

Predictive Analytics for Health and Care

Niels Peek

*Professor of Health Informatics
The University of Manchester
Manchester, UK*



This talk

1. The Data Revolution in Health
2. Predictive Analytics
3. Is Machine Learning better than Statistical Modelling?
4. Causal Prediction Models
5. Prediction with Electronic Healthcare Records
6. Summary & Conclusion

The data revolution in health

Wearable
sensors

Traditional
research data

- trials
- case-control studies
- cohort studies

Omics data

- genome
- proteome
- metabolome
- ...

Routine care data

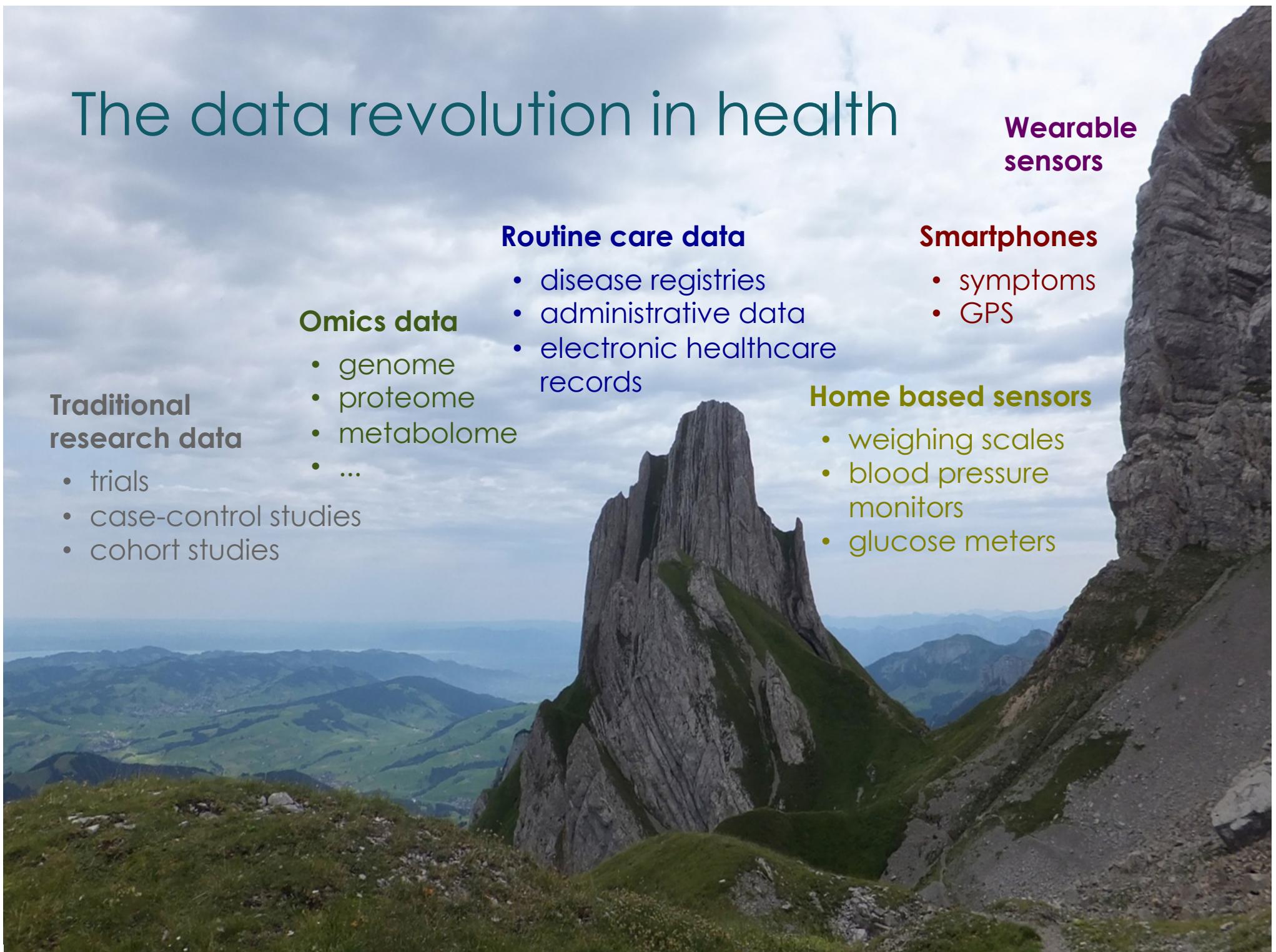
- disease registries
- administrative data
- electronic healthcare records

Smartphones

- symptoms
- GPS

Home based sensors

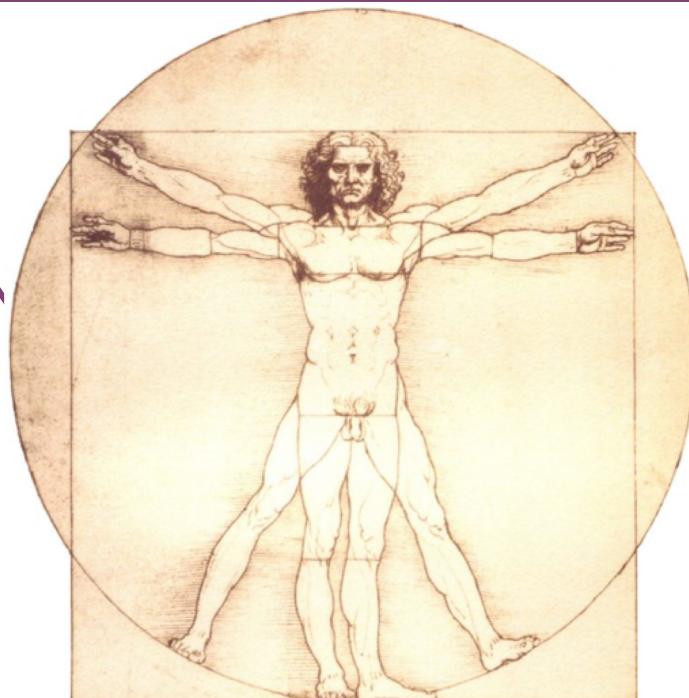
- weighing scales
- blood pressure monitors
- glucose meters



A more complete picture

biology
genome
(transcriptome
proteome, ...)

environment,
behaviour,
treatment
exposome

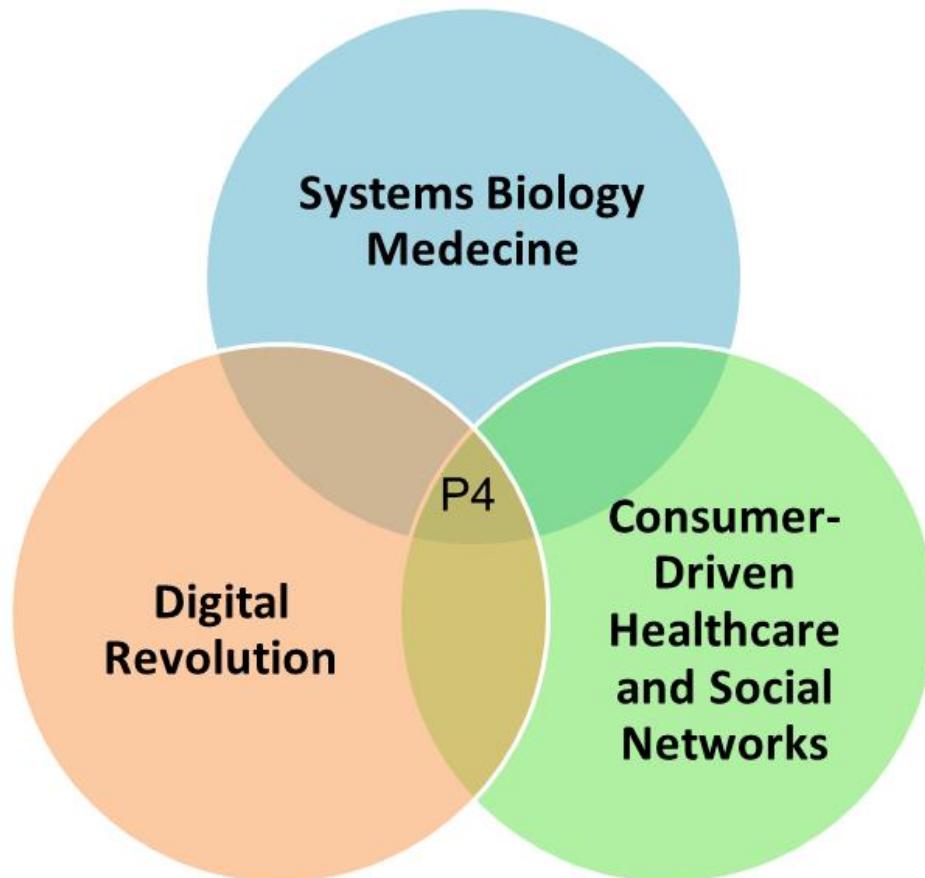


health & disease
phenome

Martin Sanchez et al.
J Am Med Inform
Assoc 2014;21:386–390.

P4 Medicine

● PREDICT ● PREVENT ● PERSONALIZE ● PARTICIPATE



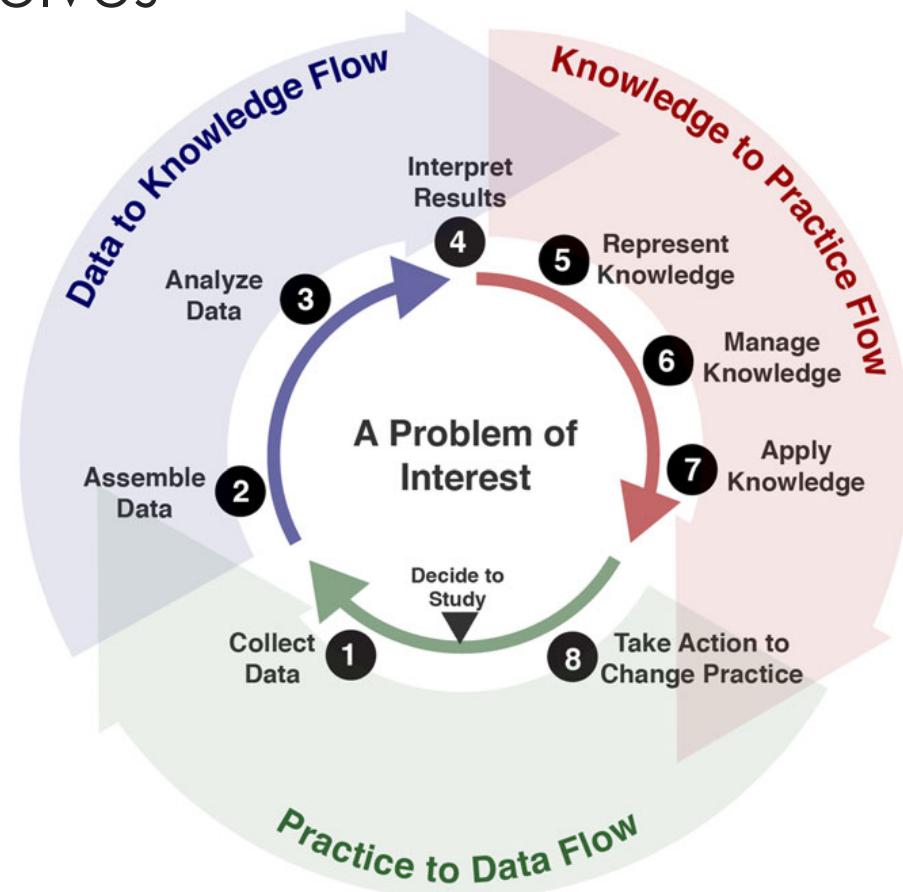
Leroy Hood

Learning Health Systems

Health systems become learning systems when they can, continuously and routinely, study and improve themselves



Charles P.
Friedman



The Alan Turing Institute

- Established by **five founding universities** (Cambridge, Edinburgh, Oxford, UCL and Warwick) in **2015**
- **Eight new universities** (incl. Manchester) joined in **2018**
- **Headquarters** at the British Library in London
- **Goals**
 1. Advance world-class research and apply it to real-world problems
 2. Train the leaders of the future in data science and AI
 3. Lead the public conversation about data science and AI
- **Programs:** e.g. Data Science at scale; Defence and Security; Finance and Economics; Health and Medical Sciences

