-19 between the beginning of the outbreak and April 11th, 2020[18].

One limitation of our study is the use of provisional or not fully updated data[18]. Arguably this limitation was unavoidable because of the timing of the analysis and the nature of the COVID-19 emergency situation in Italy.

Overall, our results were consistent with previous findings. The dramatic increase in all-cause deaths within the first months of the COVID-19 pandemic in Italy that were not fully accounted for in the number of official COVID-19 deaths was also highlighted in some previous reports: a local investigation on all-cause deaths in Bergamo[43], a study analysing more than 1000 Italian cities with increased mortality[44], and a report of the Italian Ministry of Health and the Italian National Center for Prevention and Control of Disease analysing all-cause mortality of 18 major Italian cities[45].

The mismatch between the increase in all-cause mortality and the number of official COVID-19 confirmed deaths in the first months of the pandemic may have been due to a combination of several factors. Firstly, the lack of testing due to resource shortages may have led to the misclassification of several cause of death (in Nembro it was not possible to test all individuals with symptoms or with COVID-19 contacts[46]). Secondly, the occurrence of indirect deaths of individuals affected by other conditions who did not receive appropriate care due to the crisis of the healthcare system (in the province of Bergamo the healthcare system was overwhelmed and close to collapse[24,47]). Lastly, the delay in obtaining and communicating COVID-19 testing results may explain some of the disparity between the disproportionate increase in all-cause deaths and official COVID-19 deaths[18].

In Nembro, we observed a decrease in all-cause mortality in the month of April 2020[18]. This was probably mostly due to the strict measures introduced by the Italian government to contain the spread of the virus, social and hygienic preventive measures, reallocation of healthcare resources, and increased number of immune individuals[18].

More recently, a report from the Italian National Institute of Health and the Italian National Institute of Statistics analysing mortality data of 95% of the Italian resident population confirmed that 15,133 deaths from COVID-19 positive individuals occurred in the month of March, while in the same month 27,195 more deaths compared to the average 2015-2019 were recorded[48].

These findings show that all-cause mortality is an important metric to quantify the consequences of a pandemic. Despite being only a general indicator of the health of the population, all-cause mortality has several advantages. Mortality data are systematically collected, death is an end-point whose definition is universally accepted, collection of mortality data is usually of high quality, and mortality data collection does not depend on the specific testing strategies[18]. Moreover, all-cause mortality allows researchers to consider not only the deaths directly induced by the virus but also the ones caused by the crisis in the healthcare system due to the pandemic[18]. This indirect effect is a crucial consequence of the pandemic[49] that needs to be taken into account by policy makers. For this reason, we believe that all-cause mortality should be reported along with more traditional metrics in the first stages of a pandemic[18].

8.2 Prediction task: external validation of the SCORE OP risk prediction model

Prediction focuses on “forecasting” the value of an outcome variable for an individual whose outcome status is unknown. When the outcome is binary, as often happens in medicine, the prediction task consists of estimating a probability. In this scenario, the main challenge is making sure that the developed risk prediction model has good performance on individuals outside of the development dataset[10,12]. Therefore, conducting external validations is crucial to assess the performance in estimating probabilities for individuals from different settings[10,12].

This is especially important in cardiovascular medicine, in which the use of risk prediction models to target high-risk individuals is widespread and recommended by the majority of global guidelines[50]. A recent systematic review by Damen et al. found that, from 1967 until 2013, 363 risk prediction models for future cardiovascular outcomes were developed to be used in the general population[51]. However, only 36% of all these risk prediction models had ever been externally validated[51].

In our application, we conducted an external validation in the Berlin population of a recently developed risk score, the SCORE OP, for predicting 5- and 10-year risk of fatal cardiovascular events in European older persons[19].

Our validation study found that the SCORE OP equations substantially overestimated the true risk of a fatal cardiovascular event[19].

SCORE equations, specifically the SCORE for high-risk populations, showed better calibration than the SCORE OP models in our analysis based on 10-year projections[19].

These results are important both from a practical and theoretical point of view. Indeed, the use in clinical practice of a risk score that overestimates the true risk could lead to overtreatment[19]. On the other hand, it is also interesting that the SCORE OP was originally developed to correct the SCORE's suspected overestimation in older persons but provided higher estimates in our population[28].

It is indeed worth further exploring the underlying rationale for the SCORE OP development. The SCORE model was developed using information about mostly middle-aged Europeans. Since the association between the traditional risk factors and cardiovascular outcomes decreases with age[52,53], SCORE regression coefficients were thought to reflect too strong of an association, which resulted in being too great in magnitude for older persons[28,52–54]. The inappropriately large coefficients combined with the high risk factor levels typical of older persons were, therefore, thought to lead the SCORE to overestimate the true risk in this population[28]. However, in our study, we found that the SCORE OP predicted higher risks compared to the SCORE[19], and this result was confirmed in a previously published external validation conducted in the UK[55]. Conversely, in two previously published cross-sectional studies, the SCORE OP was found to predict lower risks than the SCORE in a population of individuals aged 65 to 69[56,57]. We believe that this inconsistency can be explained by the different age distributions[19]. Comparing the values obtained by the two different risk assessment systems for six hypothetical risk profiles, we found that the SCORE OP tends to provide lower risk estimates compared to the SCORE for individuals in the age group 65-68 but tends to provide higher risk estimates for older individuals[19].

In general, the lower generalizability of the SCORE OP to the Berlin older persons population could be explained by a different unmeasured risk factors distribution, a different baseline cardiovascular risk, or a difference in the underlying hazard ratios compared to the SCORE OP development dataset[19]. A factor that could have played a role in determining the low transportability is the different definition of the risk factors[19]. Specifically, the large difference in the prevalence of diabetes in the SCORE OP development dataset compared to the BIS dataset could be explained by different ways of assessing or defining this important risk factor[19].

Despite the large sample size[58], high-quality data, and the robust methods used in this study, some limitations have to be considered when interpreting the results. Firstly, the cause of death reported on death certificates is known not to be completely reliable for older persons; secondly, the comparison of 10-year SCORE and SCORE OP equations needs to be interpreted carefully since it is based on projected endpoints rather than observed ones; and finally, we excluded individuals based on self-reported information about myocardial infarction history rather than medical records[19].

The only previously published external validation of the SCORE OP found that this risk prediction model had very good calibration but poor discrimination ability in the Norfolk older persons population (UK)[55]. Despite reporting very different performances from those observed in our study, both this and our external validations suggested the use of the SCORE OP in clinical practice in these two populations was not appropriate[19,55].

Life expectancy has continued to increase over the last decades, and currently less than one-fifth of the overall cardiovascular deaths in Europe occur in individuals younger than 65[59]. For these reasons, developing a prediction tool to identify older persons at high cardiovascular risk with good performance is of crucial importance[28,60]. Conducting external validations is fundamental for understanding the settings in which it is appropriate to use a particular clinical risk prediction model. Additionally, as in our application, external validations are valuable in identifying possible threats to transportability and problems in the theoretical framework underlying the risk score development while also providing useful information for improving clinical risk prediction model development.

8.3 Counterfactual Prediction task: DAGs and causal thinking in probability estimation

Similarly to Prediction, the Counterfactual Prediction task consists of estimating a probability when the outcome of interest is binary[4]. Relying on previous causal knowledge, Counterfactual Prediction aims at estimating the probability of the outcome in different worlds in which different interventions are applied[3,4]. In our study, we showed how counterfactual reasoning and causal knowledge could be of paramount importance when the aim is to develop a model to estimate the probability of the outcome in different settings[20]. Assessment of transportability across different settings can be conceptualized as the evaluation of a model's ability to predict the outcome in different counterfactual worlds. These various counterfactual worlds are characterized by interventions that alter, to some extent, the joint distribution of the variables. For example, other populations characterized by different distributions of an exogenous variable can be thought of as "counterfactual" worlds in which an intervention on the exogenous variable was applied. In our study, we investigated the important consequences that this reasoning can have on statistical modelling using causal inference tools such as DAGs. We showed that, in two specific clinical examples, a diagnostic clinical risk prediction model in the “anticausal direction” (i.e. one which uses a consequence of the disease as a predictor) is less transportable compared to one in the “causal direction”[20]. We provided a theoretical framework to support the idea[61] that a prediction model, which includes causes of the outcome, is more transportable[20]. We also demonstrated how the causal knowledge summarized in a DAG can be useful for predictor selection[20].

By borrowing concepts from causal discovery and causal inference fields, we demonstrated that strong causal knowledge of the underlying data generation process could provide support to clinical risk prediction modeling strategies[20]. Previous work detailing the importance of causal knowledge in prediction tasks focused on prognostic clinical risk prediction modeling strategies in situations of treatment initiation after the baseline[62,63]. This was also the topic of a commentary by Dickerman and Hernán published shortly after our study[64]. The authors discuss the importance of causal knowledge in clinical risk prediction and the practical trade-off between counterfactual and factual prediction in this setting[64]. Since then several pieces illustrating the importance of causal knowledge and counterfactual thinking when dealing with clinical prediction tasks, especially in machine learning techniques, have been published[65–67]. This increasing body of literature demonstrates the way that knowledge and methods from the Counterfactual Prediction field have an influence on Prediction model research in epidemiology. This was probably a consequence of the recent emergence of causal discovery, a research field that uses prediction tools to infer the causal structure underlying the data[15]. This exchange across the border separating the two Health Data Science tasks and the ensuing contamination of each field creates a new lively, active, and young area of research with an extremely large potential.

9 Conclusion

In this PhD project, I focused on three different relevant applications of the statistical problem of probability estimation, each involving one of the three Health Data Science tasks.

I demonstrated how Description tasks are characterized by specific statistical challenges determined by the nature of the analysed data, often originally collected for other purposes. Relying on this type of data collection represents, on the one hand, a limitation from a statistical perspective but allows, on the other, for cost-effective and quick analyses, which unquestionably provides an advantage during new public health emergencies. Description tasks focus on estimating probabilities only conditioned on a few variables (for example, the city of residence, the month, and the year) to provide general information on the phenomenon. Prediction tasks, meanwhile, generally aim to estimate probabilities conditioned on several variables. Indeed, clinical risk prediction models try to provide as "personalized" predictions as possible. The crucial statistical challenge in Prediction is producing models capable of providing valid predictions for individuals outside the development dataset. This is essential for clinical practice as only valid prediction tools allow for correct estimation of individual risk and subsequent decision on preventive strategies. The transportability of a clinical risk prediction model across different settings and populations is assessed in external validations. I showed how transportability can be interpreted as an overlapping concept between Prediction and Counterfactual Prediction, since different settings can be interpreted as the result of different interventions. Counterfactual Prediction answers "what if" questions and is deeply rooted in causal thinking. I showed how causal knowledge can be compactly summarized in a DAG and how useful the information entailed in such a causal graph can be for risk prediction problems.

Each of the three Health Data Science tasks is of fundamental importance for health research. They are each characterized by their own statistical challenges, methods, and objectives. For practical and didactical reasons, avoiding confusion by keeping the three tasks separate is important. Maintaining a clear separation of these concepts could prevent common errors, such as the interpretation of regression coefficients in prediction models as causal effect estimates. However, it is undeniable that there are no well-defined borders between the three tasks. In the last years, new synergies have developed between Prediction and Counterfactual Prediction fields. The exchange of tools, statistical techniques, and theoretical concepts across these two research areas made important scientific advancement possible and opened several tracks for future research in biomedicine.

10 References

1. T. LeVasseur M, D. Goldstein N, A. McClure L. On the Convergence of Epidemiology, Biostatistics, and Data Science. Harvard Data Science Review. 2020. doi:10.1162/99608f92.9f0215e6

2. Centre for Big Data Research in Health. What is Health Data Science? Available: https://cbdrh.med.unsw.edu.au/what-health-data-science

3. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. Chance . 2019;32: 42–49.

4. Hernan MA, Robins JM. Causal Inference. Taylor & Francis; 2019.

5. Porta M. A dictionary of epidemiology 5th edition. Oxford University Press Inc; 2008.

6. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Lippincott Williams & Wilkins; 2008.

7. Logroscino G, Kurth T, Piccininni M. The Reconstructed Cohort Design: A Method to Study Rare Neurodegenerative Diseases in Population-Based Settings. Neuroepidemiology. 2020; 1–9.

8. Noordzij M, Dekker FW, Zoccali C, Jager KJ. Measures of disease frequency: prevalence and incidence. Nephron Clin Pract. 2010;115: c17–20.

9. Vandenbroucke JP, Pearce N. Incidence rates in dynamic populations. Int J Epidemiol. 2012;41: 1472–1479.

10. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer International Publishing; 2019.

11. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012. pp. 683–690. doi:10.1136/heartjnl-2011-301246

12. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98: 691–698.

13. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media; 2009.

14. Shmueli G. To Explain or To Predict? SSRN Electronic Journal. doi:10.2139/ssrn.1351252

15. Peters J, Janzing D, Schölkopf B. Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press; 2017.

16. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10: 37–48.

17. Greenland S, Pearl J. Causal Diagrams. Wiley StatsRef: Statistics Reference Online. 2014. doi:10.1002/9781118445112.stat03732

18. Piccininni M, Rohmann JL, Foresti L, Lurani C, Kurth T. Use of all cause mortality to quantify the consequences of covid-19 in Nembro, Lombardy: descriptive study. BMJ. 2020;369: m1835.

19. Piccininni M, Rohmann JL, Huscher D, Mielke N, Ebert N, Logroscino G, Schäffner E, Kurth T. Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal. PLoS One. 2020;15: e0231097.

20. Piccininni M, Konigorski S, Rohmann JL, Kurth T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. BMC Med Res Methodol. 2020;20: 179.

21. Cornetti R, Dordi L. Dati statistici e demografici. Comune di Nembro; Available: http://comune.nembro.bg.it/export/sites/default/.galleries/documenti/allegati-eventi-news/2020/Comune-di-Nembro-anno-2019-dati-statistici.pdf

22. ISTAT. Demo-Geodemo. - Maps, Population, Demography of ISTAT - Italian Institute of Statistics. [cited 2 May 2020]. Available: http://demo.istat.it/index_e.html

23. Piano Integrato Locale di promozione della Salute 2019 - ATS Bergamo - Agenzia di Tutela della Salute della provincia di Bergamo. [cited 2 May 2020]. Available: http://www.ats-bg.it/servizi/gestionedocumentale/ricerca_fase03.aspx?ID=27872

24. Fagiuoli S, Lorini FL, Remuzzi G, Covid-19 Bergamo Hospital Crisis Unit. Adaptations and Lessons in the Province of Bergamo. N Engl J Med. 2020;382: e71.

25. Coronavirus, sindaco Nembro: "Governo e Regione non hanno seguito Iss" | Sky TG24. [cited 2 May 2020]. Available: https://tg24.sky.it/milano/2020/03/27/coronavirus-bergamo-sindaco-nembro.html

26. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetière P, Jousilahti P, Keil U, Njølstad I, Oganov RG, Thomsen T, Tunstall-Pedoe H, Tverdal A, Wedel H, Whincup P, Wilhelmsen L, Graham IM, SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24: 987–1003.

27. Authors/Task Force Members:, Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, Cooney M-T, Corrà U, Cosyns B, Deaton C, Graham I, Hall MS, Hobbs FDR, Løchen M-L, Löllgen H, Marques-Vidal P, Perk J, Prescott E, Redon J, Richter DJ, Sattar N, Smulders Y, Tiberi M, Bart van der Worp H, van Dis I, Verschuren WMM. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Atherosclerosis. 2016;252: 207–274.

28. Cooney MT, Selmer R, Lindman A, Tverdal A, Menotti A, Thomsen T, DeBacker G, De Bacquer D, Tell GS, Njolstad I, Graham IM, SCORE and CONOR investigators. Cardiovascular risk estimation in older persons: SCORE OP. Eur J Prev Cardiol. 2016;23: 1093–1103.

29. Støvring H, Harmsen CG, Wisløff T, Jarbøl DE, Nexøe J, Nielsen JB, Kristiansen IS. A competing risk approach for the European Heart SCORE model based on cause-specific and all-cause mortality. Eur J Prev Cardiol. 2013;20: 827–836.

30. Schaeffner ES, van der Giet M, Gaedeke J, Tölle M, Ebert N, Kuhlmann MK, Martus P. The Berlin initiative study: the methodology of exploring kidney function in the elderly by combining a longitudinal and cross-sectional approach. Eur J Epidemiol. 2010;25: 203–210.

31. Schoelkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. On Causal and Anticausal Learning. arXiv [cs.LG]. 2012. Available: http://arxiv.org/abs/1206.6471

32. Koller D, Sahami M. Toward Optimal Feature Selection. 1996 [cited 22 Sep 2019]. Available: http://ilpubs.stanford.edu:8090/208/

33. Brown LE, Tsamardinos I. Markov blanket-based variable selection in feature space. Technical Report DSL TR-08-01. Department Biomedical Informatics, Vanderbilt University; 2008. Available: https://www.researchgate.net/profile/Ioannis_Tsamardinos/publication/268340454_Markov_Blanket-Based_Variable_Selection_in_Feature_Space/links/546f4e3a0cf2d67fc0310b1e.pdf

34. Fu S, Desmarais MC. Markov blanket based feature selection: a review of past decade. Proceedings of the world congress on engineering. Newswood Ltd; 2010. pp. 321–328.

35. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann; 1988.

36. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997;97: 273–324.

37. Tsamardinos I, Aliferis CF. Towards principled feature selection: relevancy, filters and wrappers. AISTATS. 2003. Available: https://www.researchgate.net/profile/Ioannis_Tsamardinos/publication/2478803_Towards_Principled_Feature_Selection_Relevancy_Filters_and_Wrappers/links/00b49520e1396d8ec1000000.pdf

38. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med. 2019. doi:10.1002/sim.8281

39. Uddin MS, Kabir MT, Al Mamun A, Abdel-Daim MM, Barreto GE, Ashraf GM. APOE and Alzheimer’s Disease: Evidence Mounts that Targeting APOE4 may Combat Alzheimer’s

Pathogenesis. Molecular Neurobiology. 2019. pp. 2450–2465. doi:10.1007/s12035-018-1237-z

40. Lee JC, Kim SJ, Hong S, Kim Y. Diagnosis of Alzheimer's disease utilizing amyloid and tau as fluid biomarkers. Experimental & Molecular Medicine. 2019. doi:10.1038/s12276-019-0250-2

41. Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. Lancet. 2002;359: 145–149.

42. onData | Associazione per la promozione di trasparenza e open data. [cited 2 May 2020]. Available: https://ondata.it/

43. Coronavirus, the real death toll: 4.500 victims in one month in the province of Bergamo. [cited 24 Sep 2020]. Available: https://www.ecodibergamo.it/stories/bergamo-citta/coronavirus-the-real-death-tool-4500-victims-in-one-month-in-the-province-of_1347414_11/

44. La crescita della mortalità ai tempi del Covid-19 (IT EN) - Istituto Cattaneo. 1 Apr 2020 [cited 24 Sep 2020]. Available: https://www.cattaneo.org/2020/04/01/gli-effetti-del-covid-19-sulla-mortalita/

45. De Sario Paola Michelozzi Fiammetta Noccioli Daniela Orrù Pasqualino Rossi Matteo Scortichini MDFDDM. Andamento della Mortalità Giornaliera (SiSMG) nelle città italiane in relazione all'epidemia di Covid-19. Ministero della Salute, Centro Nazionale Prevenzione e Controllo Malattie; Available: http://www.epiprev.it/sites/default/files/SISMG_COVID19_28032020-2.pdf

46. Relazione al Consiglio Comunale EMERGENZA covid-19. Available: http://nembro.net/export/sites/default/.galleries/documenti/allegati-eventi-news/2020/Relazione-al-Consiglio-Comunale-EMERGENZA-covid-19.pdf

47. Nacoti M, Ciocca A, Giupponi A, Brambillasca P, Lussana F, Pisano M, Goisis G, Bonacina D, Fazzi F, Naspro R, Longhi L, Cereda M, Montaguti C. At the epicenter of the Covid-19 pandemic and humanitarian crises in Italy: changing perspectives on preparation and mitigation. NEJM Catalyst Innovations in Care Delivery. 2020;1. Available: https://catalyst.nejm.org/doi/abs/10.1056/CAT.20.0080

48. Mortality of the resident population. 10 Jul 2020 [cited 24 Sep 2020]. Available: https://www.istat.it/en/archivio/245457

49. Fisher D, Heymann D. Q&A: The novel coronavirus outbreak causing COVID-19. BMC Med. 2020;18: 57.

50. Collins DRJ, Tompson AC, Onakpoya IJ, Roberts N, Ward AM, Heneghan CJ. Global cardiovascular risk assessment in the primary prevention of cardiovascular disease in adults: systematic review of systematic reviews. BMJ Open. 2017;7: e013650.

51. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, Schlüssel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KGM. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353: i2416.

52. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol. 2009;54: 1209–1227.

53. Cooney MT, Dudina A, D'agostino R, Graham IM. Cardiovascular risk-estimation systems in primary prevention: do they differ? Do they make a difference? Can we see the future? Circulation. 2010. Available: https://www.ahajournals.org/doi/abs/10.1161/circulationaha.109.852756

54. Catapano AL, Graham I, De Backer G, Wiklund O, Chapman MJ, Drexel H, Hoes AW, Jennings CS, Landmesser U, Pedersen TR, Reiner Ž, Riccardi G, Taskinen M-R, Tokgozoglu L, Verschuren WMM, Vlachopoulos C, Wood DA, Zamorano JL. 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias: The Task Force for the Management of Dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Atherosclerosis. 2016;253: 281–344.

55. Verweij L, Peters RJG, Scholte Op Reimer WJM, Boekholdt SM, Luben RM, Wareham NJ, Khaw K-T, Latour CHM, Jorstad HT. Validation of the Systematic COronary Risk Evaluation - Older Persons (SCORE-OP) in the EPIC-Norfolk prospective population study. Int J Cardiol. 2019;293: 226–230.

56. Brotons C, Moral I, Fernández D, Cuixart L, Soteras A, Puig M. Assessment of the New SCORE OP Cardiovascular Risk Charts in Patients Older Than 65 Years. Rev Esp Cardiol . 2016;69: 981–983.

57. Brotons C, Moral I, Fernández D, Cuixart L, Muñoz A, Soteras A, Puig M, Joaniquet X, Casasa A. [Clinical consequences of using the new cardiovascular risk tables SCORE OP in patients aged over 65 years]. Med Clin . 2016;147: 381–386.

58. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35: 214–226.

