Original research

Systolic blood pressure, chronic obstructive pulmonary disease and cardiovascular risk

Shishir Rao ,[1,2] Milad Nazarzadeh,[1,2] Yikuan Li,[1,2] Dexter Canoy,[3] Mohammad Mamouei ,[1,2] Gholamreza Salimi-Khorshidi,[1,2] Kazem Rahimi[1,2,4]

ABSTRACT

Objective In individuals with complex underlying health problems, the association between systolic blood pressure (SBP) and cardiovascular disease is less well recognised. The association between SBP and risk of cardiovascular events in patients with chronic obstructive pulmonary disease (COPD) was investigated.

Methods and analysis In this cohort study, 39 602 individuals with a diagnosis of COPD aged 55–90 years between 1990 and 2009 were identified from validated electronic health records (EHR) in the UK. The association between SBP and risk of cardiovascular end points (composite of ischaemic heart disease, heart failure, stroke and cardiovascular death) was analysed using a deep learning approach.

Results In the selected cohort (46.5% women, median age 69 years), 10 987 cardiovascular events were observed over a median follow-up period of 3.9 years. The association between SBP and risk of cardiovascular end points was found to be monotonic; the lowest SBP exposure group of <120 mm Hg presented nadir of risk. With respect to reference SBP (between 120 and 129 mm Hg), adjusted risk ratios for the primary outcome were 0.99 (95% CI 0.93 to 1.05) for SBP of <120 mm Hg, 1.02 (0.97 to 1.07) for SBP between 130 and 139 mm Hg, 1.07 (1.01 to 1.12) for SBP between 140 and 149 mm Hg, 1.11 (1.05 to 1.17) for SBP between 150 and 159 mm Hg and 1.16 (1.10 to 1.22) for SBP ≥160 mm Hg.

Conclusion Using deep learning for modelling EHR, we identified a monotonic association between SBP and risk of cardiovascular events in patients with COPD.

INTRODUCTION

Systolic blood pressure (SBP) is a well-known risk factor for cardiovascular diseases.[1–3] However, in subgroups with complex underlying health conditions, the association of SBP with cardiovascular outcomes is less well understood. Often, in these patient groups, a so-called J-shaped association is reported, where the association between SBP and risk of cardiovascular events has an optimum, above and below which the risk increases.[4 5]

In patients with chronic obstructive pulmonary disease (COPD), the relationship remains unclear. Independently, SBP and COPD have both been associated with a higher risk of cardiovascular disease (CVD).[2 3 6 7] However, there is a dearth of evidence when it comes to conclusively understanding the relationship between SBP and risk of cardiovascular end points in patients with COPD. A J-shaped

association between SBP and risk of cardiovascular events was found in a previous observational analysis using traditional statistical modelling in patients with COPD who were at risk of developing CVD.[4] However, observational studies using conventional statistical modelling might be limited in investigating this question. The adjusted variables need to be manually chosen and their relationship assumed by researchers, naturally exposing models to issues of residual confounding. Additionally, in subgroups of patients with multiple comorbidities at baseline and a large number of complicated factors of risk and prevention, confounding factors are lesser understood; as a result, conventional statistical

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ In patients with chronic obstructive pulmonary disease (COPD), the relationship between systolic blood pressure and risk of cardiovascular events is poorly understood.

⇒ One observational study using conventional statistical modelling, which requires manual confounder selection and is found to be inadequate in modelling high-risk cohorts, has shown a J-shaped association, with increased risk above and below an apparently optimal systolic blood pressure value.

WHAT THIS STUDY ADDS

⇒ With recent access to comprehensive, routine electronic health records data and developments in causal deep learning modelling capable of extracting and adjusting for confounders both known and latent in medical history, our longitudinal cohort study on about 40 000 patients with COPD with and without prior cardiovascular disease captured a monotonic relationship between systolic blood pressure and risk of cardiovascular events rejecting the J-shaped curve hypothesis.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ While current guidelines recommend a systolic blood pressure target of 130 mm Hg (140 mm Hg in the elderly) in patients with COPD, our study demonstrates a nadir of risk at <120 mm Hg for cardiovascular outcomes in line with established knowledge concerning cardiovascular risk.

models with insufficient adjustment can result in confounded or spurious J-shaped associations.[2 8–10]

With the availability of comprehensive electronic health records (EHR) and the advancement of deep learning (DL) causal modelling, the opportunity for more accurate modelling of associations among subgroups with poorer health has arisen.[10–12] While traditional modelling requires manual confounder selection, DL approaches such as Targeted Bidirectional EHR Transformer (T-BEHRT) automatically extract latent features that are confounding the association and more accurately estimate risk ratio (RR) in observational settings.[10 12]

In this study, we applied the T-BEHRT model to evaluate the association between SBP and risk of cardiovascular outcomes in a cohort of 39 602 patients with COPD.

