Machine learning and healthcare professionals' job

Decision making and role of prediction

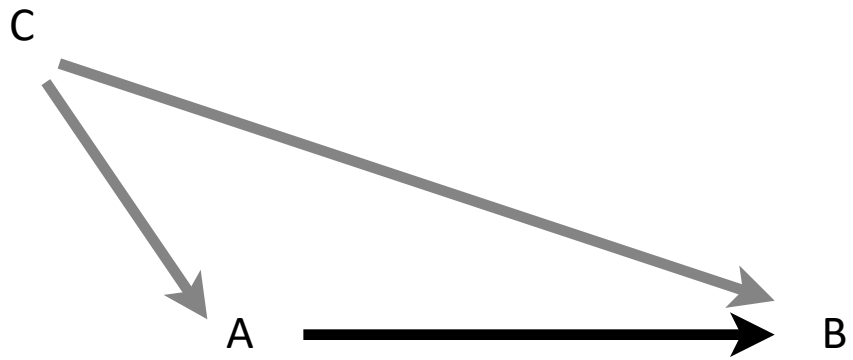
DEEP MEDICINE

UNIVERSITY OF OXFORD

- **Diagnosis** (predict what problem the patient currently has – pattern recognition)
- **Prognosis** (predict the course of disease for a patient – forecasting)
- **Intervention** (predict how a patient will respond to an intervention – causal prediction)

- Prediction is the core of decision making (which also requires judgement)
- Machine learning is all about prediction

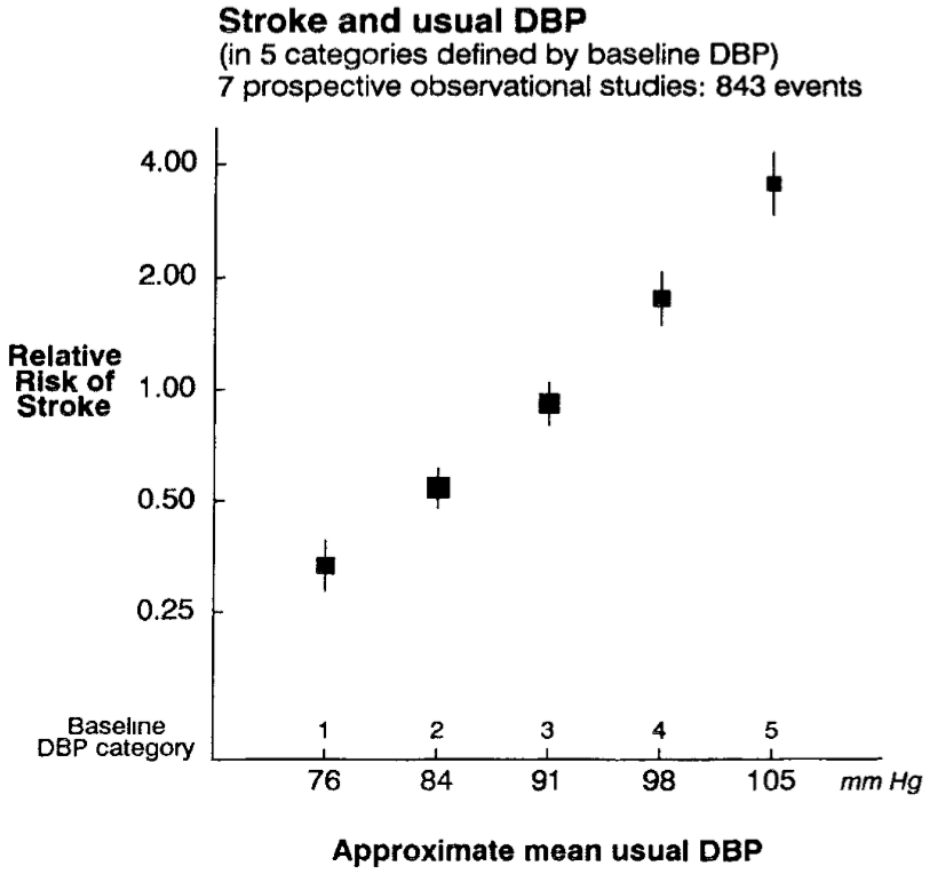
Causation in epidemiological terms



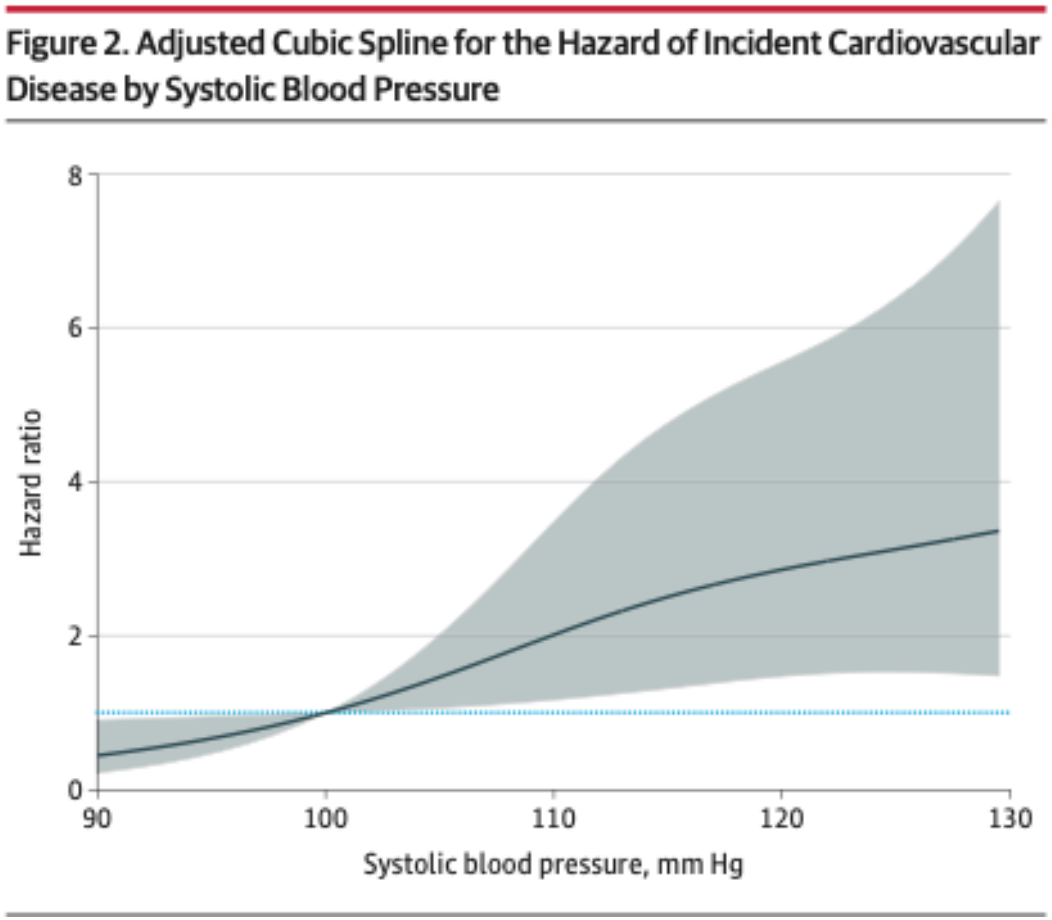
A: Exposure, e.g., blood pressure, or a drug
B: Outcome, e.g., cardiovascular disease
C: Confounder (common cause), e.g., body weight, or another condition or drug

Evidence from cohort studies with conventional statistical models

People with the lowest blood pressure have the lowest risk of cardiovascular disease



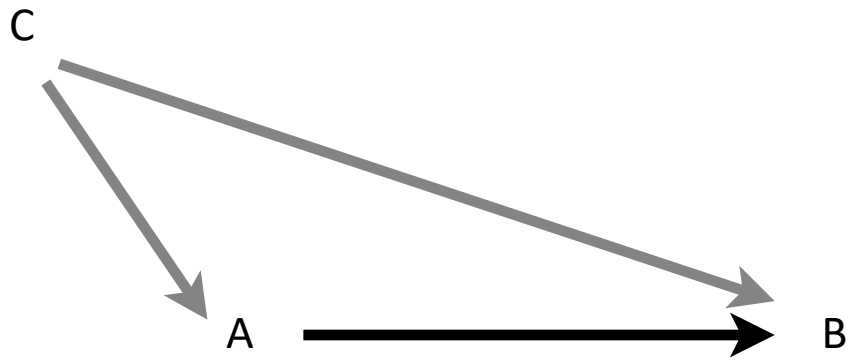
MacMahon et al. Lancet 1990



Whelton SP et al. *JAMA Cardio* 2020

Residual or uncontrolled confounding?

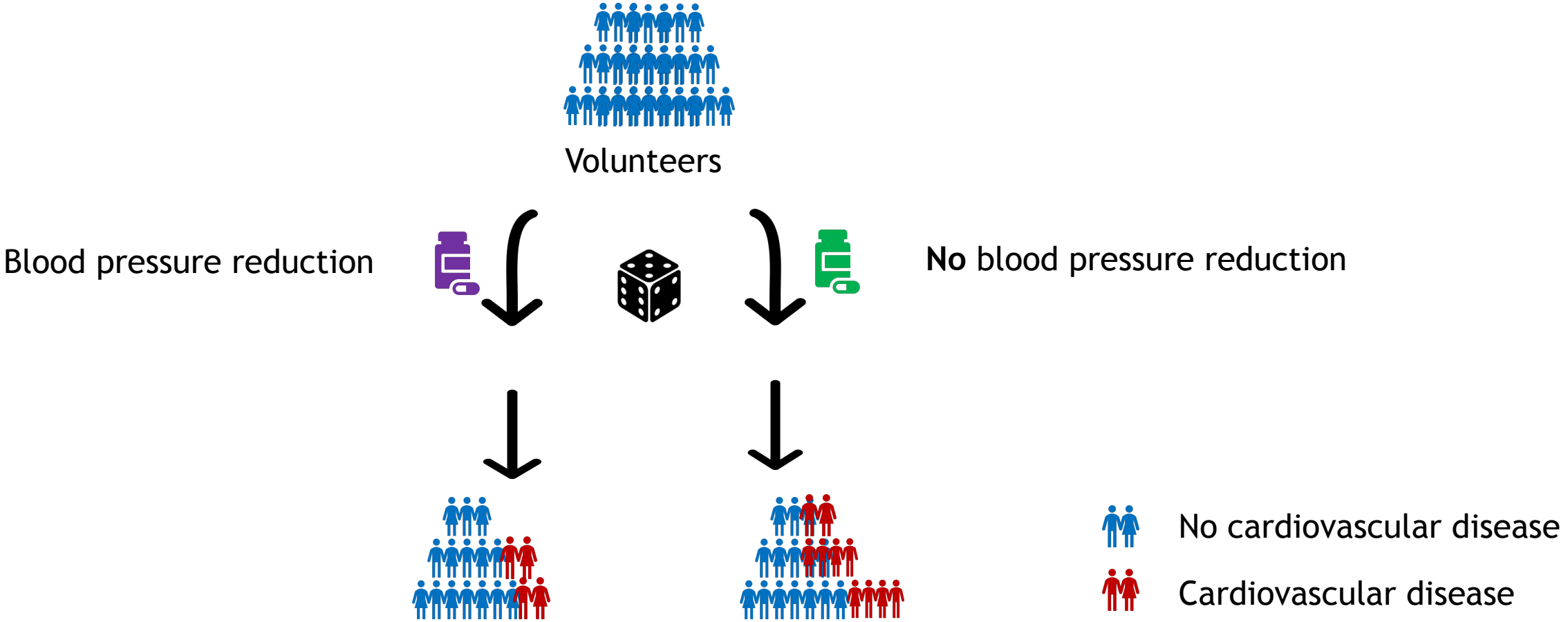
Are all confounders known, observed (measured), and included in the models?
I.e., are we comparing like with like?