59. Mortensen MB, Afzal S, Nordestgaard BG, Falk E. The high-density lipoprotein-adjusted SCORE model worsens SCORE-based risk classification in a contemporary population of 30 824 Europeans: the Copenhagen General Population Study. Eur Heart J. 2015;36: 2446–2453.

60. Koller MT, Steyerberg EW, Wolbers M, Stijnen T, Bucher HC, Hunink MGM, Witteman JCM. Validity of the Framingham point scores in the elderly: results from the Rotterdam study. Am Heart J. 2007;154: 87–93.

61. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Second edition. Springer; 2019.

62. Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Statistics in Medicine. 2018. pp. 4142–4154. doi:10.1002/sim.7913

63. van Geloven N, Swanson SA, Ramspek CL, Luijken K, van Diepen M, Morris TP, Groenwold RHH, van Houwelingen HC, Putter H, le Cessie S. Prediction meets causal inference: the role of treatment in clinical prediction models. Eur J Epidemiol. 2020;35: 619–630.

64. Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. Eur J Epidemiol. 2020;35: 615–617.

65. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, Rich S, Wang M, Buchan IE, Bian J. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence. 2020;2: 369–375.

66. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun. 2020;11: 3673.

67. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun. 2020;11: 3923.

11 Statutory Declaration

I, Marco Piccininni, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic "*The probability estimation problem in the Health Data Science framework: methodological considerations and applications*" ("*Herausforderungen der Wahrscheinlichkeitsschätzung in Health Data Science: methodische Überlegungen und Anwendungen*"), independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.

Date Signature

12 Declaration of your own contribution to the publications

Marco Piccininni contributed the following to the below listed publications:

Publication 1: **Piccininni M**, Rohmann JL, Foresti L, Lurani C, Kurth T. Use of all cause mortality to quantify the consequences of covid-19 in Nembro, Lombardy: descriptive study. *BMJ*. 2020;369: m1835.

Impact factor (2018): 27.604

Contribution: Marco Piccininni, together with his supervisor, conceived the study. He led the communications with the research partners (Centro Medico Santagostino, Italy). He planned all statistical analyses by himself. He downloaded the National Institute of Statistics and the OnData datasets. Once data were available, he conducted all steps of statistical analyses described in the article by himself. He created the original figures and tables, which were later edited by the journal graphic design team. He contributed to the interpretation of the results and to create the preliminary version of the manuscript draft, together with his supervisors. He further independently coordinated the journal submission and the revision process and compiled input from the coauthors.

Publication 2: **Piccininni M***, Rohmann JL, Huscher D, Mielke N, Ebert N, Logroscino G, Schäffner E, Kurth T. Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal. *PLoS One*. 2020;15: e0231097.

Impact factor (2018): 2.776

Contribution: Marco Piccininni conceptualized the statistical elements of the study by himself to match his supervisors' epidemiological conceptualization. He drafted a first version of the analysis plan by himself, which was then modified after detailed discussion about the methodology with his supervisors. He ran all statistical analyses described in the article using the R software by himself. He also created sample code for purposes of transpancy and reproducibility included in the Supplementary material. He, himself, created the original version of the tables included in the manuscript (table 1 was later modified by a supervisor for aesthetic reasons). He created the original panel figures of calibration assessment that are presented in the manuscript (then, a supervisor merged them to create Figure 2, Figure 3, Figure 4 and Figure S2, also modifying the labels for aesthetic reasons). He created Figure 5 by himself. He interpreted the results of the calibration and discrimination assessments. He contributed to drafting the first version of the manuscript together with his supervisor. He adjusted statistical analysis according to reviewers' comments during the revision process.

* Marco Piccininni and the supervisor Jessica L. Rohmann are joint first authors of this work

Publication 3: **Piccininni M**, Konigorski S, Rohmann JL, Kurth T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC Med Res Methodol*. 2020;20: 179.

Impact factor (2018): 2.509

Contribution: Marco Piccininni conceptualized the study by himself. He designed the simulation plan together with Stefan Konigorski. He conceived the clinical examples described in the article by himself. He wrote the first version of the simulation R code by himself (the final code is available in the Supplementary material). He created all the original figures included in the manuscript by himself. He created the table included in the manuscript by himself. He drafted the first version of the manuscript together with a supervisor. He further coordinated the journal submission and the revision process.

Signature, date and stamp of first supervising university professor / lecturer

Signature of doctoral candidate

13 List of selected publications

13.1 Publication I: Use of all cause mortality to quantify the consequences of covid-19 in Nembro, Lombardy: descriptive study

Piccininni M, Rohmann JL, Foresti L, Lurani C, Kurth T. Use of all cause mortality to quantify the consequences of covid-19 in Nembro, Lombardy: descriptive study. *BMJ*. 2020;369: m1835.

DOI: https://doi.org/10.1136/bmj.m1835

Journal Data Filtered By: **Selected JCR Year: 2018** Selected Editions: SCIE,SSCI
Selected Categories: **"MEDICINE, GENERAL and INTERNAL"**
Selected Category Scheme: WoS
Gesamtanzahl: 160 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	NEW ENGLAND JOURNAL OF MEDICINE	344,581	70.670	0.686700
2	LANCET	247,292	59.102	0.427870
3	JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	156,350	51.273	0.300810
4	Nature Reviews Disease Primers	4,339	32.274	0.019740
5	BMJ-British Medical Journal	112,901	27.604	0.152760
6	JAMA Internal Medicine	15,215	20.768	0.095580
7	ANNALS OF INTERNAL MEDICINE	57,057	19.315	0.096020
8	PLOS MEDICINE	30,689	11.048	0.071200
9	Journal of Cachexia Sarcopenia and Muscle	2,799	10.754	0.005870
10	BMC Medicine	13,630	8.285	0.045220
11	Cochrane Database of Systematic Reviews	67,607	7.755	0.158690
12	MAYO CLINIC PROCEEDINGS	14,695	7.091	0.025750
13	CANADIAN MEDICAL ASSOCIATION JOURNAL	15,351	6.938	0.016500
14	JOURNAL OF INTERNAL MEDICINE	10,547	6.051	0.015700
15	Journal of Clinical Medicine	2,315	5.688	0.007210
16	MEDICAL JOURNAL OF AUSTRALIA	11,134	5.332	0.012600
17	PALLIATIVE MEDICINE	5,682	4.956	0.009860
18	AMYLOID-JOURNAL OF PROTEIN FOLDING DISORDERS	1,335	4.919	0.003270

Selected JCR Year: 2018; Selected Categories: "MEDICINE, GENERAL und INTERNAL"

FAST TRACK

Use of all cause mortality to quantify the consequences of covid-19 in Nembro, Lombardy: descriptive study

Marco Piccininni,[1] Jessica L Rohmann,[1] Luca Foresti,[2] Caterina Lurani,[2] Tobias Kurth[1]

[1]Institute of Public Health, Charité - Universitätsmedizin Berlin, 10117 Berlin, Germany

[2]Centro Medico Santagostino, 20127 Milan, Italy

Correspondence to: M Piccininni marco.piccininni@charite.de (ORCID 0000-0002-1397-0060)

Cite this as: *BMJ* 2020;369:m1835
http://dx.doi.org/10.1136/bmj.m1835

Accepted: 5 May 2020

ABSTRACT

OBJECTIVE

To quantify the impact of coronavirus disease 2019 (covid-19) on all cause mortality in Nembro, an Italian city severely affected by the covid-19 pandemic.

DESIGN

Descriptive study.

SETTING

Nembro, in the Bergamo province of Lombardy, northern Italy.

POPULATION

Residents of Nembro.

MAIN OUTCOME MEASURES

Monthly all cause mortality between January 2012 and April 2020 (data to 11 April), number of confirmed deaths from covid-19 to 11 April 2020, and weekly absolute number of deaths between 1 January and 4 April across recent years by age group and sex.

RESULTS

Nembro had 11 505 residents as of 1 January 2020. Monthly all cause mortality between January 2012 and February 2020 fluctuated around 10 per 1000 person years, with a maximum of 21.5 per 1000 person years. In March 2020, monthly all cause mortality reached a peak of 154.4 per 1000 person years. For the first 11 days in April, this rate decreased to 23.0 per 1000 person years. The observed increase in mortality was driven by the number of deaths among older people (≥65 years), especially men. From the outbreak onset until 11 April 2020, only 85 confirmed deaths from covid-19 in Nembro were recorded, corresponding to about half of the 166 deaths from all causes observed in that period.

CONCLUSIONS

The study findings show how covid-19 can have a considerable impact on the health of a small community. Furthermore, the results suggest that the full implications of the covid-19 pandemic can only be completely understood if, in addition to confirmed deaths related to covid-19, consideration is also given to all cause mortality in a given region and time frame.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The global spread of coronavirus disease 2019 (covid-19) has severely affected northern Italy

The consequences of covid-19 are generally assessed using the number of confirmed covid-19 related deaths

WHAT THIS STUDY ADDS

The covid-19 pandemic had a substantial impact on the health of the small community of Nembro city (Lombardy, Italy) based on comparisons of monthly all cause mortality since 2012

All cause mortality represents an important metric to quantify the burden of a pandemic

Introduction

The global spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the resulting coronavirus disease 2019 (covid-19)[1] quickly escalated into a critical situation for healthcare systems worldwide and continues to pose a major threat to population health. In Italy, more than 28 700 people have died from covid-19, the highest number of officially reported covid-19 related deaths in Europe (as of 2 May 2020).[2-4] The steep increase in the number of people with symptoms of covid-19 led to a sudden and catastrophic overload of Italian healthcare capacities.[5 6]

The Lombardy region of northern Italy, an area with almost 17% of the Italian population (2019 data[7]), rapidly became the most severely affected area, and by late March the region comprised 60% of all covid-19 related deaths in Italy and 40% of the confirmed covid-19 cases.[8] Considerable media coverage of Bergamo, one of the first Italian cities in Lombardy to be severely affected by covid-19, showed military vehicles carrying coffins to other cities because of the lack of space and morgue staff.[9]

Although the reported number of covid-19 related deaths in Bergamo is high,[10] the real figure could be even higher according to all cause mortality data. A local investigation raised initial doubts about the accuracy of confirmed case and death counts, indicating a substantial underestimation of the magnitude of the burden.[11] Such underreporting is not surprising given the state of emergency in many of the hospitals, the large number of patients needing immediate intensive care, the enormous time and emotional pressure on medical teams, the shortage of materials and human resources, and the clinical complexity of the disease.[5 12 13] However, underreporting is not the only possible explanation for the considerable difference between the number of covid-19 specific deaths and the increase in all cause deaths.

The impact of covid-19 on all cause mortality is especially noticeable when data are analysed from small cities characterised by stable age-sex structures over time and low mobility. This metric is sensitive to small increases in absolute numbers of deaths in small cities. In an effort to accurately determine the consequences of covid-19 on mortality, we describe the change in all cause mortality over time in Nembro, a small city in the province of Bergamo (Lombardy) that has been severely affected by the covid-19 pandemic.

Methods

Setting

Nembro, located in the Bergamo province of Lombardy, has a population of 11 505 (2020 data[14]). In 2018, life expectancy at birth in the province was 81.2 years for men and 85.5 years for women, similar to those in the region.[7] Between 2009 and 2015, cancer was the leading cause of death in men in the province (cause specific mortality 328.7 per 100 000 person years) and cardiovascular disease was the leading cause of death in women (317.8 per 100 000 person years).[15] Among cancer related deaths, bronchial and lung cancers were the most common causes in men and the second most common causes in women.[15,16] Between 2009 and 2015, the mortality rate for all cancers in the province was higher than the rate in the region of Lombardy, while the mortality rates for cardiovascular diseases and bronchial and lung cancers did not differ.[15,16]

In a provincial survey between 2011 and 2014 among adults aged 18 to 69 years, 24% were current smokers, 19% were former smokers, and 57% were non-smokers.[16] Given the advanced age of the population, the province of Bergamo has a high prevalence of chronic conditions (especially hypertension, diabetes, and hypercholesterolaemia).[16] When requiring medical care, most residents in the province of Bergamo were treated in hospitals close to home. The province ranked first among Italian provinces for having the lowest number of residents (1.85%) discharged from hospitals outside of the region.[17]

Indeed, the healthcare system in Lombardy is characterised by high standards and plentiful resources, with more than 200 accredited hospitals employing about 130 000 skilled healthcare workers.[18] In this region, the capacity of intensive care units before the pandemic was about 720 beds (typically operating at 85-90% occupancy during winter).[6,13]

By the end of February 2020, Nembro was one of the first Italian cities to report patients with covid-19 outside the original red zone around Lodi city. The first community isolation measures in Lombardy were implemented on 23 February, such as school closures, reduced commercial activity, and the cancellation of events and large gatherings. On 2 March, as a result of the emerging numbers of confirmed cases of covid-19, the Italian National Institute of Health recommended the creation of a red zone in the area, including Nembro.[19,20] These recommendations, however, were first implemented on 8 March; thereafter, no one could enter or leave the region and residents could not leave their homes, except for certain types of essential work or necessities such as groceries.[13,21] At the end of March, the mayor of Nembro publicly reported about 200 confirmed cases of covid-19 in the city.[20]

Data sources

Our study data integrated information from multiple sources. Firstly, we retrieved publicly available information from the Italian National Institute of Statistics (ISTAT), a public organisation that provides official statistics for Italian citizens and policy makers. The ISTAT data we used are freely available.[7] We extracted information on the number of Nembro residents at the beginning of each month from January 2012 to December 2019 and the number of residents who died from all causes each month from January 2012 to November 2019.

As a second source of information, we used data from Nembro's official registration office. We obtained special authorisation from the mayor to receive anonymised information from this registry on the number of residents who died from all causes between 1 January 2015 and 11 April 2020. Because local authorities are rapidly informed about deaths, both in and out of hospitals, we believe this registry to be an accurate and direct information source for December 2019 and early 2020. Using a report issued from this office, we further extracted the number of residents in Nembro as of 1 January 2020.[14]

The number of confirmed covid-19 deaths in the city was obtained from a public repository[22] provided by OnData,[23] an association that promotes transparency and open data. OnData reported extracting this information from the official Lombardy region covid-19 map dashboard.[22] From this source we obtained information from 21 February (the date of receipt of the first positive laboratory sample) until 11 April for those people living in Nembro who tested positive for covid-19. The dates from this source indicate when the biological sample was received by the laboratory, and the geographical reference listed is the city in which those people lived at the time they were tested (we did not anticipate this to differ meaningfully from the registered city of residence).[22] From the OnData repository, we obtained information on the vital status of people living in the city who tested positive for

the**bmj** Visual Abstract

Quantified consequences of covid-19

All cause mortality data from Nembro, Lombardy, Italy

Summary

A substantial increase in all cause mortality was observed during March 2020. Only around half of the deaths recorded from 21 February to 11 April were categorised as confirmed covid-19 deaths

Study design

Descriptive epidemiological study

Data sources

ISTAT (istat.it)

Nembro's registration office

Population

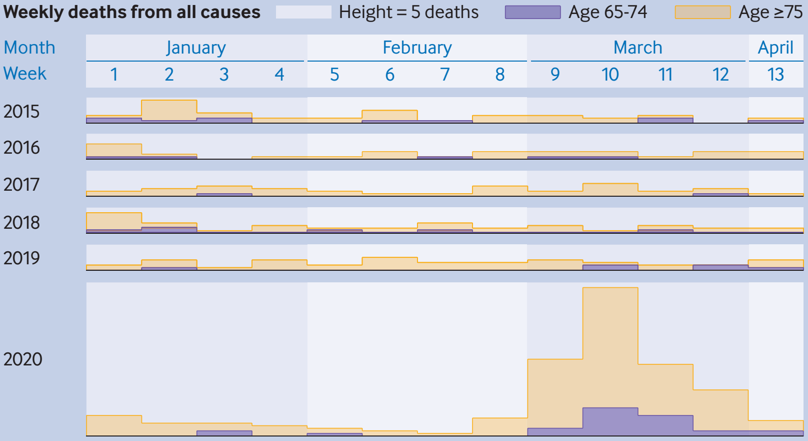
11 505 residents of Nembro

Outcomes

Monthly all cause mortality

Weekly all cause deaths in older people

Weekly deaths from all causes



https://bit.ly/BMJnembro

covid-19; however, we were unable to discern whether those who tested positive actually died from the disease of interest or from another cause.

We also used data from the recently published ISTAT mortality dataset for covid-19 emergency, available for selected municipalities.[24] From these data we extracted information about the weekly absolute number of all cause deaths by age group and sex in Nembro from 1 January to 4 April, for each year from 2015 to 2020.

Statistical analysis

To estimate the amount of total person time Nembro residents spent at risk, we combined two approaches. Firstly, between 1 January 2012 and 1 January 2020, we interpolated values between the recorded population size at the beginning of each month, assuming the change between the two consecutive time points was constant within that interval. To do this we used a spline regression model with the population size on the first day of the month as the dependent variable and the time in days as the independent variable, which was transformed using linear splines with knots set to the first day of each month.

Since no data were available on the number of residents after 1 January 2020, for each day thereafter we estimated the number of residents using a weighted average of projections obtained from two models: the prolongation of the last segment of the spline regression and a third order polynomial linear regression fitted over the entire observed interval. To estimate the daily number of residents after 1 January 2020, we used a convex combination of the two projections with weight equal to the reciprocal of the square root of the elapsed days since 1 January 2020 to avoid unnatural jump discontinuity in the function and to account more for long term trends than for short term trends.

Using this strategy, we estimated the number of residents across the entire study period (1 January 2012 to 11 April 2020). We estimated the total person years spent at risk each month by summing the estimated number of residents each day of the month divided by 365.25. Our approach to compute person time between 1 January 2012 and 1 January 2020 is about equivalent to the established practise of estimating the person time for each month as the product between the month's length and the average of the number of residents at the beginning and end of the month.

As a sensitivity analysis, we also estimated the monthly mortality rates under the hypothetical scenario of a large decrease in contributed person years during the final months of the study period (that is, to reflect many deaths or emigration, or both as a result of the pandemic). We projected the number of residents after 1 January 2020 in an alternative way by prolonging the last segment of the spline regression until 20 February 2020. From that day onwards, we estimated the number of residents by subtracting one tenth of the square of the elapsed number of days since 20 February from the estimated population size at 20 February.

We computed monthly all cause mortality rates by dividing the number of deaths by the estimated number of person years, expressed per 1000 person years. The April 2020 mortality rate was computed using data only from 1 to 11 April. We additionally compared the number of deaths captured by both the ISTAT covid-19 emergency dataset and the city registration office of Nembro from 1 January 2020 to 4 April 2020. Analyses were conducted using R version 3.6.0 and RStudio version 1.1.456.

Patient and public involvement

No patients were directly involved in this study. After the study was conceived, additional data were obtained from the mayor of Nembro, and he is interested in the wider dissemination of these results.

Results

Between 1 January 2012 and 1 January 2020, the monthly number of residents in Nembro ranged from 11 498 to 11 712 (fig 1). Information on population size was available to 1 January 2020, at which time the number of residents was 11 505. Thereafter the number of residents projected using our approach reached 11 525 on 11 April 2020 (fig 1).

Between 1 January 2012 and 11 April 2020, a total of 1116 people in Nembro died of all causes. Of these deaths, 112 (10.0%) occurred in 2012, 112 (10.0%) in 2013, 95 (8.5%) in 2014, 119 (10.7%) in 2015, 126 (11.3%) in 2016, 109 (9.8%) in 2017, 128 (11.5%) in 2018, 121 (10.8%) in 2019, and 194 (17.4%) in the first months of 2020 (until 11 April). Of the 194 deaths in the first months of 2020, 151 occurred in March alone. Between 21 February and 11 April 2020, a total of 166 deaths were recorded among the residents.

Of the biological samples received by the regional laboratory between 21 February and 11 April, 218 people later tested positive for covid-19. Of these positive tests, 85 were documented to belong to people who had died (last updated 16 April). Overall, the data source contained 64 135 confirmed cases of covid-19 in the Lombardy region; only 2% did not have recorded information on location.

Monthly all cause mortality between January 2012 and February 2020 fluctuated around 10 per 1000 person years (range 1.0 to 21.5 per 1000 person years) (fig 2). In March 2020, monthly all cause mortality reached a peak of 154.4 per 1000 person years—the corresponding rate for the same month in 2019 was 14.3 per 1000 person years. In April 2020, based on data from the first 11 days, all cause mortality decreased to 23.0 per 1000 person years (fig 2).

Results from the sensitivity analysis accounting for a potential sudden decrease in the population size showed monthly all cause mortality rates of 15.3, 155.7, and 23.5 per 1000 person years in February, March, and April, respectively.

The number of deaths in 2020 began to rapidly increase during the week of 23 February, peaked during the week of 8 March, and subsequently declined until 4 April. Of the 161 people who died

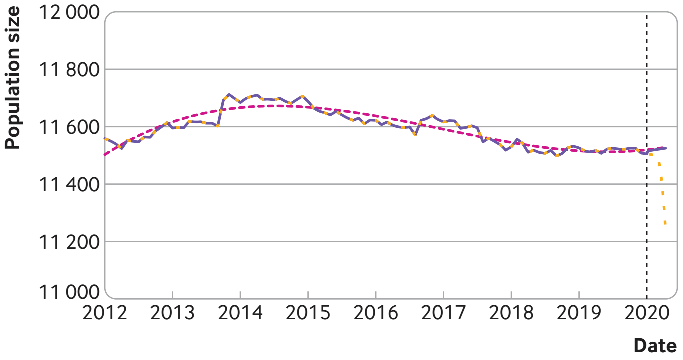


Fig 1 | Population size of Nembro from 1 January 2012 to 11 April 2020. Actual numbers of residents were recorded for the first day of each month between 1 January 2012 and 1 January 2020. Beyond this date (black vertical dashed line), a convex combination of two projections was used to estimate the number of residents to 11 April 2020. The two projections were obtained from the last segment of the spline regression (purple line until 1 January 2020) and a third order polynomial linear regression fitted on the whole observed time period (dotted pink line). The purple line represents the population size used to estimate person years in the main analysis, while the yellow dashed line represents the population size under the possible scenario of large numbers of deaths or emigration after 20 February 2020 (used in the sensitivity analysis)

during this period, none were aged 14 years or younger and 14 (8.7%) were aged between 15 and 64 years. The deviation in weekly all cause deaths compared with previous years was largely driven by the increase in deaths among older people (≥65 years) and men (table 1). Among those aged 75 years and older, 47 deaths were observed during the week of 8 March alone, 33 of which were in men.

No differences in weekly death counts were observed for 1 January to 4 April 2020 between the ISTAT covid-19 emergency dataset and the one used to compute mortality rates in our analysis.