METHODS
Study setting and participants

We used retrospective anonymised EHR data from Clinical Practice Research Datalink (CPRD), an EHR database representative of the UK population that has been validated for epidemiological research.[13 14] We used EHR from two data sources within CPRD to identify a cohort of 39 602 individuals with COPD: primary care and secondary care (Hospital Episode Statistics (HES)). Those between 55 and 90 years of age with at least one blood pressure measurement taken between the years 1990 and 2009 were included in this study with index date (baseline) being defined as the date of the first SBP measurement in this time period (online supplemental figure S1). COPD was identified at
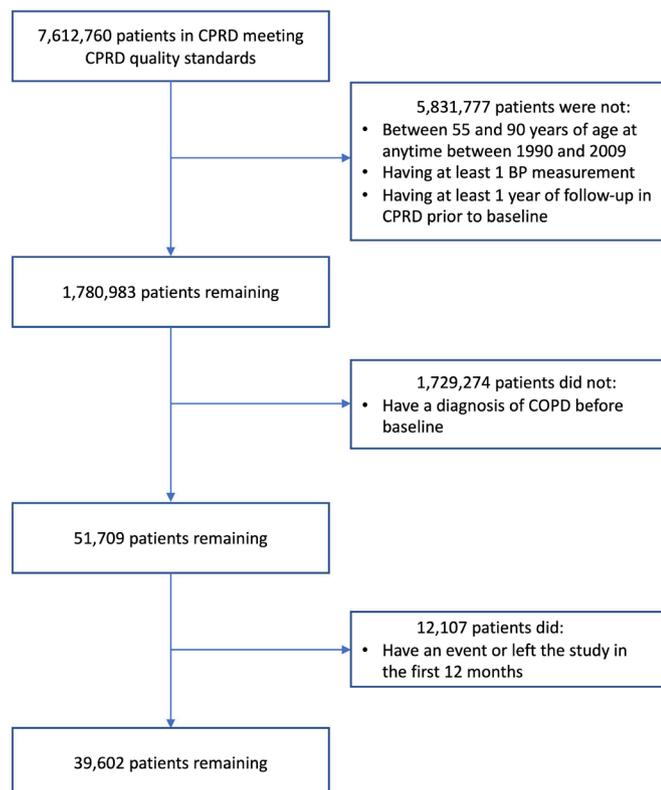

7,612,760 patients in CPRD meeting CPRD quality standards

→ 5,831,777 patients were not:
- Between 55 and 90 years of age at anytime between 1990 and 2009
- Having at least 1 BP measurement
- Having at least 1 year of follow-up in CPRD prior to baseline

1,780,983 patients remaining

→ 1,729,274 patients did not:
- Have a diagnosis of COPD before baseline

51,709 patients remaining

→ 12,107 patients did:
- Have an event or left the study in the first 12 months

39,602 patients remaining

Figure 1 Cohort selection flow chart. Process for selecting cohort used in the study of the association between systolic blood pressure (BP) and risk of cardiovascular events in patients with chronic obstructive pulmonary disease (COPD) using observational data from the Clinical Practice Research Datalink (CPRD) database.

baseline using phenotyping methods validated for use on CPRD data.[15]

This cohort study followed the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines.

Exposures

The exposure variable in this study was SBP and was derived from the CPRD measurements dataset. Blood pressure measurement data are recorded by staff at the general practice (GP) during a visit/consultation.[14] In our study, we extracted SBP values and excluded measurements <50 and >300 mm Hg as recommended by previously published methods to clean measurements data.[16] Next, the exposure status for a patient was calculated as mean of the SBP measurements in the first 12 months after baseline (ie, exposure period). Patients were categorised into six exposure categories of this averaged measure of SBP over the course of the exposure period: <120 mm Hg, 120–129 mm Hg (reference), 130–139 mm Hg, 140–149 mm Hg, 150–159 mm Hg and ≥160 mm Hg.