University of Manchester Turing Health programme

- Predictive healthcare

Matthew Sperrin, Niels Peek & Lijing Lin
School of Health Sciences



- Patient subgroup discovery

Thomas House & Timothy Kinyanjui
School of Mathematics



- Human-centred systems

Caroline Jay & Manuele Reani
School of Computer Science



Goran
Nenadic



Angelo
Cangelosi



Ian Hall



Ken Muir



Hui Guo



Tjeerd
van Staa



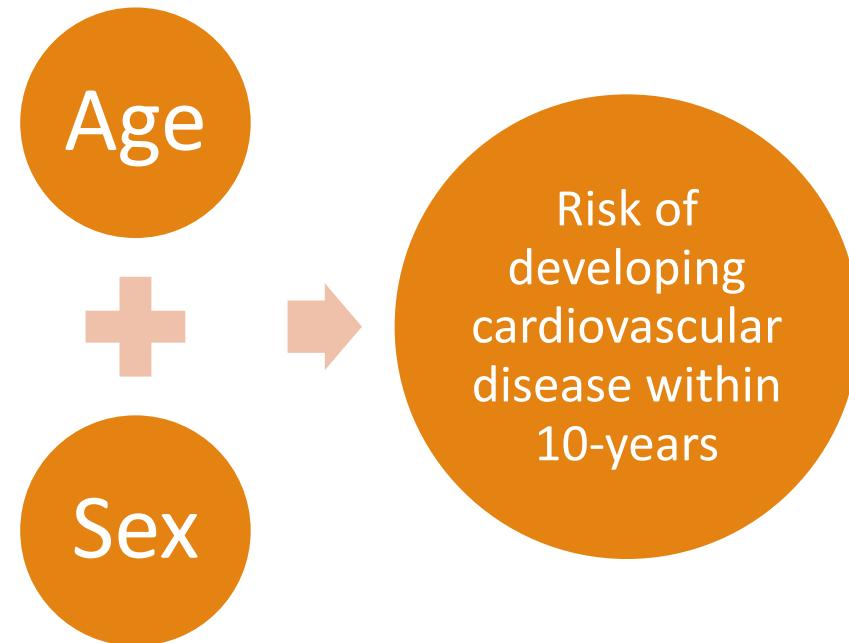
The University of Manchester

Clinical Prediction Models

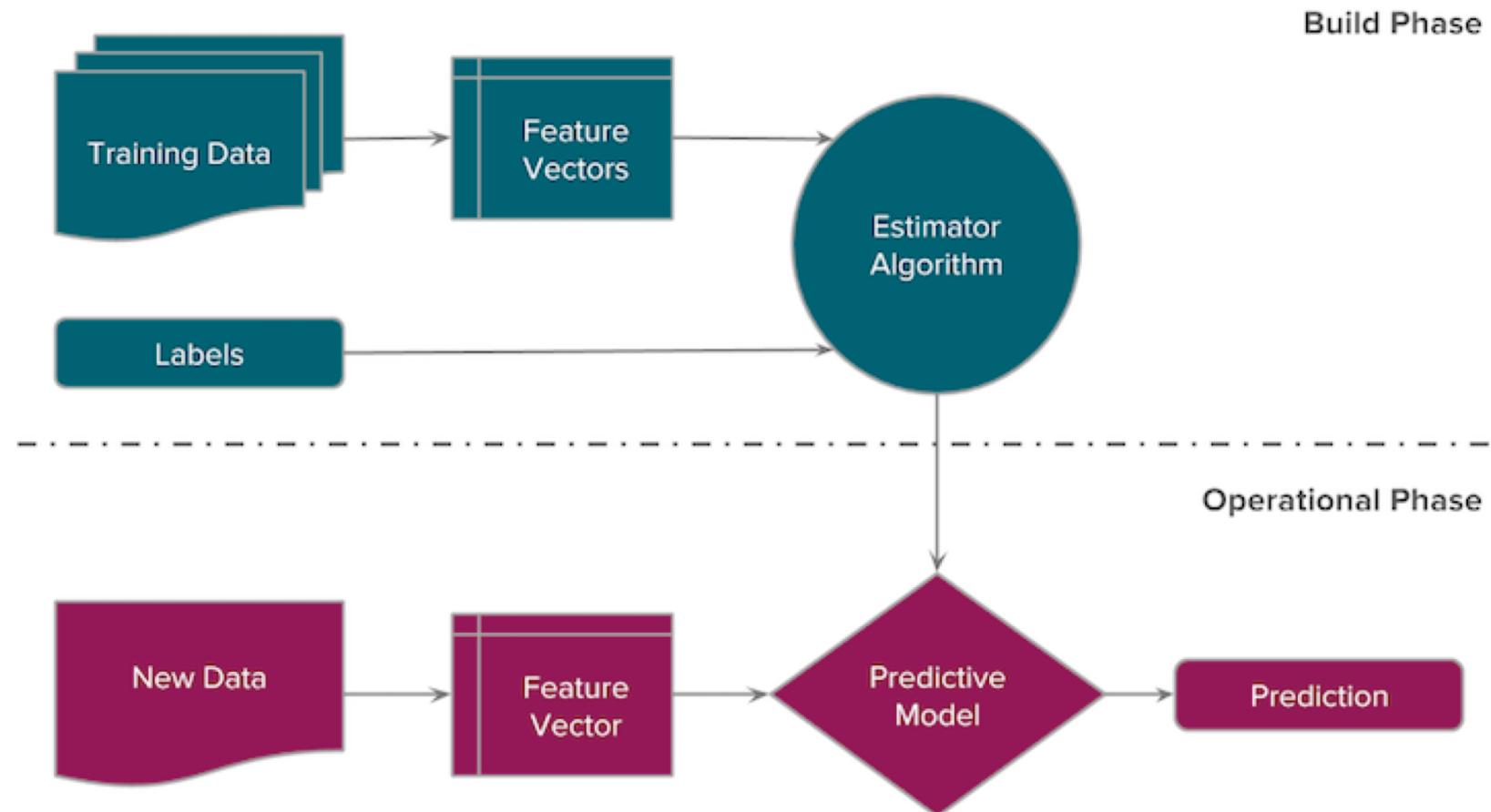
Clinical Prediction Models (CPMs)

Clinical prediction models (CPMs) are mathematical equations or algorithms, which take information about a patient as inputs, and output a prediction (usually a probability) regarding a patient's diagnosis or prognosis.

For example:



CPMs are a form of supervised learning



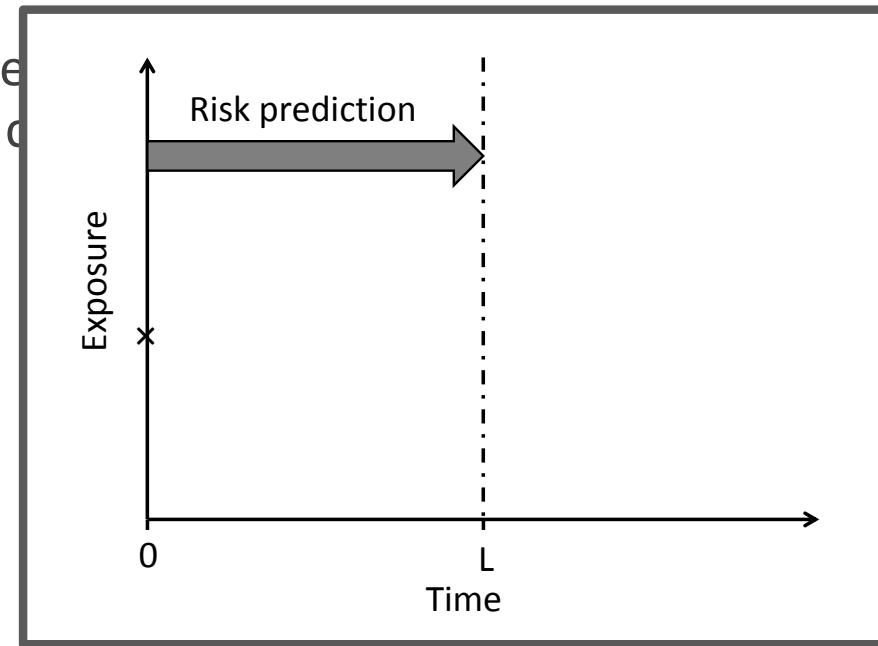
Types of CPMs

- **Diagnostic**

Predicts current presence of a disease or condition of interest, based on observed characteristics

- **Prognostic**

Predicts the likelihood of a future progression, based on observed c



Applications

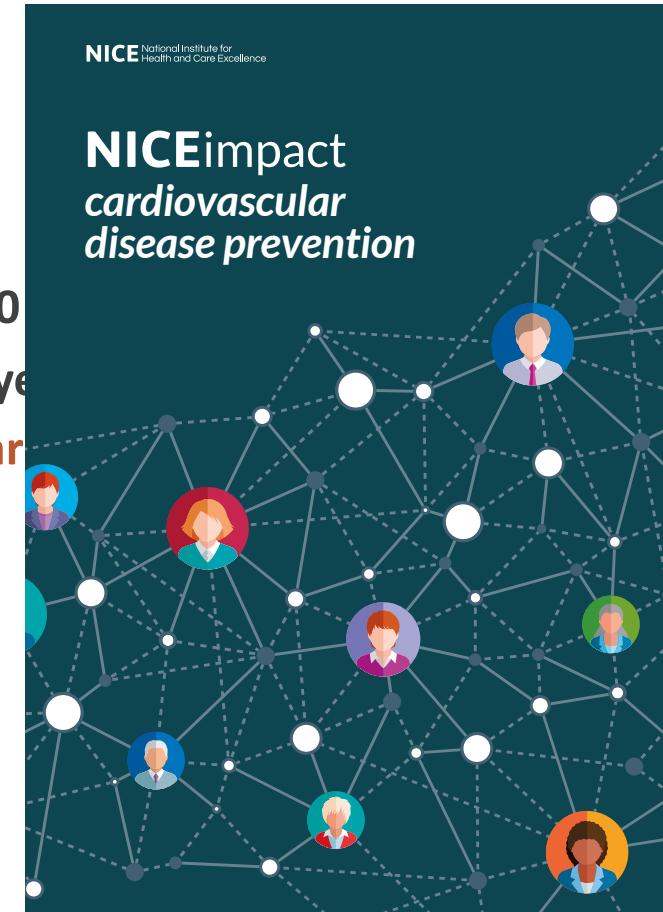
- Screening diagnostic
 - Triage diagnostic
 - Prevention prognostic
 - acute care
 - cancer recurrence
 - chronic disease
 - Benchmarking of clinical performance prognostic

Applications

- Screening
- Triage
- Prevention
 - acute care **short term (up to 30 days)**
 - cancer recurrence **medium term (1-5 years)**
 - chronic disease **long term (5-10 years)**
- Benchmarking of clinical performance

Applications

- Screening
 - Triage
 - Prevention
 - acute care
 - cancer recurrence
 - chronic disease
 - Benchmarking of clinical performance
- short term (up to 30 days)**
- medium term (1-5 years)**
- long term (5-10 years)**



This calculator is only valid if you do not already have a diagnosis of coronary heart disease (including angina or heart attack) or stroke/transient ischaemic attack.

[Reset](#) [Information](#) [Publications](#) [About](#) [Copyright](#) [Contact Us](#) [Algorithm](#) [Software](#) [CE](#)

About you

Age (25-84):

Sex: Male Female

Ethnicity:

UK postcode: leave blank if unknown

Postcode:

Clinical information

Smoking status:

Diabetes status:

Angina or heart attack in a 1st degree relative < 60?

Chronic kidney disease (stage 3, 4 or 5)?

Atrial fibrillation?

On blood pressure treatment?

Do you have migraines?

Rheumatoid arthritis?

Systemic lupus erythematosus (SLE)?

Severe mental illness?