Discussion

This study found a steep increase in all cause mortality in Nembro in early 2020 compared with the rather stable mortality rate observed over the past eight years in this city. More Nembro residents died in March 2020 than in the entire previous year or in any single year since 2012, with the all cause mortality rate in that month almost 11 times that observed in March 2019. After accounting for a potential sudden decrease in population size in a sensitivity analysis, this deviation was even more pronounced. The increase in mortality was mostly driven by an increase in deaths of older people (≥65 years), especially men. Since the population of Nembro had been relatively stable across recent years, we conclude that this rapid increase in deaths is attributable to the covid-19 pandemic. Only about half of the deaths observed since the pandemic onset (21 February to 11 April 2020), however, were categorised as confirmed covid-19 deaths.

Strengths and limitations of this study

The information used in our study was obtained from various sources. We acknowledge that some of the data might be provisional or not fully updated. Given the state of emergency and rapid development of covid-19, however, this limitation was unavoidable, and we did not observe any meaningful mismatches between the data sources. The source for confirmed covid-19 deaths is not official and does not include date of death, but rather the date the laboratory received the biological sample. This means the number of confirmed covid-19 deaths reported in our study is likely to be slightly higher than the official number in the same period because we included deaths of those who might have died after the 11 of April (although their sample was sent to the laboratory earlier). Furthermore, we know these deceased individuals tested positive for covid-19, but we cannot be absolutely certain whether the disease was a contributing cause of death. This means the difference between covid-19 specific deaths and the increase in all cause deaths could be even more extreme. The OnData repository we used in our study was the only data source available at the municipality level.

We did not estimate age-sex specific all cause mortality because information on the age-sex structure for Nembro was only available yearly, and the last update was on the 1 January 2019. Therefore, we preferred to avoid unreliable projections of person time at risk that would be based on strong, likely unreasonable assumptions.

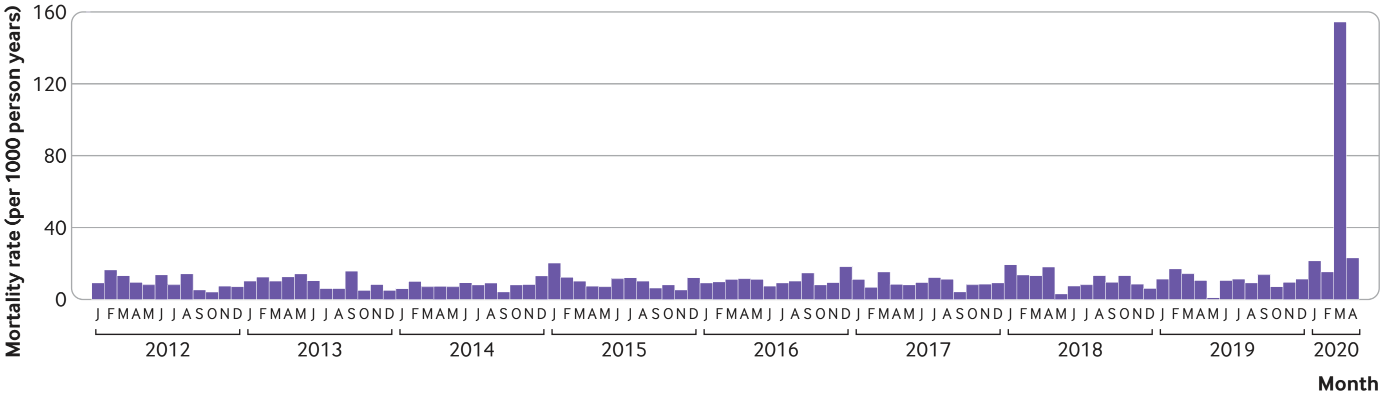


Fig 2 | Monthly all cause mortality per 1000 person years in Nembro between January 2012 and April 2020 (data only available to 11 April). Initials represent the months

Table 1 | Number of weekly all cause deaths among Nembro residents aged 65 and older between 1 January and 4 April during six years (2015-20) by sex and age group according to the ISTAT covid-19 emergency dataset

Week	65-74 years						≥75 years					
	2015	2016	2017	2018	2019	2020	2015	2016	2017	2018	2019	2020
Women												
01/01-11/01	0	0	0	0	0	0	1	1	1	3	1	6
12/01-18/01	1	0	0	1	0	0	5	1	2	2	3	4
19/01-25/01	1	0	0	0	0	1	2	0	3	0	1	2
26/01-01/02	0	0	0	0	0	0	1	0	2	3	2	2
02/02-08/02	0	0	0	1	0	0	0	0	2	0	0	1
09/02-15/02	1	0	0	0	0	0	3	2	1	1	4	2
16/02-22/02	0	1	0	1	0	0	0	0	0	3	1	1
23/02-29/02	0	0	0	0	0	0	2	1	3	1	3	2
01/03-07/03	0	1	0	0	0	2	2	1	1	3	3	16
08/03-14/03	0	1	0	0	1	2	2	1	4	1	1	14
15/03-21/03	0	0	0	1	0	0	1	0	1	0	1	10
22/03-28/03	0	0	1	0	1	0	0	2	2	2	0	12
29/03-04/04	0	0	0	0	1	0	0	2	1	1	2	3
Men												
01/01-11/01	2	1	0	1	0	0	0	4	1	4	1	2
12/01-18/01	0	1	0	1	1	0	3	0	1	0	0	1
19/01-25/01	1	0	1	0	0	1	0	0	0	1	0	1
26/01-01/02	0	0	0	0	0	0	1	1	1	0	2	2
02/02-08/02	0	0	0	0	0	1	2	1	0	1	2	1
09/02-15/02	0	0	0	0	0	0	1	1	0	1	1	0
16/02-22/02	1	0	0	0	0	0	0	0	1	0	2	0
23/02-29/02	0	0	0	0	0	0	1	2	1	1	0	5
01/03-07/03	0	0	0	0	0	1	1	1	1	0	1	11
08/03-14/03	0	0	0	0	1	9	0	1	1	0	0	33
15/03-21/03	2	0	0	0	0	8	0	1	1	2	1	10
22/03-28/03	0	0	0	0	1	2	0	1	0	0	0	4
29/03-04/04	1	0	0	0	0	2	1	1	0	1	1	1
Total	10	5	2	6	6	29	29	25	31	31	33	146

ISTAT=Italian National Institute of Statistics.[24]

Comparison with other studies

Our findings corroborate results from a large study by the Istituto di studi e ricerche Carlo Cattaneo.[25] In that investigation encompassing more than 1000 Italian cities selected because of an increase in mortality compared with previous years, the authors compared the overall number of deaths between 21 February 2020 and 21 March 2020 with the number of deaths in the same period averaged over the previous five years.[25] The study concluded that even under the best case scenario, in which all other Italian municipalities (about 7000) showed no deviation from the average mortality measured in previous years, the number of deaths attributable to covid-19 in Italy would still be twice as high as the number of confirmed deaths from covid-19 reported by the Italian authorities.[25] This study sheds light on the scale of the problem—that many deaths are erroneously not being attributed to covid-19 and that many of those who die outside of a hospital and have the disease are not being tested.[25] The authors also noticed large increases in mortality in regions not considered to be key Italian SARS-CoV-2 hot spots.[25] Another report, issued by the Italian Ministry of Health and the Italian National Center for Prevention and Control of Disease, about the daily surveillance mortality project (SiSMG) involving 18 major Italian cities, found similar results.[8] According to this report, the two included Lombardy cities (Milan and Brescia) showed a large increase in the number of all cause deaths in the period from the beginning of the outbreak to 18 March 2020 compared with the average number of deaths in the same period across the previous five years.

Conclusions and implications

Across Italian cities, all cause mortality has notably increased because of the covid-19 pandemic, but this increase is not being completely captured by officially reported statistics on confirmed covid-19 deaths. We believe several factors might have contributed to the discrepancy between the burden described by the confirmed death counts for covid-19 and that described by the increase in all cause mortality.

Firstly, covid-19 related deaths are generally counted as such if people test positive for the disease. Given the higher case fatality of covid-19 among older people with comorbidities, as well as the shortage of healthcare resources, many who actually died from covid-19 were likely never tested; therefore, the cause of death in these people was misclassified. For example, a shortage of tests prevented the assessment of covid-19 in people with symptoms and confirmed contacts in Nembro.[26]

A second explanation for the mismatch between these two death counts could lie in the group who did not have covid-19 but experienced other serious medical conditions and died from causes indirectly related to covid-19. During this period, this group might have experienced restricted access to healthcare owing to shortages in capacity, limited human resources for such a large patient influx (10% of patients with confirmed covid-19 in Italy worked in healthcare[27]), or fear of seeking hospital care during the pandemic. In Nembro, signs of the burden on the healthcare system and logistical challenges were noticeable.[26] A recent article describes the challenges and difficulties of the provincial healthcare system to provide even basic healthcare services.[28]

Thirdly, the known delay between administering and processing the test and the availability of results, especially in overwhelmed settings, might have exacerbated this difference.

Our results describe the impact of the covid-19 pandemic on the health of a small community. On a larger scale, the consequence of an uncontrolled SARS-CoV-2 outbreak in Italy would be the collapse of the healthcare system,[13] which, in turn, would have a substantial negative impact on the health of the entire population. We emphasise that measures of lethality are hardly interpretable solely as characteristics of the disease but also depend on the continuous availability and quality of care. The consequences of a pandemic are not only limited to covid-19 related deaths but rather contribute in an indirect way to the potentially avoidable deaths due to extreme triage of limited resources in crisis situations.[29]

Despite being weakened by a substantial reduction in public funding during the past decade,[6] the Italian healthcare system's overall performance still ranks high in international comparisons.[30] However, in the face of the unprecedented challenge from covid-19, policy makers are only equipped with the ability to introduce social distancing measures to slow down the spread of the virus and protect vulnerable groups and simultaneously strengthen the healthcare system to ensure high quality care for all patients.[31 32] Our results showed a decrease in all cause mortality in early April 2020, possibly attributable to reduced spread of the virus and reduced case fatality. Potential explanations for reduced spread of the virus include the implemented stringent community isolation measures, the promotion of preventive behaviours, as well as a growing number of immune people. For the reduced case fatality rate, possible contributing factors include a smaller pool of vulnerable people as well as boosted healthcare capacities from reallocation and optimisation of resources.

Our findings imply that the reporting of confirmed covid-19 specific deaths represents, at least for some Italian regions, a substantial underestimation of the actual number of deaths from the disease. As a consequence, we believe data on all cause mortality should be considered along with traditionally reported measures as an important metric to evaluate and compare the consequences of the covid-19 pandemic within and between settings. Although all cause mortality can only be interpreted as an approximation of the health status of the population under study, it is more often systematically collected under high quality standards, relies on universally accepted classification, and is not influenced by testing strategies or shortages of tests. [25] Furthermore, this metric captures indirect deaths, such as those related to a healthcare system under crisis, yielding a more complete picture of the pandemic's effects on population health. As we have outlined, this metric has several advantages and overcomes major drawbacks of other statistics for quantifying the impact of the covid-19 pandemic.

We thank the mayor of Nembro, Claudio Cancelli, for his indispensable collaboration.

Contributors: MP and TK conceived the study. LF and CL acquired and managed the data. MP analysed the data and created the figures. MP, JLR, and TK interpreted the results, and LF and CL provided important local insights. MP, JLR, and TK drafted the first version of the manuscript. TK supervised the work. MP is the guarantor and accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish. All coauthors provided critical comments and approved the final version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: No specific funding.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare the following competing interests: no support from any organisation for the submitted work; JLR is funded by a grant from the Else-Kröner-Fresenius Foundation (www.ekfs.de, GSO/EKFS-17; TK was the grant recipient). She further reports receiving a research grant from the Soyka Award for Pain Research 2019 sponsored by the German Migraine and Headache Society (DMKG) and Novartis Pharma for work outside this project. LF and CL are employed by Centro Medico Santagostino, a polyclinic that offers private healthcare services in Lombardy. Centro Medico Santagostino is planning to conduct scientific research on the impact of serological testing on covid-19 detection in the coming months, independent of this project. TK reports outside the submitted work to have contributed to an advisory board of CoLucid and a research project funded by Amgen, for which the Charité - Universitätsmedizin Berlin received unrestricted compensation. He further reports having received honorariums from Lilly, Newsenselab, and Total for providing methodological advice, from Novartis and Daiichi Sankyo for providing a lecture on neuroepidemiology and research methods, and from *The BMJ* for editorial services.

Ethical approval: Not required. Most data sources used were fully publicly available. Data obtained with permission from the Nembro city registration office were fully anonymised.

Data sharing: Most of the data used in this study are freely available from the sources cited. The most recent data obtained from the Nembro city registration office is available upon reasonable request from Caterina Lurani at caterina.lurani@cmsantagostino.it. The analysis script is available from the corresponding author at marco.piccininni@charite.de.

Dissemination to participants and related patient and public communities: Upon request, the data will be available for policy makers and government bodies. We are sharing the results with the mayor of Nembro, for further dissemination within the city.

MP (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

1 Callaway E.Time to use the p-word? Coronavirus enters dangerous new phase. *Nature* 2020; published online 25 February.

2 Paterlini M. On the front lines of coronavirus: the Italian response to covid-19. *BMJ* 2020;368:m1065. doi:10.1136/bmj.m1065

3 Livingston E, Bucher K. Coronavirus Disease 2019 (COVID-19) in Italy. *JAMA* 2020.

4 COVID-19 Map. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html (accessed 2 May 2020).

5 Rosenbaum L. Facing Covid-19 in Italy - Ethics, Logistics, and Therapeutics on the Epidemic's Front Line. *N Engl J Med* 2020; published online 18 March. doi:10.1056/NEJMp2005492.

6 Armocida B, Formenti B, Ussai S, Palestra F, Missoni E. The Italian health system and the COVID-19 challenge. *Lancet Public Health* 2020;5:e253.

7 ISTAT. Demo-Geodemo. - Maps, Population, Demography of ISTAT - Italian Institute of Statistics. http://demo.istat.it/index_e.html (accessed 2 May 2020).

8 Davoli M, dè Donato F, De Sario M, et al. Andamento della Mortalità Giornaliera (SiSMG) nelle città italiane in relazione all'epidemia di Covid-19. Ministero della Salute, Centro Nazionale Prevenzione e Controllo Malattie http://www.epiprev.it/sites/default/files/SISMG_COVID19_28032020-2.pdf

9 ANSA. Army takes Bergamo coffins away - English. ANSA.it. 2020. https://www.ansa.it/english/news/general_news/2020/03/19/army-takes-bergamo-coffins-away_a6e09e12-ba62-4dc9-834b-c227c6bd1fa3.html (accessed 2 May 2020).

10 Bucciarelli F, Horowitz J. 'We Take the Dead From Morning Till Night'. The New York Times. 2020. https://www.nytimes.com/interactive/2020/03/27/world/europe/coronavirus-italy-bergamo.html (accessed 2 May 2020).

11 Coronavirus, the real death toll: 4.500 victims in one month in the province of Bergamo. https://www.ecodibergamo.it/stories/bergamo-citta/coronavirus-the-real-death-tool-4500-victims-in-one-month-in-the-province-of_1347414_11/ (accessed 2 May 2020).

12 Guan W-J, Liang W-H, Zhao Y, et al, China Medical Treatment Expert Group for Covid-19. Comorbidity and its impact on 1590 patients with Covid-19 in China: A Nationwide Analysis. *Eur Respir J* 2020;2000547.

13 Grasselli G, Pesenti A, Cecconi M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. *JAMA* 2020.

14 Cornetti R, Dordi L. Dati statistici e demografici. Comune di Nembro http://comune.nembro.bg.it/export/sites/default/.galleries/documenti/allegati-eventi-news/2020/Comune-di-Nembro-anno-2019-dati-statistici.pdf

15 Sampietro G, Zucchi A. INCIDENZA E MORTALITÁ ONCOLOGICA IN PROVINCIA DI BERGAMO. ATS Bergamo http://www.ats-bg.it/upload/asl_bergamo/gestionedocumentale/testointero_3_784_13161.pdf

16 Piano Integrato Locale di promozione della Salute 2019 - ATS Bergamo - Agenzia di Tutela della Salute della provincia di Bergamo. http://www.ats-bg.it/servizi/gestionedocumentale/ricerca_fase03.aspx?ID=27872 (accessed 2 May 2020).

17 Qualità della vita del Sole 24 Ore / Da Isernia a Bergamo, chi emigra di più (e di meno) per curarsi. Il Sole 24 ORE. 2017. https://www.ilsole24ore.com/art/qualita-vita-sole-24-ore-isernia-bergamo-chi-emigra-piu-e-meno-curarsi-AEZacVKD (accessed 2 May 2020).

18 Lombardy Healthcare System. https://www.en.regione.lombardia.it/wps/portal/site/en-regione-lombardia/health/lombardy-healthcare-system (accessed 2 May 2020).

19 Coronavirus, zone rosse in Lombardia. Scontro Conte-Calderoli-Fontana. Sky TG24. https://tg24.sky.it/politica/2020/04/06/coronavirus-zona-rossa-alzano-nembro-conte-calderoli.html (accessed 2 May 2020).

20 Coronavirus, sindaco Nembro: 'Governo e Regione non hanno seguito Iss'. Sky TG24. https://tg24.sky.it/milano/2020/03/27/coronavirus-bergamo-sindaco-nembro.html (accessed 2 May 2020).

21 Gazzetta ufficiale. https://www.gazzettaufficiale.it/eli/id/2020/03/08/20A01522/sg (accessed 2 May 2020).

22 Borruso A. *covid19italia*. Github https://github.com/ondata/covid19italia (accessed 2 May 2020).

23 OnData. Associazione per la promozione di trasparenza e open data. https://ondata.it/ (accessed 2 May 2020).

24 Istat during the Covid-19 emergency. 2020. https://www.istat.it/en/archivio/240106 (accessed 2 May 2020).

25 Colombo AD, Impicciatore R. Gli effetti della pandemia da Covid-19 sulla mortalità. https://www.cattaneo.org/2020/04/01/gli-effetti-del-covid-19-sulla-mortalita/

26 Relazione al Consiglio Comunale EMERGENZA covid-19 . http://nembro.net/export/sites/default/.galleries/documenti/allegati-eventi-news/2020/Relazione-al-Consiglio-Comunale-EMERGENZA-covid-19.pdf

27 Daily infographic, May 1st. Istituto Superiore di Sanità. https://www.epicentro.iss.it/en/coronavirus/bollettino/Infografica_1maggio%20ENG.pdf

28 Nacoti M, Ciocca A, Giupponi A, et al. At the epicenter of the Covid-19 pandemic and humanitarian crises in Italy: changing perspectives on preparation and mitigation. *NEJM Catalyst Innovations in Care Delivery* 2020;1.https://catalyst.nejm.org/doi/abs/10.1056/CAT.20.0080

29 Fisher D, Heymann D. Q&A: The novel coronavirus outbreak causing COVID-19. *BMC Med* 2020;18:57. doi:10.1186/s12916-020-01533-w

30 Monasta L, Abbafati C, Logroscino G, et al, GBD 2017 Italy Collaborators. Italy's health performance, 1990-2017: findings from the Global Burden of Disease Study 2017. *Lancet Public Health* 2019;4:e645-57. doi:10.1016/S2468-2667(19)30189-6

31 Saglietto A, D'Ascenzo F, Zoccai GB, De Ferrari GM. COVID-19 in Europe: the Italian lesson. *Lancet* 2020;395:1110-1.

32 Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? *Lancet* 2020;395:1225-8. doi:10.1016/S0140-6736(20)30627-9.

13.2 Publication II: Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal

Piccininni M, Rohmann JL, Huscher D, Mielke N, Ebert N, Logroscino G, Schäffner E, Kurth T. Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal. *PLoS One*. 2020;15: e0231097.

DOI: https://doi.org/10.1371/journal.pone.0231097

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
19	FRACTALS-COMPLEX GEOMETRY PATTERNS AND SCALING IN NATURE AND SOCIETY	1,429	2.971	0.001120
20	Journal of Radiation Research and Applied Sciences	860	2.963	0.001860
21	MIT Technology Review	929	2.893	0.001910
22	JOURNAL OF KING SAUD UNIVERSITY SCIENCE	1,120	2.835	0.001670
23	PROCEEDINGS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES	18,683	2.818	0.018940
24	PLoS One	650,727	2.776	1.706770
25	COMPLEXITY	2,753	2.591	0.003890
26	Royal Society Open Science	4,118	2.515	0.017150
27	PeerJ	11,911	2.353	0.045900
28	SCIENCE AND ENGINEERING ETHICS	1,719	2.275	0.003450
29	INTERNATIONAL JOURNAL OF BIFURCATION AND CHAOS	7,008	2.145	0.007390
30	Symmetry-Basel	2,097	2.143	0.002590
31	SCIENTIFIC AMERICAN	6,609	1.946	0.003540
32	Science of Nature	508	1.839	0.002000
33	PROCEEDINGS OF THE JAPAN ACADEMY SERIES B-PHYSICAL AND BIOLOGICAL SCIENCES	1,532	1.833	0.001960
34	Journal of Taibah University for Science	779	1.640	0.001240
35	Frontiers in Life Science	241	1.622	0.000500
36	ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING	3,838	1.518	0.005840
37	SCIENCE PROGRESS	521	1.500	0.000400

Selected JCR Year: 2018; Selected Categories: “MULTIDISCIPLINARY SCIENCES”

PLOS ONE

RESEARCH ARTICLE

Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal

Marco Piccininni[1,2,3‡], Jessica L. Rohmann[1‡*], Dörte Huscher[1,4], Nina Mielke[1], Natalie Ebert[1], Giancarlo Logroscino[2,3], Elke Schäffner[1☯], Tobias Kurth[1☯]

1 Institute of Public Health, Charité –Universitätsmedizin Berlin, Berlin, Germany, **2** Department of Basic Medical Sciences, Neuroscience and Sense Organs, University of Bari Aldo Moro, Bari, Italy, **3** Center for Neurodegenerative Diseases and the Aging Brain, Department of Clinical Research in Neurology, University of Bari Aldo Moro, Pia Fondazione Cardinale G Panico, Tricase, Italy, **4** Institute of Biometry and Clinical Epidemiology, Charité –Universitätsmedizin Berlin, Berlin, Germany

☯ These authors contributed equally to this work.
‡ MP and JLR are joint first authors of this work.
* jessica.rohmann@charite.de

Abstract

Background

European guidelines recommend the use of the Systematic COronary Risk Evaluation (SCORE) to assess 10-year risk of fatal cardiovascular events in people aged 40 to 65. The SCORE Older Persons (SCORE OP, 5-year and 10-year versions) was recently developed for people aged 65 or older. We assessed the performance of these risk scores in predicting fatal cardiovascular events in older persons in Berlin.