Outcomes

The primary outcome was fatal/non-fatal CVDs defined as a composite of ischaemic heart disease (IHD), heart failure, stroke and cardiovascular-related death. Secondary outcomes investigated in this study were individual components of the defined primary outcome: (1) IHD, (2) heart failure and (3) stroke. We identified cardiovascular events using three data sources in CPRD: (1) primary care, (2) secondary care (HES) and (3) the Office of National Statistics (cause-specific mortality) using previously published phenotyping algorithms.[15] Read codes were used to identify the conditions in the primary care setting while International Classification of Diseases 10th Revision codes were used to identify cases in the secondary care and mortality setting. Follow-up period started 1 year from baseline (ie, following the exposure period). Events within 5 years of the follow-up period (ie, between 1 year and 6 years after baseline) were captured for analysis; this feature of study design was incorporated to avoid conducting association estimation in the time period overlapping with the exposure period (ie, the first 12 months following baseline). Those who had events or left the study within the first 12 months following baseline were removed from the analysis as consistent with similar past studies.[10]

Statistical and deep learning analyses

For analyses of the primary and secondary outcomes, the DL model, T-BEHRT was implemented.[12] The T-BEHRT model is a DL approach that uses minimally processed EHR to estimate RR more accurately than other statistical and DL benchmark models.[12] The model incorporates EHR records, specifically diagnoses and medications—longitudinal in nature—along with few static attributes of the patient (ie, sex, smoking status) and adjusts for confounding features in the medical history of the patient (online supplemental figure S2).[12] In addition to adjusting for confounders and estimating risk of outcome, the T-BEHRT model estimates probability of being assigned to a particular exposure status (propensity score).[12 17] By conducting both outcome and propensity score prediction, the DL framework offers the opportunity to conduct doubly robust estimation using propensity score modelling in order to limit issues of selection bias (further information in online supplemental methods).[17]

In order to compare our DL approach against established statistical modelling, we implemented logistic regression (LR)

Table 1 Characteristics of patients by systolic blood pressure categories at index date

SBP categories	<120 mm Hg	120–129 mm Hg	130–139 mm Hg	140–149 mm Hg	150–159 mm Hg	≥160 mm Hg
No. (%)	3943 (10.0)	5870 (14.8)	8097 (20.4)	9050 (22.9)	6101 (15.4)	6541 (16.5)
Follow-up, years (IQR)	3.4 (1.1–5.0)	3.0 (1.5–5.0)	4.0 (1.6–5.0)	4.0 (1.6–5.0)	4.0 (1.5–5.0)	3.7 (1.5–5.0)
Age, years (IQR)	66.0 (58.0–75.0)	67.0 (58.0–75.0)	68.0 (59.0–76.0)	69.0 (61.0–77.0)	70.0 (62.0–77.0)	72.0 (65.0–78.0)
Women (%)	1892 (48.0)	2695 (45.9)	3774 (46.6)	4179 (46.2)	2765 (45.3)	3142 (48.0)
YOB (IQR)	1937 (1927–1945)	1936 (1927–1945)	1935 (1926–1944)	1933 (1925–1942)	1931 (1924–1940)	1927 (1921–1935)
BMI* kg/m² (IQR)	25.7 (24.0–27.0)	26.0 (24.4–27.2)	26.0 (24.7–27.4)	25.9 (24.7–27.2)	25.9 (24.8–27.1)	25.6 (24.7–26.8)
LDL*, mmol/L (IQR)	3.1 (2.9–3.2)	3.1 (2.9–3.2)	3.1 (2.9–3.2)	3.1 (3.0–3.2)	3.1 (3.0–3.2)	3.1 (3.0–3.2)
TG*, mmol/L (IQR)	1.6 (1.4–1.8)	1.6 (1.4–1.8)	1.6 (1.4–1.8)	1.6 (1.4–1.8)	1.6 (1.4–1.8)	1.6 (1.4–1.7)
TC*, mmol/L (IQR)	5.3 (5.0–5.7)	5.3 (5.0–5.6)	5.3 (5.0–5.6)	5.3 (5.0–5.6)	5.3 (5.0–5.6)	5.3 (5.0–5.6)
Smoking status*						
Current smoker (%)	1960 (49)	2831 (48)	3627 (44)	4074 (45)	2728 (44)	3049 (46)
Former smoker (%)	1453 (36)	2131 (36)	3148 (38)	3490 (38)	2336 (38)	2307 (35)
Never smoker (%)	530 (13)	908 (15)	1322 (16)	1486 (16)	1037 (16)	1185 (18)
Disease at baseline						
IHD (%)	711 (18.0)	872 (14.9)	1131 (14.0)	1085 (12.0)	650 (10.7)	670 (10.2)
CKD (%)	41 (1.0)	36 (0.6)	38 (0.5)	51 (0.6)	30 (0.5)	34 (0.5)
Diabetes (%)	268 (6.8)	477 (8.1)	684 (8.4)	618 (6.8)	390 (6.4)	300 (4.6)
Severe mental illness (%)	47 (1.2)	62 (1.1)	54 (0.7)	46 (0.5)	29 (0.5)	33 (0.5)
Atrial fibrillation (%)	290 (7.4)	319 (5.4)	397 (4.9)	386 (4.3)	220 (3.6)	225 (3.4)
Medications at baseline						
Antihypertensive (%)	1283 (32.5)	1898 (32.3)	2851 (35.2)	3273 (36.2)	2337 (38.3)	2444 (37.4)
IC (%)	2221 (56.3)	3214 (54.8)	4557 (56.3)	5280 (58.3)	3617 (59.3)	3874 (59.2)
LABA (%)	637 (16.2)	873 (14.9)	1271 (15.7)	1263 (14.0)	756 (12.4)	602 (9.2)