(this includes schizophrenia, bipolar disorder and moderate/severe depression)

On atypical antipsychotic medication?

Are you on regular steroid tablets?

A diagnosis of or treatment for erectile dysfunction?

Leave blank if unknown

Cholesterol/HDL ratio:

Systolic blood pressure (mmHg):

Standard deviation of at least two most recent systolic blood pressure readings (mmHg):

Body mass index

Height (cm):

Weight (kg):

This calculator is only valid if you do not already have a diagnosis of coronary heart disease (including angina or heart attack) or stroke/transient ischaemic attack.

[Reset](#) [Information](#) [Publications](#) [About](#) [Copyright](#) [Contact Us](#) [Algorithm](#) [Software](#) [CE](#)
About you
Age (25-84): Sex: Male Female

Ethnicity: White or not stated

UK postcode: leave blank if unknown

Postcode:
Clinical information

Smoking status: non-smoker

Diabetes status: type 2

Angina or heart attack in a 1st degree relative < 60? Chronic kidney disease (stage 3, 4 or 5)? Atrial fibrillation? On blood pressure treatment? Do you have migraines? Rheumatoid arthritis? Systemic lupus erythematosus (SLE)?

Severe mental illness?

(this includes schizophrenia, bipolar disorder and moderate/severe depression)

On atypical antipsychotic medication? Are you on regular steroid tablets? A diagnosis of or treatment for erectile dysfunction?

Leave blank if unknown

Cholesterol/HDL ratio: Systolic blood pressure (mmHg): Standard deviation of at least two most recent systolic blood pressure readings (mmHg):

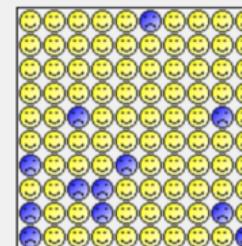
Body mass index

Height (cm): Weight (kg):
Your results

Your risk of having a heart attack or stroke within the next 10 years is:

12.4%

In other words, in a crowd of 100 people with the same risk factors as you, 12 are likely to have a heart attack or stroke within the next 10 years.



Risk of
a heart attack or stroke

Your score has been calculated using estimated data, as some information was left blank.

Your body mass index was calculated as 29.98 kg/m².

How does your 10-year score compare?
Your score

Your 10-year QRISK®3 score	12.4%
----------------------------	--------------

The score of a healthy person with the same age, sex, and ethnicity* 3.8%

Relative risk** 3.3

Your QRISK®3 Healthy Heart Age*** 66

* This is the score of a healthy person of your age, sex and ethnic group, i.e. with no adverse clinical indicators and a cholesterol ratio of 4.0, a stable systolic blood pressure of 125, and BMI of 25.

** Your relative risk is your risk divided by the healthy person's risk.

*** Your QRISK®3 Healthy Heart Age is the age at which a healthy person of your sex and ethnicity has your 10-year QRISK®3 score.

Logistic regression

Single predictor:

$$P(y=1 | x) = \frac{1}{1 + e^{-\alpha - \beta \cdot x}}$$

Multiple predictors:

$$P(y=1 | x_1, \dots, x_m) = \frac{1}{1 + e^{-\beta_0 - \sum_{i=1}^m \beta_i x_i}}$$

... or equally

Single predictor:

$$\ln \frac{P(y=1 | x)}{1 - P(y=1 | x)} = \alpha + \beta \cdot x$$

Multiple predictors:

$$\ln \frac{P(y=1 | x_1, \dots, x_m)}{1 - P(y=1 | x_1, \dots, x_m)} = \beta_0 + \sum_{i=1}^m \beta_i x_i$$

Example: APACHE II model

Model covariates

- s = Apache II severity-of-illness score (integer 0-71)
- d = diagnosis (54 categories)
- u = urgent surgery, $u \in \{0,1\}$.

Logistic regression model

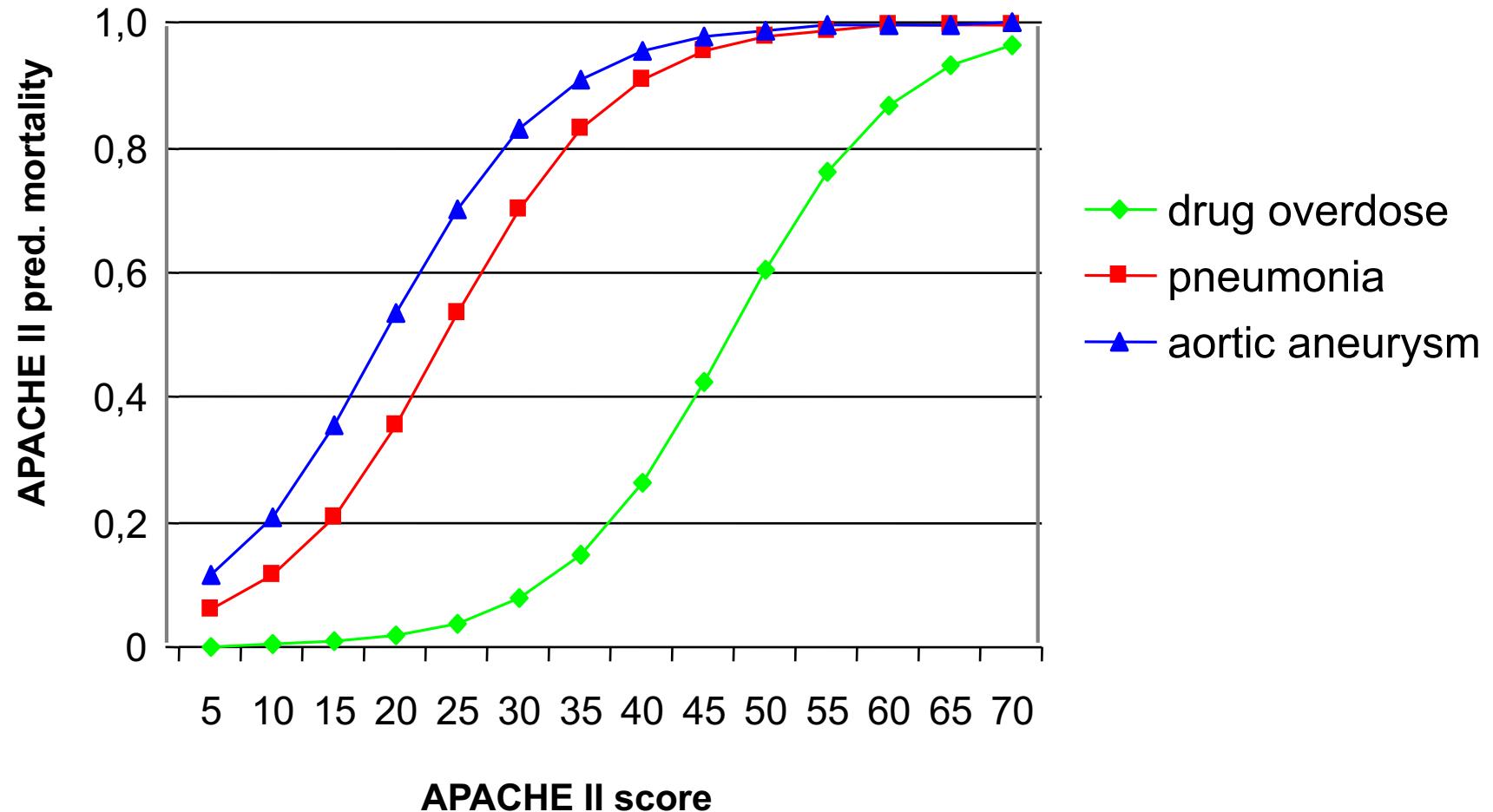
$$\ln \frac{P(y=1 | s, d, u)}{1 - P(y=1 | s, d, u)} = -3.517 + \beta_d + 0.146 \cdot s + 0.603 \cdot u$$

(conditional) log odds of death

57 coefficients,
estimated from data

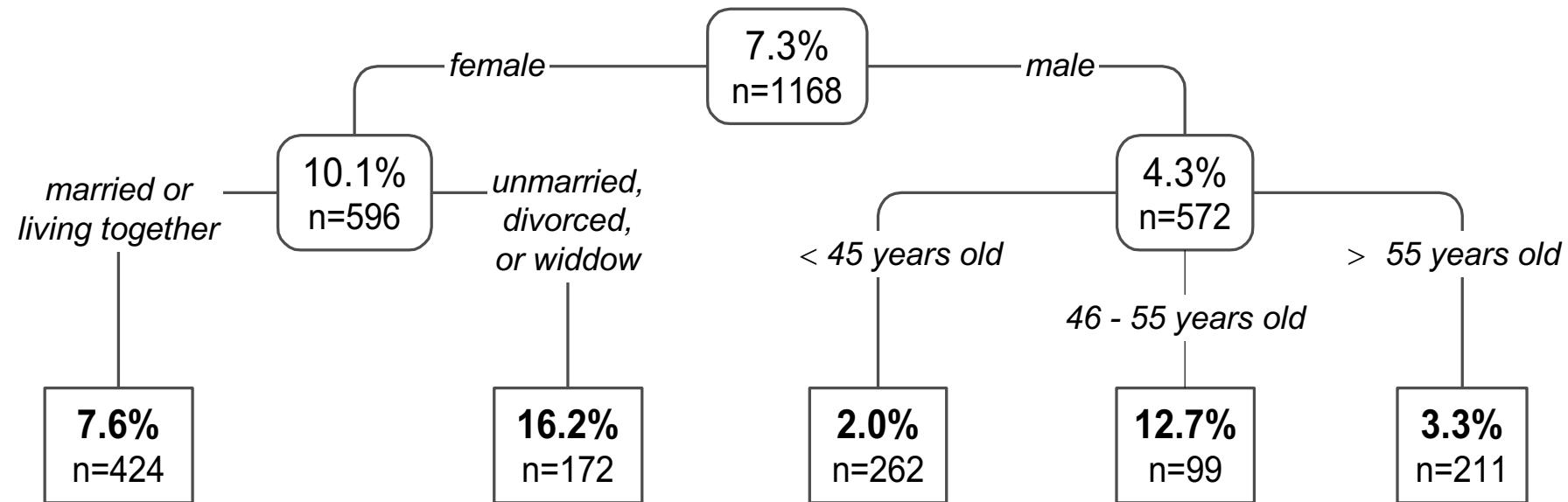
```
graph LR; A["ln P(y=1 | s, d, u) / (1 - P(y=1 | s, d, u))"] --> B["= -3.517 + β_d + 0.146 · s + 0.603 · u"]; B --> C["57 coefficients, estimated from data"]
```

APACHE II risk functions

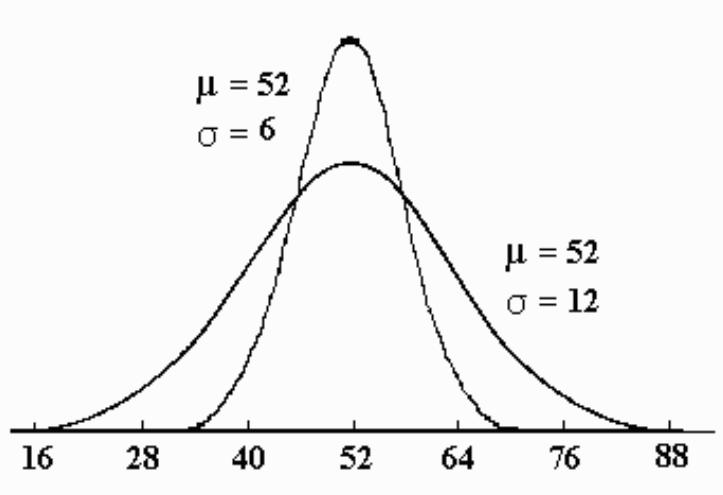


Is Machine Learning better than Statistical Modelling for predicting health risks?