Methods and findings

Data from the Berlin Initiative Study (BIS), a prospective, population-based study of older persons recruited from a German public health insurance company database were used. 1,657 participants aged 70 or older without reported previous myocardial infarction were included. We assessed calibration by comparing predicted risks to observed (for 5-year versions, 5y) or projected (for 10-year versions) probabilities. During follow-up (median: 4.8 years), 118 cardiovascular deaths occurred. The calibration assessment of the SCORE OP-H 5y and SCORE OP-L 5y equations revealed 2.1- and 1.5-fold overestimation. Comparing 10-year versions, the SCORE OP showed better discrimination ability compared to the SCORE (C-indices of around 0.80 compared to 0.72) and the SCORE for high-risk regions showed the best calibration (chi-square = 29.68). The SCORE OP overestimated the true risk; 519 and 677 events were predicted using the low-risk and high-risk region SCORE OP equations compared to 397 to 399 events projected based on BIS follow-up data (predicted/actual ratios of 1.3 and 1.7).

OPEN ACCESS

Citation: Piccininni M, Rohmann JL, Huscher D, Mielke N, Ebert N, Logroscino G, et al. (2020) Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal. PLoS ONE 15(4): e0231097. https://doi.org/10.1371/journal.pone.0231097

Editor: Bart Ferket, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: October 27, 2019

Accepted: March 16, 2020

Published: April 9, 2020

Data Availability Statement: R code for the equations externally validated in this paper as well as a sample calibration assessment with simulated data are included in the Supporting Information. All proposals requesting data access from the Berlin Initiative Study will need to specify how it is planned to use the data and will need to be approved by the Berlin Initiative Study principal investigator (ES). Further details about the study and contact information can be found on the Berlin Initiative Study website: bis.charite.de/en/. We are unable to place data in a public repository, as such sharing of individual participant data was not included in the informed consent of the study.

Funding: The Berlin Initiative Study was supported by the Kuratorium für Dialyse und Nierentransplantation (KfH) Foundation of Preventive Medicine (https://www.kfh-stiftung-praeventivmedizin.de/content/startseite), the Dr. Werner Jackstädt Foundation (https://www.jackstaedt-stiftung.de/) as well as the DDnÄ - Institut für Disease Management e.V. (https://www.dnev.de) (all granted to ES). JLR's research position is funded by a grant from the Else-Kröner-Fresenius Foundation (www.ekfs.de, GSO/EKFS-17, granted to TK). We further acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of the Charité - Universitätsmedizin Berlin. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: We have read the journal's policy and the authors of this manuscript have the following competing interests: We declare no support or financial relationships with any organisation for the submitted work. Outside of the submitted work: JLR, MP and NM have nothing further to disclose. DH received travel reimbursement from Actelion Switzerland and Böhringer-Ingelheim Germany for analysis project meetings. NE has received honoraria from Siemens Healthcare for presenting data on kidney function. GL serves as associate editor for the Neuroepidemiology journal, Karger. He received funding from Italian Ministry of Health, Italian Ministry of University and Research and Regione Puglia Special program for research. He further received speaking honoraria from Lilly, GE healthcare and Lundbeck. ES has received honoraria from Siemens Healthcare and Fresenius Medical Care for giving a lecture on kidney function assessment as well as from Fresenius Kabi for providing lectures and chairing a workshop on critiquing the literature. TK reports having contributed to an advisory board of CoLucid and a research project funded by Amgen, for which the Charité – Universitätsmedizin Berlin received an unrestricted compensation. He further reports having received honoraria from Lilly, Newsenselab, and Total for providing methodological advice, from Novartis and from Daiichi Sankyo for providing a lecture on neuroepidemiology and research methods, and from the BMJ for editorial services. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Conclusions

Given the low transportability of the SCORE OP observed in our population, we caution against its use in routine clinical practice until further information is available to avoid possible overtreatment among older persons in Berlin.

Introduction

Cardiovascular (CV) diseases are a leading cause of morbidity and are responsible for approximately 40% of deaths in the European Union[1] and 30% of deaths worldwide[2]. As most major CV risk factors are modifiable and can be targeted in preventive strategies[3–6], application of risk stratification tools remain of increasing importance and is suggested in the majority of CV disease guidelines[7,8].

More than 363 prediction models for CV disease have been developed over the last decades [9]. Generally, most known are the risk scores developed in the Framingham Heart Study[10–12]. The Systematic COronary Risk Evaluation (SCORE) was developed in 2003 to predict the 10-year risk of CV disease mortality in European populations using data from twelve European cohort studies[13]. The original SCORE is recommended for individuals aged 40 to 65 by the European guidelines on CV disease prevention from the European Society of Cardiology[14], and has been re-calibrated in various European countries to account for differences in mortality rates and risk factor distributions[15–18].

Since the majority of risk scores have been developed using data from primarily middle-aged populations, it remains unknown how well they perform in older populations[19,20]. With steadily increasing life expectancy, the need for valid CV risk assessment tools for older individuals is becoming more pressing[21,22]. Indeed, fewer than 20% of all fatal CV events occur between ages 40 and 65 in Europe[23]. Furthermore, since relationships between individual risk factors and CV events are known to change with age[19], it seems unlikely that fatal CV event risk scores based on coefficients estimated from a mostly middle-aged population provide reliable estimates of the actual probabilities in older persons[20,22].

Based on this rationale, an updated version of SCORE was developed in 2015 to predict the risk of fatal CV events specifically for persons aged ≥65 (SCORE Older Persons, SCORE OP) [22].

Until recently, the SCORE OP had not been externally validated; the newly published study calls for further validations of this risk tool[24]. Our objective was to assess calibration and discrimination of SCORE OP in a prospective cohort study of individuals aged ≥70 in Berlin. Our secondary aim was to compare the 10-year predictive performance of SCORE OP and SCORE in this cohort.

Methods

Study population

We used data from the ongoing Berlin Initiative Study, a longitudinal, population-based cohort study of adults aged ≥70 with biennial follow-up visits. Details of the study design have been previously described[25]. In brief, starting in November 2009, participants were selected using age and sex-stratified random sampling from a database of one of the largest German health insurance companies in the Berlin region (Allgemeine Ortskrankenkasse (AOK) Nordost) that covers about 50% of older persons in the Berlin region. Oversampling was conducted to increase participation among women and the highest age strata[26]. At the end of the recruitment in June 2011, a total of 2,069 individuals (52.6% female, mean age 80.4 years) completed baseline assessment at one of 13 clinical centers across Berlin. The response rate of the random sample was 8.1%, and the distribution of common comorbidities, including myocardial infarction and stroke, was found to be representative of the Allgemeine Ortskrankenkasse Nordost source population[26]. An administrative censorship date of September 30th, 2015 was used to ensure endpoint information completeness. In accordance with the eligibility criteria used to select participants in the development of the original SCORE risk scores, all participants with follow-up information who self-reported no previous history of myocardial infarction at baseline were included. Written informed consent was obtained from all participants prior to recruitment. The Berlin Initiative Study was approved by the ethics committee of Charité –Universitätsmedizin Berlin, Germany (EA2/009/08).

Risk scores

The following risk scores were selected for external validation:

- SCORE OP-H 5y: European risk score for 5-year fatal CV events among older persons in high-risk regions[22]
- SCORE OP-L 5y: European risk score for 5-year fatal CV events among older persons in low-risk regions[22]
- SCORE OP-H: European risk score for 10-year fatal CV events among older persons in high-risk regions[22]
- SCORE OP-L: European risk score for 10-year fatal CV events among older persons in low-risk regions[22]
- SCORE-H: European risk score for 10-year fatal CV events among adults in high-risk regions[13]
- SCORE-L: European risk score for 10-year fatal CV events among adults in low-risk regions [13]

Since Germany is considered to be between a high- and low-risk region[27], we validated both versions (for definitions, see [14]).

The published formulas, parameters, and coefficients were used for the computation of the risk scores. Of the two original parallel SCORE estimation models, we used the simpler version (with total cholesterol), since no superiority was demonstrated by the model including total cholesterol to the high-density lipoprotein cholesterol ratio.[13] Additionally, we used the corrected version of the original SCORE formula, which was characterized by a mathematical inconsistency.[28] In the 2003 publication, the overall risk for fatal CV events (***R***) was estimated as the sum of two risks; the risk for coronary heart disease death (***RCHD***) and risk for non-coronary CV disease death (***Rnon-CHD CVD***) as follows:

$$\boldsymbol{R = R_{CHD} + R_{non-CHD\ CVD}}$$

Defining the risk of CV mortality in this way is not suitable for participants with high-risk profiles since the sum of the two specific risks can exceed one. For this reason, we used a simple correction to estimate the overall risk, assuming cause-specific risk independence:

$$\boldsymbol{R = 1 - (1 - R_{CHD}) \cdot (1 - R_{non-CHD\ CVD})}$$

This correction was initially proposed by Støvring *et al*. as the first step in their approach aiming to account for competing events in the SCORE model.[28]

Since we are validating these scores in an increased risk population due to advanced age, we have applied this approach to avoid implausible probabilities exceeding 100%.

Risk factors

The following variables used in the models were assessed at baseline during a face-to-face interview: age as an integer, sex, the mean value of two consecutive blood pressure measurements, HDL and total cholesterol levels (converted from mg/dl to mmol/l dividing by 38.67), self-reported current smoking status, diabetes mellitus status (as determined by self-reported use of antidiabetic medication and/or measured HbA1c level >6.5%), and self-reported use of antihypertensive medication.

Endpoint definition

In the Berlin Initiative Study, information about participant deaths was obtained from several sources. When no reply to a follow-up invitation letter was received and the participant could not be reached, the general practitioner on file for the participant was contacted to inquire about vital status (and date of death, if deceased). Occasionally, participants' relatives contacted the study team directly about death events. If no information could be obtained from these sources, AOK Nordost records were used to determine vital status. Berlin Initiative Study staff additionally obtained records from the Berlin death certificates archive to validate all deaths, confirm dates, and obtain information on causes of death. Independent of the death certificate information, the study team attempted to obtain archived medical discharge letters for all in-hospital deaths as supplemental information to help correctly classify the cause of death. CV death was defined as death due to fatal myocardial infarction, fatal coronary heart disease, sudden cardiac death, death due to other cardiac diseases such as heart failure, fatal cerebrovascular disease (ischemic stroke, subarachnoid hemorrhage, intracerebral hemorrhage), and death due to peripheral occlusive arterial disease (complications of an aortic aneurysm, organ ischemia, ischemia). Ambiguous cause of death information were discussed and final coding decisions were made by two medical doctors without knowledge of individuals' risk factor profiles. In cases where mortality was confirmed but cause of death could not be ascertained due to insufficient information, we assumed death by non-CV cause in the primary analysis.

Statistical analyses

Participants' baseline risk factors are reported as frequencies and percentages or using means and standard deviations for categorical and continuous variables. Person-years were calculated from the date of recruitment to the individual's death date (as confirmed by a death certificate), or the date of the last available study visit in cases of loss to follow-up, or the date of administrative censorship (September 30, 2015).

Associations between the predicted risks from each of the risk estimation systems were assessed using pairwise Spearman's rank correlation coefficients. We then computed overall and CV-specific mortality rates as ratios of the number of observed events to the total amount of observed person-years with 95% Poisson exact confidence intervals.

For each prognostic model, and for both high- and low-risk region versions, we computed predicted risks for all individuals. In a second step, we grouped individuals according to deciles of the predicted risk and assessed calibration for each predicted risk decile group. The "actual" probability of an event was compared to the mean of the predicted risks in each decile group.

This comparison was conducted both graphically using a calibration plot and, to allow for comparability with previous studies, using the Nam-D'Agostino chi-square test[29]. We further compared the overall number of "actual" and predicted events across the decile groups to assess calibration-in-the-large for each risk score.

Given that the Berlin Initiative Study follow-up was shorter than 10 years, we performed calibration assessment using five years of the observed follow-up data for the SCORE OP risk equations (SCORE OP-H 5y and SCORE OP-L 5y) as the primary analysis. In this analysis, the calibration assessment was performed using the Kaplan-Meier estimates as "actual" probabilities. The original SCORE-L and SCORE-H risk equations were unfortunately only reported as 10-year versions. For calibration assessment of the SCORE risk equations, it was therefore necessary to project probabilities of the endpoints beyond the observed follow-up. Specifically, after creating the predicted risk decile groups, we ran Weibull regression survival models treating the predicted risk groups as a categorical covariate and the 10-year probabilities for each decile group projected by the Weibull models were considered "actual" probabilities. In this secondary analysis, assessment of 10-year calibration was also performed for the SCORE OP risk equations.

The discrimination ability of each risk score was assessed using the concordance index (C-index)[30] on the entire observed follow-up.

We computed projected probability estimates and the C-index along with their 95% bias-corrected and accelerated bootstrapped confidence intervals with 2,000 bootstrap replications.

To check for robustness of our primary results treating unknown causes of death as non-CV, we re-ran all analyses making the most extreme assumption that all deaths of unknown cause were of CV nature. Moreover, to confirm that our external validation study findings are applicable to the general older Berlin population as a whole, we performed an additional sensitivity analysis. All analytical steps described above were repeated in five resampled datasets (by age and sex strata of the BIS study population), which were generated to create a pseudopopulation exactly representative in terms of size and demographic structure of the 2010 Berlin population aged 70 or older (data from [31]; see S1 File).

To explore the possible impact on clinical decision making, we applied different "very-high-risk" thresholds, above which prevention and/or treatment intervention strategies would be indicated among older persons. Since no such single very-high-risk threshold has been established, we compared five different hypothetical thresholds (≥10%, ≥15%, ≥20%, ≥25%, and ≥30%) and calculated the number of very-high-risk persons based on 10-year predicted risks using the various risk scores. We also computed the percentage of participants classified as very-high-risk at the various thresholds based on a full Weibull model (with sex, systolic blood pressure, total cholesterol, HDL cholesterol, smoking status, diabetes and age as covariates and cardiovascular fatal events as the outcome) fitted on BIS data to give a near description of the projected reality at 10 years for our cohort.

All statistical analyses were performed using R v3.4.3 (https://www.R-project.org/) and RStudio v1.0.153 (https://www.rstudio.com/).

Results

Of the original 2,069 Berlin Initiative Study participants, 412 people who self-reported a history of myocardial infarction or had missing information on past myocardial infarction, on one or more risk factors included in the SCORE OP equations, or lacked any follow-up were excluded (Fig 1). Baseline characteristics of the 1,657 remaining included participants, as well as mean values for predicted risks estimated by SCORE OP-H 5y, SCORE OP-L 5y, SCORE-H, SCORE-L, SCORE OP-H, and SCORE OP-L are displayed in Table 1. During the observed

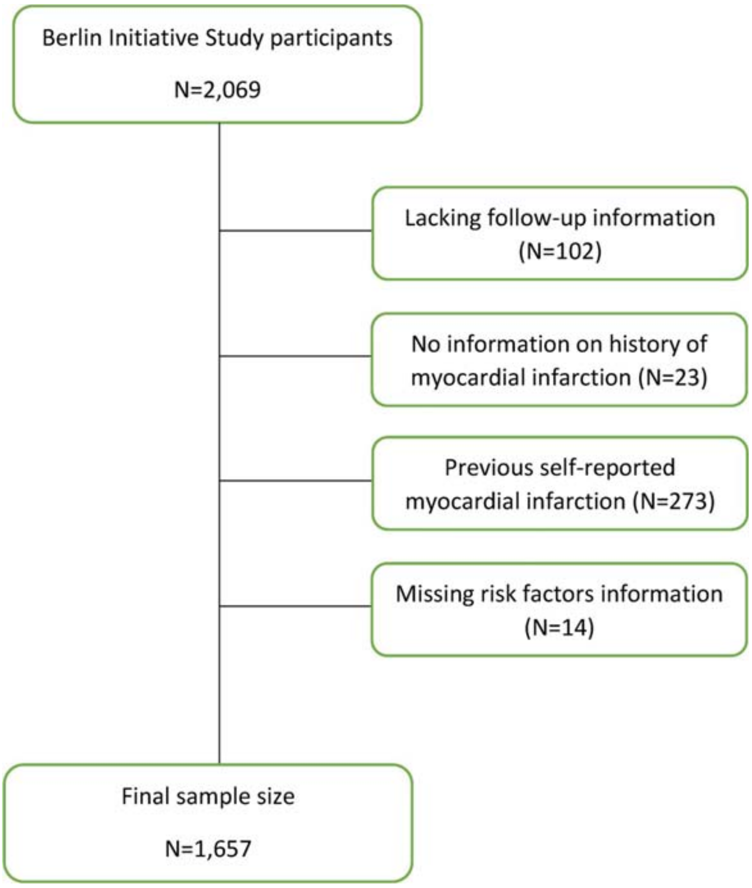

Berlin Initiative Study participants
N=2,069
Lacking follow-up information (N=102)
No information on history of myocardial infarction (N=23)
Previous self-reported myocardial infarction (N=273)
Missing risk factors information (N=14)
Final sample size
N=1,657

Fig 1. Flow chart showing berlin initiative study participant inclusion/exclusion criteria for this external validation study.

https://doi.org/10.1371/journal.pone.0231097.g001

follow-up period (median: 4.8 years), a total of 324 deaths were recorded, of which 118 (36.4%) were CV deaths. In total, participants contributed 7,370.3 person-years, and the overall mortality and CV-specific fatal event rates in the cohort were 44.0 (39.3 to 49.0) and 16.0 (13.3 to 19.2) per 1,000 person-years.

A correlation plot illustrates the distribution of the participants' predicted probabilities of the various prognostic models and correlations between the individual scores (S1 Fig in S1 File). As expected, both high-risk and low-risk region pairs from the same models were highly correlated (Spearman's rho 0.99–1.00), as well as the 5- and 10-year versions of the SCORE OP (rho = 0.99–1.00). Overall, we observed moderate correlation between all SCORE and SCORE OP equations (0.76–0.78).

The calibration of both the SCORE OP-H 5y and the SCORE OP-L 5y equations were assessed using observed probabilities (Fig 2). In total, 302 fatal CV events were predicted by the SCORE OP-H 5y while 142 fatal CV events were estimated without accounting for competing risks, showing an overestimation (predicted/actual ratio = 2.13). This score systematically overestimated the true risk (chi-square = 139.2, Fig 2A and 2B). The SCORE OP-L 5y also showed overestimation, but to a lesser extent, with 215 predicted compared to 142 observed events (ratio = 1.51, chi-square = 39.7, Fig 2C and 2D). Again, a systematic overestimation was observed across most of the risk score decile groups.

Table 1. Baseline characteristics of the study population[a].

	Total (n = 1,657)		Males (n = 734)		Females (n = 923)	
Mean age (SD), years	79.7	(6.7)	80.2	(6.7)	79.2	(6.6)
Current smoking, N (%)	86	(5.2%)	47	(6.4%)	39	(4.2%)
Diabetes, N (%)	414	(25.0%)	200	(27.2%)	214	(23.2%)
Hypertensive treatment, N (%)	1254	(75.9%)	551	(75.3%)	703	(76.3%)
Cholesterol (SD), mmol/l	5.6	(1.2)	5.1	(1.1)	5.9	(1.2)
HDL cholesterol (SD), mmol/l	1.5	(0.5)	1.3	(0.4)	1.7	(0.4)
Mean systolic blood pressure, mmHg (SD)	147.0	(22.8)	147.3	(22.8)	146.7	(22.7)
Mean diastolic blood pressure, mmHg (SD)	82.0	(14.5)	82.4	(14.6)	81.6	(14.4)
Risk scores[b]						
SCORE OP-H 5y risk (SD)	0.18	(0.17)	0.20	(0.16)	0.17	(0.17)
SCORE OP-L 5y risk (SD)	0.13	(0.13)	0.16	(0.14)	0.10	(0.12)
SCORE-H risk (SD)	0.22	(0.15)	0.27	(0.15)	0.19	(0.13)
SCORE-L risk (SD)	0.16	(0.11)	0.17	(0.11)	0.14	(0.11)
SCORE OP-H risk (SD)	0.41	(0.27)	0.45	(0.25)	0.37	(0.27)
SCORE OP-L risk (SD)	0.31	(0.24)	0.36	(0.23)	0.28	(0.24)

[a] All participants were enrolled between 2009–2011.

[b] SCORE OP[22] and SCORE[13] risk scores are described in detail in the Methods section. H and L indicate high- and low- cardiovascular risk regions. Unless otherwise specified, 10-year risk equations were used. 5y indicates 5-year risk equations.

https://doi.org/10.1371/journal.pone.0231097.t001

For the secondary analysis, after grouping individuals according to deciles of predicted risk, we ran Weibull regression survival models. Weibull regression assumptions were fulfilled in all models (see diagnostic plots; S2 Fig in S1 File). Actual and predicted SCORE-H probabilities are displayed in Fig 3A. The corresponding calibration plot is shown in Fig 3B. Globally, SCORE-H predicted 372 expected events over ten years, and 382 actual events were projected (ratio = 0.97, chi-square = 29.7). The discrimination ability of SCORE-H as measured by the C-index was 0.72 (0.67 to 0.76).

The corresponding probabilities for SCORE-L are illustrated in Fig 3C. In eight decile groups, predicted probabilities were underestimated (Fig 3D). In total, the number of actual events projected was 384 while only 258 events were expected based on the SCORE-L (predicted/actual ratio = 0.67, chi-square = 117.6). The C-index for discrimination was found to be 0.72 (0.67 to 0.77).

Compared to SCORE-H, the 10-year SCORE OP-H equation designed for older persons had a higher discrimination ability in our study population (C-index = 0.79, 0.75 to 0.83). However, the SCORE OP-H overestimated risk in each decile group (Fig 4A; chi-square = 327.9). This systematic overestimation is visible in the calibration plot (Fig 4B). The SCORE OP-H predicted 677 events, while only 399 actual events were projected, a considerable overestimation (ratio = 1.70). Similarly, the 10-year version of the SCORE OP-L, despite its good discrimination ability (C-index = 0.80, 0.75 to 0.83), overestimated the risk for fatal CV events in eight decile groups (see Fig 4C and 4D). As illustrated by the calibration plot, this overestimation by SCORE OP-L was to a lesser extent than for SCORE OP-H (chi-square = 76.3). In total, SCORE OP-L predicted 519 events compared with 397 projected actual events (ratio = 1.31).