A: Exposure, e.g., blood pressure, or a drug
B: Outcome, e.g., cardiovascular disease
C: Confounder (common cause), e.g., body weight, or another condition or drug

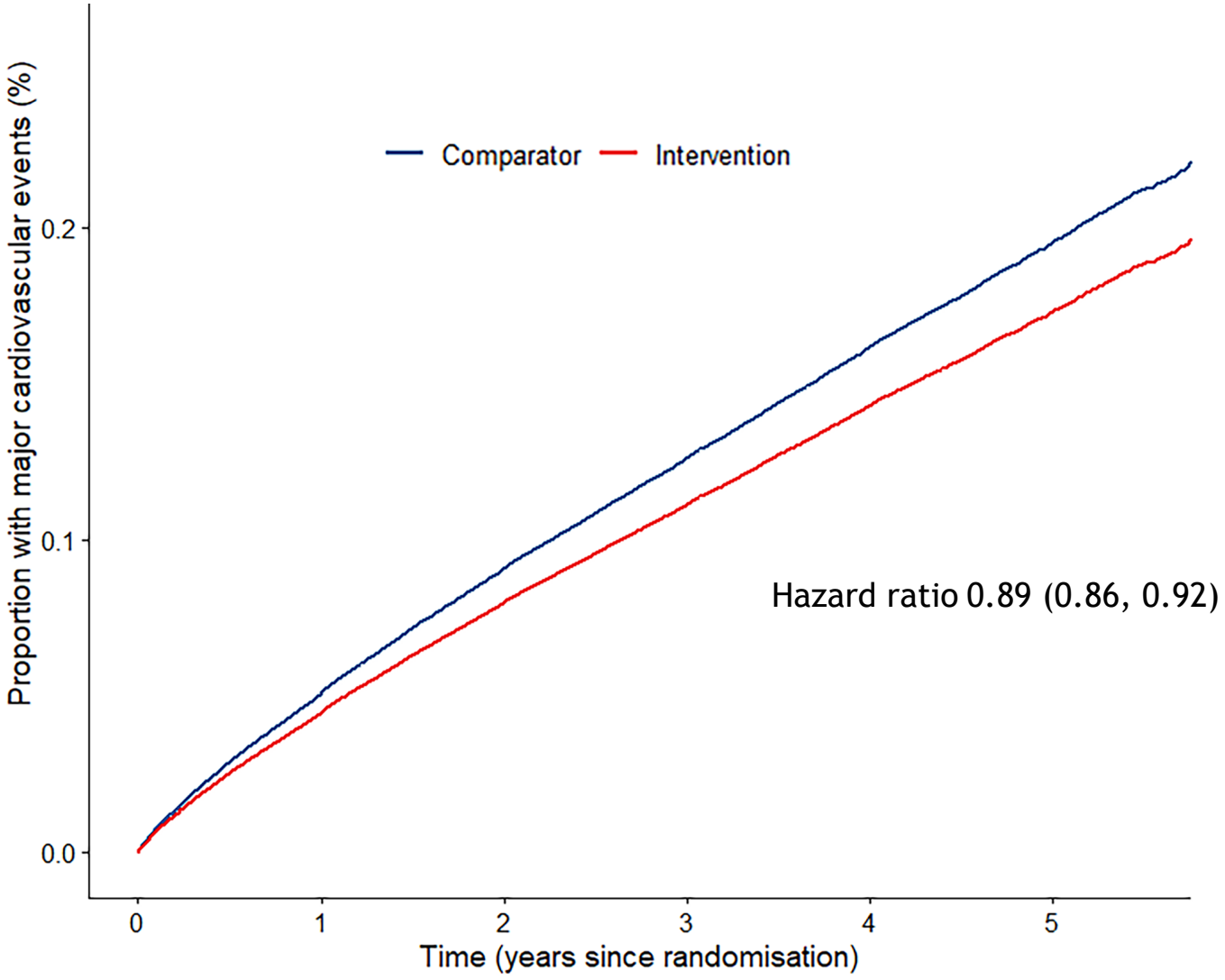
Randomised controlled trials (RCTs)

Causation that takes account of known & unknown confounding



Randomised controlled trials (RCTs)

Causation that takes account of known & unknown confounding



Trials are expensive, time-consuming
Sometimes unethical or infeasible
Under-representative of certain groups

BPLTTC *Lancet* 2021

Can use of EHR (routine, administrative data) help?

Historical use of administrative data

Elevated blood pressure and risk of death

First reports of associations between elevated blood pressure and death:

- Life insurance data in 1910!
- 20,000 insured individuals in the first report, subsequently extended to >10M individuals

Conclusion: “mortality increases rapidly with the increase in blood pressure over the average”

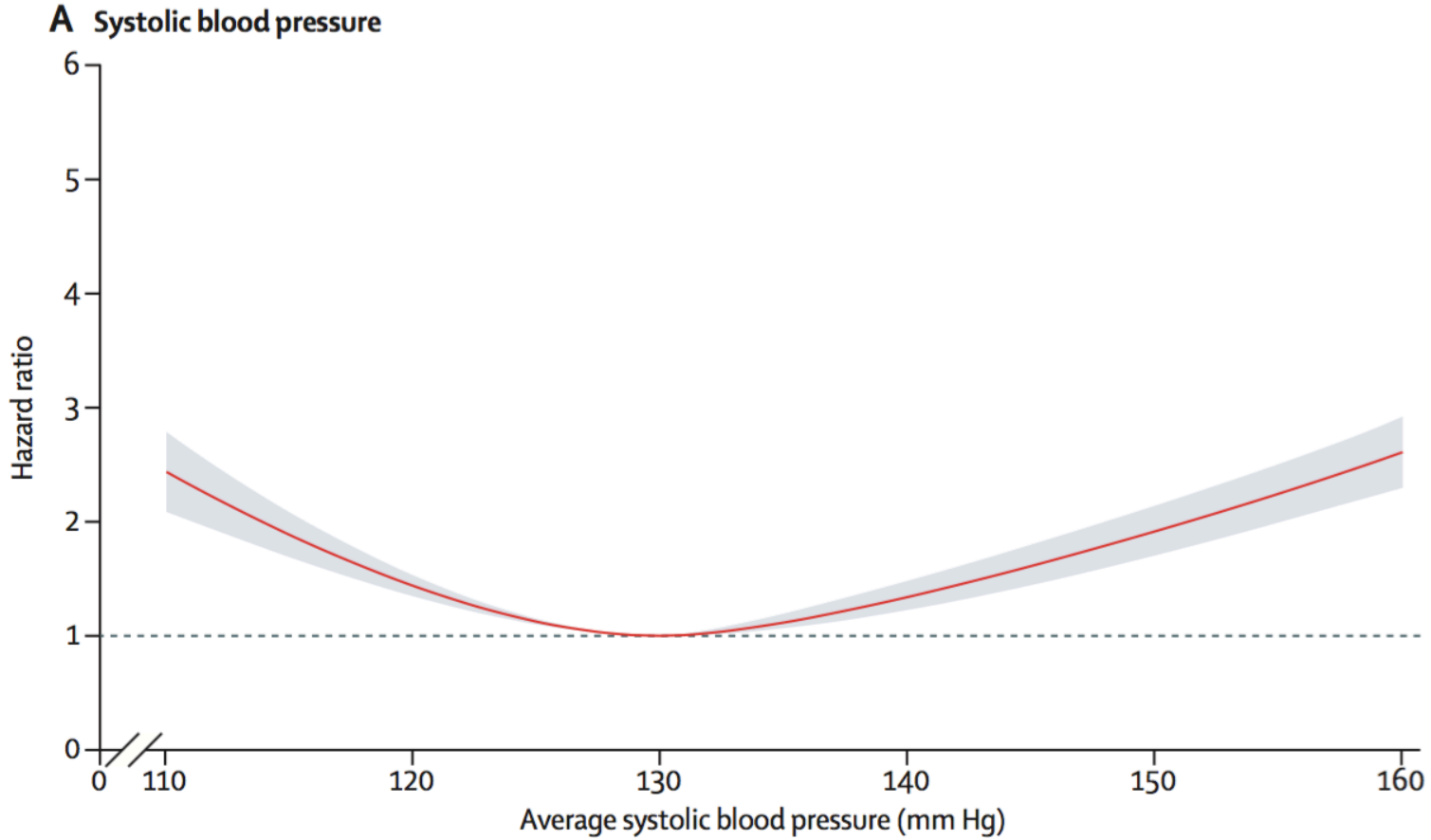
Society of Actuaries. Blood Pressure: Report of the Joint Committee on Mortality of the Association of Life Insurance Medical Directors and the Actuarial Society of America. New York, NY: 1925.

Can use of EHR (routine, administrative data) help?

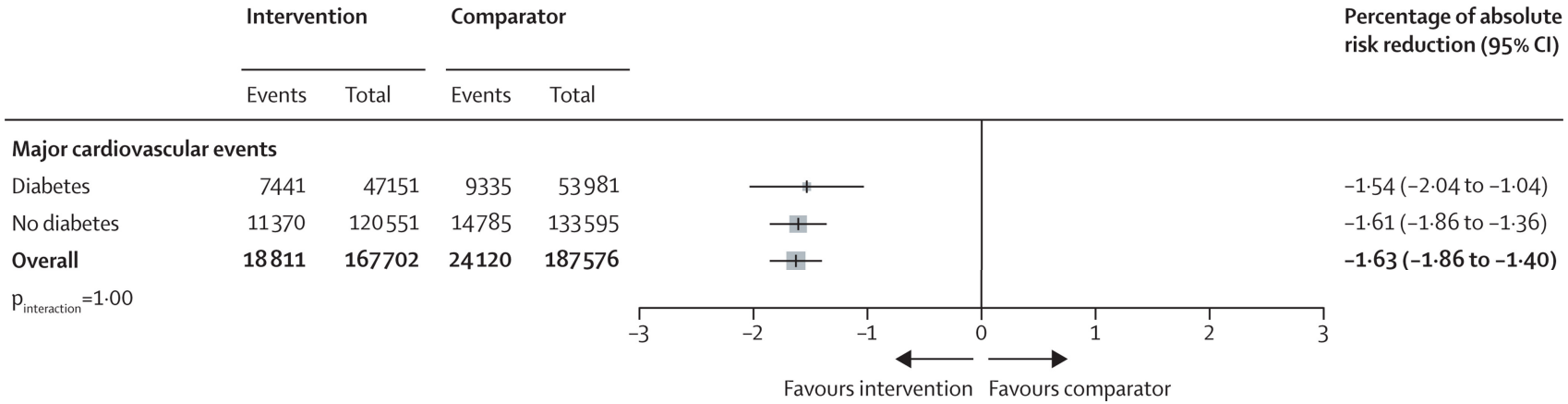
Yes, under certain circumstances even simple statistical models work

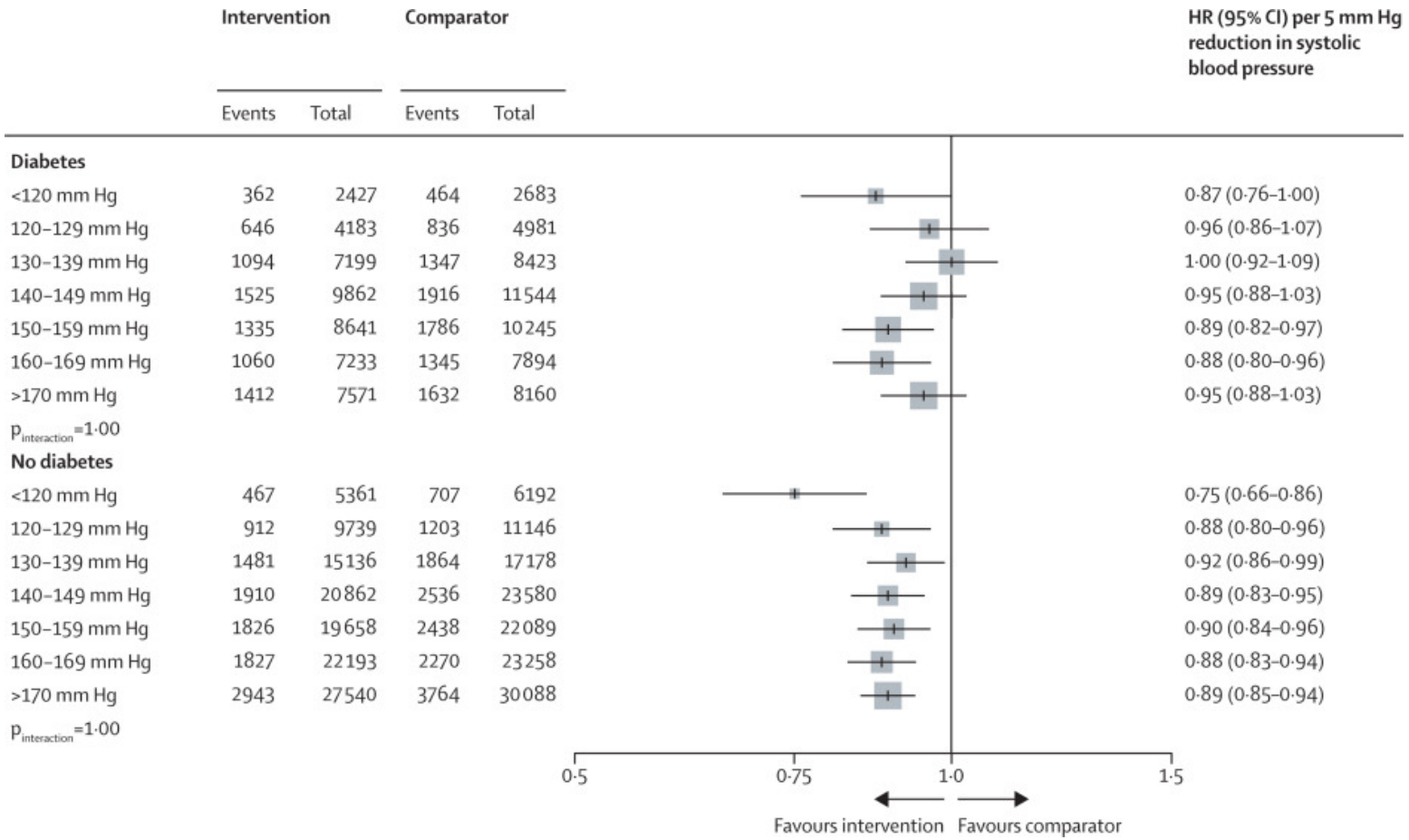
J-shaped association in cohort studies of people with disease

Uncontrolled confounding or causal?