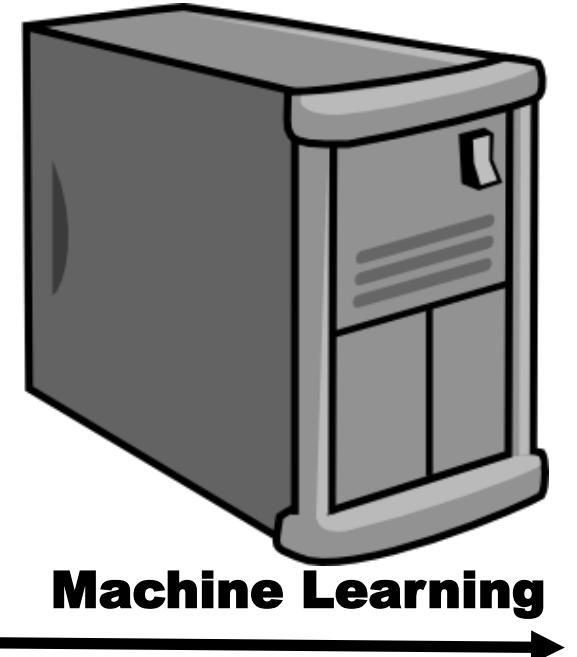
Decision tree for predicting risk of PTSD



Bronner MB, et al. J Trauma Stress. 2009 Feb;22(1):74-8.



Statistics
↔



Machine Learning

T-Test

Logistic Regression

Elastic Net

Gradient Boosting

Deep Learning

Machine Learning

network, graphs

focus on prediction

weights

learning

generalization

supervised learning

unsupervised learning

large grant = \$1,000,000

nice place to have a meeting:
Snowbird, Utah, French Alps

Statistics

model

focus on inference

parameters

fitting

test set performance

regression/classification

density estimation, clustering

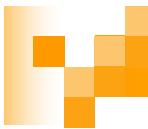
large grant = \$50,000

nice place to have a meeting:
Las Vegas in August

<http://statweb.stanford.edu/~tibs/stat315a/>

Robert Tibshirani





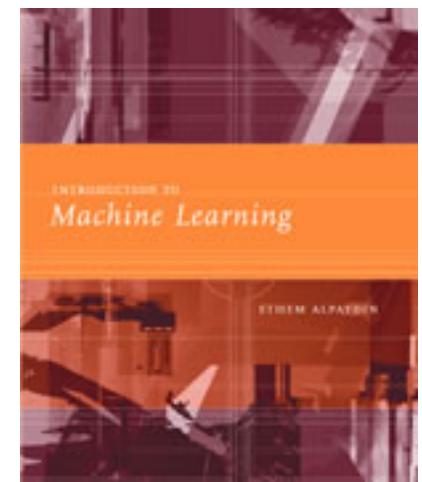
Machine Learning vs. Statistics

■ Statistical Modelling

- Field is defined by solution (**regression**)
- Starting point: **stochastic functions** (probability distributions)
- Important role for **analyst**
- Attention for **bias/variance tradeoff**

■ Machine Learning

- Field is defined by **problem type**
- Starting point: **deterministic functions**
- Goal is **fully automated learning**
- Focus on **representational expressiveness**
- Heavy use of **search algorithms**



Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches



Neophytos Stylianou ^{a,*}, Artur Akbarov ^a, Evangelos Kontopantelis ^a,
Iain Buchan ^a, Ken W. Dunn ^b

^a Centre for Health Informatics, Institute of Population Health, University of Manchester, UK

^b University Hospital South Manchester, Greater Manchester, UK

ARTICLE INFO

Article history:

Accepted 28 March 2015

Keywords:

Machine learning

Burn

Mortality

Clinical prediction

ABSTRACT

Introduction: Predicting mortality from burn injury has traditionally employed logistic regression models. Alternative machine learning methods have been introduced in some areas of clinical prediction as the necessary software and computational facilities have become accessible. Here we compare logistic regression and machine learning predictions of mortality from burn.

Methods: An established logistic mortality model was compared to machine learning methods (artificial neural network, support vector machine, random forests and naïve Bayes) using a population-based (England & Wales) case-cohort registry. Predictive evaluation used: area under the receiver operating characteristic curve; sensitivity; specificity; positive predictive value and Youden's index.

Results: All methods had comparable discriminatory abilities, similar sensitivities, specificities and positive predictive values. Although some machine learning methods performed marginally better than logistic regression the differences were seldom statistically significant and clinically insubstantial. Random forests were marginally better for high positive predictive value and reasonable sensitivity. Neural networks yielded slightly better prediction overall. Logistic regression gives an optimal mix of performance and interpretability.

Discussion: The established logistic regression model of burn mortality performs well against more complex alternatives. Clinical prediction with a small set of strong, stable, independent predictors is unlikely to gain much from machine learning outside specialist research contexts.

Predicting urinary tract infections in the emergency department with machine learning

R. Andrew Taylor*, Christopher L. Moore, Kei-Hoi Cheung, Cynthia Brandt

Department of Emergency Medicine, Yale University School of Medicine, New Haven CT, United States of America

Background

Urinary tract infection (UTI) is a common emergency department (ED) diagnosis with reported high diagnostic error rates. Because a urine culture, part of the gold standard for diagnosis of UTI, is usually not available for 24–48 hours after an ED visit, diagnosis and treatment decisions are based on symptoms, physical findings, and other laboratory results, potentially leading to overutilization, antibiotic resistance, and delayed treatment. Previous research has demonstrated inadequate diagnostic performance for both individual laboratory tests and prediction tools.

Objective

Our aim, was to train, validate, and compare machine-learning based predictive models for UTI in a large diverse set of ED patients.

Methods

Single-center, multi-site, retrospective cohort analysis of 80,387 adult ED visits with urine culture results and UTI symptoms. We developed models for UTI prediction with six machine learning algorithms using demographic information, vitals, laboratory results, medications, past medical history, chief complaint, and structured historical and physical exam findings. Models were developed with both the full set of 211 variables and a reduced set of 10 variables. UTI predictions were compared between models and to proxies of provider judgment (documentation of UTI diagnosis and antibiotic administration).

Results

The machine learning models had an area under the curve ranging from 0.826–0.904, with extreme gradient boosting (XGBoost) the top performing algorithm for both full and reduced models. The XGBoost full and reduced models demonstrated greatly improved specificity when compared to the provider judgment proxy of UTI diagnosis OR antibiotic administration with specificity differences of 33.3 (31.3–34.3) and 29.6 (28.5–30.6), while also demonstrating superior sensitivity when compared to documentation of UTI diagnosis with sensitivity differences of 38.7 (38.1–39.4) and 33.2 (32.5–33.9). In the admission and discharge cohorts using the full XGboost model, approximately 1 in 4 patients (4109/15855) would be re-categorized from a false positive to a true negative and approximately 1 in 11 patients (1372/15855) would be re-categorized from a false negative to a true positive.

Conclusion

The best performing machine learning algorithm, XGBoost, accurately diagnosed positive urine culture results, and outperformed previously developed models in the literature and several proxies for provider judgment. Future prospective validation is warranted.

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou^a, Jie Ma^b, Gary S. Collins^{b,c}, Ewout W. Steyerberg^d,
Jan Y. Verbakel^{a,e,f}, Ben Van Calster^{a,d,*}

^a*Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium*

^b*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK*

^c*Oxford University Hospitals NHS Foundation Trust, Oxford, UK*

^d*Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands*

^e*Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium*

^f*Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK*

Accepted 5 February 2019; Published online 11 February 2019

Abstract

Objectives: The objective of this study was to compare performance of logistic regression (LR) with machine learning (ML) for clinical prediction modeling in the literature.

Study Design and Setting: We conducted a Medline literature search (1/2016 to 8/2017) and extracted comparisons between LR and ML models for binary outcomes.

Results: We included 71 of 927 studies. The median sample size was 1,250 (range 72–3,994,872), with 19 predictors considered (range 5–563) and eight events per predictor (range 0.3–6,697). The most common ML methods were classification trees, random forests, artificial neural networks, and support vector machines. In 48 (68%) studies, we observed potential bias in the validation procedures. Sixty-four (90%) studies used the area under the receiver operating characteristic curve (AUC) to assess discrimination. Calibration was not addressed in 56 (79%) studies. We identified 282 comparisons between an LR and ML model (AUC range, 0.52–0.99). For 145 comparisons at low risk of bias, the difference in logit(AUC) between LR and ML was 0.00 (95% confidence interval, −0.18 to 0.18). For 137 comparisons at high risk of bias, logit(AUC) was 0.34 (0.20–0.47) higher for ML.

Conclusion: We found no evidence of superior performance of ML over LR. Improvements in methodology and reporting are needed for studies that compare modeling algorithms. © 2019 Elsevier Inc. All rights reserved.

Keywords: Clinical prediction models; Logistic regression; Machine learning; AUC; Calibration; Reporting



The University of Manchester

Causal Prediction Models



Matthew
Sperrin





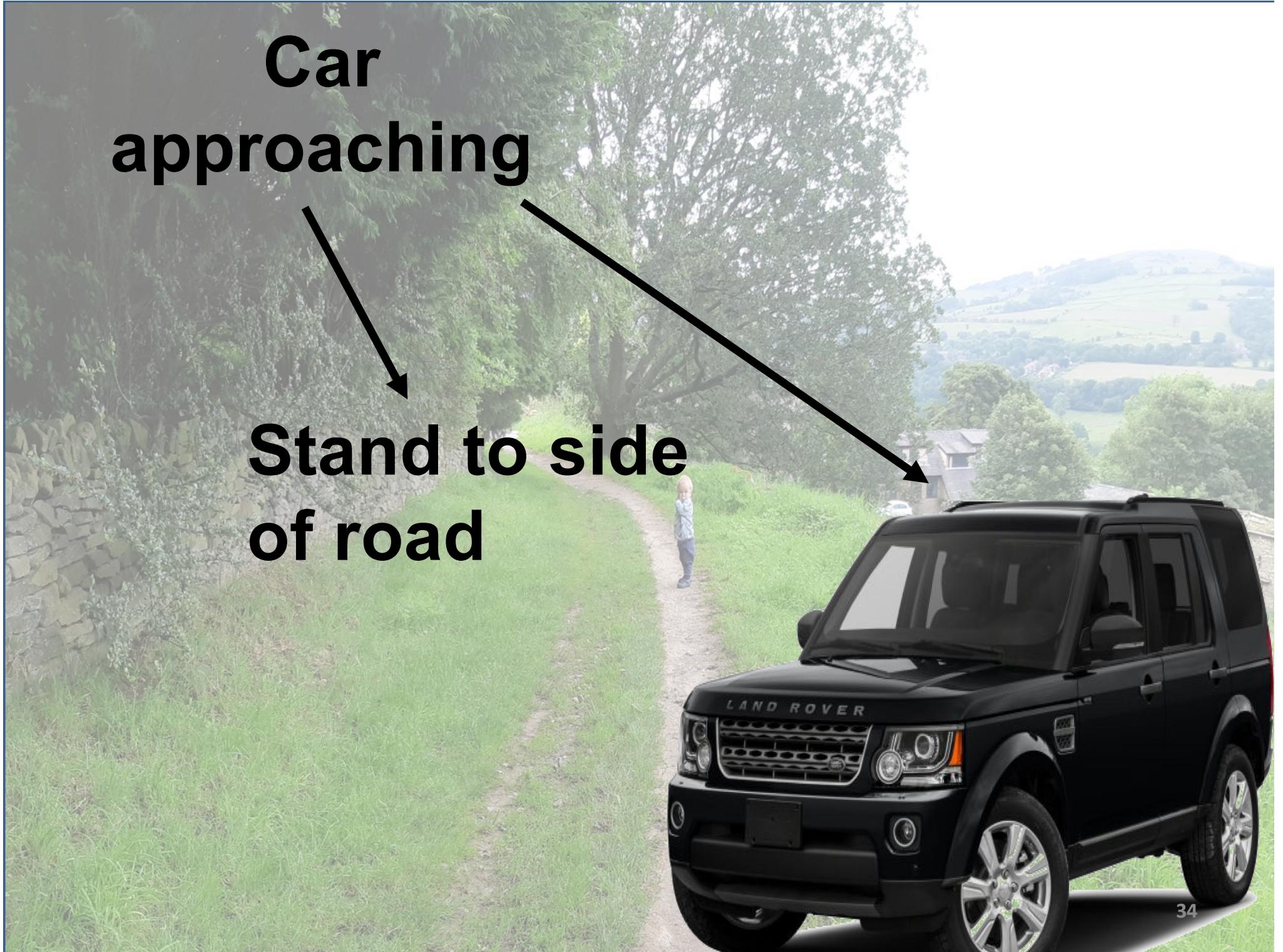


**Stand to side
of road**



**Car
approaching**

**Stand to side
of road**



Sometimes it's obvious that we should not interpret a risk model causally

- Risk of in-hospital death among patients with pneumonia (Caruana et al., 2015)
- Patients with asthma less likely to die from pneumonia
- Why?
- Existing policy: asthma patients with pneumonia go to ICU

This calculator is only valid if you do not already have a diagnosis of coronary heart disease (including angina or heart attack) or stroke/transient ischaemic attack.