We have provided a summary of all the results previously described in Table 2.

To determine the impact of potential misclassification due to 44 fatalities with unknown cause of death information, we performed all analyses again under a 'worst-case' scenario assuming that all 44 individuals died due to CV reasons. In this sensitivity analysis, the SCORE-H still underestimated risk (predicted/actual ratio = 0.88), and the SCORE OP-L

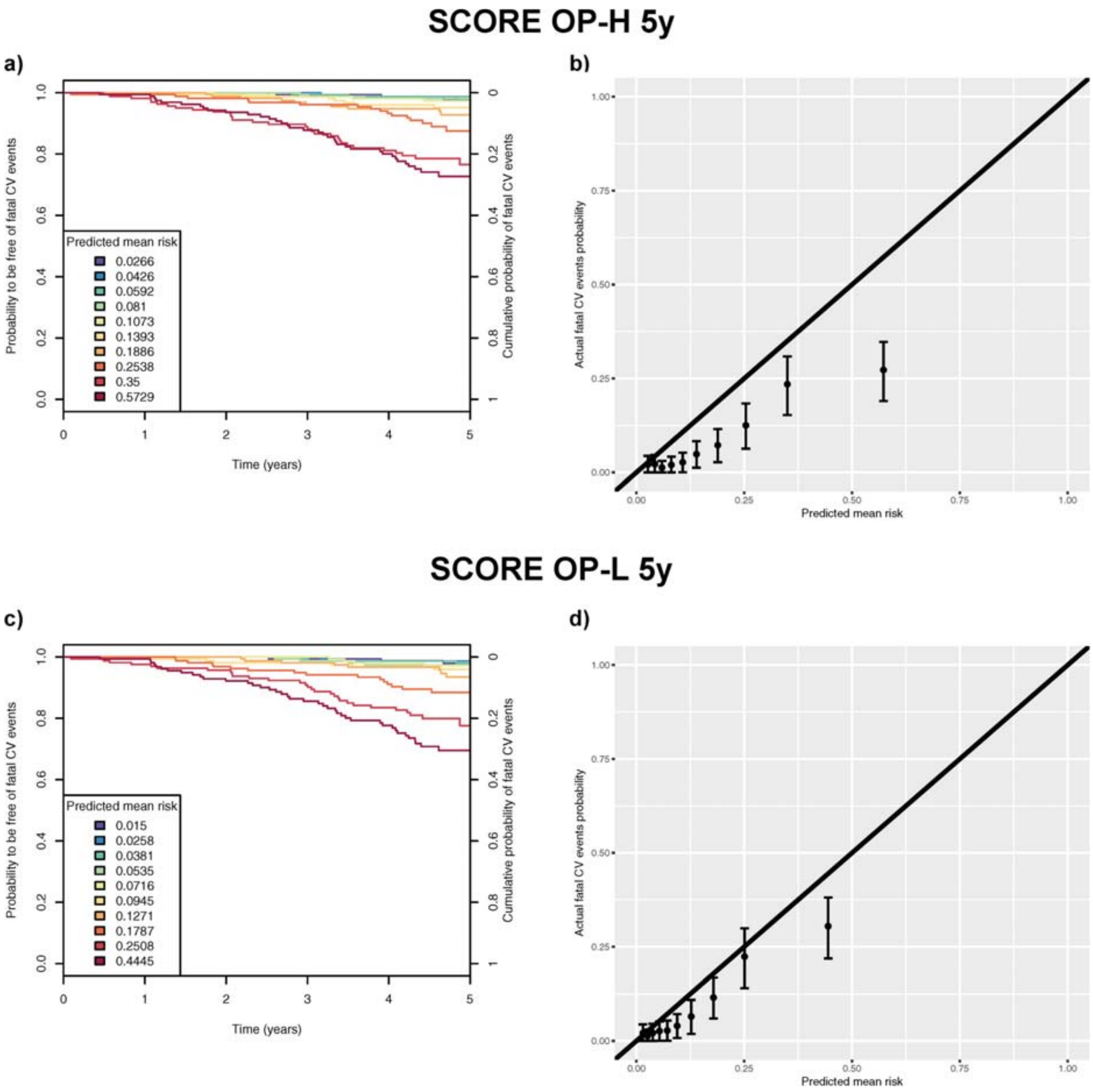


Fig 2. Panel a) shows observed Kaplan-Meier probabilities to be free of fatal cardiovascular (CV) events at a given time point grouped by deciles of risk as predicted by SCORE OP-H 5y (5-year risk equation for high-risk regions). The right y-axis scale shows the probability of the complementary event: occurrence of a fatal CV event before a given time point in the counterfactual scenario of no competing events (if we assume independent competing risks). The legend indicates the average predicted risk of having a fatal CV event within each decile group of risk. Panel b) shows the calibration plot for SCORE OP-H 5y comparing the predicted mean risk (corresponding to ones in the legend of Panel a)) to the actual fatal CV event probabilities within five years (corresponding to the intersection between the curves and the right Y-axis in Panel a)) for each decile group. We report 95% confidence intervals. Panels c) and d) show the results as described in Panels a) and b) for SCORE OP-L 5y (5-year risk equation for low-risk regions).

https://doi.org/10.1371/journal.pone.0231097.g002

overestimated risk (ratio = 1.22) (S1 Table in S1 File). Consistent results were obtained upon repeating all analyses in the pseudopopulation dataset with the same size and age-sex structure of the 2010 Berlin older population, obtained by resampling (see S1 File).

We found that using SCORE OP compared to SCORE led to more individuals classified as "very-high-risk" at hypothetical thresholds beyond 10% (Table 3). For example, using a 20%

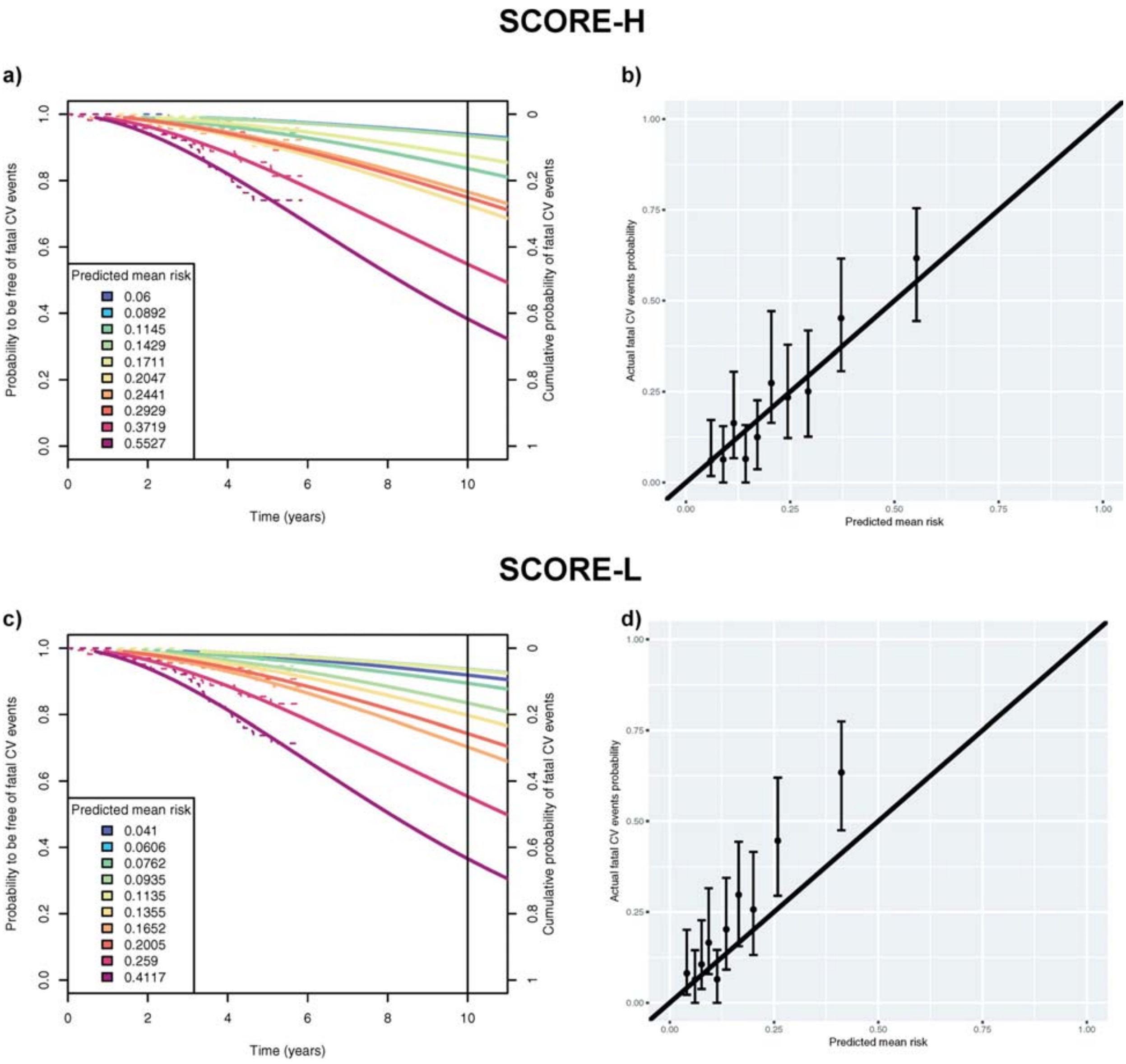


Fig 3. Panel a) shows both observed (Kaplan-Meier, dotted lines) and projected (Weibull regression model, solid lines) probabilities to be free of fatal cardiovascular (CV) events at a given time point grouped by deciles of risk as predicted by SCORE-H (for high-risk regions). The right y-axis scale shows the probability of the complementary event: occurrence of a fatal CV event before a given time point in the counterfactual scenario of no competing events (if we assume independent competing risks). The legend indicates the average predicted risk of having a fatal CV event within each decile group of risk. Panel b) shows the calibration plot for SCORE-H comparing the predicted mean risk (corresponding to ones in the legend of Panel a)) to the actual fatal CV event probabilities within ten years (corresponding to the intersection between the curves and the right Y-axis in Panel a)) for each decile group. We report 95% bias-corrected and accelerated bootstrapped confidence intervals. Panels c) and d) show the results as described in Panels a) and b) for SCORE-L (for low-risk regions).

https://doi.org/10.1371/journal.pone.0231097.g003

cut-off, which would be reasonable given the European Society of Cardiology's recommendation to select a threshold higher than 10% in older persons[14], results in the following percentage of participants being classified as "very-high-risk" per score: SCORE-H: 45.9% SCORE-L 25.0%, SCORE OP-H 71.2%, and SCORE OP-L 56.5%. According to the full Weibull model fitted on BIS data, 40.9% of BIS participants should be classified as "very-high-risk".

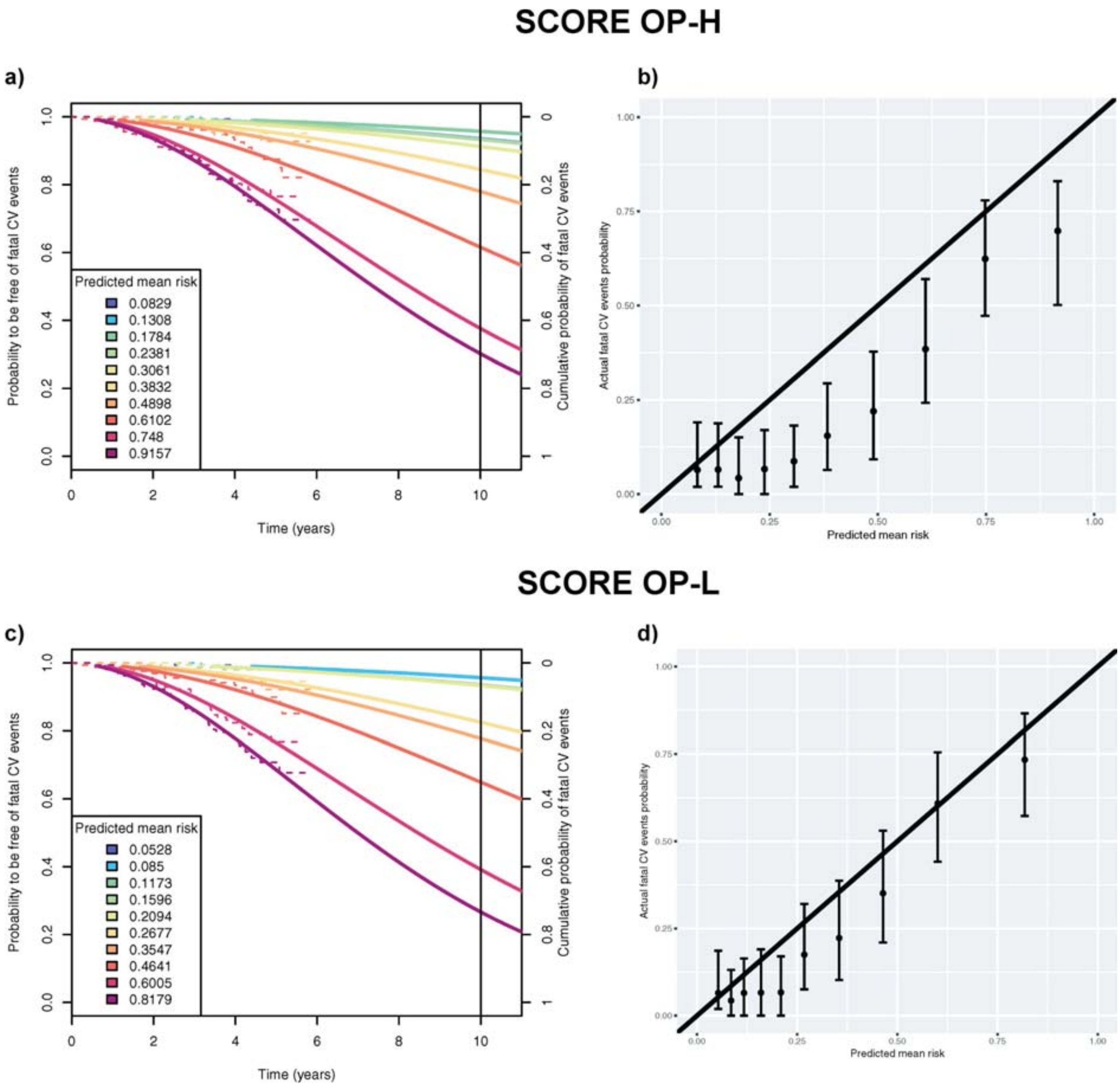


Fig 4. Panel a) shows both observed (Kaplan-Meier, dotted lines) and projected (Weibull regression model, solid lines) probabilities to be free of fatal cardiovascular (CV) events at a given time point grouped by deciles of risk as predicted by SCORE OP-H (for high-risk regions). The right y-axis scale shows the probability of the complementary event: occurrence of a fatal CV event before a given time point in the counterfactual scenario of no competing events (if we assume independent competing risks). The legend indicates the average predicted risk of having a fatal CV event within each decile group of risk. Panel b) shows the calibration plot for SCORE OP-H comparing the predicted mean risk (corresponding to ones in the legend of Panel a)) to the actual fatal CV event probabilities within ten years (corresponding to the intersection between the curves and the right Y-axis in Panel a)) for each decile group. We report 95% bias-corrected and accelerated bootstrapped confidence intervals. Panels c) and d) show the results as described in Panels a) and b) for SCORE OP-L (for low-risk regions).

https://doi.org/10.1371/journal.pone.0231097.g004

To further explore differences in performance of the two score systems among older persons, we compared fatal CV event risk predicted by the two risk scores for six hypothetical risk profiles based on risk factors. We created high-, medium-, and low-risk examples for females and males across the entire age spectrum from 60 to 100 years (see Fig 5). We found that SCORE OP predicted higher risks compared to SCORE in female individuals aged ≥75 and in male individuals aged ≥78.

Table 2. Risk scores: Measures of validity.

Risk score[a]	Predicted number of fatal cardiovascular events	Actual number of fatal cardiovascular events[b]	Predicted/ Actual ratio	Nam-D'Agostino chi-square (p-value)	C-index[c] (95% CI)
SCORE OP high-risk Regions					0.79 (0.75 to 0.83)
SCORE OP-H 5y	302	142	2.13	139.16 (p<0.001)	
SCORE OP-H	677	399	1.70	327.9 (p<0.001)	
SCORE OP low-risk Regions					0.80 (0.75 to 0.83)
SCORE OP-L 5y	215	142	1.51	39.68 (p<0.001)	
SCORE OP-L	519	397	1.31	76.29 (p<0.001)	
SCORE-H	372	382	0.97	29.68 (p = 0.001)	0.72 (0.67 to 0.76)
SCORE-L	258	384	0.67	117.63 (p<0.001)	0.72 (0.67 to 0.77)

[a] SCORE OP[22] and SCORE[13] systems have been previously described elsewhere. H and L indicate high- and low- cardiovascular risk regions. 5y indicates 5-year risk equations. All other scores listed are 10-year versions.

[b] Weibull regression model projections beyond the observed follow-up are reported for 10-year risk scores, leading to small differences in the number of actual events. 5-year risk scores use observed Berlin Initiative Study data only using the Kaplan-Meier estimator.

[c] Risk score discrimination capability was assessed using the entire observed follow-up data.

https://doi.org/10.1371/journal.pone.0231097.t002

Discussion

In this prospective, population-based study of Berlin older individuals, the SCORE OP 5y substantially overestimated the true risk of fatal CV events. In the 10-year comparison, interestingly, the SCORE equation recommended for persons aged from 40 to 65, showed better

Table 3. Percentage of berlin initiative study participants classified as very-high-risk[a] based on various hypothetical thresholds of cardiovascular mortality predicted risk.

Prognostic model[b]	Predicted 10-year risk threshold				
	≥10%	≥15%	≥20%	≥25%	≥30%
SCORE-H	80.0%	62.8%	45.9%	33.2%	23.5%
SCORE-L	60.9%	39.4%	25.0%	16.3%	10.4%
SCORE OP-H	91.4%	80.4%	71.2%	62.8%	55.8%
SCORE OP-L	79.9%	67.2%	56.5%	47.6%	40.7%
Full Weibull model fitted on BIS data	58.1%	48.2%	40.9%	33.7%	28.7%

[a] In this hypothetical example, only individuals with a predicted risk higher than the threshold are considered very-high-risk persons. This simplification is only intended to illustrate possible clinical implications of the use of prognostic tools.

[b] SCORE[13] and SCORE OP[22] risk scores have been previously described elsewhere. H and L indicate high- and low- cardiovascular risk regions. The full Weibull model fitted on BIS data includes sex, systolic blood pressure, total cholesterol, HDL cholesterol, smoking status, diabetes and age as covariates and cardiovascular fatal events as the outcome to give a near description of the projected reality at 10 years for our cohort.

https://doi.org/10.1371/journal.pone.0231097.t003

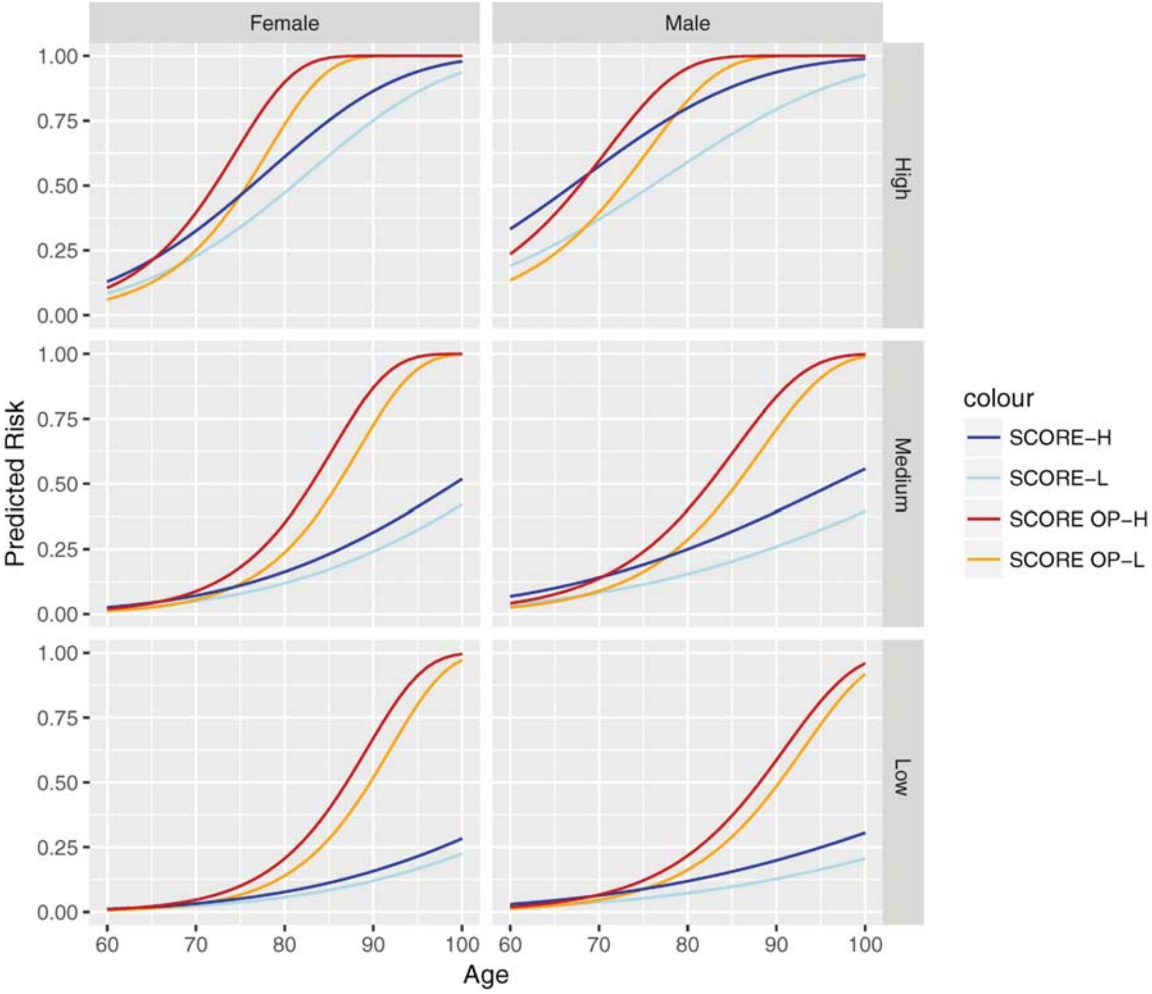


Fig 5. We compared SCORE-L, SCORE-H, SCORE OP-H and SCORE OP-L predicted 10-year risks for six risk profiles; high-, medium- and low-risk for females and males. The high-risk profile was constructed using data from a hypothetical diabetic, current smoker with a systolic blood pressure of 180 mmHg, total cholesterol level of 8 mmol/l, and HDL cholesterol level of 1 mmol/l. The medium-risk profile was created using mean values for all variables based on the baseline characteristics of Berlin Initiative Study participants (see Table 1). The low-risk profile was constructed using data from a hypothetical non-diabetic, current non-smoker with a systolic blood pressure of 120 mmHg, total cholesterol level of 4 mmol/l, and HDL cholesterol level of 2 mmol/l. The high- and low-risk reference values for systolic blood pressure and total cholesterol were taken from the minimum and maximum values reported in the SCORE chart.

https://doi.org/10.1371/journal.pone.0231097.g005

calibration than the SCORE OP, developed specifically for persons aged from 65 to 80[22]. The SCORE OP did, however, demonstrate slightly superior discrimination capabilities as assessed by the C-index, likely attributable to the higher number of included risk factors.