Values presented are median with IQR or percentage (%).
*Percentage of missing variables—BMI (56.3%), smoking status (24.4%), TC (71.7%), TG (80.7%), LDL (85.6%).
BMI, body mass index; CKD, chronic kidney disease; IC, inhaled corticosteroids; IHD, ischaemic heart disease; LABA, long-acting beta-agonists; LDL, low-density lipoprotein; TC, total cholesterol; TG, triglycerides; YOB, year of birth.;

modelling to investigate the association between SBP and risk of cardiovascular outcomes in those with COPD. The SBP exposure group was included as a categorical variable. Since we motivated our work with findings from the research conducted by Byrd *et al*, we adjusted for the same variables as those chosen in their research: sex, age, body mass index (BMI), smoking status (current, former, never a smoker), beta-blocker use, long-acting beta-agonist (LABA) use and inhaled corticosteroid use.[4] In a second LR model with an expanded set of predictors including known cardiovascular risk factors, we additionally adjusted for triglycerides (TG), low-density lipoprotein (LDL), total cholesterol (TC), atrial fibrillation, rheumatoid arthritis, severe mental illness (psychosis, schizophrenia or bipolar disorder), chronic kidney disease and diabetes. Diagnoses and medication use were identified using validated phenotyping algorithms.[15 18 19] For BMI, TC, TG and LDL, average of the measurements recorded in the 36 months before baseline were computed to minimise

issues of random measurement error.[2 20] We conducted imputations on missing variables to ensure fairer comparison with the DL approach. Multiple imputations using chained equations were implemented (15 imputations) to impute the continuous and categorical missing variables: BMI, TC, TG, LDL and smoking status. Estimation of RR was conducted using the direct standardisation method (further elaboration in online supplemental methods).[21]

Five sensitivity analyses were pursued in our studies using the T-BEHRT model. First, we investigated the association of SBP and cardiovascular risk in patients who had not taken antihypertensives during the follow-up period. Antihypertensives are established medications for lowering high blood pressure, thereby potentially attenuating cardiovascular risk; hence, we conducted this sensitivity analysis in order to investigate the undiluted association between SBP and risk of cardiovascular outcomes in patients with COPD.[22] Second, to investigate the

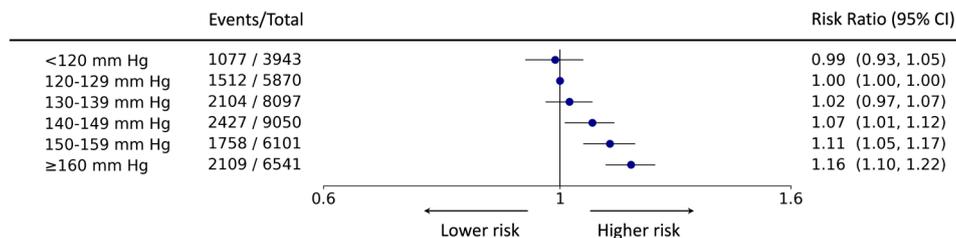|	Events/Total		Risk Ratio (95% CI)
<120 mm Hg	1077 / 3943		0.99 (0.93, 1.05)
120-129 mm Hg	1512 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	2104 / 8097		1.02 (0.97, 1.07)
140-149 mm Hg	2427 / 9050		1.07 (1.01, 1.12)
150-159 mm Hg	1758 / 6101		1.11 (1.05, 1.17)
≥160 mm Hg	2109 / 6541		1.16 (1.10, 1.22)

Figure 2 Forest plot of risk ratio estimates of the Targeted Bidirectional EHR Transformer model with 95% CIs for the association of systolic blood pressure and the primary outcome. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to the reference class, 120–129 mm Hg. The effect size is plotted on a logarithmic scale. For the reference category, there is no CI.