Vidal-Petiot et al. Lancet 2016





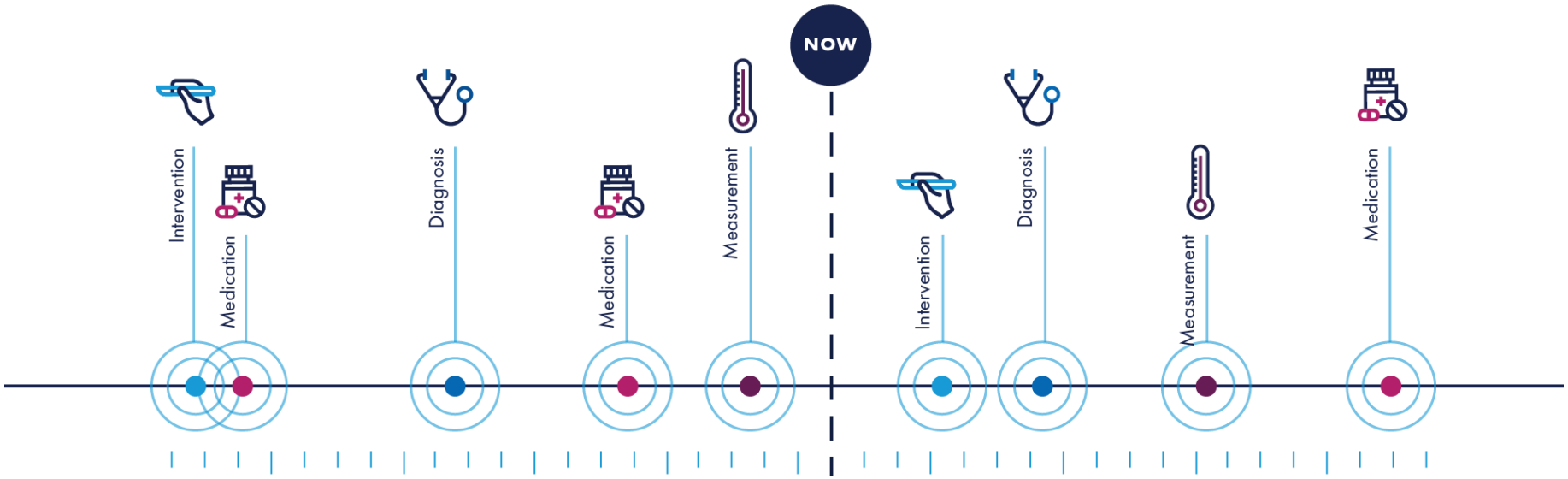
Stratification of results of several RCTs suggests that the results of the cohort studies were confounded

BPLTTC *Lancet* 2021

DEEP MEDICINE

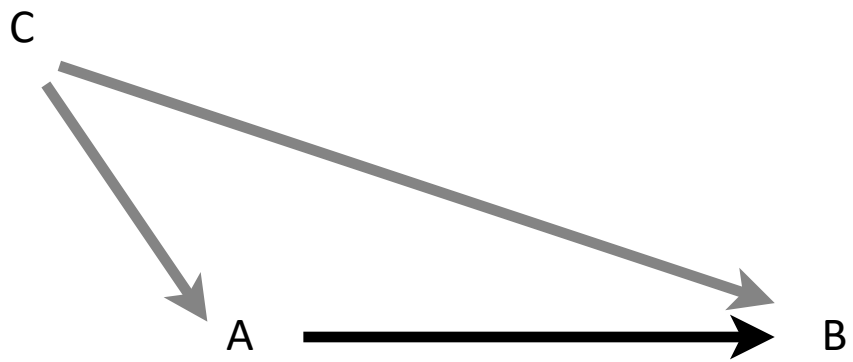
UNIVERSITY OF OXFORD

Can use of EHR (routine, administrative data) together with DL help?



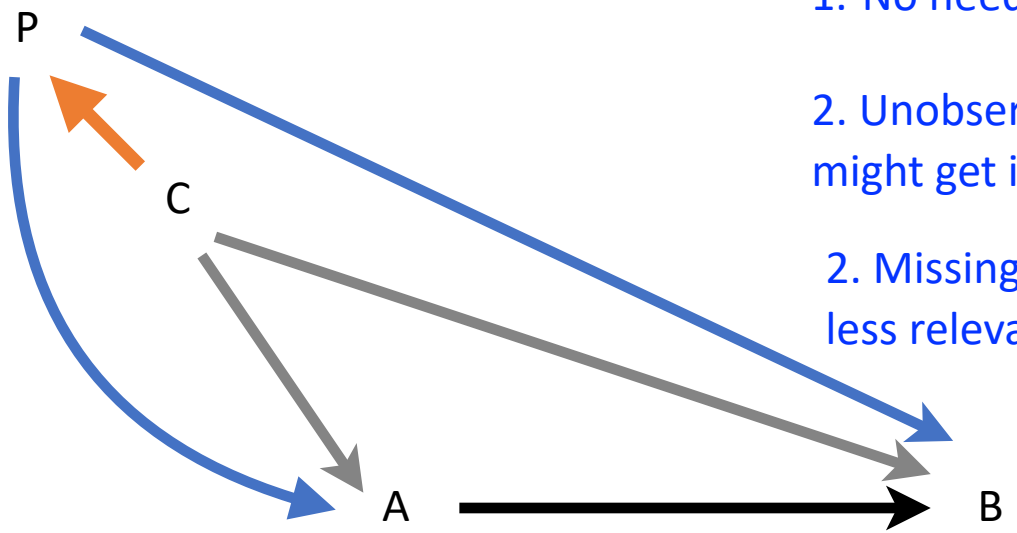
Residual or uncontrolled confounding

Are all confounders known, observed (measured) and included in the models?
I.e., are we comparing like with like?



A: Exposure, e.g., blood pressure, or a drug
B: Outcome, e.g., cardiovascular disease
C: Confounder (common cause), e.g., body weight, or another condition or drug

Theoretical advantage of DL models



1. No need for expert specification of confounders

2. Unobserved, unmeasured or unknown confounders might get indirectly captured in the model

2. Missingness of individual co-variates becomes less relevant

A: Exposure, e.g., blood pressure, or a drug
B: Outcome, e.g., cardiovascular disease
C: Confounder (common cause), e.g., body weight, or another condition or drug
P: Proxy for an unobserved confounder: e.g., regular doctor visits and tests as a proxy for "chronically sick" status

ELSEVIER

Journal of Clinical Epidemiology 125 (2020) 183–187

Journal of Clinical Epidemiology

COMMENTARY

Missing data should be handled differently for prediction than for description or causal explanation

Matthew Sperrin*, Glen P. Martin, Rose Sisk, Niels Peek

Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

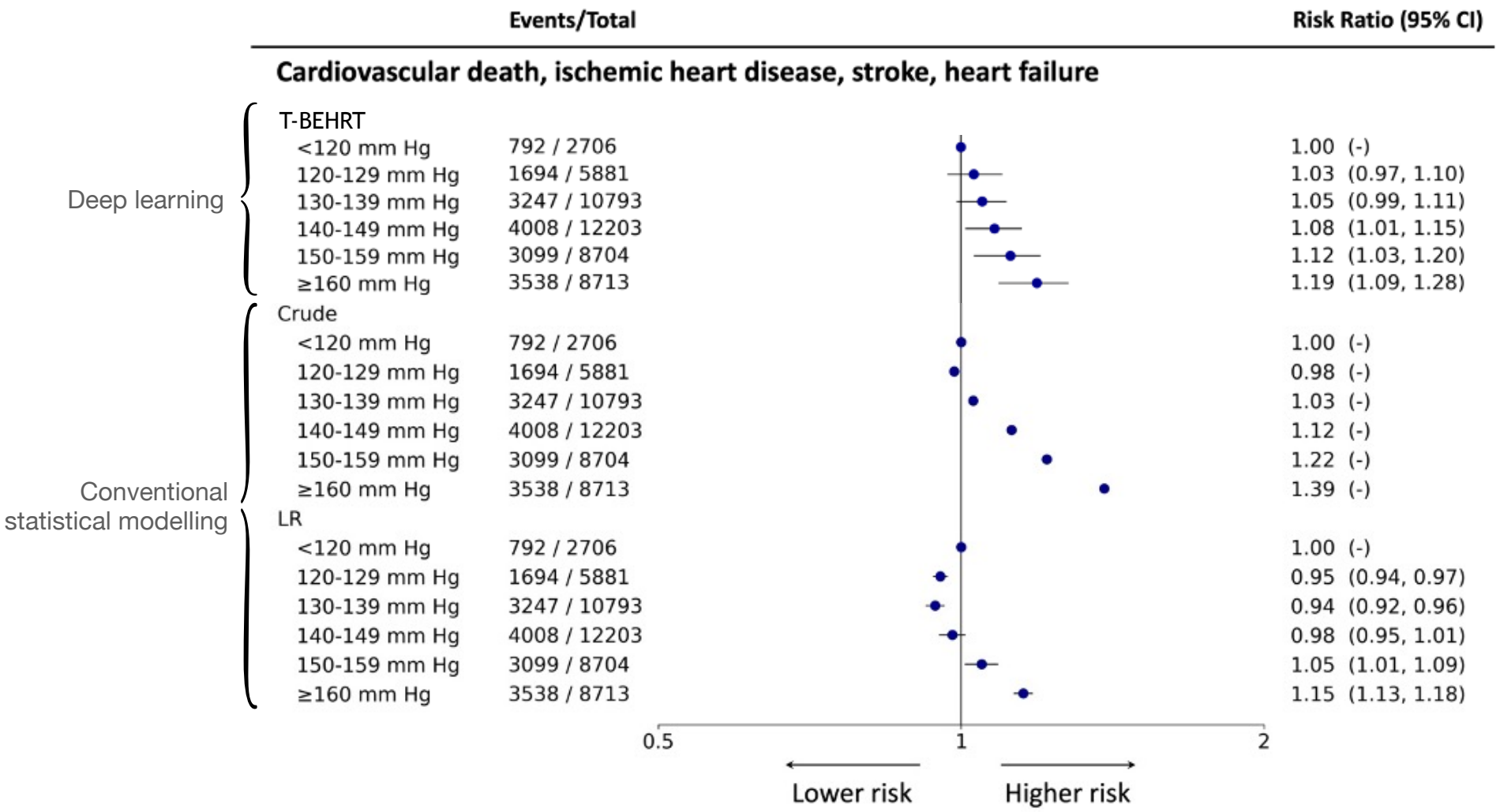
Accepted 18 March 2020; Published online 12 June 2020

Abstract

Missing data are much studied in epidemiology and statistics. Theoretical development and application of methods for handling missing data have mostly been conducted in the context of prospective research data and with a goal of description or causal explanation. However, it is now common to build predictive models using routinely collected data, where missing patterns may convey important information, and one might take a pragmatic approach to optimizing prediction. Therefore, different methods to handle missing data may be preferred. Furthermore, an underappreciated issue in prediction modeling is that the missing data method used in model development may not match