Because the prevalence of CV risk factors and their estimated effects on CV outcomes are known to change with increasing age[20,32], the SCORE OP was developed to correct for a suspected overestimation of fatal CV event risk as predicted by the original SCORE in persons aged ≥65[22]. The SCORE development dataset was comprised largely of middle-aged individuals (only three cohorts out of 12 included participants aged ≥70, and no cohort included individuals aged >80). Because of this, it was hypothesized that the estimated prediction model beta coefficients were inappropriately large in magnitude, likely overemphasizing the contribution of these risk factors to the predicted risk in older persons[20,22,32,33]. Since the levels of traditional CV risk factors are known to be elevated in older persons, the use of inappropriately large coefficients for this population was expected to result in substantial overestimation of the true fatal CV event risk using the SCORE[22].

The SCORE OP is based on re-estimated beta coefficients using data from people aged ≥65 also including additional parameters (diabetes and HDL) in an attempt to correct the expected overestimation. Based on this rationale, the use of SCORE OP is currently suggested as an alternative by the European Society of Cardiology and European Society of Hypertension guidelines for the management of arterial hypertension in older individuals (aged ≥65)[34].

The idea that the SCORE gives higher estimates of risk compared to the SCORE OP among older persons was reinforced by findings from a cross-sectional study in Spain[35,36] and the charts comparison presented in the SCORE OP development paper (p.8)[22]. However, these two comparisons were based solely on the 65–69 age group. No attempt was made to compare predictions across *all* ages 65 and above.

In our cohort comprised of older persons, we found that SCORE OP yielded higher risk estimates than SCORE (with mathematical correction). The same result was found in the only published external validation of the SCORE OP[24]. While the authors attributed this inconsistency with the previous cross-sectional study[35,36] to a difference in the methods used, we think it may have arisen because of a difference in age composition of the participants. Upon comparing the scores designed for low-risk regions, we confirm that in the age group 65–68, the SCORE-L yields higher risk estimates than the SCORE OP-L for medium- and high-risk individuals (see Fig 5). However, this is not true for any of the risk profiles in individuals aged ≥70; in these individuals, the risk estimated by SCORE OP-L consistently exceeds that of SCORE-L. We observed a similar behavior upon comparing the high-risk region risk scores (Fig 5).

Moreover, we found a lower transportability of the SCORE OP compared to the original SCORE. This is likely explained by a difference in the distribution of unmeasured risk factors and baseline CV fatal risk in our cohort compared to the SCORE OP development dataset. The SCORE OP development dataset was composed of 85% individuals from Norway and included no German cohorts, while the SCORE development dataset included a German cohort and a balanced distribution of individuals from several European countries. Differences in the definition of CV risk factors included in the models or in the true underlying hazard ratios may have led to the unsuitability of the SCORE OP coefficients. For example, in the SCORE OP development dataset, included cohort prevalences of diabetes mellitus ranged from 6% to 7% in females and 4% to 7% in males[22]. These prevalences are considerably lower than the diabetes mellitus prevalences observed in the Berlin Initiative Study (23% in females and 27% in males, which align with German prevalence figures among older persons [37]) and it is likely a difference in the definition of diabetes contributed to the observed low transportability of the SCORE OP.

We found the best performing risk score among older persons in Berlin was the SCORE-H developed for high-risk regions despite that Germany's fatal CV risk is considered to be between that of high- and low-risk regions, with a tendency towards the latter[27]. The fact that the SCORE-L seems more appropriate in middle-aged persons while the SCORE-H seems better suited among older persons suggests that in the Berlin population, the observed difference in risk prediction is likely explained by differences in baseline survivorship functions (determined by age, sex, unmeasured risk factors) of these age groups rather than by differences in the coefficients of classical risk factors such as cholesterol, blood pressure and smoking. In fact, SCORE developers used the same coefficients in both regional versions.

In general, the performance of the SCORE equations was surprisingly good, especially considering that this risk score was developed in 2003 using data from cohort studies with recruitment periods between 1967 and 1991 and that the incidence and treatment strategies for CV diseases changed substantially over the last decades[38].

Recently, the SCORE OP was subjected to external validation for the first time in a cohort of 6,590 older individuals aged between 65 and 79 living in Norfolk, UK[24]. Their results about the transportability of the SCORE OP and SCORE among older persons were divergent compared to ours. In this UK population, the SCORE OP showed "excellent calibration", performing better than the SCORE, despite showing low discrimination ability[24]. Interestingly, in this UK cohort, diabetes prevalence was extremely low, around 3%[24], likely because this information was only self-reported. This observation provides additional support for our aforementioned argument that the definition of diabetes plays a crucial role in the transportability of the SCORE OP. Regional differences and the age of the cohort, overall younger than our study participants, may also have contributed to the discrepancies.

Finally, the SCORE and SCORE-OP were developed without accounting for possible competing events, thus, our calibration used a consistent approach; neither the Weibull model projections nor the Kaplan-Meier estimates accounted for competing events. We acknowledge, however, that competing risks do pose a large problem for practical use among older persons, in whom competing fatal events are common. For this reason, methods have been suggested for the development of local, updated, recalibrated scores that can be used to inform regional risk prediction accounting for mortality due to other causes[28].

Study strengths and limitations

Strengths include the prospective design, population-based setting, and availability of comprehensive health-related information, providing unique insights into the health of the very old, a population often excluded from larger studies. Death information is considered to be complete and was obtained from the Berlin death certificates archive and supplemented with information from medical records. Specific cause of death information was consistently extracted when available. Furthermore, a total of 118 fatal CV endpoints were recorded during observed follow-up, exceeding the minimum amount needed to properly validate a 10-year prognostic model over the entire time span (at least 100)[39].

Some limitations should be considered when interpreting our findings. First, we compared predicted probabilities to projected ones since the Berlin Initiative Study follow-up data were available for less than 10 years. However, the Weibull diagnostic plots indicate fulfillment of the assumptions of all projection models, and our analyses predicting 5-year risk using observed data were consistent with the findings using 10-year projected probabilities and thus confirm our results. A minor drawback to our approach is that the number of "actual" events is not constant across calibration assessments because it depends on decile groupings of participants, which differed for each risk score. However, these "actual" event numbers did not differ substantially (range: 382 to 399).

Second, the reliability of cause of death information on death certificates is known to be error-prone, especially among older persons. However, we believe this potential misclassification is similar for most population-based settings and was also present in the risk score development studies[22].The impact of unknown or unavailable cause of death information as demonstrated by our "worst-case" sensitivity analysis was negligible.

Furthermore, all of the original SCORE endpoints are well-represented in our definition, with the exception of non-aortic aneurysms, which are very rare events.

Finally, the exclusion of subjects with previous history of myocardial infarction was based on self-reported information. However, the self-reported nature of this information in our study is unlikely to be problematic, since this exclusion criterion was introduced during the

SCORE development only to ensure overall CV health in the considered sample and not for the purpose of excluding participants not at risk for the outcome.

Nonetheless, we suggest against excluding people with a predefined CV event from future prognostic models developed to predict CV mortality risk, as this limits prediction score usability. This is a particularly important consideration among older persons, in whom prevalent non-fatal CV events are common.

Implications and suggestions for future research

This external validation study shows that the SCORE OP overestimates CV mortality risk among Berlin older persons, which ultimately leads to the classification of more individuals to higher levels of CV mortality risk at most hypothetical very-high-risk thresholds, above which intervention strategies would be indicated. We believe our results may have important implications since overestimation of risk in these individuals may lead to overtreatment in a potentially vulnerable population already prone to polypharmacy[40], which is known to lead to adverse drug events or interactions and increase health care costs[41]. Our results underscore the importance of external validation of prediction tools before clinical use. Mass medicalization may result from an overestimation due to lack of transportability of these tools or from setting too low of a risk threshold[42]. This potential danger is illustrated by the higher number of BIS participants who would be classified as very-high-risk for 10-year fatal CV risk according to the SCORE OP compared to the reality.

In our external validation study, the original SCORE developed for high-risk regions performed best in older persons living in Berlin. Our findings are very different from the ones obtained in the first SCORE OP external validation study conducted among older adults in the UK; however, both studies do not support the use of SCORE OP in clinical practice. Therefore, the challenge of finding a valid tool for profiling risk among older European individuals may not yet be solved.

Supporting information

S1 File.
(PDF)

Acknowledgments

We are indebted to the participants in the Berlin Initiative Study for their commitment and cooperation and to the entire Study staff for their expertise and assistance. We also thank the health insurance fund AOK Nordost-Die Gesundheitskasse for its continued cooperation and technical support of the Berlin Initiative Study. We are grateful to Dr. Patrick Larscheid, director of the Berlin death certificates archive, and his colleagues for their support.

Author Contributions

Conceptualization: Marco Piccininni, Jessica L. Rohmann, Giancarlo Logroscino, Tobias Kurth.

Data curation: Dörte Huscher, Nina Mielke, Natalie Ebert, Elke Schäffner.

Formal analysis: Marco Piccininni, Jessica L. Rohmann.

Funding acquisition: Elke Schäffner, Tobias Kurth.

Investigation: Marco Piccininni, Jessica L. Rohmann, Tobias Kurth.

Methodology: Marco Piccininni, Jessica L. Rohmann, Tobias Kurth.

Project administration: Marco Piccininni, Jessica L. Rohmann, Elke Schäffner, Tobias Kurth.

Resources: Elke Schäffner, Tobias Kurth.

Software: Marco Piccininni.

Supervision: Elke Schäffner, Tobias Kurth.

Validation: Jessica L. Rohmann.

Visualization: Marco Piccininni, Jessica L. Rohmann.

Writing – original draft: Marco Piccininni, Jessica L. Rohmann.

Writing – review & editing: Marco Piccininni, Jessica L. Rohmann, Dörte Huscher, Nina Mielke, Natalie Ebert, Giancarlo Logroscino, Elke Schäffner, Tobias Kurth.

References

1. Nichols M, Townsend N, Scarborough P, Rayner M. European cardiovascular disease statistics. European Heart Network; 2012 [cited 29 Oct 2018]. Available: http://www.ehnheart.org/component/attachments/attachments.html?task=download&folder=publications&id=1435
2. WHO. cardiovascular diseases (CVDs). 2016 [cited 29 Oct 2018]. Available: http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
3. Vartiainen E, Puska P, Pekkanen J, Tuomilehto J, Jousilahti P. Changes in risk factors explain changes in mortality from ischaemic heart disease in Finland. BMJ. 1994; 309: 23–27. https://doi.org/10.1136/bmj.309.6946.23 PMID: 8044063
4. WHO Europe. cardiovascular disease data and statistics. 1 Aug 2018 [cited 29 Oct 2018]. Available: http://www.euro.who.int/en/health-topics/noncommunicable-diseases/cardiovascular-diseases/data-and-stat istics
5. Vartiainen E, Sarti C, Tuomilehto J, Kuulasmaa K. Do changes in cardiovascular risk factors explain changes in mortality from stroke in Finland? BMJ. 1995; 310: 901–904. https://doi.org/10.1136/bmj.310.6984.901 PMID: 7719179
6. Menotti A, Puddu PE, Kromhout D, Kafatos A, Tolonen H. Coronary heart disease mortality trends during 50 years as explained by risk factor changes: The European cohorts of the Seven Countries Study. Eur J Prev Cardiol. 2019; 2047487318821250.
7. Stewart J, Manmathan G, Wilkinson P. Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. JRSM Cardiovasc Dis. 2017; 6: 2048004016687211. https://doi.org/10.1177/2048004016687211 PMID: 28286646
8. Collins DRJ, Tompson AC, Onakpoya IJ, Roberts N, Ward AM, Heneghan CJ. Global cardiovascular risk assessment in the primary prevention of cardiovascular disease in adults: systematic review of systematic reviews. BMJ Open. 2017; 7: e013650. https://doi.org/10.1136/bmjopen-2016-013650 PMID: 28341688
9. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016; 353: i2416. https://doi.org/10.1136/bmj.i2416 PMID: 27184143
10. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol. 1976; 38: 46–51. https://doi.org/10.1016/0002-9149(76)90061-8 PMID: 132862
11. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008; 117: 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579 PMID: 18212285
12. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97: 1837–1847. https://doi.org/10.1161/01.cir.97.18.1837 PMID: 9603539
13. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003; 24: 987–1003. https://doi.org/10.1016/s0195-668x(03)00114-3 PMID: 12788299
14. Authors/Task Force Members:, Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of

the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Atherosclerosis. 2016; 252: 207–274. https://doi.org/10.1016/j.atherosclerosis.2016.05.037 PMID: 27664503

15. Panagiotakos DB, Fitzgerald AP, Pitsavos C, Pipilis A, Graham I, Stefanadis C. Statistical modelling of 10-year fatal cardiovascular disease risk in Greece: the HellenicSCORE (a calibration of the ESC SCORE project). Hellenic J Cardiol. 2007; 48: 55–63. PMID: 17489342

16. Marques-Vidal P, Rodondi N, Bochud M, Pécoud A, Hayoz D, Paccaud F, et al. Predictive accuracy and usefulness of calibration of the ESC SCORE in Switzerland. Eur J Cardiovasc Prev Rehabil. 2008; 15: 402–408. https://doi.org/10.1097/HJR.0b013e3282fb040f PMID: 18677163

17. van Dis I, Kromhout D, Geleijnse JM, Boer JMA, Verschuren WMM. Evaluation of cardiovascular risk predicted by different SCORE equations: the Netherlands as an example. Eur J Cardiovasc Prev Rehabil. 2010; 17: 244–249. https://doi.org/10.1097/HJR.0b013e328337cca2 PMID: 20195155

18. De Bacquer D, De Backer G. Predictive ability of the SCORE Belgium risk chart for cardiovascular mortality. Int J Cardiol. 2010; 143: 385–390. https://doi.org/10.1016/j.ijcard.2009.03.101 PMID: 19386372

19. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, et al. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. PLoS One. 2012; 7: e34287. https://doi.org/10.1371/journal.pone.0034287 PMID: 22470551

20. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol. 2009; 54: 1209–1227. https://doi.org/10.1016/j.jacc.2009.07.020 PMID: 19778661

21. Koller MT, Steyerberg EW, Wolbers M, Stijnen T, Bucher HC, Hunink MGM, et al. Validity of the Framingham point scores in the elderly: results from the Rotterdam study. Am Heart J. 2007; 154: 87–93. https://doi.org/10.1016/j.ahj.2007.03.022 PMID: 17584559

22. Cooney MT, Selmer R, Lindman A, Tverdal A, Menotti A, Thomsen T, et al. Cardiovascular risk estimation in older persons: SCORE OP. Eur J Prev Cardiol. 2016; 23: 1093–1103. https://doi.org/10.1177/2047487315588390 PMID: 26040999

23. Mortensen MB, Afzal S, Nordestgaard BG, Falk E. The high-density lipoprotein-adjusted SCORE model worsens SCORE-based risk classification in a contemporary population of 30 824 Europeans: the Copenhagen General Population Study. Eur Heart J. 2015; 36: 2446–2453. https://doi.org/10.1093/eurheartj/ehv251 PMID: 26082084

24. Verweij L, Peters RJG, Scholte Op Reimer WJM, Boekholdt SM, Luben RM, Wareham NJ, et al. Validation of the Systematic COronary Risk Evaluation—Older Persons (SCORE-OP) in the EPIC-Norfolk prospective population study. Int J Cardiol. 2019; 293: 226–230. https://doi.org/10.1016/j.ijcard.2019.07.020 PMID: 31324398

25. Schaeffner ES, van der Giet M, Gaedeke J, Tölle M, Ebert N, Kuhlmann MK, et al. The Berlin initiative study: the methodology of exploring kidney function in the elderly by combining a longitudinal and cross-sectional approach. Eur J Epidemiol. 2010; 25: 203–210. https://doi.org/10.1007/s10654-010-9424-x PMID: 20094758

26. Ebert N, Jakob O, Gaedeke J, van der Giet M, Kuhlmann MK, Martus P, et al. Prevalence of reduced kidney function and albuminuria in older adults: the Berlin Initiative Study. Nephrol Dial Transplant. 2017; 32: 997–1005. https://doi.org/10.1093/ndt/gfw079 PMID: 27190381

27. Rücker V, Keil U, Fitzgerald AP, Malzahn U, Prugger C, Ertl G, et al. Predicting 10-Year Risk of Fatal Cardiovascular Disease in Germany: An Update Based on the SCORE-Deutschland Risk Charts. PLoS One. 2016; 11: e0162188. https://doi.org/10.1371/journal.pone.0162188 PMID: 27612145

28. Støvring H, Harmsen CG, Wisløff T, Jarbøl DE, Nexøe J, Nielsen JB, et al. A competing risk approach for the European Heart SCORE model based on cause-specific and all-cause mortality. Eur J Prev Cardiol. 2013; 20: 827–836. https://doi.org/10.1177/2047487312445425 PMID: 22498473

29. D'Agostino RB, Nam B-H. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures. Handbook of Statistics. Elsevier; 2003. pp. 1–25.

30. Therneau T, Atkinson E. Concordance. The Comprehensive R Archive Network; 2019 Mar. Available: https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf

31. Berlin Brandenburg Amt Für Statistik. Statistisches Informationssystem Berlin Brandenburg (StatIS-BBB). [cited 19 Mar 2019]. Available: www.statistik-berlin-brandenburg.de

32. Cooney MT, Dudina A, D'agostino R, Graham IM. Cardiovascular risk-estimation systems in primary prevention: do they differ? Do they make a difference? Can we see the future? Circulation. 2010. Available: https://www.ahajournals.org/doi/abs/10.1161/circulationaha.109.852756

33. Catapano AL, Graham I, De Backer G, Wiklund O, Chapman MJ, Drexel H, et al. 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias: The Task Force for the Management of Dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Atherosclerosis. 2016; 253: 281–344. https://doi.org/10.1016/j.atherosclerosis.2016.08.018 PMID: 27594540

34. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J. 2018; 39: 3021–3104. https://doi.org/10.1093/eurheartj/ehy339 PMID: 30165516

35. Brotons C, Moral I, Fernández D, Cuixart L, Soteras A, Puig M. Assessment of the New SCORE OP Cardiovascular Risk Charts in Patients Older Than 65 Years. Rev Esp Cardiol. 2016; 69: 981–983. https://doi.org/10.1016/j.rec.2016.04.049 PMID: 27474480

36. Brotons C, Moral I, Fernández D, Cuixart L, Muñoz A, Soteras A, et al. [Clinical consequences of using the new cardiovascular risk tables SCORE OP in patients aged over 65 years]. Med Clin. 2016; 147: 381–386.

37. Tamayo T, Brinks R, Hoyer A, Kuß OS, Rathmann W. The Prevalence and Incidence of Diabetes in Germany. Dtsch Arztebl Int. 2016; 113: 177–182. https://doi.org/10.3238/arztebl.2016.0177 PMID: 27118665

38. Mortensen MB, Falk E. Limitations of the SCORE-guided European guidelines on cardiovascular disease prevention. Eur Heart J. 2017; 38: 2259–2263. https://doi.org/10.1093/eurheartj/ehw568 PMID: 27941016

39. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016; 35: 214–226. https://doi.org/10.1002/sim.6787 PMID: 26553135

40. O Riordan D, Aubert CE, Walsh KA, Van Dorland A, Rodondi N, Du Puy RS, et al. Prevalence of potentially inappropriate prescribing in a subpopulation of older European clinical trial participants: a cross-sectional study. BMJ Open. 2018; 8: e019003. https://doi.org/10.1136/bmjopen-2017-019003 PMID: 29567842

41. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. Expert Opin Drug Saf. 2014; 13: 57–65. https://doi.org/10.1517/14740338.2013.827660 PMID: 24073682

42. Lancet T. Statins for millions more? Elsevier; 2014. https://doi.org/10.1016/s0140-6736(14)60240-3

13.3 Publication III: Directed acyclic graphs and causal thinking in clinical risk prediction modeling

Piccininni M, Konigorski S, Rohmann JL, Kurth T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC Med Res Methodol*. 2020;20: 179.

DOI: https://doi.org/10.1186/s12874-020-01058-z

22	Journal of Managed Care & Specialty Pharmacy	1,221	3.024	0.004750
23	BMC Palliative Care	1,522	2.922	0.003880
24	HEALTH EXPECTATIONS	3,199	2.847	0.007740
25	MEDICAL DECISION MAKING	5,281	2.793	0.009000
26	ADVANCES IN HEALTH SCIENCES EDUCATION	2,697	2.761	0.005400
27	SUPPORTIVE CARE IN CANCER	11,975	2.754	0.024130
28	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	4,765	2.731	0.006720
29	HEALTH POLICY AND PLANNING	5,401	2.717	0.010110
30	HEALTH SERVICES RESEARCH	8,061	2.706	0.013670
30	MEDICAL TEACHER	7,977	2.706	0.010530
32	Patient-Patient Centered Outcomes Research	1,008	2.673	0.003090
33	Applied Health Economics and Health Policy	1,126	2.664	0.003350
34	MEDICAL CARE RESEARCH AND REVIEW	2,431	2.577	0.004060
35	BMC Medical Research Methodology	9,832	2.509	0.021050
36	International Journal of Integrated Care	1,137	2.489	0.002010
37	QUALITY OF LIFE RESEARCH	13,192	2.488	0.019050
38	JOURNAL OF PALLIATIVE MEDICINE	5,938	2.477	0.010540
39	JOURNAL OF RURAL HEALTH	1,729	2.471	0.002630
40	EUROPEAN JOURNAL OF CANCER CARE	3,149	2.421	0.005380
41	JOURNAL OF MEDICAL SYSTEMS	4,680	2.415	0.006220
42	STATISTICAL METHODS IN MEDICAL RESEARCH	4,156	2.388	0.012230
43	Health and Quality of Life Outcomes	8,070	2.318	0.012120
44	Health Informatics Journal	691	2.297	0.001450
45	Risk Management and Healthcare Policy	416	2.283	0.001270

Selected JCR Year: 2018; Selected Categories: “HEALTH CARE SCIENCES and SERVICES”

Piccininni *et al. BMC Medical Research Methodology* (2020) 20:179
https://doi.org/10.1186/s12874-020-01058-z

BMC Medical Research Methodology

TECHNICAL ADVANCE Open Access

Directed acyclic graphs and causal thinking in clinical risk prediction modeling

Marco Piccininni[1*], Stefan Konigorski[2,3], Jessica L. Rohmann[1] and Tobias Kurth[1]

Abstract

Background: In epidemiology, causal inference and prediction modeling methodologies have been historically distinct. Directed Acyclic Graphs (DAGs) are used to model a priori causal assumptions and inform variable selection strategies for causal questions. Although tools originally designed for prediction are finding applications in causal inference, the counterpart has remained largely unexplored. The aim of this theoretical and simulation-based study is to assess the potential benefit of using DAGs in clinical risk prediction modeling.