Figure 3 Forest plot of risk ratio estimates of the Targeted Bidirectional EHR Transformer model with 95% CIs for the association of systolic blood pressure and the secondary outcomes. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to the reference class, 120–129 mm Hg. The effect size is plotted on a logarithmic scale. For the reference category, there is no CI.
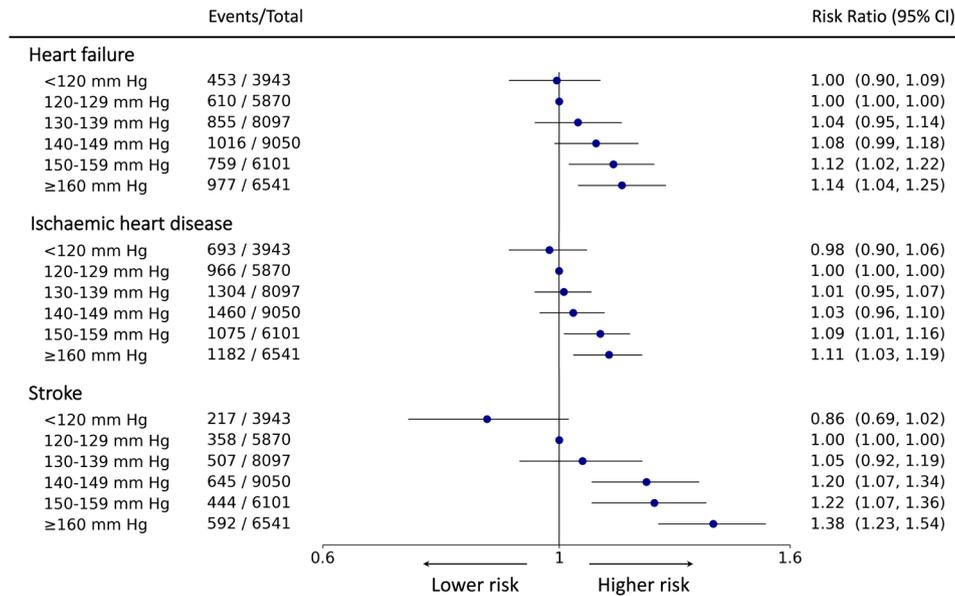
effects of time period, we limited the investigation to only include those with baseline after 1 January 2001. Third and fourth, to mitigate issues of reverse causality, we investigated the primary outcome excluding individuals who had cardiovascular events in the first 12 and 24 months of the follow-up period, respectively. Fifth, in order to investigate the association in smokers, we limited the analysis to only include current and former smokers in the cohort.

Patient and public involvement

Patients were not involved in this research for the development of the research question, exposure definition or the outcome definition. They were not involved in any form for any possible recruitment, design or implementation of this study. There are no current plans to involve patients in the dissemination stage of this study.

RESULTS
Population statistics

A total of 39 602 individuals with COPD at baseline were included in our analysis (figure 1). The median follow-up time was 3.9 years (IQR 1.5–5.0) with 10 987 events, and the median age at baseline, 69 years (IQR 60–76) (table 1). Patients with lower SBP had a higher percentage of atrial fibrillation, chronic kidney disease and IHD and were more likely to be current smokers at baseline. Also, patients with lower SBP had more clinical encounters (medications and diagnoses) recorded in GP/secondary care (online supplemental figure S3). However, individuals with a higher SBP had a higher percentage of antihypertensive usage.

Association of systolic blood pressure and risk of cardiovascular events

The T-BEHRT model estimated a monotonic relationship between SBP and the primary outcomes in patients with COPD (figure 2). By contrast, the crude and adjusted LR estimates of RR both demonstrate a nadir of risk at SBP between 130 and

139 mm Hg (online supplemental figure S4). The adjusted LR model with expanded set of predictors demonstrated similar trends as compared with the base-adjusted LR model (ie, predictors defined in Byrd et al) for the analysis of the primary outcome (online supplemental figure S5A).[4] All models found that ≥160 mm Hg demonstrated greatest risk of cardiovascular events.

In analyses of the components of the primary outcome, the T-BEHRT model showed a monotonic association between SBP and risk of individual cardiovascular end points with lowest risk at <120 mm Hg in comparison with the reference category (figure 3). Additionally, for end points of heart failure and IHD, the crude and adjusted LR estimates of RR found SBP between 130 and 150 mm Hg to contribute to the lowest risk of secondary outcomes (online supplemental figure S6) with little deviation in findings from the adjusted LR approach using the expanded predictor set (online supplemental figure S5B). All four approaches found <120 mm Hg is associated with the lowest risk of stroke. Lastly, the trends discovered in the five sensitivity analyses demonstrated little deviation from the patterns found in the main analysis (figure 4).