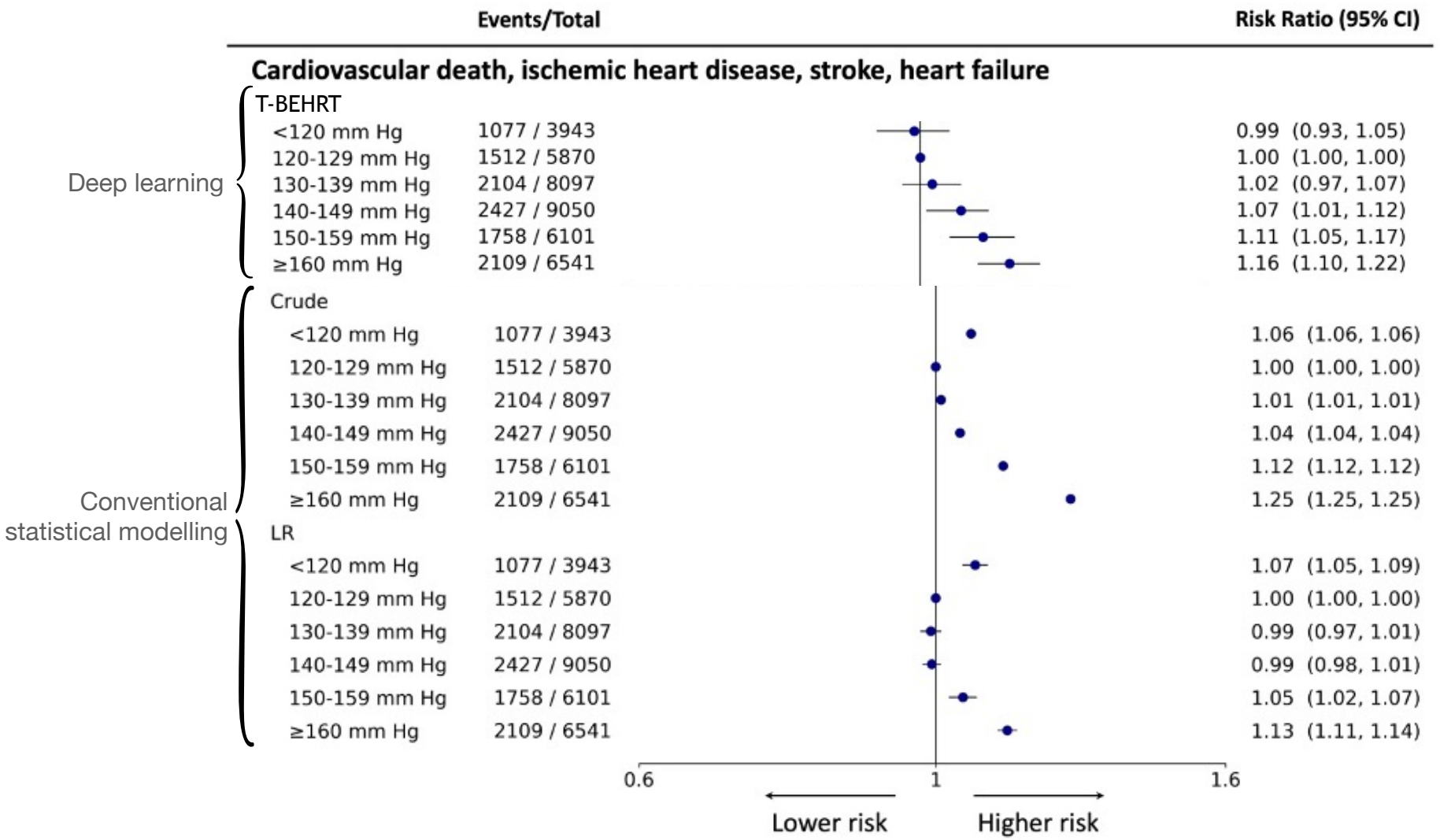
Examples of T-BEHRT in action

Blood pressure and risk of cardiovascular disease in people with diabetes



Rao et al. Unpublished

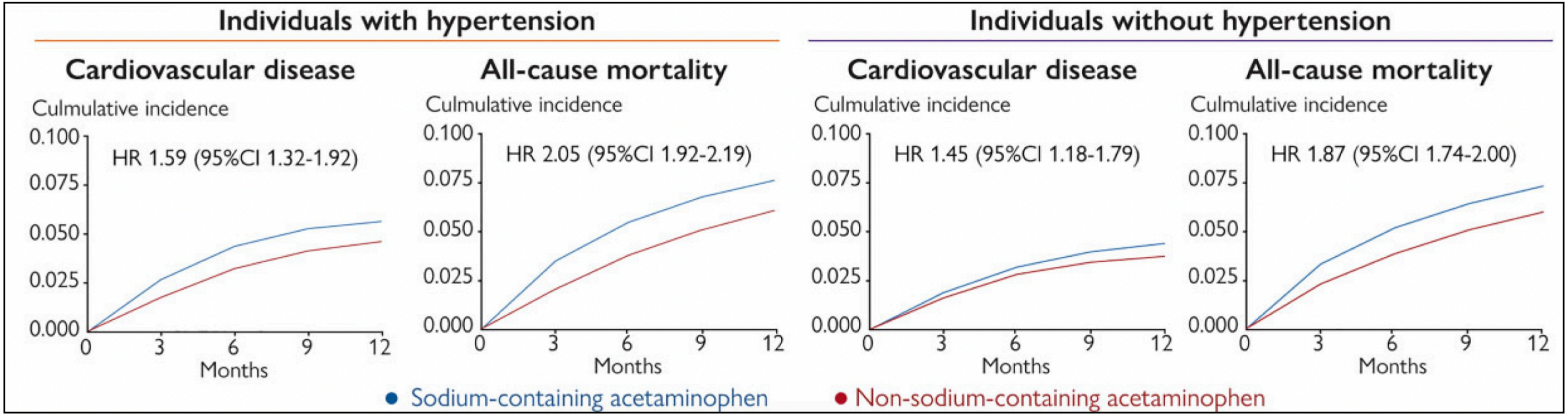
Blood pressure and risk of cardiovascular disease in people with COPD



Rao et al. Unpublished

Treatment effects with drug

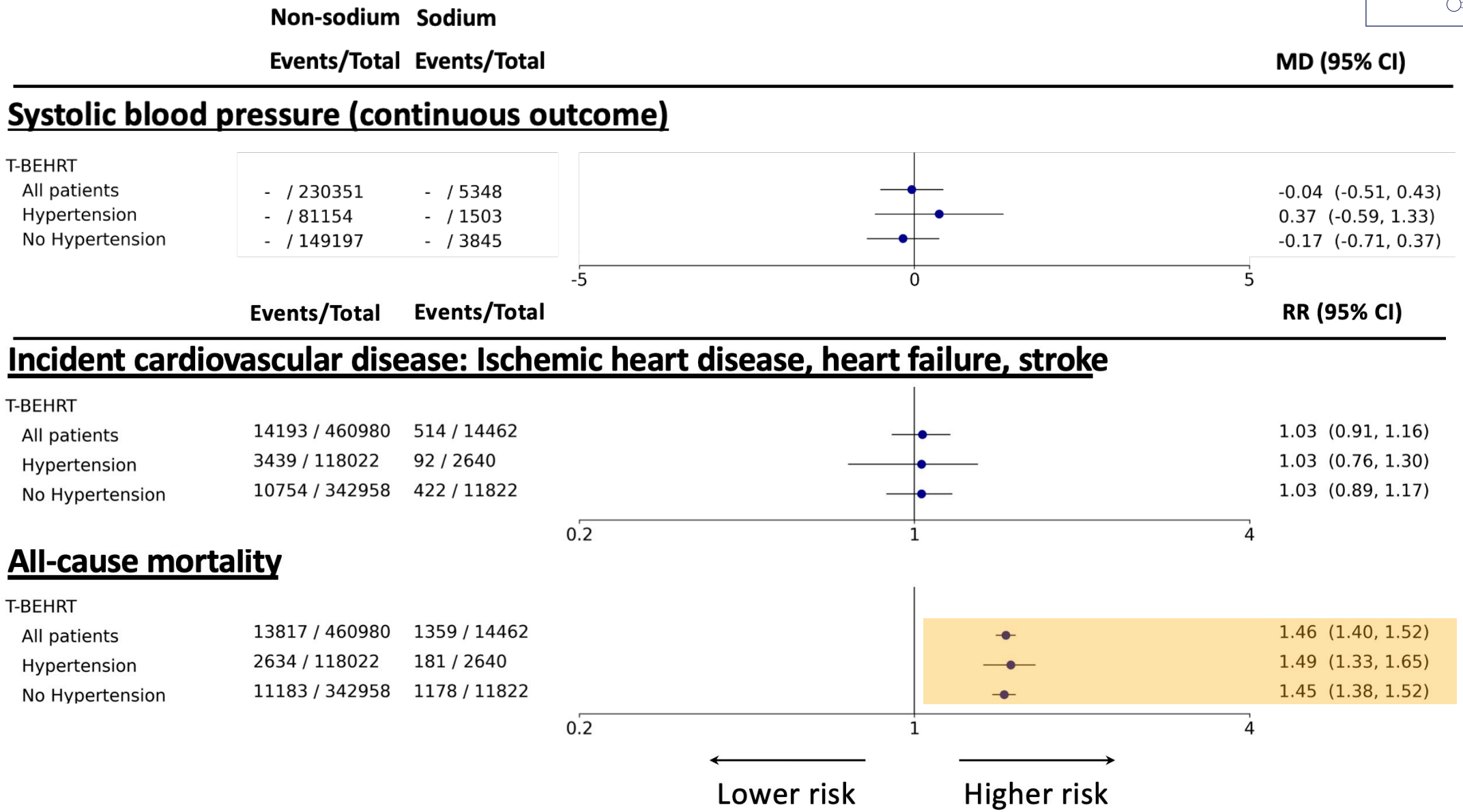
Case of sodium-containing paracetamol on death



- The sodium in the soluble/dispersible formulations contain **1.5x the recommended daily limit of sodium consumption → blood pressure increase → CVD increases**
- However, **no known pathways of blood pressure and all-cause death** (other than CVD related death)