Methods: We explore how incorporating knowledge about the underlying causal structure can provide insights about the transportability of diagnostic clinical risk prediction models to different settings. We further probe whether causal knowledge can be used to improve predictor selection in clinical risk prediction models.

Results: A single-predictor model in the causal direction is likely to have better transportability than one in the anticausal direction in some scenarios. We empirically show that the Markov Blanket, the set of variables including the parents, children, and parents of the children of the outcome node in a DAG, is the optimal set of predictors for that outcome.

Conclusions: Our findings provide a theoretical basis for the intuition that a diagnostic clinical risk prediction model including causes as predictors is likely to be more transportable. Furthermore, using DAGs to identify Markov Blanket variables may be a useful, efficient strategy to select predictors in clinical risk prediction models if strong knowledge of the underlying causal structure exists or can be learned.

Keywords: Causality, Clinical risk prediction, Prediction models, Markov blanket, Directed acyclic graph, Transportability, Predictor selection

Background

In modern epidemiology, prediction modeling and causal inference are generally considered separate branches with unique sets of methods and aims. However, recently, the emerging field of "causal learning" or "causal discovery" has led to the introduction of prediction modelling and machine learning techniques as tools to generate causal structures based on data-driven procedures [1]. Despite some specific implementations [2], movement in the other direction has been less explored; namely, the application of causal inference principles and graph theory in clinical risk prediction modeling strategies.

Diagrams and graphs are intuitive, visual tools used to inform analytic methods to answer causal questions [3]. The increasing use of causal graphs and the need for automated procedures to assess causal effects given the combination of previous structural knowledge and new data led to the development of a compact, formal theory free of parametric assumptions to transparently model causal relationships [3]. Directed Acyclic Graphs (DAGs) are used to rigorously map all a priori assumptions surrounding a causal question of interest [3] and to graphically describe the underlying data generating process. In

* Correspondence: marco.piccininni@charite.de
[1]Institute of Public Health, Charité - Universitätsmedizin Berlin, Berlin, Germany
Full list of author information is available at the end of the article

DAGs, each node represents a random variable, and directed causal paths are represented by arrows. The causal graph structure thus provides qualitative information about the conditional independencies of the variables of interest. DAGs are used as a tool in causal inference to illustrate potential sources of confounding and selection bias and ultimately identify suitable strategies to address them [3, 4]. We assume the reader is familiar with DAGs; for those not yet familiar, several accessible introductions have been published elsewhere [3, 5].

The aim of this work is to investigate the potential benefits of using DAGs and causal thinking in clinical risk prediction problems. Specifically, we describe the use of causal knowledge in assessing transportability and selecting predictors for a clinical risk prediction model.

Methods

Transportability and the principle of independent mechanisms

A causal concept that could be useful in clinical risk prediction modeling is the *principle of independent mechanisms* [1]. This fundamental assumption was formalized to justify the inference of causal structure from observed data [1, 6] and was later suggested as a useful hypothesis to drive machine learning-based prediction approaches [7].

This principle of independent mechanisms states that the "causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other" [1]. This means that a causal process can be interpreted as a chain of independent mechanisms, in which each causal mechanism takes the state output from the previous mechanism as input and "feeds" the next mechanism with its own state output. Each causal mechanism on the chain can be conceptualized as a physical mechanism invariant to the input it receives [1]. The idea of the autonomy of the mechanisms is actually more intuitive than it seems. In fact, it is how we justify all clinical interventions: we assume that artificially changing one mechanism or its input will not affect any of the other mechanisms [1].

Let's consider two variables with an unconfounded causal relationship. For simplicity, we will call these two variables "Cause" and "Effect". The joint probability distribution of these two variables $\mathbb{P}$(Cause,Effect) can be factorized in two ways [1, 7]:

$$\mathbb{P}(\text{Cause}, \text{Effect}) = \mathbb{P}(\text{Effect}|\text{Cause})\mathbb{P}(\text{Cause}) = \mathbb{P}(\text{Cause}|\text{Effect})\mathbb{P}(\text{Effect})$$

The principle of independent mechanisms states that the marginal distribution of the variable Cause, $\mathbb{P}$(Cause), and the conditional distribution of the variable Effect on the variable Cause, $\mathbb{P}$(Effect|Cause), contain no information about each other [1, 7]. Indeed, $\mathbb{P}$(Effect|Cause) is the distribution of the variable Effect for each given value of the variable Cause. It represents the physical mechanism that transforms the input (Cause) into an output (Effect), while $\mathbb{P}$(Cause) represents the state of the input. Under the principle of independent mechanisms, $\mathbb{P}$(Cause) and $\mathbb{P}$(Effect|Cause) change independently of each other across different joint distributions [1].

This independence constraint in the first factorization induces a dependency between the conditional distribution of Cause on Effect, $\mathbb{P}$(Cause|Effect), and the marginal distribution of the Effect, $\mathbb{P}$(Effect), shown in the second mathematical factorization in the anticausal direction [1, 7]. Therefore, $\mathbb{P}$(Effect) and $\mathbb{P}$(Cause|Effect) often change in a dependent way across different joint distributions [1]. Since this concept of independence involves mechanisms rather than variables, it cannot be simply defined, tested, or quantified like the concept of statistical independence in probability theory [1].

In this work, we present two hypothetical, simplified clinical examples from the field of neurodegenerative disease to illustrate the consequences of the *principle of independent mechanisms* in the context of diagnostic clinical risk prediction models. Specifically, we describe the transportability of two clinical risk prediction models for Alzheimer's disease diagnosis using different predictors. In the first example, the disease is the effect of the predictor (allele APOE ε4 status, which is a known cause of Alzheimer's disease), while in the second example, the disease is the cause of the predictor (concentration of tau protein in cerebrospinal fluid, which is described as an effect of the Alzheimer's disease pathological process).

Predictor selection and the Markov blanket

There is another causal concept that may be useful for the first and arguably most important step in building clinical risk prediction models: predictor selection. Here, we focus on the main challenge of selecting the smallest possible subset of all available variables that provide enough information to predict the outcome of interest with good validity in terms of calibration.

There are many well-known reasons to limit the number of predictors used to build a risk prediction model: (i) to reduce problems due to the high number of variables in the model, thereby increasing performance, (ii) to reduce the costs, time and effort associated with data collection and storage, model development or training, (iii) to enable easier use of the model in different settings, and (iv) to increase the interpretability of the mechanisms behind the generation of the probability estimates [8, 9]. The last reason is particularly important in the context of clinical risk prediction models. Indeed, medical doctors are reluctant to use prediction models without a certain degree of interpretability [10], since the output probabilities are used to support clinical decisions about treatments and prevention strategies.

Intuitively, the predictor selection problem can be interpreted as how to choose the smallest subset of variables excluding all variables that do not provide additional information on the outcome of interest.

By operationalizing the lack of additional information using the notion of conditional independence [11], the entire problem of predictor selection is analogous to identifying the so-called "Markov Blanket" of the outcome variable.

We define Y as the random variable for the outcome of interest and **X** as the set of all available candidate predictor variables of Y. We assume that **X** is a superset of the variables relevant to the causal processes in which Y is involved. The Markov Blanket of Y, MB(Y), is the minimal subset of **X**, conditioned on which, all other variables of **X** *not* included in MB(Y) are independent of Y [8, 9]:

$$\forall \mathrm{V} \in \mathbf{X}-\mathrm{MB}(\mathrm{Y}): \ \Pr(\mathrm{Y}|\mathrm{MB}(\mathrm{Y}), \mathrm{V}) = \Pr(\mathrm{Y}|\mathrm{MB}(\mathrm{Y})),$$

where **X** - MB(Y) denotes the set of variables which are contained in **X** but not in MB(Y). The concept of the Markov Blanket was first introduced by Pearl in 1988 in his work on Bayesian networks [12]. Years later, it was first used to identify the theoretical optimal set of variables for prediction tasks [11].

According to the definition above, given MB(Y), the other variables contained in **X** are independent of the outcome Y. This means that they do not provide any further information about Y, and all the information to predict the behavior of the outcome is already contained in the Markov Blanket MB(Y) [1, 13].

If the technique used to build the prediction model for Y can fully describe the underlying true probabilities Pr(Y|MB(Y)), and a model with fewer variables is preferred, then the variables included in the Markov Blanket of the outcome Y are the only variables needed for an optimal prediction in terms of calibration [8]. Therefore, in an idealized regression setting, to fit the appropriate model, the predictor selection task consists of finding the Markov Blanket of the outcome variable [1, 9]. This concept can be used to link variable selection in clinical risk prediction modeling to the underlying causal structure of the data [14].

Let's consider a DAG *G* and a set of variables S described by a joint distribution $\mathbb{P}_S$ with a density. The distribution $\mathbb{P}_S$, is said to be Markovian with respect to *G* if each variable is conditionally independent of its non-descendants (i.e. variables it does not affect), given its parents (i.e. its direct causes) [1, 9]. This Markov property creates a link between $\mathbb{P}_S$ and *G*, ensuring that all the conditional independencies entailed by the DAG are also present in the probability distribution [1, 15].

A further condition makes this link stronger; "faithfulness" implies that the only conditional independencies to hold in the joint distribution $\mathbb{P}_S$ are the ones entailed in *G* [14].

The previous intuition can be formalized; it has been demonstrated that if the joint distribution of the variables is faithful and Markovian with respect to the DAG, a predictor is strongly relevant (see [16, 17] for a definition) for predicting the outcome if and only if it is part of the Markov Blanket of the outcome [17]. Under these conditions, the Markov Blanket of the outcome is unique and has a particular constitution: it includes all parents of the outcome node, all of its children, and all parents of its children [1, 8, 9, 12].

As shown in Fig. 1, these nodes "shield" the outcome variable Y from all the remaining variables in the DAG [13]. Therefore, the information contained in these nodes is sufficient to describe the outcome variable's status.

These results are appealing for researchers tasked with selecting predictors for clinical risk prediction modeling. According to a 2010 review, at least 8 different algorithms have been developed to identify the Markov Blanket for an outcome variable using data-driven procedures [9]. In the field of causal learning, algorithms that learn the entire causal structure [14] and the local causal structure [18] based on the identification of Markov Blankets have been developed. Given this theoretical line of argumentation, we believe that a knowledge of the underlying causal processes behind the data generation can help to identify the best predictors to be included in a clinical risk prediction model.

As proof of concept, we conducted a series of simulations using R version 3.6.3 (R code can be found in the Supplementary file). We simulated 100,000 datasets with

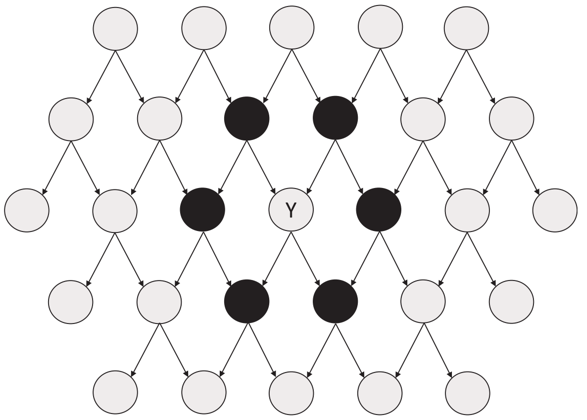


Fig. 1 Example of the Markov Blanket (in black) of outcome Y in a simple Directed Acyclic Graph (DAG) with many nodes

25 variables and 10,000 observations each. Each dataset was simulated according to a randomly generated DAG (using the randomDAG function in the dagitty R package). The DAG included 25 ordered nodes corresponding to 25 variables. Each node was given a probability of 0.1 of receiving a directed arrow from each of the individual previous nodes. One of the nodes was then randomly selected as the binary outcome of interest, all other 24 variables were assumed to be continuous. Any exogenous variables (i.e. variables without any parent nodes) were generated as normally distributed variables with a mean of 0 and variance of 1, or, if the outcome was exogenous, as a Bernoulli random variable with an event probability of 0.2.

When the outcome was an endogenous variable (i.e., with at least one parent node), each observation was drawn from a Bernoulli distribution with a defined probability parameter. This was set as the inverse-logit function evaluated at the linear combination of the outcome node's parent variables, with randomly drawn coefficients. Specifically, the coefficients (including the intercept) for the outcome endogenous variable were drawn from a uniform distribution on (− 1,1).

Similarly, the observations of the continuous endogenous variables were randomly drawn from a normal distribution with unit variance and with the mean equal to the linear combination of randomly drawn coefficients and the values of the node's parent variables. Here, the coefficients (including the intercept) for each endogenous variable were drawn from a uniform distribution on (− 2,2). The choice of the regression coefficients was therefore not restricted in order to satisfy the faithfulness assumption by design.

For each of the 100,000 datasets, eight prediction tools were developed to predict the probability that the binary outcome equals 1:

(i) a logistic regression model including only variables in the Markov Blanket of the outcome as predictors,

(ii) a logistic regression model including all 24 variables as predictors,

(iii) a logistic regression model including any variable with a path leading to the outcome node (regardless of arrow direction on the path) as predictors,

(iv) a logistic regression model including only the outcome node's parent variables as predictors,

(v) a logistic lasso regression model inputting all 24 variables,

(vi) a logistic ridge regression model inputting all 24 variables,

(vii) a logistic elastic net regression model with mixing parameter alpha of 0.5 inputting all 24 variables, and

(viii) a random forest algorithm inputting all 24 variables.

In all regression models, all included variables were modeled as being linearly related to the logit of the outcome. Lasso, ridge, and elastic net models were computed using the glmnet function in the glmnet R package with default settings. The regularization parameter, lambda, that minimized the 10-fold cross-validated error based on the deviance for logistic regression with the cv.glmnet function (glmnet package) was selected. Random forests were built using the randomForest function in the randomForest R package with 1000 trees and default settings.

For each dataset, the calibration of each prediction tool was measured using the Integrated Calibration Index [19] (ICI) based on 10-fold cross-validation. Lower ICI indicates better model calibration. The ICI estimation relies on a non-parametric regression between the outcome variable and the predicted risk estimated by the prediction tool. Therefore, if the non-parametric regression fails in one or more of the 10 cross-validation sets, it is not possible to compute the ICI. This happens if an intercept-only model or a model with variables' regression coefficients very close to 0 is evaluated. We also compared the variable sets included in the Markov Blanket-based logistic models with the ones selected by the lasso and elastic net regression models. We considered a variable to be selected by the model if the absolute value of its estimated regression coefficient was nonzero, which we operationalized as a value higher than 10^{-10}.

Results

Transportability and the principle of independent mechanisms

The potential benefit gained from applying the *principle of independent mechanisms* to the assessment of transportability of clinical risk prediction models is presented using two simplified clinical examples from the field of neurodegenerative disease.

Example 1

Say that we are interested in building a diagnostic clinical risk prediction model for the presence of Alzheimer's disease (Y = 1), using the APOE ε4 allele status (X = 1, presence; X = 0, absence) as the sole predictor of the outcome in the general population of older persons. Y = 0 indicates disease absence.

Since APOE ε4 is a known cause for Alzheimer's disease [20], we could draw the DAG shown in Fig. 2. Note that we are assuming a direct, unconfounded causal relationship (a strong assumption). By convention, each variable in the DAG is affected by a "noise" variable, which

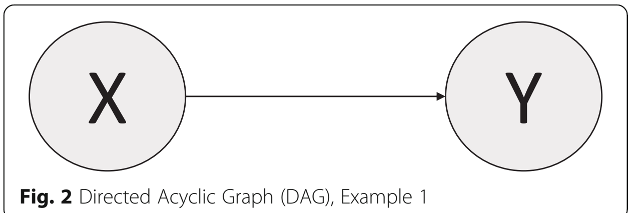


Fig. 2 Directed Acyclic Graph (DAG), Example 1

are assumed to be independent of other noise variables and modeled as random variables. These are usually not explicitly depicted because they are not of relevance to the causal relationship under study. However, it is worth noting that the noise variable affecting X determines the prevalence of the APOE ε4 allele, while the noise variable affecting Y contributes to the definition of the causal mechanism between the APOE ε4 allele status and Alzheimer's disease [7].

Assume we collect cross-sectional data about Alzheimer's disease and APOE ε4 allele status in a population A. Using this data, we can develop a simple diagnostic clinical risk prediction model using logistic regression to predict the presence of Alzheimer's disease. The regression equation would be:

$$\log_e(\Pr(Y=1|X=x)/\Pr(Y=0|X=x)) = \beta_0 + \beta_1 x$$

Using the logistic regression equation it's possible to estimate the four conditional probabilities $\Pr(Y = 1|X = 0)$, $\Pr(Y = 1|X = 1)$, $\Pr(Y = 0|X = 0)$, and $\Pr(Y = 0|X = 1)$, which define the conditional distribution $\mathbb{P}(Y|X)$. We will assume that the logistic regression is able to fully describe this conditional distribution, while the prevalence of the APOE ε4 allele ($\Pr(X = 1)$) defines the marginal distribution $\mathbb{P}(X)$ of this predictor.

Next, say we want to use our newly developed risk prediction model as a diagnostic tool for Alzheimer's disease in another population B in which we know there is a different prevalence of the APOE ε4 allele. The new distribution of the predictor X in population B can be denoted as $\mathbb{P}^*(X)$.

According to the principle of independent mechanisms, the fact that the original distribution of X, $\mathbb{P}(X)$, has been changed to $\mathbb{P}^*(X)$ does not give any information on the mechanism $\mathbb{P}(Y|X)$ in population B [1, 7]. This is because X causes Y, and $\mathbb{P}$(Cause) is independent of $\mathbb{P}$(Effect|Cause).

If the underlying causal mechanism is not altered ($\mathbb{P}(Y|X)$ is the same in the two populations), the diagnostic clinical risk prediction model developed in population A will produce valid estimates also in population B. On the other hand, if the causal mechanism changed, knowing the predictor distribution $\mathbb{P}^*(X)$ does not give us any information about how the mechanism changed [1, 7]. In this case, the logistic regression model developed in population A for modeling $\mathbb{P}(Y|X)$ is still our best diagnostic tool candidate [1, 7].

In this example, knowledge of the underlying causal structure suggests that using the same diagnostic clinical risk prediction model in the new population is a reasonable choice [1, 7].

Example 2

Next, say we are still interested in building a diagnostic clinical risk prediction model for the presence of Alzheimer's disease, but instead choose to use a different variable as the sole predictor, which indicates whether the concentration of tau protein in cerebrospinal fluid (CSF-tau) is above a predefined threshold. As before, $Y = 1$ and $Y = 0$ indicate presence and absence of Alzheimer's disease. $K = 1$ indicates high tau protein concentration, and $K = 0$ indicates low tau protein concentration.

It is known that high CSF-tau levels are associated with the presence of Alzheimer's disease. Specifically, as a consequence of the deposition of proteins in the brain that characterizes Alzheimer's disease, the concentration of tau protein is altered in the cerebrospinal fluid [21]. Therefore, the high level of tau protein in the cerebrospinal fluid can be interpreted as a consequence of Alzheimer's disease, leading to the DAG shown in Fig. 3.

In this example, we define Alzheimer's disease by its underlying pathological process instead of based on diagnostic criteria. However, in the real world, direct effects are usually incorporated as part of the diagnostic criteria of the disease for practical clinical purposes. We further assume a direct effect of Y on K without confounding, even though we acknowledge direct effects of a disease are typically also caused by risk factors for the disease (introducing confounding in the Y → K causal relationship depicted in Fig. 3). These strong assumptions are needed to create a simplified, illustrative example.

As before, assume we have collected cross-sectional data about Alzheimer's disease and CSF-tau concentration in a new population C. Using population C data, we can develop another simple diagnostic clinical risk prediction model to predict Alzheimer's disease using logistic regression. The estimated regression equation would be:

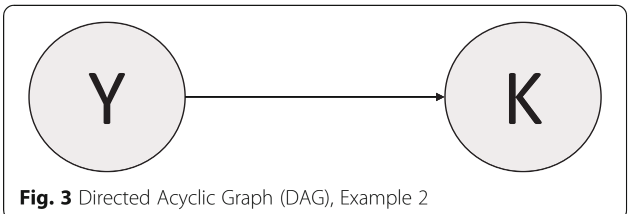


Fig. 3 Directed Acyclic Graph (DAG), Example 2

$$\log_e(\Pr(Y=1|K=k)/\Pr(Y=0|K=k)) = \gamma_0 + \gamma_1 k$$

Assuming that logistic regression is suitable, its equation fully describes the underlying conditional distribution $\mathbb{P}(Y|K)$, while the prevalence of the high CSF-tau ($\Pr(K=1)$) defines the marginal distribution $\mathbb{P}(K)$ of the predictor.

Say that we now want to apply this diagnostic clinical risk prediction model developed in population C to detect the presence of Alzheimer's disease in a population D with a different prevalence of high CSF-tau concentration. However, we are now in an anticausal scenario in which we are trying to use the effect, CSF-tau concentration, to detect the cause, Alzheimer's disease. Therefore, $\mathbb{P}(Y|K)$ does *not* represent a causal mechanism and is *not* independent of $\mathbb{P}(K)$.