DISCUSSION

Using a DL approach for longitudinal EHR, we found that SBP was monotonically associated with cardiovascular risk in 39 602 patients with COPD. Individuals with SBP <120 mm Hg were found to have the lowest risk of both the primary and secondary outcomes with little material deviation in the trends found in the sensitivity analyses.

SBP is established to be log-linearly associated with cardiovascular risk in the general population and in fact, naturally below average blood pressure values in industrialised communities.[3 23 24] However, in groups with prior CVDs and associated risk factors, the relationship remains insufficiently described. In this context of high-risk patients—such as those with diabetes, IHD and other risk factors at study entry—many observational studies reject the monotonic relationship between SBP and cardiovascular risk,
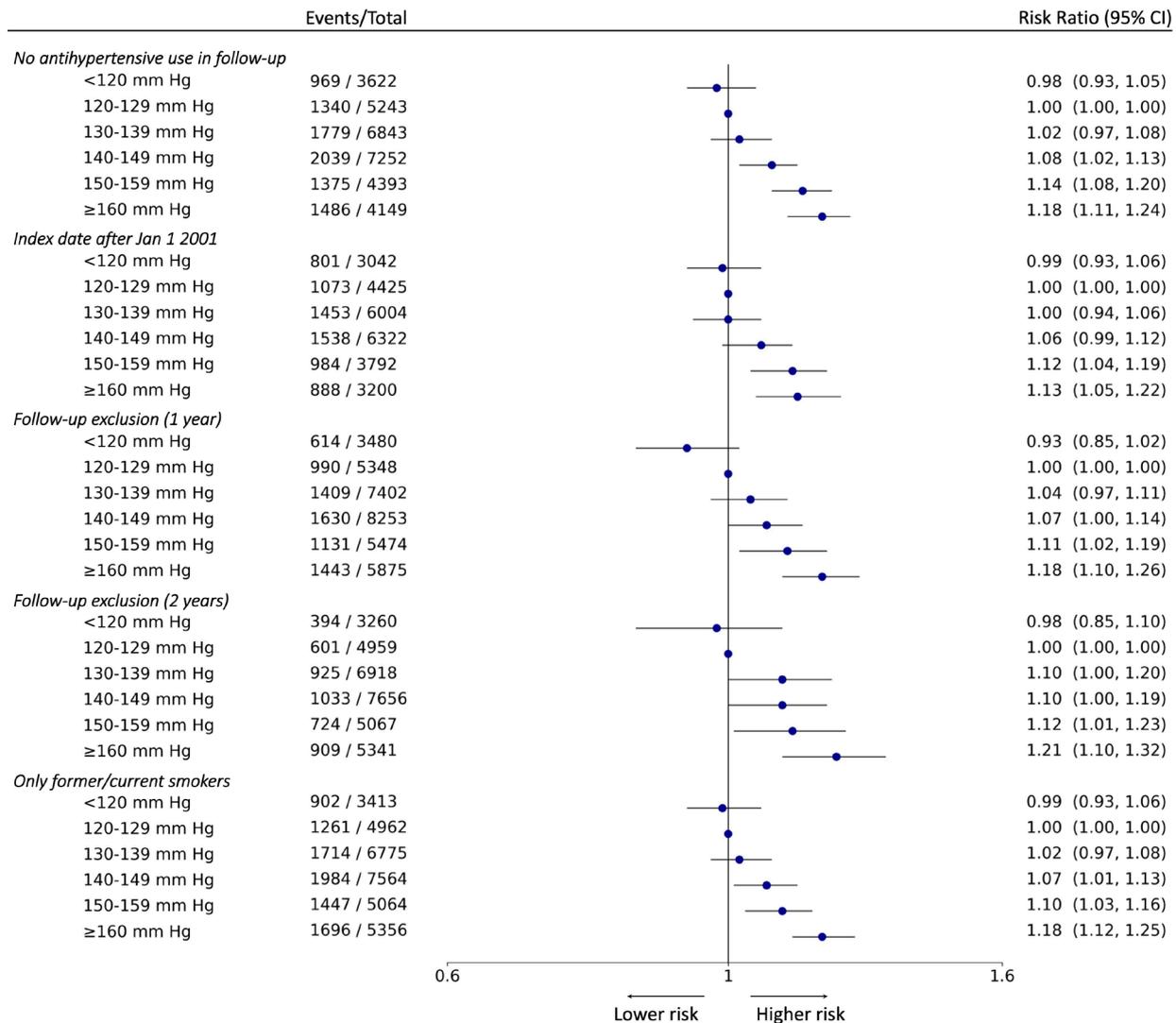
	Events/Total		Risk Ratio (95% CI)
No antihypertensive use in follow-up			
<120 mm Hg	969 / 3622		0.98 (0.93, 1.05)
120-129 mm Hg	1340 / 5243		1.00 (1.00, 1.00)
130-139 mm Hg	1779 / 6843		1.02 (0.97, 1.08)
140-149 mm Hg	2039 / 7252		1.08 (1.02, 1.13)
150-159 mm Hg	1375 / 4393		1.14 (1.08, 1.20)
≥160 mm Hg	1486 / 4149		1.18 (1.11, 1.24)
Index date after Jan 1 2001			
<120 mm Hg	801 / 3042		0.99 (0.93, 1.06)
120-129 mm Hg	1073 / 4425		1.00 (1.00, 1.00)
130-139 mm Hg	1453 / 6004		1.00 (0.94, 1.06)
140-149 mm Hg	1538 / 6322		1.06 (0.99, 1.12)
150-159 mm Hg	984 / 3792		1.12 (1.04, 1.19)
≥160 mm Hg	888 / 3200		1.13 (1.05, 1.22)
Follow-up exclusion (1 year)			
<120 mm Hg	614 / 3480		0.93 (0.85, 1.02)
120-129 mm Hg	990 / 5348		1.00 (1.00, 1.00)
130-139 mm Hg	1409 / 7402		1.04 (0.97, 1.11)
140-149 mm Hg	1630 / 8253		1.07 (1.00, 1.14)
150-159 mm Hg	1131 / 5474		1.11 (1.02, 1.19)
≥160 mm Hg	1443 / 5875		1.18 (1.10, 1.26)
Follow-up exclusion (2 years)			
<120 mm Hg	394 / 3260		0.98 (0.85, 1.10)
120-129 mm Hg	601 / 4959		1.00 (1.00, 1.00)
130-139 mm Hg	925 / 6918		1.10 (1.00, 1.20)
140-149 mm Hg	1033 / 7656		1.10 (1.00, 1.19)
150-159 mm Hg	724 / 5067		1.12 (1.01, 1.23)
≥160 mm Hg	909 / 5341		1.21 (1.10, 1.32)
Only former/current smokers			
<120 mm Hg	902 / 3413		0.99 (0.93, 1.06)
120-129 mm Hg	1261 / 4962		1.00 (1.00, 1.00)
130-139 mm Hg	1714 / 6775		1.02 (0.97, 1.08)
140-149 mm Hg	1984 / 7564		1.07 (1.01, 1.13)
150-159 mm Hg	1447 / 5064		1.10 (1.03, 1.16)
≥160 mm Hg	1696 / 5356		1.18 (1.12, 1.25)