[Reset](#) [Information](#) [Publications](#) [About](#) [Copyright](#) [Contact Us](#) [Algorithm](#) [Software](#) [CE](#)

About you

Age (25-84):
 Sex: Male Female
 Ethnicity: White or not stated
 UK postcode: leave blank if unknown

Clinical information

Smoking status:
 Diabetes status:
 Angina or heart attack in a 1st degree relative < 60?
 Chronic kidney disease (stage 3, 4 or 5)?
 Atrial fibrillation?
 On blood pressure treatment?
 Do you have migraines?
 Rheumatoid arthritis?
 Systemic lupus erythematosus (SLE)?
 Severe mental illness?
 (this includes schizophrenia, bipolar disorder and moderate/severe depression)
 On atypical antipsychotic medication?
 Are you on regular steroid tablets?
 A diagnosis of or treatment for erectile dysfunction?
 Leave blank if unknown

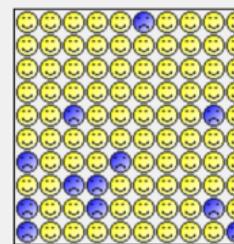
 Systolic blood pressure (mmHg):
 Standard deviation of at least two most recent systolic blood pressure readings (mmHg):
 Body mass index
 Height (cm):
 Weight (kg):

Your results

Your risk of having a heart attack or stroke within the next 10 years is:

12.4%

In other words, in a crowd of 100 people with the same risk factors as you, 12 are likely to have a heart attack or stroke within the next 10 years.



Risk of
a heart attack or stroke

Your score has been calculated using estimated data, as some information was left blank.

Your body mass index was calculated as 29.98 kg/m².

How does your 10-year score compare?

Your score	
Your 10-year QRISK®3 score	12.4%
The score of a healthy person with the same age, sex, and ethnicity*	3.8%
Relative risk**	3.3
Your QRISK®3 Healthy Heart Age***	66

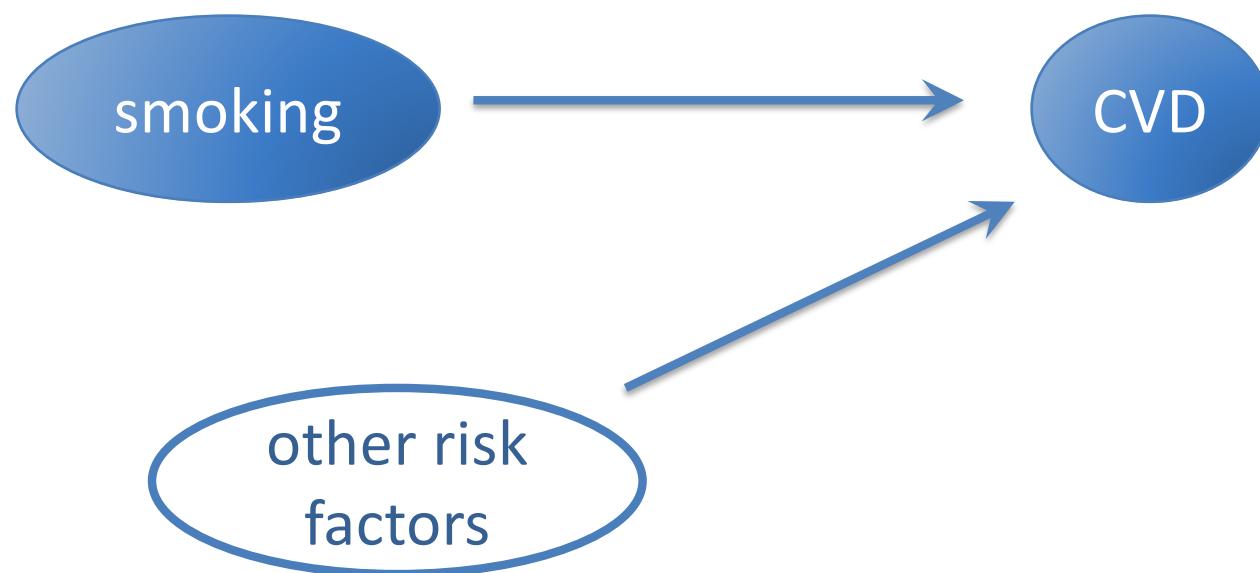
* This is the score of a healthy person of your age, sex and ethnic group, i.e. with no adverse clinical indicators and a cholesterol ratio of 4.0, a stable systolic blood pressure of 125, and BMI of 25.

** Your relative risk is your risk divided by the healthy person's risk.

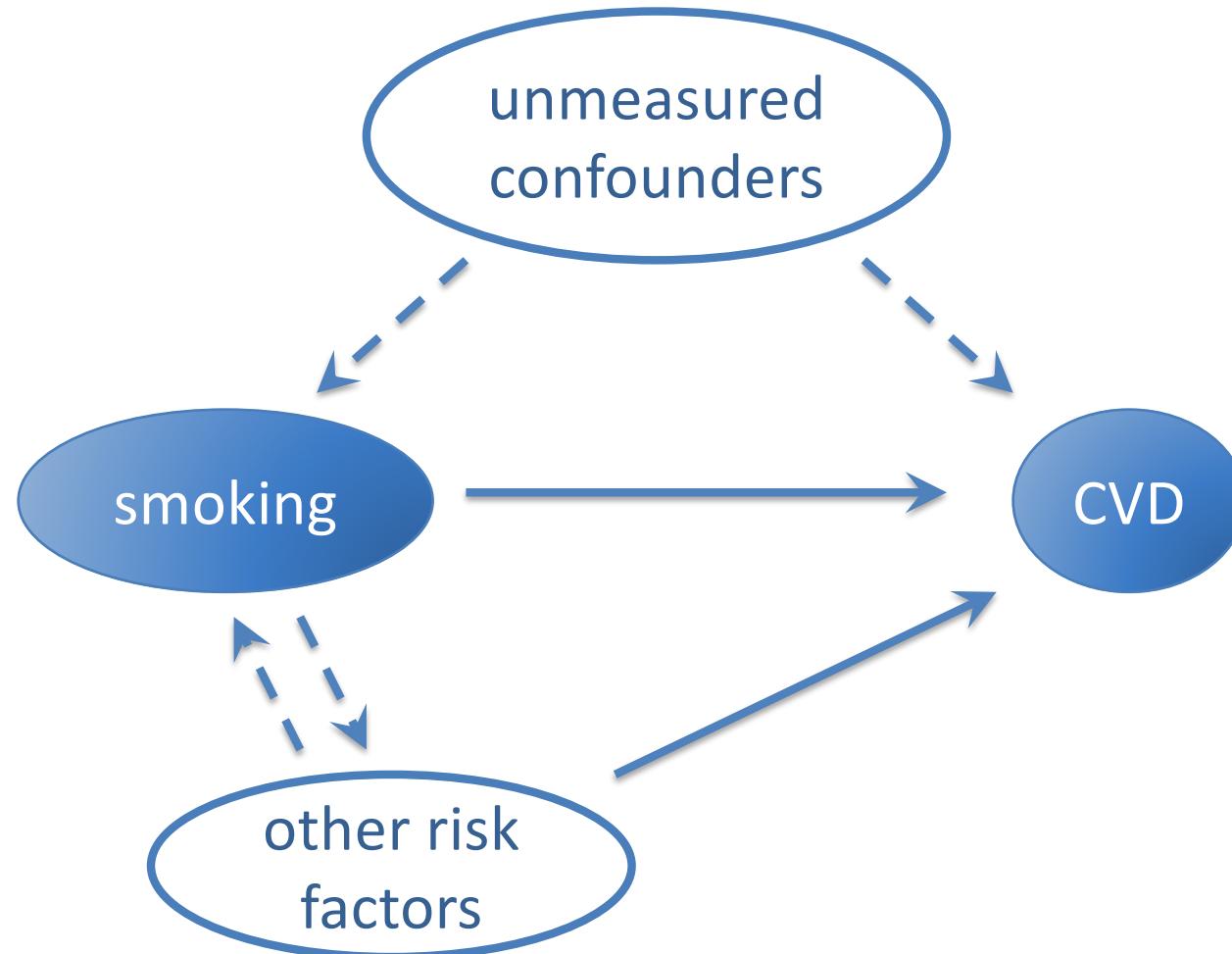
*** Your QRISK®3 Healthy Heart Age is the age at which a healthy person of your sex and ethnicity has your 10-year QRISK®3 score.

[Calculate risk](#)

Model structure



Causal structure



REVIEW

Open Access



Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies

Romin Pajouheshnia^{1*} , Johanna A. A. G. Damen^{1,2}, Rolf H. H. Groenwold¹, Karel G. M. Moons^{1,2} and Linda M. Peelen¹

Abstract

Background: Ignoring treatments in prognostic model development or validation can affect the accuracy and transportability of models. We aim to quantify the extent to which the effects of treatment have been addressed in existing prognostic model research and provide recommendations for the handling and reporting of treatment use in future studies.

Methods: We first describe how and when the use of treatments by individuals in a prognostic study can influence the development or validation of a prognostic model. We subsequently conducted a systematic review of the handling and reporting of treatment use in prognostic model studies in cardiovascular medicine. Data on treatment use (e.g. medications, surgeries, lifestyle interventions), the timing of their use, and the handling of such treatment use in the analyses were extracted and summarised.

Results: Three hundred two articles were included in the review. Treatment use was not mentioned in 91 (30%) articles. One hundred forty-six (48%) reported specific information about treatment use in their studies; 78 (26%) provided information about multiple treatments. Three articles (1%) reported changes in medication use ("treatment drop-in") during follow-up. Seventy-nine articles (26%) excluded treated individuals from their analysis, 80 articles (26%) modelled treatment as an outcome, and of the 155 articles that developed a model, 86 (55%) modelled treatment use, almost exclusively at baseline, as a predictor.