Since the marginal distribution of CSF-tau levels changes from $\mathbb{P}(K)$ in population C to $\mathbb{P}^*(K)$ in population D, a change in the conditional distribution, $\mathbb{P}(Y|K)$, is likely to occur because we are in an anticausal direction [1, 7]. The model developed in population C to describe $\mathbb{P}(Y|K)$ will probably not be well calibrated for use in the population D because the underlying conditional distribution of Y on K is different in the two populations. This would also hold if the causal mechanism that leads from Alzheimer's disease to the high CSF-tau concentration was the same in the two populations, as the equation describing the conditional distribution of Y on K is purely a mathematical artefact and does not describe the causal process.

Predictor selection and the Markov blanket

The results of the simulation study investigating whether a strong knowledge of the causal structure underlying the data generation process improves predictor selection compared to other commonly implemented methods are shown in Table 1.

In 37,272 of the 100,000 simulated datasets, the outcome variable node did not have any parents, therefore it was not possible to assess the performance of logistic regression including only the outcome node's parent variables as predictors in these cases (Table 1). In 8032 simulated datasets, the outcome variable node did not have any parents or children, therefore it was not possible to assess the performance of the Markov Blanket-based logistic model and the logistic regression including all the variables with a path to the outcome as predictors (Table 1).

When the Markov Blanket set was empty, both the lasso and elastic net regression models correctly shrunk all regression coefficients to zero or very close to zero approximately 93.3% of the time, leading to an uncomputable ICI. Overall, the lasso regression selected exactly the Markov Blanket set of variables in at least one of the ten cross-validations in 14,936 (14.9%) simulated datasets. The percentage was higher when the Markov Blanket was empty (93.3%) or included only one variable (46.8%) compared to when it contained two (7.6%) or more variables. This finding supports the idea proposed by Li et al. that there is a link between the lasso regularization and selection algorithm and the identification of the Markov Blanket [22].

Overall, the average ICI of the Markov Blanket-based logistic model (0.01882) was lower compared with all other investigated prediction tools. This model also yielded the lowest average ICI (0.01956) when considering only those datasets in which all prediction tools had computable ICI values (Table 1). In head-to-head comparisons, the ICI of the various prediction tools were greater than or equal to the ICI of the Markov Blanket-based logistic model in the majority of the simulated datasets (range: 57.0 to 98.2%).

Discussion

Transportability and the principle of independent mechanisms

Through the two simple examples presented, we provide a theoretical basis for the intuition that a diagnostic clinical risk prediction model including causes as predictors may be more transportable [23]. As illustrated in Example 2, transportability in terms of calibration is likely to be lower in anticausal scenarios, in which the predicted outcome is the disease and the predictor is an effect of the outcome [1, 7].

No common causes of Y and K were included in the simplified Example 2, and we note that the transportability of the diagnostic clinical risk prediction model to different populations in similar anticausal scenarios could be higher if the predictor and the disease share one or more common cause(s). The idea that risk prediction models including the direct causes of an outcome of interest as predictors will be more transportable to different settings is also exploited in the causal learning "invariant causal prediction" method [1] and in the machine learning practice of "covariate shift" [1, 7]. In general, we think the field of diagnostic clinical risk prediction modeling could greatly benefit from the practice of incorporating knowledge of the underlying causal structure in modelling strategies. The integration of such information could provide insights into the transportability of a given diagnostic risk prediction model in different settings [7].

Predictor selection and the Markov blanket

Our results empirically demonstrated equal or superior performance of the Markov Blanket-based logistic model, corroborating the theories presented earlier. In the head-to-head comparisons with each of the other

Table 1 Simulation Results: Prediction Tools' Performance Metrics

	Logistic, Markov Blanket set (Nsim=100,000)	Logistic, all 24 variables (Nsim=100,000)	Logistic, any variables with a path to the outcome (Nsim=100,000)	Logistic, node's parent variables (Nsim=100,000)	Lasso, all 24 variables (Nsim=100,000)	Ridge, all 24 variables (Nsim=100,000)	Elastic net, all 24 variables (Nsim=100,000)	Random forest, all 24 variables (Nsim=100,000)
FULL RESULTS: Including all simulated datasets								
ICI								
N Missing	8032	0	8032	37,272	8597	0	8612	1
Mean (SD)	0.01882 (0.00445)	0.01964 (0.00495)	0.01900 (0.00461)	0.02215 (0.00421)	0.01912 (0.00451)	0.03807 (0.02058)	0.01907 (0.00456)	0.04133 (0.01779)
Median	0.01857	0.01925	0.01867	0.02242	0.01888	0.02895	0.01881	0.03636
Range	0.00290–0.03834	0.00289–0.04330	0.00287–0.04330	0.00290–0.03826	0.00287–0.03919	0.00710–0.18537	0.00340–0.04283	0.00704–0.16493
Number of input variables								
N Missing	0	0	0	0	0	0	0	0
Mean (SD)	4.0 (2.8)	24.0 (0.0)	18.9 (7.0)	1.2 (1.3)	24.0 (0.0)	24.0 (0.0)	24.0 (0.0)	24.0 (0.0)
Median	3.0	24.0	22.0	1.0	24.0	24.0	24.0	24.0
Range	0.0–19.0	24.0–24.0	0.0–24.0	0.0–9.0	24.0–24.0	24.0–24.0	24.0–24.0	24.0–24.0
Direct comparison: ICI of various methods compared to Markov Blanket-based logistic tool								
N Missing	8032	8032	8032	37,272	9140	8032	9147	8033
< ICI logistic MB, N (%)		39,354 (42.79%)	39,540 (42.99%)	4864 (7.75%)	26,514 (29.18%)	8871 (9.65%)	31,089 (34.22%)	1650 (1.79%)
≥ ICI logistic MB, N (%)		52,614 (57.21%)	52,428 (57.01%)	57,864 (92.25%)	64,346 (70.82%)	83,097 (90.35%)	59,764 (65.78%)	90,317 (98.21%)
COMPLETE CASE RESULTS: only including datasets for which ICI could be estimated for all tools								
ICI								
N Missing	37,841	37,841	37,841	37,841	37,841	37,841	37,841	37,841
Mean (SD)	0.01956 (0.00463)	0.01975 (0.00477)	0.01970 (0.00473)	0.02211 (0.00421)	0.01995 (0.00471)	0.03886 (0.02177)	0.01990 (0.00476)	0.04049 (0.02011)
Median	0.01953	0.01962	0.01960	0.02238	0.01993	0.02883	0.01987	0.03283
Range	0.00290–0.03834	0.00289–0.04330	0.00287–0.04330	0.00290–0.03826	0.00287–0.03919	0.00710–0.18537	0.00340–0.04283	0.00704–0.16493
Number of input variables								
N Missing	37,841	37,841	37,841	37,841	37,841	37,841	37,841	37,841
Mean (SD)	4.1 (2.7)	24.0 (0.0)	20.8 (3.9)	1.9 (1.1)	24.0 (0.0)	24.0 (0.0)	24.0 (0.0)	24.0 (0.0)
Median	4.0	24.0	22.0	2.0	24.0	24.0	24.0	24.0
Range	1.0–19.0	24.0–24.0	1.0–24.0	1.0–9.0	24.0–24.0	24.0–24.0	24.0–24.0	24.0–24.0
Direct comparison: ICI of various methods compared to Markov Blanket-based logistic tool								
N Missing	37,841	37,841	37,841	37,841	37,841	37,841	37,841	37,841
< ICI logistic MB, N (%)		26,872 (43.23%)	27,124 (43.64%)	4850 (7.80%)	16,887 (27.17%)	6508 (10.47%)	19,959 (32.11%)	1636 (2.63%)
≥ ICI logistic MB, N (%)		35,287 (56.77%)	35,035 (56.36%)	57,309 (92.20%)	45,272 (72.83%)	55,651 (89.53%)	42,200 (67.89%)	60,523 (97.37%)

In a series of 100,000 simulated datasets, we obtained these results for ICI and number of input variables for the eight investigated prediction tools. Full results and complete case results, including only datasets for which ICI could be estimated for all tools are presented

Abbreviations: *ICI* integrated calibration index, *MB* Markov Blanket, *Nsim* number of simulations, *SD* standard deviation approaches, the Markov Blanket-based logistic model yielded an equal or better calibration in more than 57% of all generated datasets (range 57 to 98% across the compared prediction tools). Not only did the Markov Blanket-based logistic model show good performance in terms of calibration but also required considerably fewer input variables than the number of available variables. Moreover, this approach relies explicitly on summarizing causal knowledge, which provides a high degree of interpretability in contrast to commonly encountered causally agnostic approaches.

We acknowledge that in real-world settings, it is unlikely to encounter ideal situations in which there is perfect knowledge of the underlying causal structure, all requisite variables are available and complete, and non-linear relationships and interactions are absent. Further research on deviations from these ideal conditions is needed, in particular to understand consequences of model misspecification when statistical interactions or non-linear relationships are present as well as measurement error. Nevertheless, we believe our results provide an important contribution as a theoretical basis for using a DAG that summarizes a priori knowledge of the causal structure to identify predictors in a simple and structured way in an ideal setting.

Conclusions

Through a series of theoretical examples and simulation results, we have shown that strong knowledge of the underlying causal structure can be useful for understanding potential transportability and optimizing predictor selection for a given clinical risk prediction model. In the field of clinical risk prediction model development and application, we think that a priori causal information is often ignored or used intuitively without a structured framework. We are eager to see first applications of the framework we have outlined, further theoretical development, and scientific discussion of this concept.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12874-020-01058-z.

Additional file 1.

Abbreviations

APOE ε4: Apolipoprotein E allele ε4; CSF-tau: Tau protein in cerebrospinal fluid; DAG: Directed Acyclic Graph; ICI: Integrated Calibration Index; Lasso: Least absolute shrinkage and selection operator; MB: Markov Blanket

Acknowledgements

The authors wish to thank Dr. James M. Robins for insightful comments and suggestions, which helped improve this manuscript. A preprint version of this manuscript is available on the arXiv server [24].

Authors' contributions

MP conceptualized the study. MP and SK designed and ran the simulations. MP and JR drafted the manuscript. TK and SK supervised the project. All authors critically reviewed the final version.

Funding

JLR's research position was supported by a grant from the Else-Kröner-Fresenius Foundation (www.ekfs.de, GSO/EKFS-17, granted to TK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of Charité – Universitätsmedizin Berlin.

Availability of data and materials

The dataset supporting the conclusions of this article is simulated and the code can be found in the Supplementary file. The simulation conditions are described in detail in the manuscript along with software, package names and functions for full transparency.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

We declare no conflicts of interest for the submitted work. JLR, MP and SK have nothing further to disclose. Outside of the submitted work, TK reports having contributed to an advisory board of CoLucid and a research project funded by Amgen, for which the Charité – Universitätsmedizin Berlin received an unrestricted compensation. He further reports having received honoraria from Lilly, Newsenselab, and Total for providing methodological advice, from Novartis and from Daiichi Sankyo for providing a lecture on neuroepidemiology and research methods, and from the BMJ for editorial services.

Author details

[1]Institute of Public Health, Charité - Universitätsmedizin Berlin, Berlin, Germany. [2]Digital Health & Machine Learning Research Group, Hasso Plattner Institute for Digital Engineering, Potsdam, Germany. [3]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, USA.

Received: 28 February 2020 Accepted: 19 June 2020
Published online: 02 July 2020

References

1. Peters J, Janzing D, Schölkopf B. Elements of Causal Inference: Foundations and Learning Algorithms. Cambridge: MIT Press; 2017.
2. Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Stat Med. 2018;37:4142–54. https://doi.org/10.1002/sim.7913.
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10:37–48.
4. Greenland S, Pearl J. Causal Diagrams: Wiley StatsRef. Statistics Reference Online; 2014. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03732.
5. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol. 2002;155:176–84.
6. Janzing D, Schölkopf B. Causal inference using the algorithmic Markov condition. IEEE Trans Inf Theory. 2010;56:5168–94.
7. Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. On Causal and Anticausal Learning. arXiv [cs. LG]; 2012.
8. Brown LE, Tsamardinos I. Markov blanket-based variable selection in feature space. Technical Report DSL TR-08-01; 2008.
9. Fu S, Desmarais MC. Markov blanket based feature selection: a review of past decade. In: Proceedings of the world congress on engineering. Hong Kong: Newswood Ltd; 2010;1:321–8.

10. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak. 2019;19:146.
11. Koller D, Sahami M. Toward Optimal Feature Selection. In: ICML'96 Proceedings of the Thirteenth International Conference on International Conference on Machine Learning; 1996. p. 284–92.
12. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco: Morgan Kaufmann; 1988.
13. Yaramakala S, Margaritis D. Speculative Markov blanket discovery for optimal feature selection. In: Fifth IEEE International Conference on Data Mining (ICDM'05); 2005.
14. Pellet J-P, Elisseeff A. Using Markov Blankets for Causal Structure Learning. J Mach Learn Res. 2008;9:1295–342.
15. Tsamardinos I, Aliferis CF, Statnikov AR, Statnikov E. Algorithms for large scale Markov blanket discovery. In: FLAIRS conference; 2003. p. 376–80.
16. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997; 97:273–324.
17. Tsamardinos I, Aliferis CF. Towards principled feature selection: relevancy, filters and wrappers. AISTATS: Proceedings of the ninth International workshop on artificial intelligence and statistics; 2003.
18. Yang S, Wang H, Hu X. Efficient Local Causal Discovery Based on Markov Blanket. arXiv [cs.AI]; 2019.
19. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med. 2019;38:4051–65.
20. Uddin MS, Kabir MT, Al Mamun A, Abdel-Daim MM, Barreto GE, Ashraf GM. APOE and Alzheimer's disease: evidence mounts that targeting APOE4 may combat Alzheimer's pathogenesis. Mol Neurobiol. 2019;56:2450–65.
21. Lee JC, Kim SJ, Hong S, Kim Y. Diagnosis of Alzheimer's disease utilizing amyloid and tau as fluid biomarkers. Exp Mol Med. 2019;51:1–10.
22. Li G, Dai H, Tu Y. Identifying Markov Blankets Using Lasso Estimation. In: Advances in Knowledge Discovery and Data Mining. Berlin Heidelberg: Springer; 2004. p. 308–18.
23. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Second edition. Cham: Springer; 2019.
24. Piccininni M, Konigorski S, Rohmann JL, Kurth T. Directed Acyclic Graphs and causal thinking in clinical risk prediction modeling. arXiv [stat.ME]. 2020. http://arxiv.org/abs/2002.09414. Accessed 15 June 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

14 Curriculum Vitae

My curriculum vitae does not appear in the electronic version of my work for reasons of data protection.

15 Complete list of publications

- Tortelli R, Zecca C, Piccininni M, Benmahamed S, Dell'Abate MT, Barulli MR, Capozzo R, Battista P, Logroscino G. Plasma Inflammatory Cytokines Are Elevated in ALS. Front. Neurol. 11:552295 (2020).
- Al-Hassany L, Haas J, Piccininni M, Kurth T, Maassen Van Den Brink , Rohmann JL. Giving Researchers a Headache – Sex and Gender Differences in Migraine. Front. Neurol. 11:549038 (2020).
- Meier K, Glatz T, Guijt MC, Piccininni M, van der Meulen M, Atmar K, Jolink AC, Kurth T, Rohmann JL, Zamanipoor Najafabadi AH, COVID-19 Survey Study group (2020). Public perspectives on protective measures during the COVID-19 pandemic in the Netherlands, Germany and Italy: A survey study. PLoS ONE 15(8): e0236917.
- Piccininni M, Konigorski S, Rohmann JL, Kurth T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. BMC Med Res Methodol 20, 179 (2020).
- Piccininni M, Rohmann JL, Logroscino G, Kurth T. Blockchain-Based Innovations for Population-Based Registries for Rare Neurodegenerative Diseases . Frontiers in Blockchain . 2020 May 19; 3:20.
- Piccininni M, Rohmann JL, Foresti L, Lurani C, Kurth T. Use of all cause mortality to quantify the consequences of covid-19 in Nembro, Lombardy: descriptive study. BMJ 2020; 369 :m1835.
- Piccininni M, Rohmann JL, Huscher D, Mielke N, Ebert N, Logroscino G, Schäffner E, Kurth T. Performance of risk prediction scores for cardiovascular mortality in older persons: External validation of the SCORE OP and appraisal. PLoS One. 2020 Apr 9;15(4):e0231097.
- Logroscino G, Kurth T, Piccininni M. The Reconstructed Cohort Design: A Method to Study Rare Neurodegenerative Diseases in Population-Based Settings. Neuroepidemiology. 2020 Jan 7:1-9.
- GBD 2017 Italy Collaborators. Italy's health performance, 1990-2017: findings from the Global Burden of Disease Study 2017. Lancet Public Health. 2019 Dec;4(12):e645-e657.

- Zecca C, Brescia V, Piccininni M, Capozzo R, Barone R, Barulli MR, Logroscino G. Comparative evaluation of two immunoassays for cerebrospinal fluid β-Amyloid(1-42) measurement. Clin Chim Acta. 2019 Jun;493:107-111.
- Logroscino G, Piccininni M, Binetti G, Zecca C, Turrone R, Capozzo R, Tortelli R, Battista P, Bagoj E, Barone R, Fostinelli S, Benussi L, Ghidoni R, Padovani A, Cappa SF, Alberici A, Borroni B. Incidence of frontotemporal lobar degeneration in Italy: The Salento-Brescia Registry study. Neurology. 2019 May 14;92(20):e2355-e2363.
- Depenni R, Cossu Rocca M, Ferrari D, Azzarello G, Baldessari C, Alù M, Nolé F, Codecà C, Boscolo G, Piccininni M, Cavalieri S, Bossi P. Clinical outcomes and prognostic factors in recurrent and/or metastatic head and neck cancer patients treated with chemotherapy plus cetuximab as first-line therapy in a real-world setting. Eur J Cancer. 2019 May 10;115:4-12.
- Barulli MR, Piccininni M, Di Dio C, Musarò C, Grasso A, Tursi M, Iurillo A, Lozupone M, Capozzo R, Tortelli R, Simone IL, Panza F, Logroscino G. Episodic memory and learning rates in amyotrophic lateral sclerosis without dementia. Cortex. 2019 Mar 19;117:257-265.
- GBD 2016 Neurology Collaborators. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019 May;18(5):459-480.
- Logroscino G, Piccininni M. Amyotrophic Lateral Sclerosis Descriptive Epidemiology: The Origin of Geographic Difference. Neuroepidemiology 2019;52:93-103.
- Panza F, Lozupone M, Sardone R, Battista P, Piccininni M, Dibello V, La Montagna M, Stallone R, Venezia P, Liguori A, Giannelli G, Bellomo A, Greco A, Daniele A, Seripa D, Quaranta N, Logroscino G. (2018). Sensorial frailty: age-related hearing loss and the risk of cognitive impairment and dementia in later life. Therapeutic Advances in Chronic Disease.
- GBD 2016 Motor Neuron Disease Collaborators. Global, regional, and national burden of motor neuron diseases 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2018 Nov 2. pii: S1474-4422(18)30404-6.

- Battista P, Catricalà E, Piccininni M, Copetti M, Esposito V, Polito C, Miozzo A, Gobbi E, Cuoco S, Boschi V, Picillo M, Sorbi S, Barone P, Iannaccone S, Garrard P, Logroscino G, Cappa SF. Screening for Aphasia in NeuroDegeneration for the Diagnosis of Patients with Primary Progressive Aphasia: Clinical Validity and Psychometric Properties. Dement Geriatr Cogn Disord. 2018;46(3-4):243-252.
- Lozupone M, Panza F, Piccininni M, Copetti M, Sardone R, Imbimbo BP, Stella E, D'Urso F, Rosaria Barulli M, Battista P, Grasso A, Tortelli R, Capozzo R, Coppola F, Isabel Abbrescia D, Bellomo A, Giannelli G, Quaranta N, Seripa D, Logroscino G. Social Dysfunction in Older Age and Relationships with Cognition, Depression, and Apathy: The GreatAGE Study. J Alzheimers Dis. 2018;65(3):989-1000.
- Zecca C, Tortelli R, Panza F, Arcuti S, Piccininni M, Capozzo R, Barulli MR, Barone R, Cardinali R, Abbrescia D, Seripa D, Brescia V, Logroscino G. Plasma β-amyloid1–42 reference values in cognitively normal subjects. Journal of the Neurological Sciences. 2018; 391 (2018): 120-126.
- Logroscino G, Marin B, Piccininni M, Arcuti S, Chiò A, Hardiman O, Rooney J, Zoccolella S, Couratier P, Preux PM, Beghi E; for EURALS. Referral bias in ALS epidemiological studies. PLoS One. 2018 Apr 16;13(4):e0195821.
- Panza F, Lozupone M, Solfrizzi V, Custodero C, Valiani V, D'Introno A, Stella E, Stallone R, Piccininni M, Bellomo A, Seripa D, Daniele A, Greco A, Logroscino G. Contribution of Mediterranean Diet in the Prevention of Alzheimer's Disease. Role of the Mediterranean Diet in the Brain and Neurodegenerative Diseases. 2018. p. 139–55.
- Battista P, Miozzo A, Piccininni M, Catricalà E, Capozzo R, Tortelli R, Padovani A, Cappa S, Logroscino G. Primary Progressive Aphasia: a review of Neuropsychological tests for the assessment of speech and language disorders. Aphasiology. 2017; 31 (12): 1359-1378.

16 Acknowledgments

Studying counterfactual prediction unavoidably means recognizing the difficulty, and probably the impossibility, of identifing and thanking all the people without whom this PhD dissertation would not have been possible.
Nevertheless, I believe there are some individuals who, regardless of any theoretical concerns, deserve a mention in this section.

First of all, I wish to sincerely thank Prof. Giancarlo Logroscino, who introduced me to the fascinating world of health research and guided me towards the best path for my personal and academic development. He has believed in me since I was a graduate student and has never missed an opportunity to share his passion and appreciation for epidemiology.

I wish to express my deepest gratitude to Prof. Tobias Kurth, whose knowledge, scientific integrity and experience represented and continue to represent to me an essential point of reference. He also deserves indisputable merit for having created and continuously alimented a wonderful and pleasant work environment, full of intellectual stimuli, critical thinking and friendly human interaction.

I wish to convey my heartfelt sense of gratitude to my colleague and mentor Jessica L. Rohmann, whose unconditioned altruism, empathy and continuous support over these years taught me more than any book could have done. My training would have been largely incomplete without her ability to gently push me beyond the boundaries of my comfort zone. I wish her all the best for her new path.

I want also to thank my father, my mother, and my brother Luca. Despite the physical distance, they were behind every single step I took during my doctoral studies.

Last but not least, I want to thank Ilaria, an endless source of support and tireless companion in this journey.