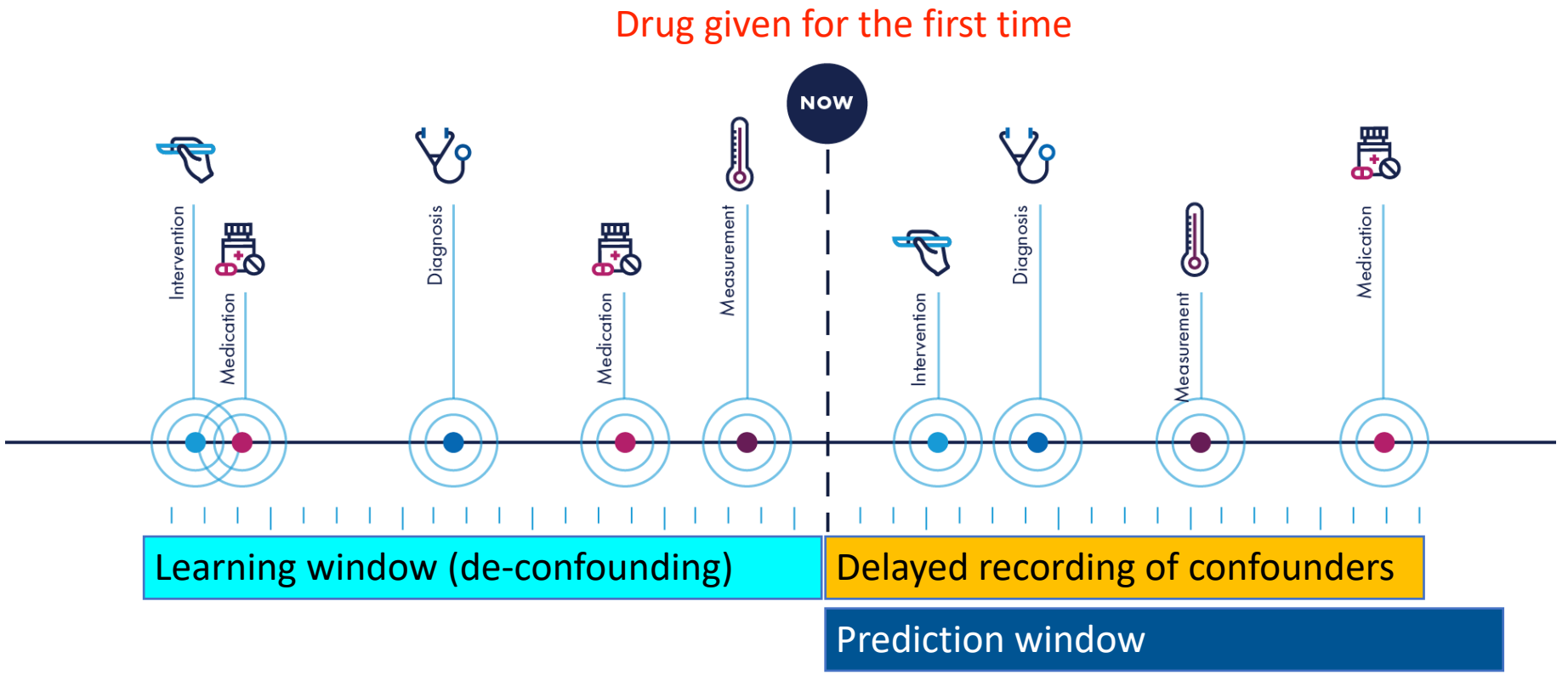
Zeng et al. *Eur Heart J* 2022

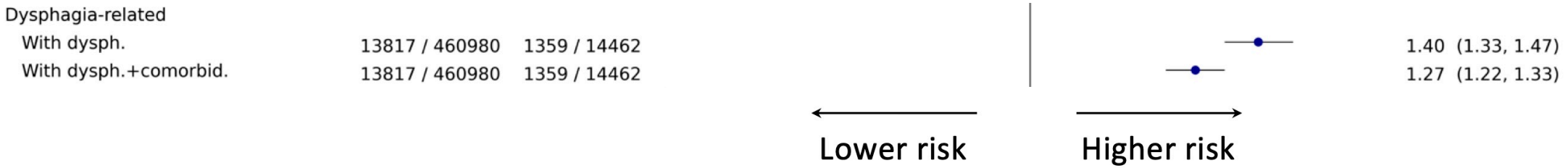
T-BEHRT



Rao et al. Unpublished

T-BEHRT and investigation of residual confounding

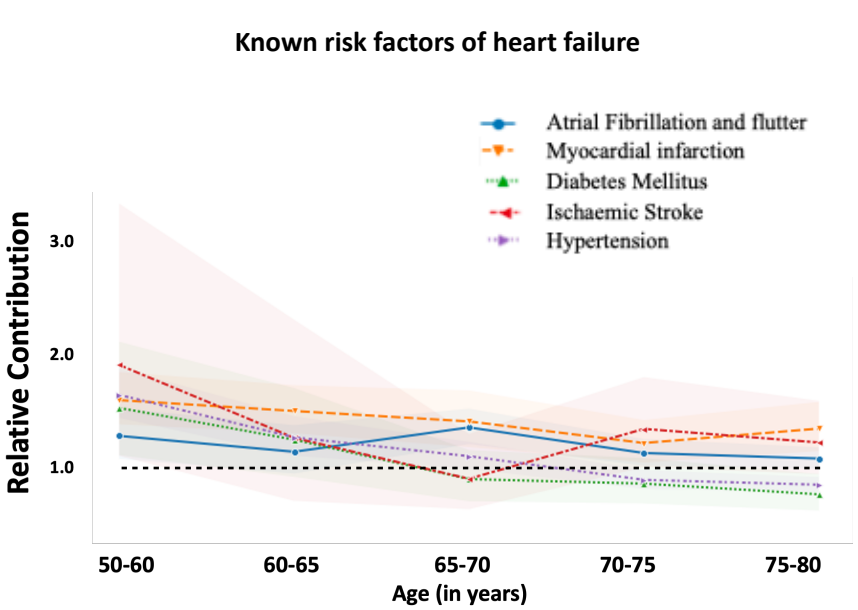


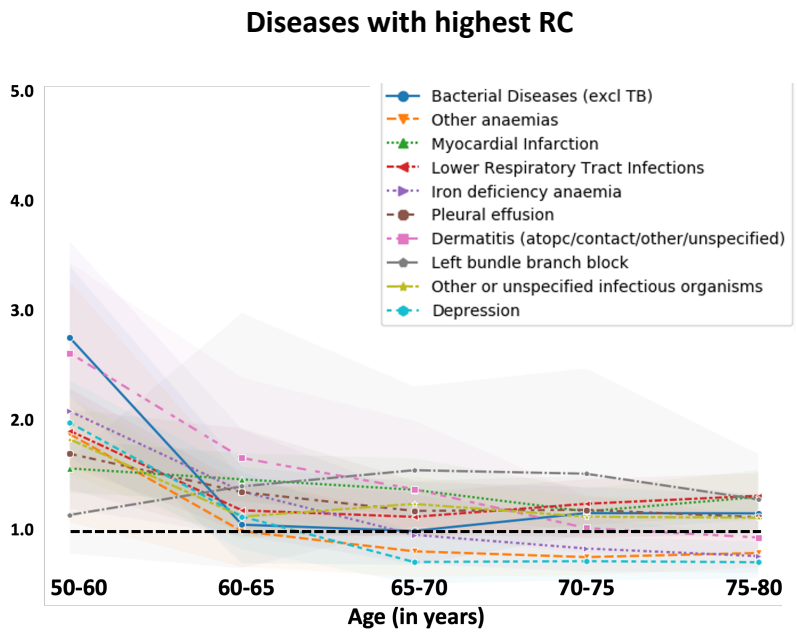


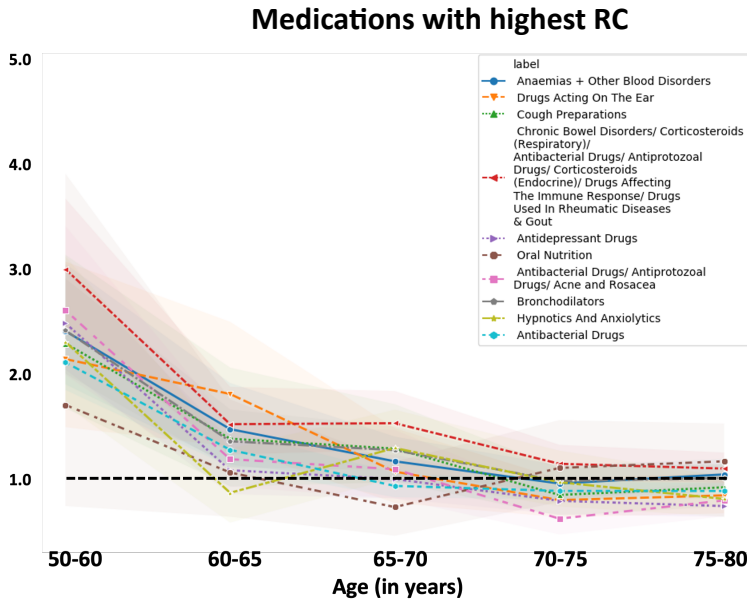
From hypothesis testing to hypothesis generation

How does BEHRT predict?

Towards explainability, trustworthiness and hypothesis generation



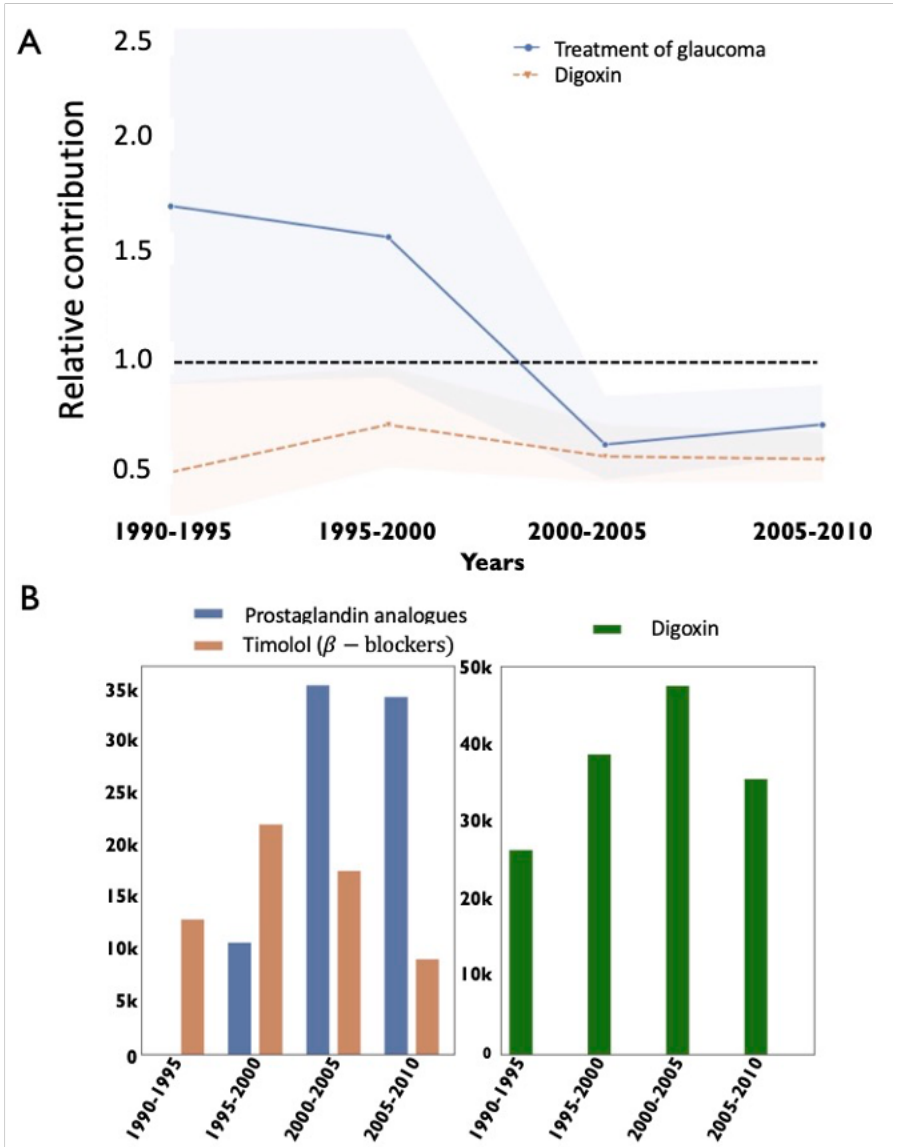




Rao et al, IEEE Journal of Biomedical Health Informatics

How does BEHRT predict?

Towards explainability, trustworthiness and hypothesis generation



Contextual information to help explain unknown or unexpected relationships

Rao et al, IEEE Journal of Biomedical Health Informatics

DEEP MEDICINE

UNIVERSITY OF OXFORD

One of the great promises of applying machine learning to clinical data is the possibility of learning optimal **per-patient treatment rules (personalized medicine)**

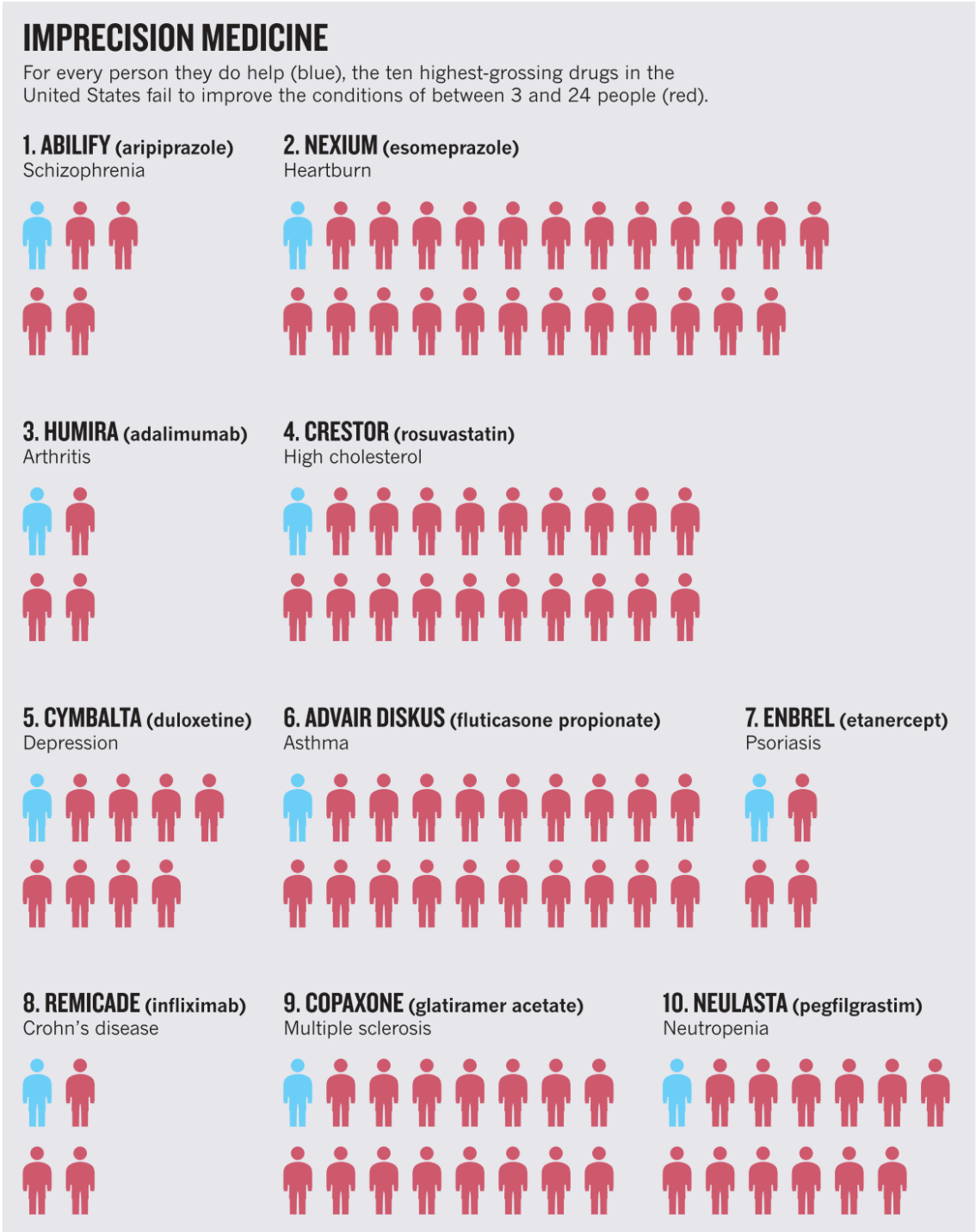
The (typical) case for personalized or precision medicine

Failure of population medicine?

- Treatments do not work for most people who have been given treatment
- Thus, we must sharpen our tool and identify those individuals for whom the treatment will work spectacularly. How can we distinguish between the 'responders' and non-responders'

Example:

- Antihypertensive treatment a widely used treatment.
- Need to treat about 100 people over 10 years treated to prevent 5 patient from suffering a cardiovascular event)
- "Treatment would not work in 95% of individuals treated."