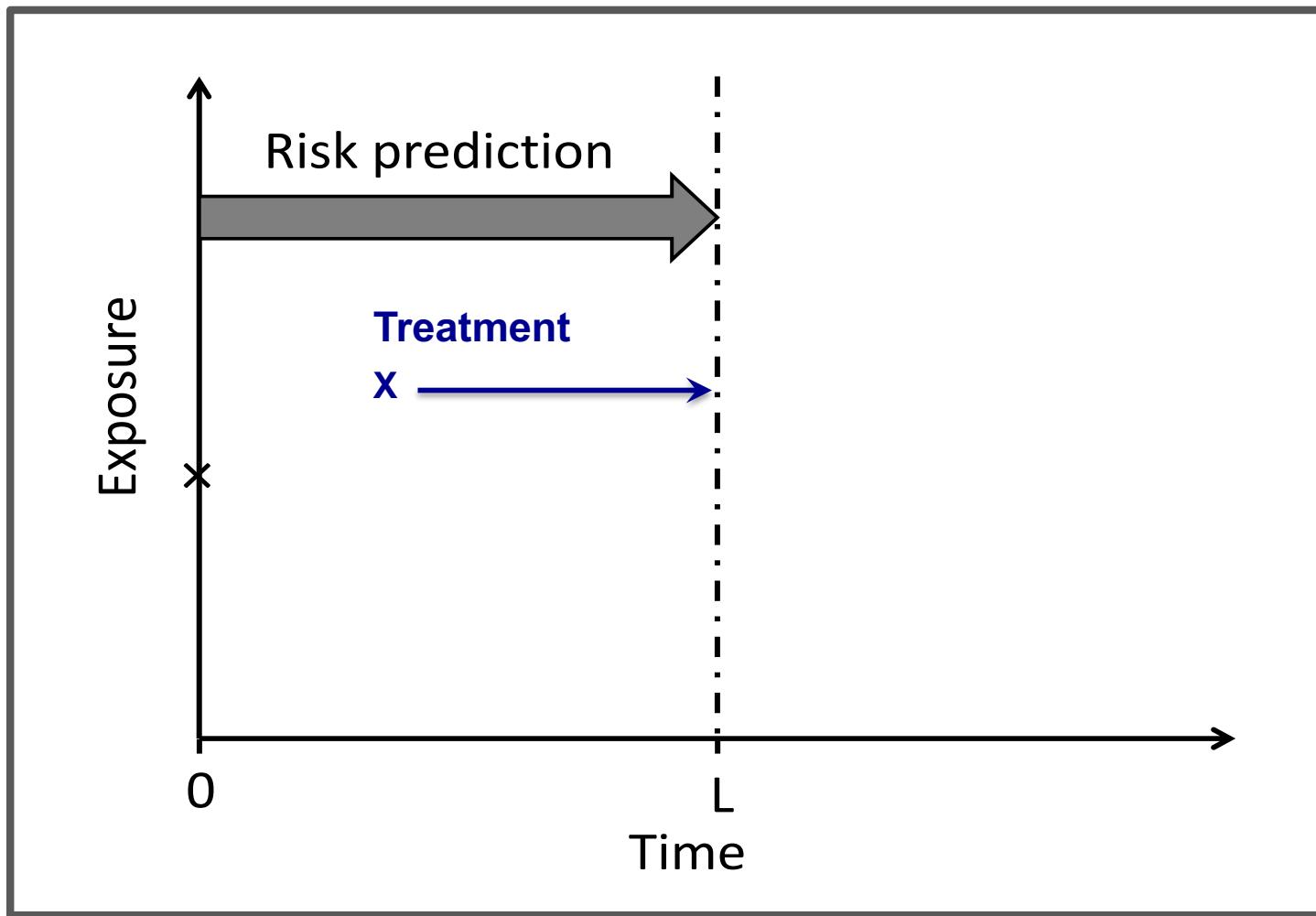
Conclusions: The use of treatments has been partly considered by the majority of CVD prognostic model studies. Detailed accounts including, for example, information on treatment drop-in were rare. Where relevant, the use of treatments should be considered in the analysis of prognostic model studies, particularly when a prognostic model is designed to guide the use of certain treatments and these treatments have been used by the study participants. Future prognostic model studies should clearly report the use of treatments by study participants and consider the potential impact of treatment use on the study findings.

So ..

- Published prediction model studies largely neglect treatments / interventions
- Existing risk prediction methods are not suitable for counterfactual causal reasoning
- Yet users are keen to explore “what if” scenarios

Risk prediction methods should be extended with methods for counterfactual causal reasoning

“Drop-in” treatment (training dataset)



Hari Seldon, QRISK3, and the Prediction Paradox

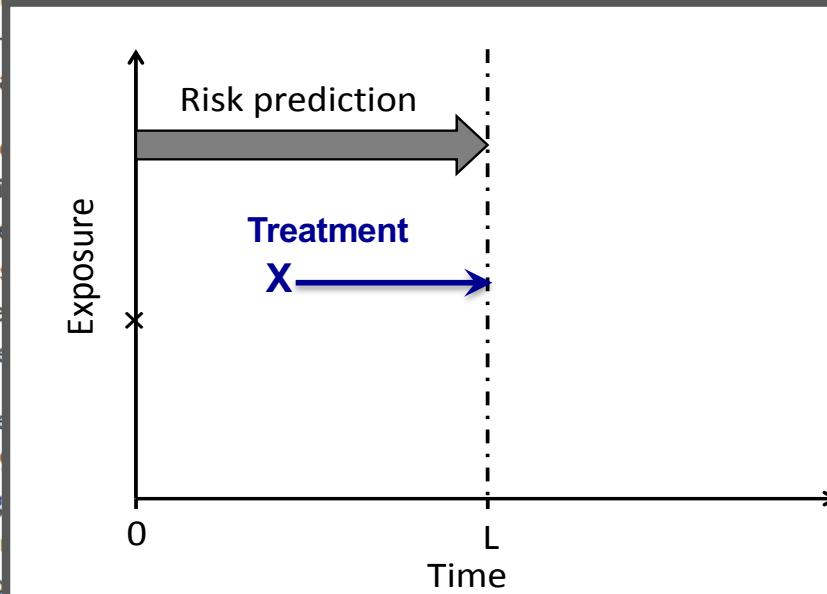
Hari Seldon was one of Isaac Asimov's science fiction characters, famed for developing Psychohistory, an algorithmic way to predict society's future through statistical 'laws' derived from 'big data'. Using his algorithms, Seldon predicted the future of the Galactic Empire, provided that two conditions were met. Firstly, the population whose behaviour was modeled had to be sufficiently large; secondly, citizens should not be told the results of their psychohistorical analyses so as to prevent "Prediction Paradox" (predictions influencing behaviours that in turn invalidate predictions). So, Seldon has become a cautionary icon of Big Data research [1].

In the real world, 'big data' are widely used to predict the risks of adverse health events over the life courses of patients. The risk models are typically developed using data from dedicated cohort studies (e.g. Framingham [2]) or naturalistic cohorts derived from electronic health records (e.g. QRISK from QResearch [3–5]). Such models are used to support decisions about: the care of individual patients; the management and funding of healthcare systems; and the prevention of disease in populations.

Last month witnessed the publication of QRISK3, the third in a series of cardiovascular risk prediction algorithms [5]. The first QRISK model was published in 2007 and was followed by an updated model (QRISK2) in 2008 which included additional risk factors. Since then, QRISK2 has been updated annually and recalibrated to the latest version of the QResearch database. QRISK2 is used across England's health service (NHS) and the National Health Service of Scotland.

The newly described disease; minor mental illness; larger data sets. While all new improvements

Patients were in January 1998. Interesting cardiovascular introduction



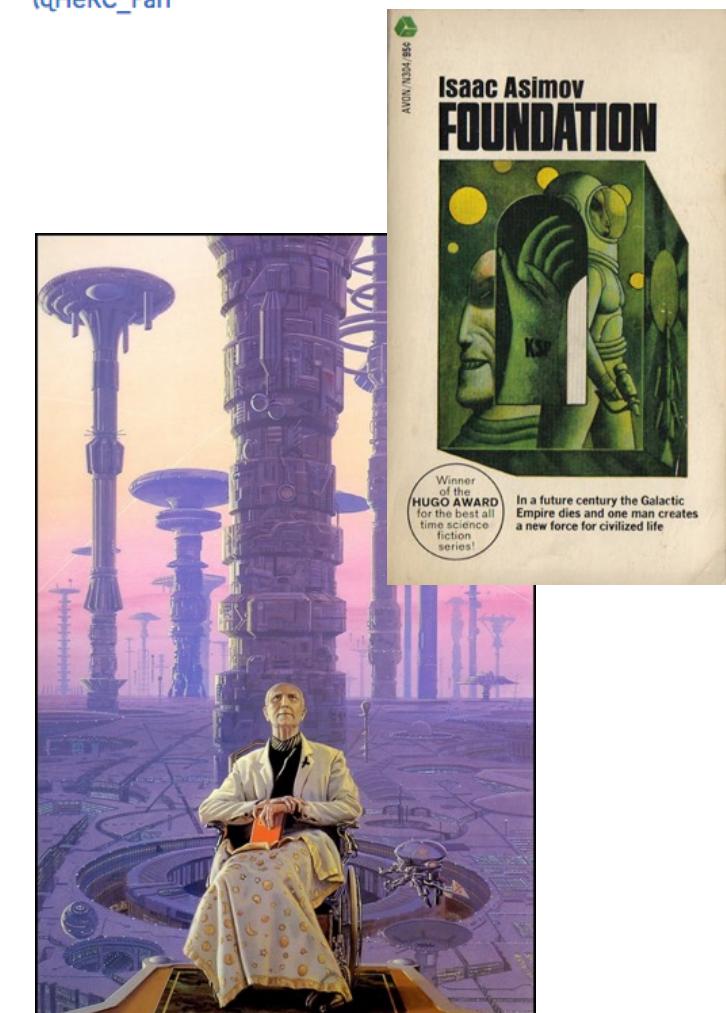
guidance from the National Institute for Clinical Excellence; severe renal disease; derived from general practices. Between 10 and 20% of secondary care; social; financial

08 June 2017

Niels Peek

Professor of Health Informatics

Matthew Sperrin, Mamas Mamas, Tjeerd van Staa, Iain Buchan
Health e-Research Centre, The University of Manchester
Vaughan House, Portsmouth St, Manchester M13 9GB, UK
[@HeRC_Farr](#)

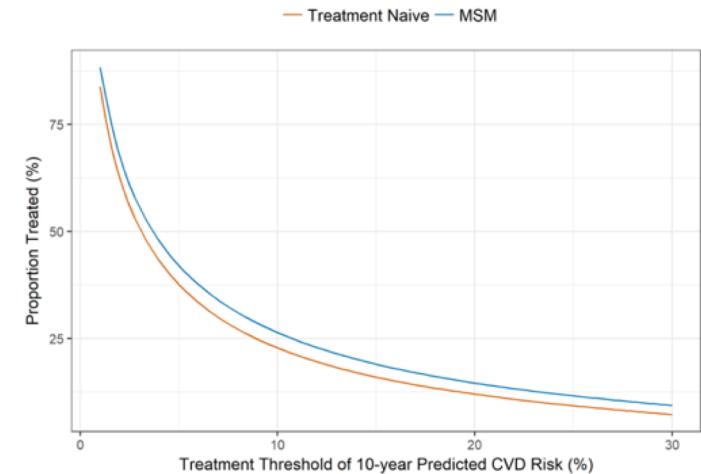
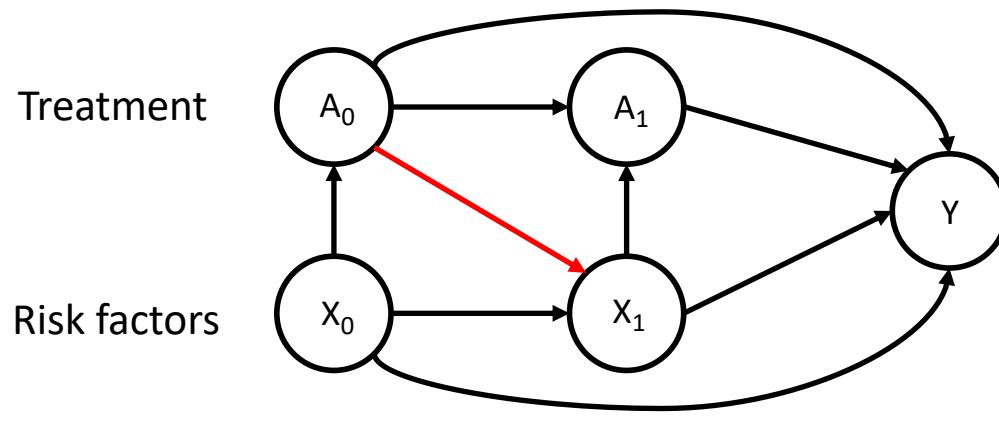


Adjusting for “drop-in” treatment



Matthew
Sperrin

- Problem: modelling change in treatment after baseline



- Failure to adjust for this leads to underestimate of risk
- Preliminary results based on simulation show that an additional 4% of population would be considered for statins after correcting for this

Sperrin et al. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Statistics in Medicine; 2018: 1-13.