0.6   1   1.6

Lower risk   Higher risk

Figure 4 Forest plot of risk ratio estimates of the Targeted Bidirectional EHR Transformer model with 95% CIs in sensitivity analyses. From the left, the specific sensitivity analysis is annotated and the six exposure groups are shown indented in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to the reference class, 120–129 mm Hg. The effect size is plotted on a logarithmic scale. For the reference category, there is no CI.

concluding a J-shaped pattern.[4 5 25] However, these observational studies are criticised for improperly dealing with manifestations of reverse causality and confounding. With cardiometabolic multimorbidity at baseline more prevalent in those with lower SBP than higher, additional variables capturing this poor baseline health and associated cardiovascular illnesses must be included for adjustment. Given an insufficient understanding of risk and protection in multimorbid patients currently, solely relying on expert selection of known confounders (eg, gender, age, BMI, known risk factors of CVD) exposes the modelling to issues of residual confounding.[26] As a result, unadjusted confounding due to multimorbidity in lower SBP groups can result in the J-shaped pattern: an optimum exists such that SBP below and above is associated with higher cardiovascular risk.[4 27]

In our own implementation of conventional regression modelling, adjusting for predictors as previously defined in Byrd *et al*, the results captured this described J-shaped pattern and rejected the established log-linear relationship between SBP and risk of cardiovascular outcomes.[3 4] Even the fully adjusted LR model with the expanded set of predictors resulted in a non-monotonic trend across analyses of both primary and secondary outcomes.