Schork Nature, 2015

Common misconceptions about ‘treatment failure’

1. Confusion of deterministic causal links with probabilistic multicause nature of most conditions
2. Misunderstanding of ‘individual risk’ and undue criticism towards average treatment effects
3. Confusion of low incidence conditions (low ‘base rate’) with low effectiveness of interventions

Misconception #2

Average treatment effects are the problem and we need personalized risk models

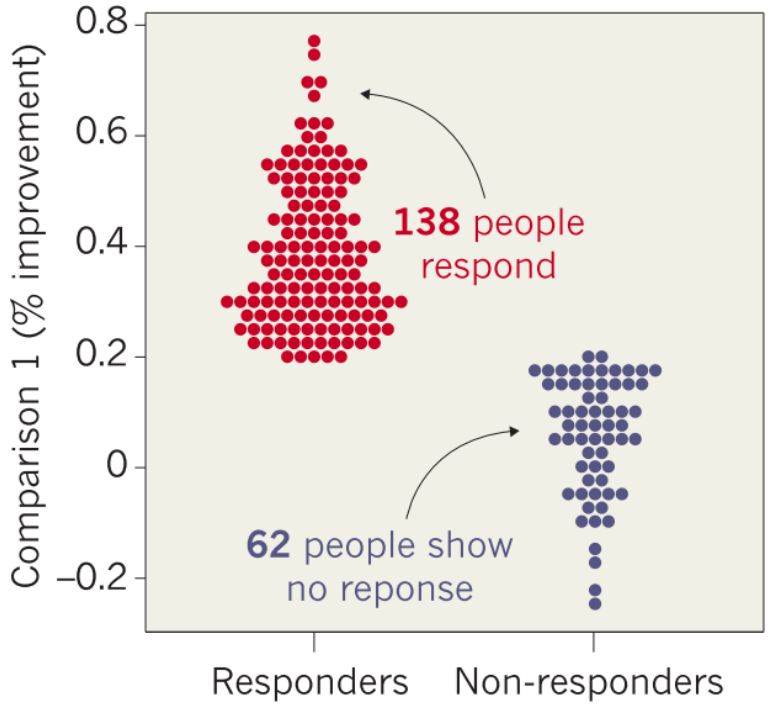
- Patients do not have an ‘individual risk’ that is unique to them and fixed
- Risk is a **conditional group-level** probability and varies depending on the predictors considered, model used and over time.
- Thus, individuals can belong to in indefinite number of groups and this also means that their predicted risk will vary according to the group they are being assigned to (the ‘reference class problem’ John Venn, 1866)
- Most variability in disease incidence is commonly due to “non-shared environment” (stochastic factors, measurement error etc.)

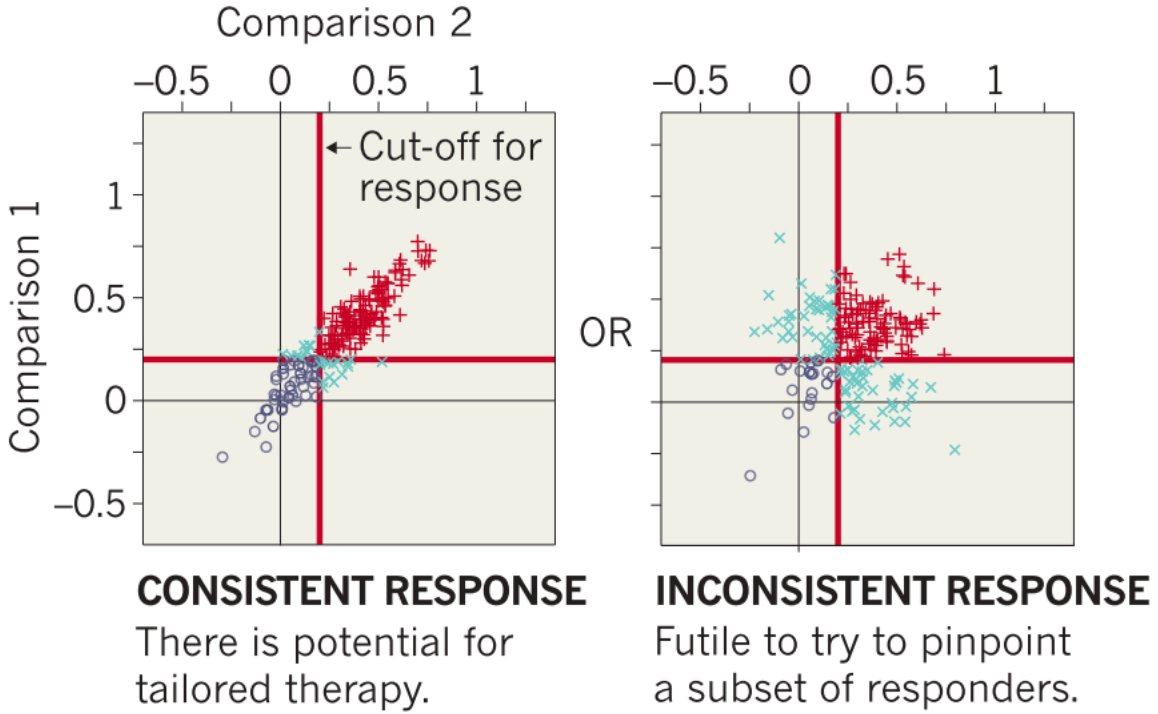
- No amount of data or modelling technique can fix this problem and provide the “true risk” of an outcome for an individual

- But some models can be better at discriminating groups of individuals: **identification of more meaningful subgroups or stratification**

Interaction!

When interactions are found they are not reliable and usually only generate a hypothesis for further testing





Senn Nature. 2018

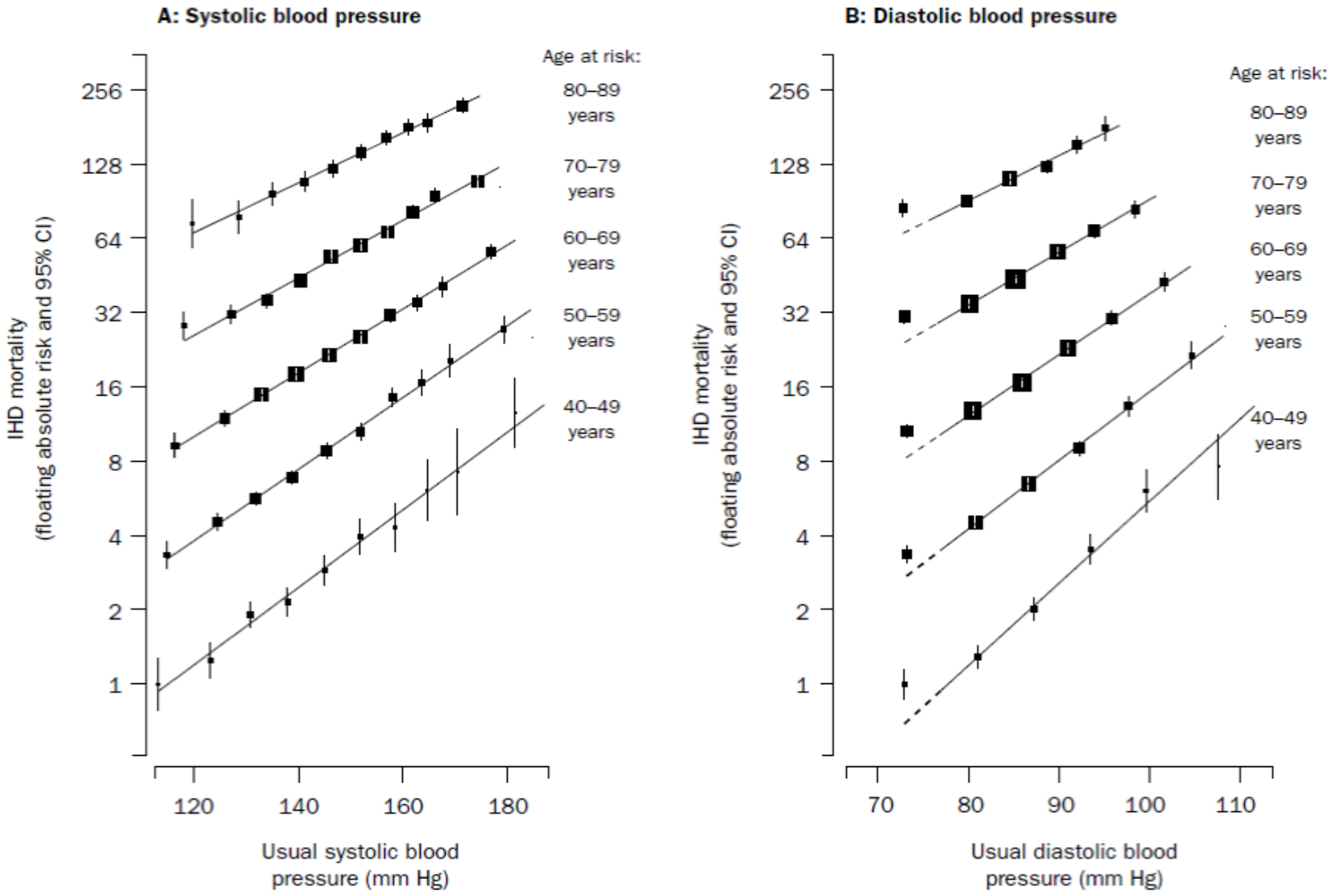
DEEP MEDICINE

UNIVERSITY OF OXFORD

How does epidemiology deal with interaction?

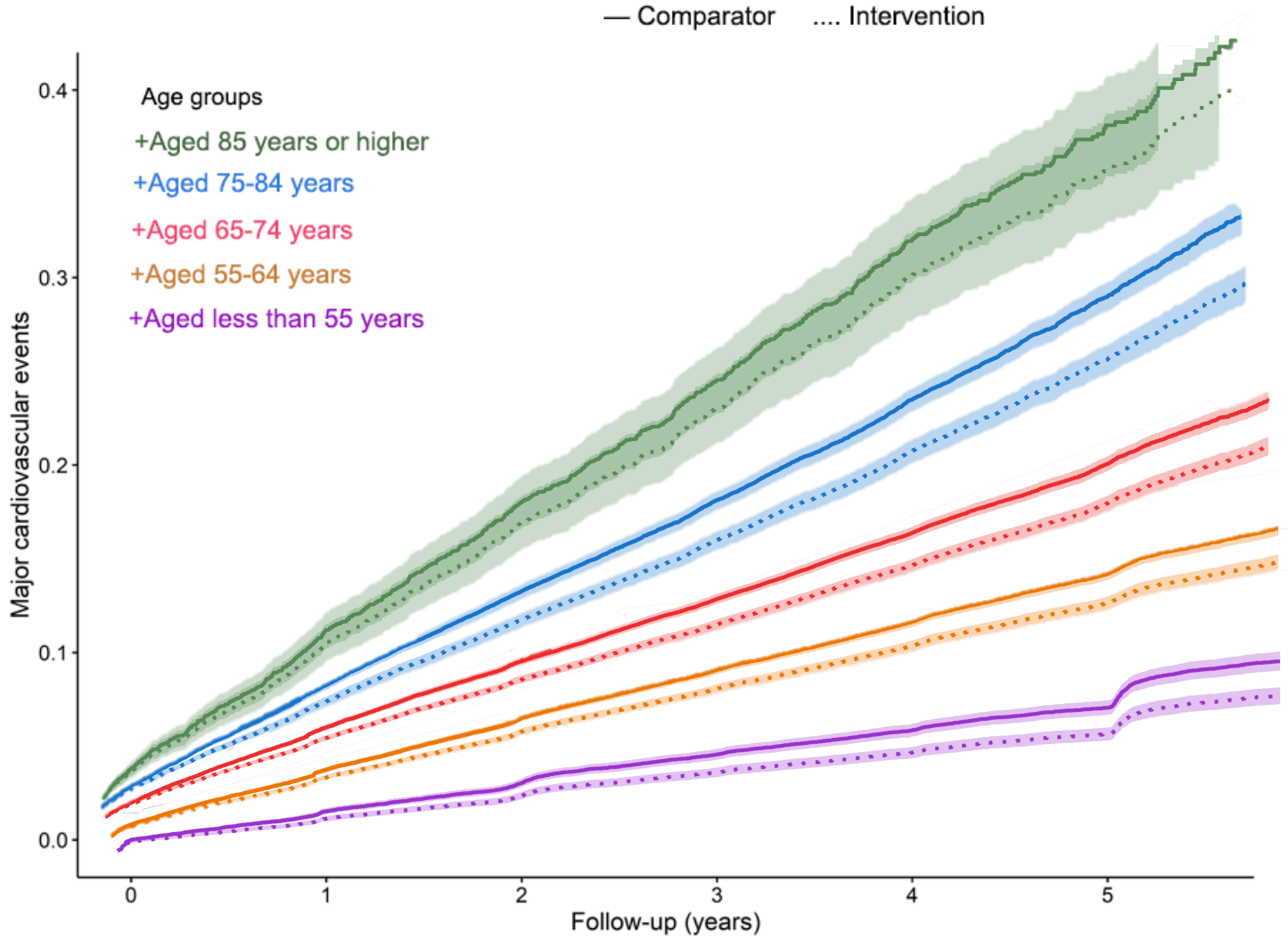
Does age matter?