Time series deconfounder



Mihaela van der
Schaar

Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders

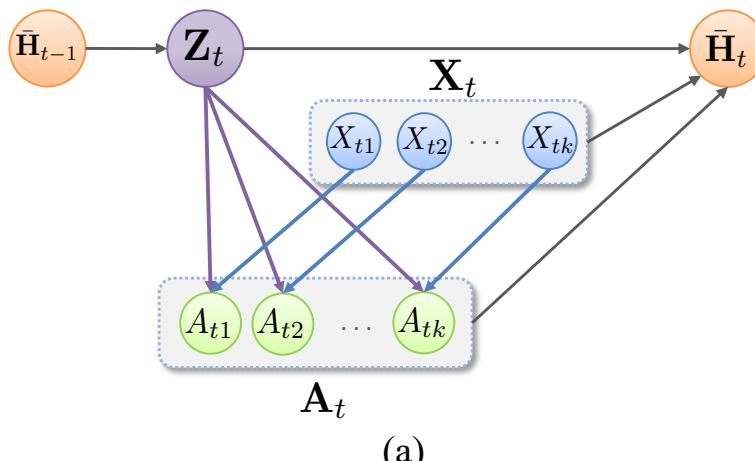
Ioana Bica^{1,2} Ahmed M. Alaa³ Mihaela van der Schaar^{2,3,4}

Abstract

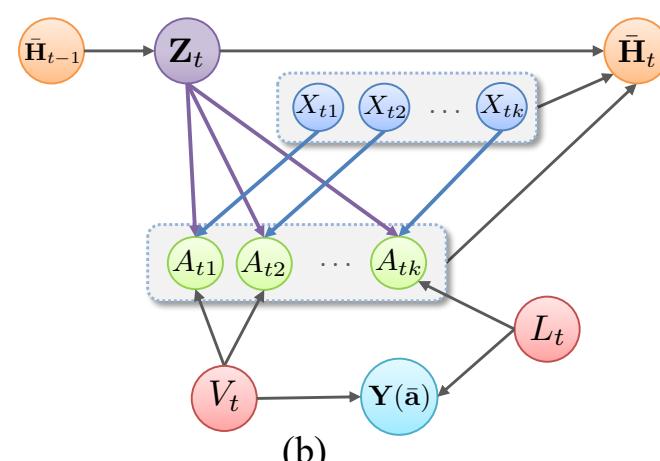
The estimation of treatment effects is a pervasive problem in medicine. Existing methods for estimating treatment effects from longitudinal observational data assume that there are no hidden confounders. This assumption is not testable in practice and, if it does not hold, leads to biased estimates. In this paper, we develop the Time Se-

methods assume that all confounders — variables affecting both the treatment assignment and the potential outcomes — are observed, an assumption which is not testable in practice¹ and probably not true in many circumstances.

To understand why the presence of hidden confounders introduces bias, consider the problem of estimating treatment effects for patients diagnosed with cancer. They are often



(a)



(b)

¹Department of Engineering Science, University of Oxford, Oxford, United Kingdom ²The Alan Turing Institute, London, United Kingdom ³Department of Electrical Engineering, UCLA, Los Angeles, USA ⁴University of Cambridge, Cambridge, United Kingdom. Correspondence to: Ioana Bica <ioana.bica@eng.ox.ac.uk>.

In this paper, we propose the Time Series Deconfounder, a method that enables the unbiased estimation of treatment responses *over time* in the presence of hidden confounders, by taking advantage of the potential sequential assignment of multiple treatments. We draw from the main idea in

¹Since counterfactuals are never observed, it is not possible to test for the existence of hidden confounders that could affect them.



The University of Manchester

Prediction with Electronic Healthcare Records

Electronic healthcare record data

- Recordings are driven by transactions with the healthcare system
- Variation in recording practice
 - Variable follow-up times
- Both structured (“coded”) and non-structured data (text, images)
- Only positive diagnostic codes
 - there is no code for “no diabetes”
- Meaningful events are often not explicitly recorded
 - e.g. stopping medication

Biases in electronic health record data due to processes within the healthcare system: retrospective observational study

Denis Agniel,¹ Isaac S Kohane,^{1,2} Griffin M Weber^{1,3}

¹Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St, Boston, MA 02115, USA

²Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

³Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

Correspondence to: G M Weber
weber@hms.harvard.edu
Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2018;360:k1479
<http://dx.doi.org/10.1136/bmj.k1479>

Accepted: 13 March 2018

ABSTRACT

OBJECTIVE

To evaluate on a large scale, across 272 common types of laboratory tests, the impact of healthcare processes on the predictive value of electronic health record (EHR) data.

DESIGN

Retrospective observational study.

SETTING

Two large hospitals in Boston, Massachusetts, with inpatient, emergency, and ambulatory care.

PARTICIPANTS

All 669 452 patients treated at the two hospitals over one year between 2005 and 2006.

MAIN OUTCOME MEASURES

The relative predictive accuracy of each laboratory test for three year survival, using the time of the day, day of the week, and ordering frequency of the test, compared to the value of the test result.

RESULTS

The presence of a laboratory test order, regardless of any other information about the test result, has a significant association ($P<0.001$) with the odds of survival in 233 of 272 (86%) tests. Data about the timing of when laboratory tests were ordered were more accurate than the test results in predicting survival in 118 of 174 tests (68%).

CONCLUSIONS

Healthcare processes must be addressed and accounted for in analysis of observational health data. Without careful consideration to context, EHR data are unsuitable for many research questions. However, if explicitly modeled, the same processes that make EHR data complex can be leveraged to gain insight into patients' state of health.

Introduction

Rapid progress is being made towards the adoption and use of electronic health record (EHR) systems, resulting in massive amounts of data being generated through

the routine delivery of healthcare.^{1–3} This, in turn, is transforming biomedical research, as individuals who now have access through informatics to their EHRs,^{4–7} link to genomic data,^{8,9} and scale up studies.¹⁰

However, there is a lack of understanding of how the naive use of Big Data can lead to erroneous findings. A full understanding of EHR data is required to interpret these findings.

An easily overlooked source of bias in observational data is the health of the patient, which is often associated with the health record. A physician made first developed the code used for recording the reason for the visit, which was elevated white blood cell count. It is known unless the test is a reference test. Hripcak and colleagues¹¹ proposed a process model, where indirect measures are recorded during the recording process.

The recording of patient factors, such as clinical tests and treatments, by provider and patient, and in that they vary by standards of care, changing population, separate studies, patient data from the following: how they report increased from by practice and incentives to screen the number of interactions. The interactions may be complex, as explained by seeking to explain during weekend the weekend effect by country. For example, antigen testing is higher in Asia,²¹ as implemented in China, use of unnecessary tests too. Dozens of studies show cancer who live in

Journal of the American Medical Informatics Association, 0(0), 2019, 1–9

doi: 10.1093/jamia/ocz148
Research and Applications



OXFORD

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz148/5573796 by University of Manchester user on 13 July 2019

Research and Applications

How and when informative visit processes can bias inference when using electronic health records data for clinical research

Benjamin A Goldstein,^{1,2} Matthew Phelan,² Neha J Pagidipati,^{2,3} and Sarah B Peskoe¹

¹Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA, ²Center for Predictive Medicine, Duke Clinical Research Institute, Durham, North Carolina, USA, and ³Department of Medicine, Duke University, Durham, North Carolina, USA

Corresponding Author: Benjamin A Goldstein, PhD, MPH, Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC 27705, USA; ben.goldstein@duke.edu

Received 13 March 2019; Revised 16 July 2019; Editorial Decision 21 July 2019; Accepted 23 July 2019

ABSTRACT

Objective: Electronic health records (EHR) data have become a central data source for clinical research. One concern for using EHR data is that the process through which individuals engage with the health system, and find themselves within EHR data, can be informative. We have termed this process informed presence. In this study we use simulation and real data to assess how the informed presence can impact inference.

Materials and Methods: We first simulated a visit process where a series of biomarkers were observed informatively and uninformatively over time. We further compared inference derived from a randomized control trial (ie, uninformative visits) and EHR data (ie, potentially informative visits).

Results: We find that only when there is both a strong association between the biomarker and the outcome as well as the biomarker and the visit process is there bias. Moreover, once there are some uninformative visits this bias is mitigated. In the data example we find, that when the “true” associations are null, there is no observed bias.

Discussion: These results suggest that an informative visit process can exaggerate an association but cannot induce one. Furthermore, careful study design can mitigate the potential bias when some noninformative visits are included.

Conclusions: While there are legitimate concerns regarding biases that “messy” EHR data may induce, the conditions for such biases are extreme and can be accounted for.

Key words: misclassification, electronic health records

WHAT IS ALREADY KNOWN ON THIS TOPIC

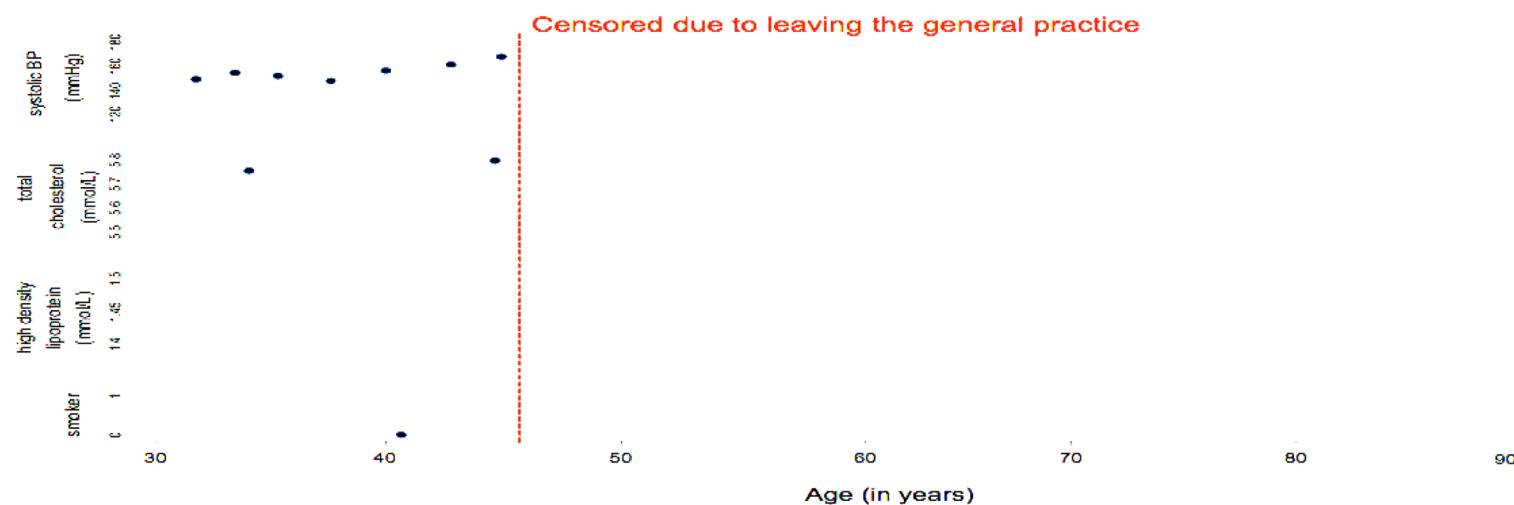
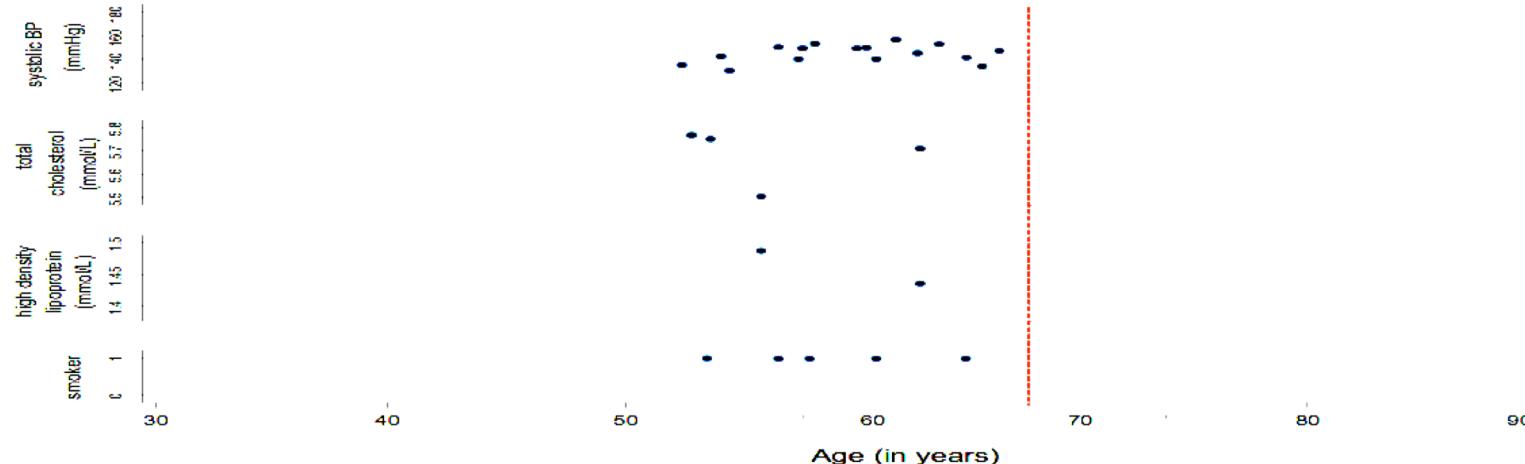
Dynamic processes within the healthcare system, such as the hours when clinics are open and when patients are scheduled to be seen, leave an imprint on electronic health record data.