Stratification by age in observational studies



Lewington et al. *Lancet* 2002

Rate of major cardiovascular events for a 5 mmHg reduction in systolic blood pressure, stratified by treatment allocation and age categories at baseline



HR 0.99 (0.87, 1.12)

HR 0.91 (0.87, 0.96)

HR 0.91 (0.88, 0.95)

HR 0.91 (0.88, 0.95)

HR 0.82 (0.76, 0.88)

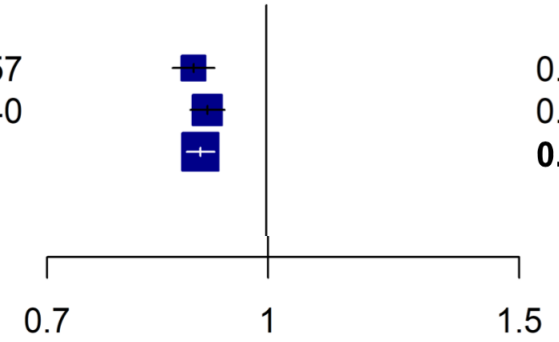
Adjusted $P_{interaction} = 0.05$

BPLTTC *Lancet* 2021

Stratified effects of pharmacological blood pressure lowering

Outcomes	Intervention Events	Intervention Total	Comparator Events	Comparator Total	HR	95% CI
Major cardiovascular events						
Prior CVD	9356	73896	11718	82657	0.89	[0.86; 0.92]
No prior CVD	8931	87033	12319	98840	0.91	[0.89; 0.94]
	18287		**24037**		**0.90**	**[0.88; 0.92]**

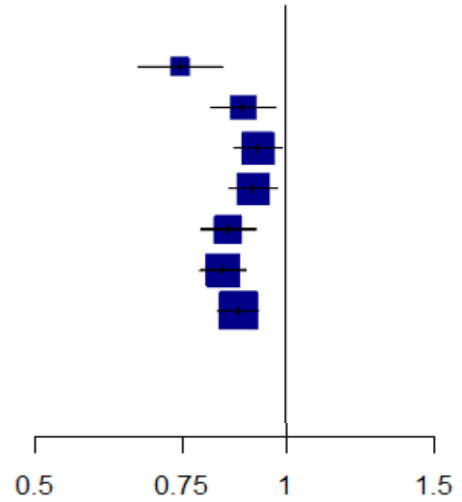
p for interaction-adj=1.00
p for interaction-unadj=0.99



Hazard ratio per 5 mmHg reduction in systolic blood pressure

Outcomes	Intervention Events	Intervention Total	Comparator Events	Comparator Total	HR	95% CI
Major cardiovascular events						
<120	830	7700	1216	8792	0.79	[0.72; 0.87]
120 to 129	1562	13571	2094	15786	0.91	[0.85; 0.98]
130 to 139	2536	21425	3273	24737	0.94	[0.89; 0.99]
140 to 149	3423	29295	4496	33889	0.93	[0.88; 0.98]
150 to 159	3130	27046	4264	31256	0.88	[0.83; 0.94]
160 to 169	2818	28443	3605	30265	0.87	[0.83; 0.92]
≥170	3979	33433	5080	36724	0.90	[0.86; 0.94]

p for interaction-adj=0.35
p for interaction-unadj=0.05

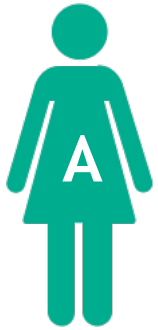


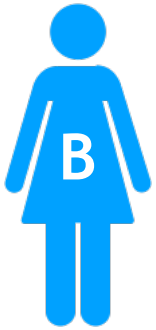
BPLTTC, *Lancet* 2021

For most treatments, no major sources of heterogeneity of (relative) effects have been found.

We seem to have biologically more in common that some people would like to see

Absolute (not relative) risk reduction matters for selection





10-year CVD risk 20%
BP 145/80 mmHg

10-year CVD risk 40%
BP 145/80 mmHg

BP 135/75 mmHg

BP 135/75 mmHg

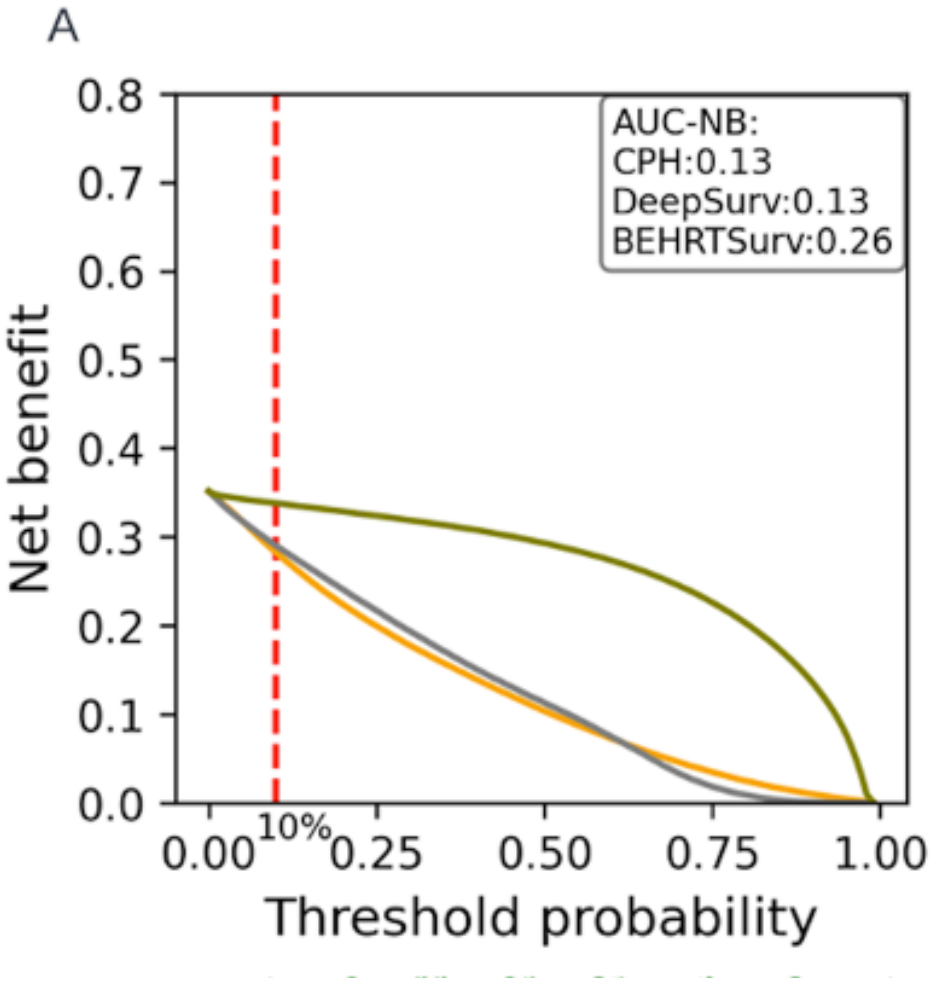
Relative RR: 20% (from RCTs)

Relative RR: 20%

Absolute RR: **4**% (from 20%)
Treat **25** people over 10 years to prevent one event

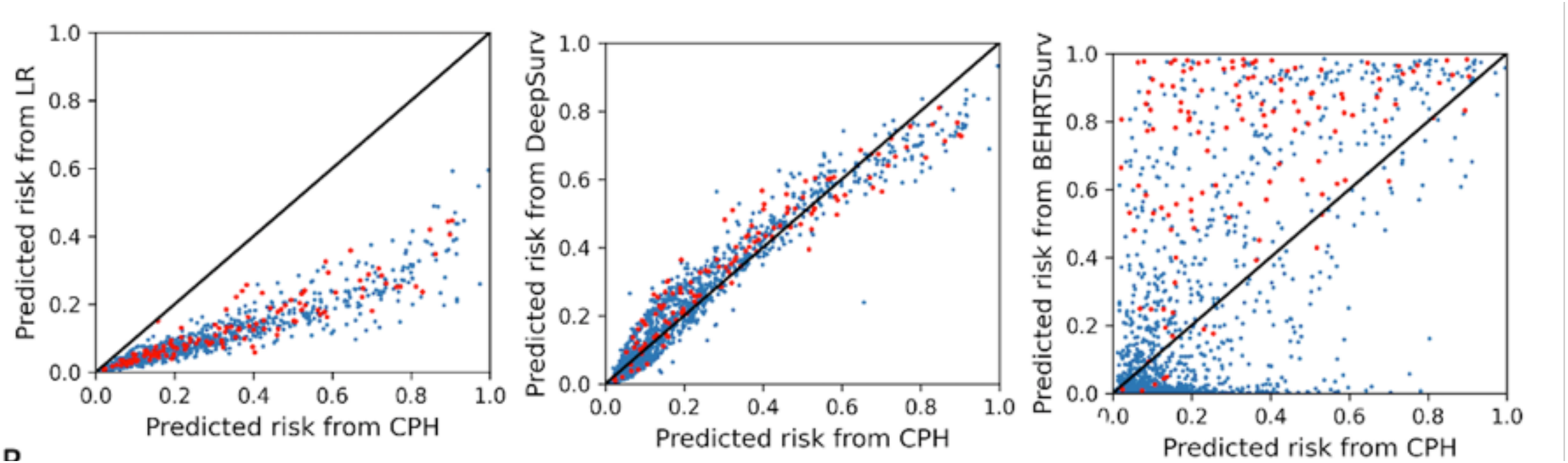
Absolute RR: 8% (from 40%)
Treat 13 people over 10 years to treat one event