WHAT THIS STUDY ADDS

An evaluation of using the effects of healthcare processes on 272 laboratory tests to predict three year survival in the full patient populations seen over a year at two large hospitals.

The hour of the day the test was ordered, the day of the week, and the amount of time between consecutive tests is more predictive of three year survival than the actual value of the test result, for most tests.

Informative observations



From a presentation by Jessica Barrett, University of Cambridge

Prediction with electronic health records

Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review

Benjamin A Goldstein^{1,2}, Ann Marie Navar^{2,3}, Michael J Pencina^{1,2}, John PA Ioannidis^{4,5}

ABSTRACT

Objective Electronic health records (EHRs) offer opportunities and challenges. We sought to evaluate prediction studies using EHR data.

Methods We searched PubMed for articles extracted by two reviewers, and we abstracted and supplemented documentation.

Results We identified 107 articles from 15 diverse array of predictors. Most used validated studies did not fully leverage the breadth of predictor variables (median = 27 variables). Less than half of the studies addressed mortality (0.84), clinical prediction (0.83), hospitalizations (0.63).

Conclusions EHR data present both opportunities and challenges for prediction studies.

RECEIVED 27 October 2015
REVISED 25 January 2016
ACCEPTED 20 February 2016



OXFORD
UNIVERSITY PRESS

- 107 studies (6 years)
- Most (n=70) studies did not consider repeated measurements.
- Only 58 studies assessed missingness, with the most common strategy being multiple imputation.
- No study assessed the role informative observations may play.
- Only 12 studies assessed the role of outcome censoring and loss to follow-up.

Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design

Eleanor M Pullenayegum¹ and Lily SH Lim²

Statistical Methods in Medical Research

2016, Vol. 25(6) 2992–3014

© The Author(s) 2014

Reprints and permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: [10.1177/0962280214536537](https://doi.org/10.1177/0962280214536537)

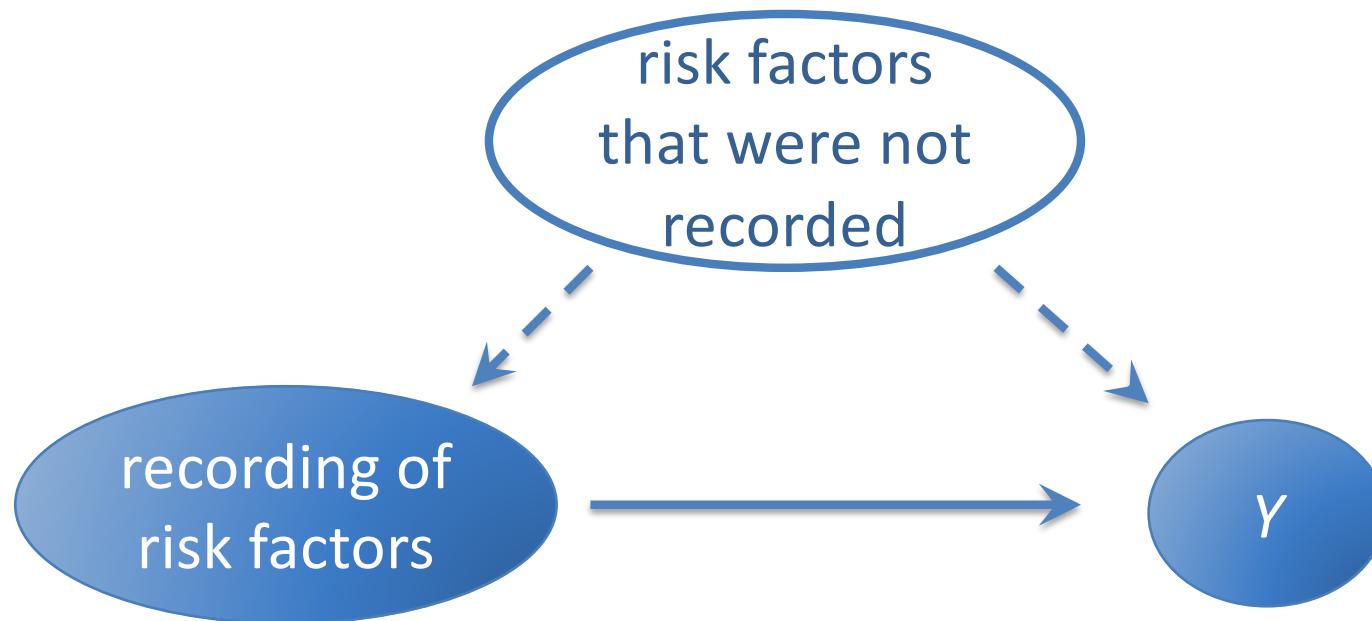
smm.sagepub.com



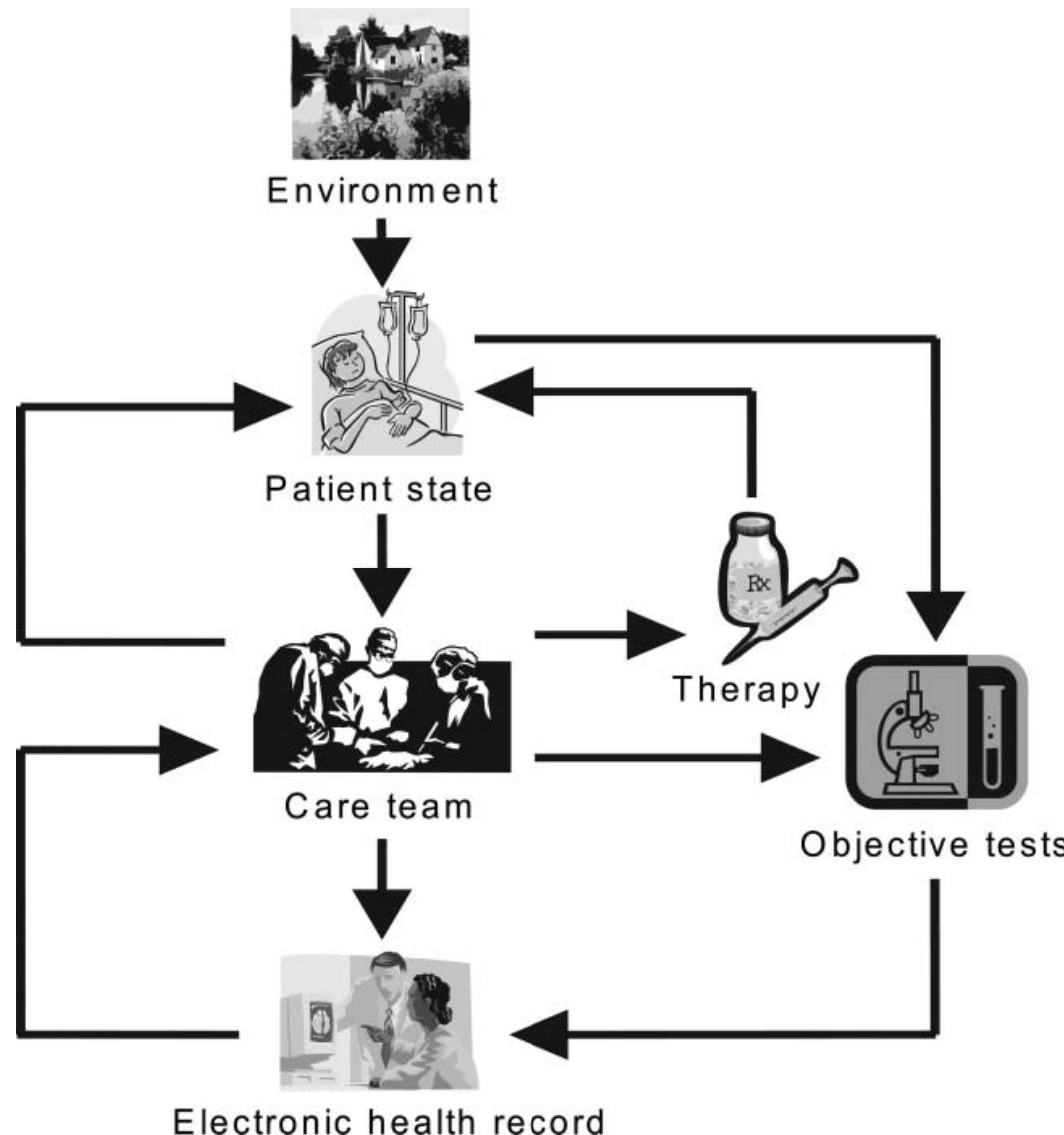
Abstract

When data are collected longitudinally, measurement times often vary among patients. This is of particular concern in clinic-based studies, for example retrospective chart reviews. Here, typically no two patients will share the same set of measurement times and moreover, it is likely that the timing of the measurements is associated with disease course; for example, patients may visit more often when unwell. While there are statistical methods that can help overcome the resulting bias, these make assumptions about the nature of the dependence between visit times and outcome processes, and the assumptions differ across methods. The purpose of this paper is to review the methods available with a particular focus on how the assumptions made line up with visit processes encountered in practice. Through this we show that no one method can handle all plausible visit scenarios and suggest that careful analysis of the visit process should inform the choice of analytic method for the outcomes. Moreover, there are some commonly encountered visit scenarios that are not handled well by any method, and we make recommendations with regard to study design that would minimize the chances of these problematic visit scenarios arising.

Causal structure



Studying the “physics” of the EHR



Hripcsak G, Albers DJ.
J Am Med Inform Assoc 2013;20:117–121.

Some further challenges

- Risk prediction using longitudinal data
Incorporating each patient's entire medical history
- Dynamic risk prediction
Making multiple predictions over time for the same individual
- Calibration drift
Adjusting to changes in the population and in medical practice over time
- Competing risks and multi-state prediction
Modelling multiple outcomes and intermediate states



The University of Manchester

Summary & Conclusion

This talk

1. The Data Revolution in Health
2. Predictive Analytics
3. Is Machine Learning better than Statistical Modelling?
4. Causal Prediction Models
5. Prediction with Electronic Healthcare Records

Summary & Conclusion

- Most (all?) prediction models that are used in healthcare practice are based on **logistic regression** or related methods
- There is **no evidence** that ML yields better predictive performance than statistical models (on structured data)
- There is a need for prediction models can that reason about the effects of interventions using **causal inference**
- Electronic healthcare records reflect **informative observation processes** that can have predictive value in themselves
- There are **further challenges** related to longitudinal data, dynamic prediction, calibration drift, and competing risks

Thank you to my collaborators

- Dr Matthew Sperrin
- Dr Glen Martin
- Ms Rose Sisk
- Dr Lijing Lin
- Mr David Jenkins
- Dr William Hulme
- Dr Jessica Barrett
- Prof Iain Buchan
- Prof Kenneth Muir
- Prof Tjeerd van Staa

Thank you

Niels Peek
Centre for Health Informatics
The University of Manchester, UK
 niels.peek@manchester.ac.uk
 @NielsPeek