Net benefit across different decision threshold for various models



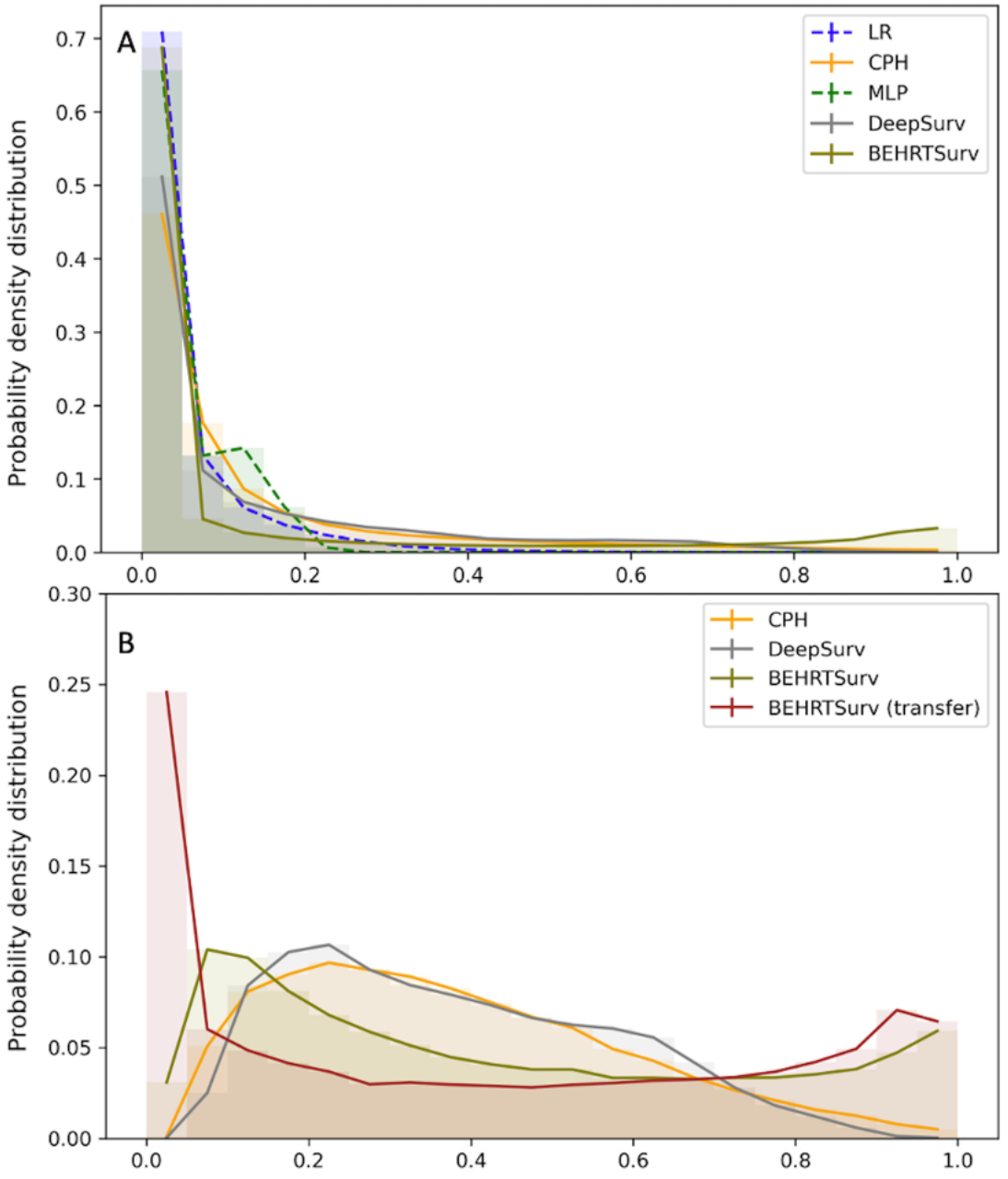
Rao, Li et al. Unpublished

Individual-level correlation of predicted risk by different models



Red dots depict true positives

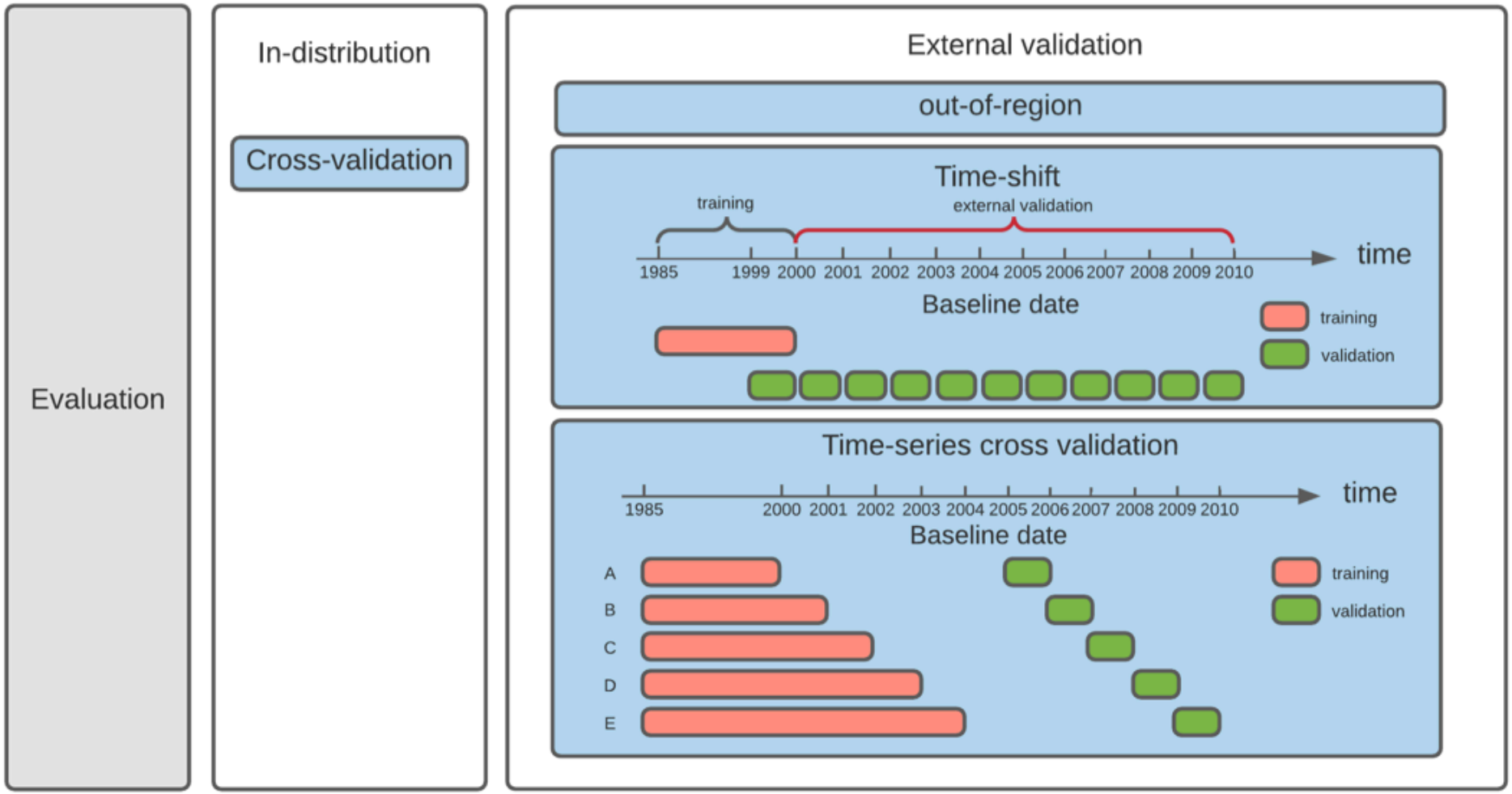
Rao, Li et al. Unpublished



Rao, Li et al. Unpublished

Does the better performance of DL model comes at the cost of limited generalisability?

External validation framework in EHR



Li et al. Unpublished

DEEP MEDICINE

UNIVERSITY OF OXFORD

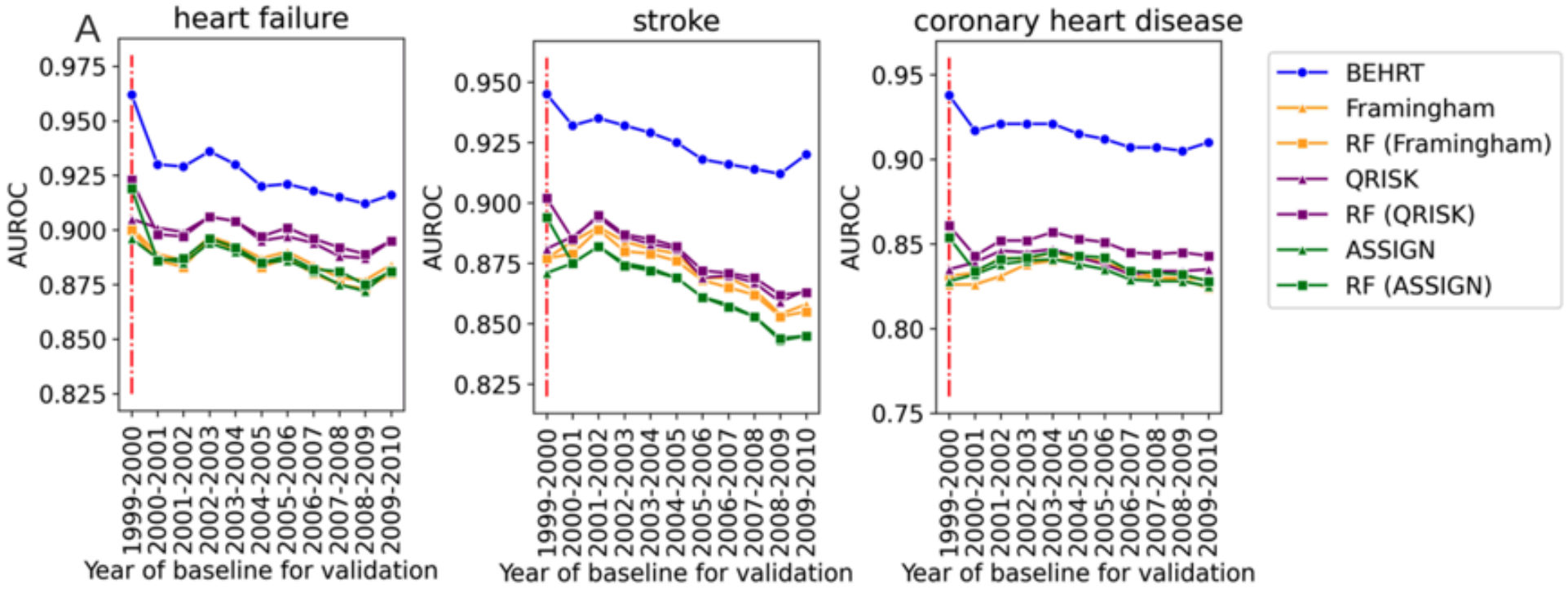
In-distribution validation

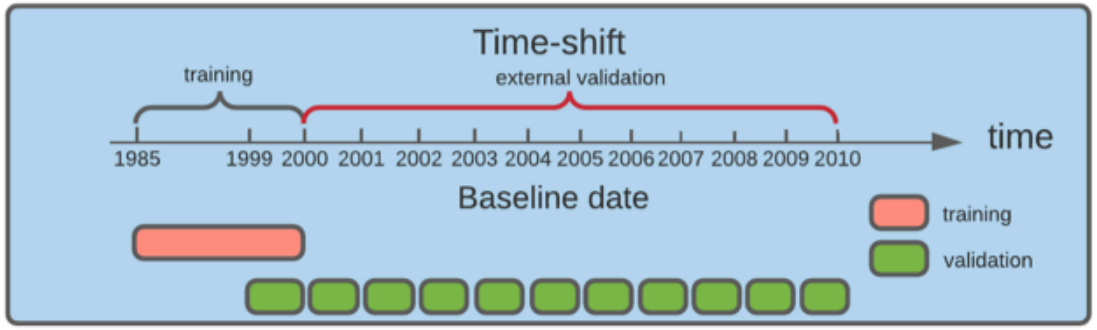
	AUROC (95% confidence interval)		
Models	HF	Stroke	CHD
BEHRT	0·954 (0·004)	0·957 (0·002)	0·951 (0·002)
QRISK3	0·895 (0·002)	0·874 (0·005)	0·838 (0·005)
RF (QRISK)	0·897 (0·005)	0·877 (0·005)	0·850 (0·003)
ASSIGN	0·885 (0·002)	0·858 (0·002)	0·829 (0·001)
RF (ASSIGN)	0·884 (0·005)	0·859 (0·003)	0·833 (0·002)
Framingham	0·883 (0·004)	0·869 (0·005)	0·831 (0·005)
RF (Framingham)	0·884 (0·002)	0·868 (0·004)	0·836 (0·003)

Out-of-region validation

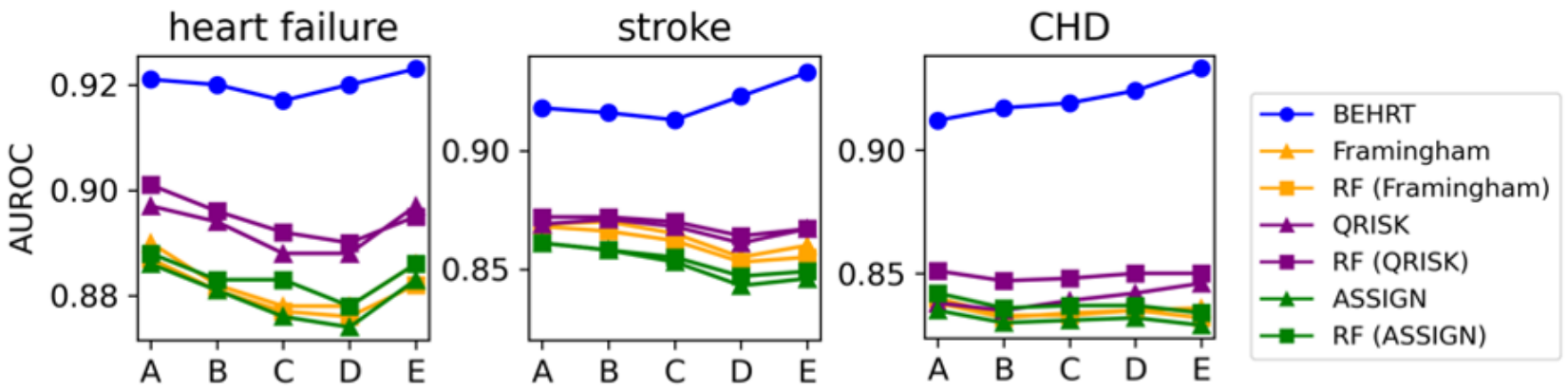
	AUROC (absolute decline compared to the internal performance)		
Models	HF	Stroke	CHD
BEHRT	0·909 (-0·044)	0·932 (-0·025)	0·929 (-0·022)
QRISK3	0·883 (-0·012)	0·865 (-0·009)	0·830 (-0·008)
RF (QRISK)	0·883 (-0·014)	0·866 (-0·011)	0·840 (-0·010)
ASSIGN	0·873 (-0·012)	0·852 (-0·003)	0·823 (-0·006)
RF (ASSIGN)	0·874 (-0·010)	0·853 (-0·006)	0·827 (-0·006)
Framingham	0·871 (-0·012)	0·862 (-0·007)	0·821 (-0·010)
RF (Framingham)	0·873 (-0·011)	0·855 (-0·013)	0·826 (-0·010)

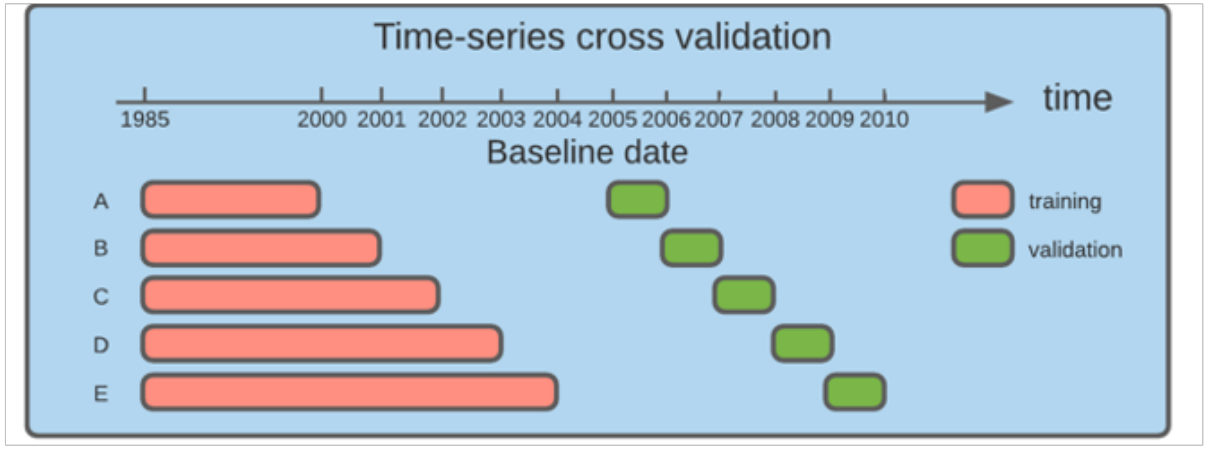
Li et al. Unpublished





Li et al. Unpublished





Li et al. Unpublished

DEEP MEDICINE

UNIVERSITY OF OXFORD

Thank you!