Implementing the DL approach for assessing the studied association directly confronted these modelling issues. By using minimally processed diagnoses and medications data in routine clinical EHR, our DL approach accounts for a breadth of risk factors potentially confounding the exposure-outcome relationship. In our cohort with COPD and cardiometabolic multimorbidity at baseline, in which traditional approaches failed to sufficiently capture confounding factors in observational data, our approach was appropriately implemented to model the association between SBP and risk of cardiovascular events.

The monotonic association concluded in this work raises important clinical questions for cardiovascular care. What is the optimal SBP in patients with COPD? Does this threshold differ from the recommendations for the general population (<120 mm Hg)? While guidelines for hypertension indeed endorse blood pressure lowering in patients with concomitant COPD and high blood pressure, the recommendations suggest a treatment target of <130 mm Hg (<140 mm Hg in the elderly).[28] Our results demonstrated an infimum of risk at SBP of <120 mm Hg—consistent with the established log-linear understanding of the association between SBP and cardiovascular risk.

Naturally, our investigation does not answer questions relating to antihypertensive treatment effects. Hence, while our study in isolation is insufficient for recommending revisions of hypertension guidelines, our investigation sheds light on the aetiological nature of SBP and CVD in those with COPD—imperative, especially since randomised evidence of blood pressure-lowering therapies in patients with COPD is unavailable and likely to remain unavailable in the near future. While (1) external validation of the studied association would be prudent and (2) in-depth investigations of the association between antihypertensive and CVD (in at least the observational capacity) are needed to comprehensively capture all facets of the relationships between blood pressure, antihypertensives and CVD risk in patients with COPD, our investigation serves as one such source of well-adjusted evidence.

Strengths and limitations

First, in terms of data, the comprehensive information provided by CPRD is a strength of our research. The linkage capabilities of CPRD allow the capture of rich health encounters (eg, diagnoses, medications, measurements, static attributes) from various sources including primary care, secondary care and mortality-based datasets. With access to rich EHR, our DL approach could better extract confounders, both known and latent in routine clinical data as shown in past investigations of SBP and CVD risk in high-risk patients.[10 12] Second, with access to repeated SBP measurements specifically, we were able to derive a summary value (mean value of multiple SBP measurements) limiting issues of measurement error.[20] Third, we were able to capture many more patients than prior studies investigating this association, and also, unlike previous studies of SBP and cardiovascular risk, we included older aged patients and those with cardiovascular multimorbidity at baseline.[4] Exclusion from our study was limited, thereby allowing understanding of the association of SBP and cardiovascular outcomes in high-risk subgroups with COPD. Fourth, rich longitudinal data in CPRD afforded us the opportunity to follow patients for a median of 3.9 years as opposed to the prior exploration of this association in patients with COPD, which reported median follow-up of 1.9 years.[4] With a longer follow-up period, potential biases in RR estimation due to issues of reverse causation are mitigated. Fifth, we explored various sensitivity analyses in order to understand the role of unforeseen biases (eg, reverse causality) and supplement the narrative of the main results. In terms of modelling, a strength of our work is the DL approach capable of extracting and adjusting for confounding factors in rich annotated EHR.[10 12] Additionally, we implemented two varieties of the conventional statistical approach with validated predictor sets allowing direct comparison with the DL approach.[4] By using superior confounding adjustment methods, we demonstrated the utility of DL modelling ultimately rejecting the evidence of a J-shaped relationship.

In terms of limitations, while EHR data in CPRD have some degree of diagnostic recording error, past studies have validated the primary care, secondary care and mortality-based sources within the CPRD database for observational research.[11 14 15] Also, SBP variability is a concern; we have attempted to ameliorate issues of random measurement error by taking an average of repeat measurements over the course of 12 months following baseline as recommended by previous works.[20] Furthermore, more accurate consideration of the outcome and censoring with time-to-event modelling is needed. Given the nascent stage of deep survival modelling for EHR, further methodological innovation is required to fuse DL-based causal models and survival framework modelling.[29] Also, methods that can interpret confounding capture conducted by T-BEHRT would be useful for fully characterising DL estimation processes. While importance of adjusted variables can be readily assessed in the conventional approach, auxiliary methods to extract and decompose the confounders captured by T-BEHRT into explicit medical history variables would lend insight into shared risk factors of blood pressure and CVD. In terms of adjustment, while overadjustment (collider variable adjustment and M-structure bias) is a theoretical concern, empirical research has shown that conditioning on all pre-exposure variables in similar types of EHR studies does not lead to biased estimates.[30] Additionally, we have attempted to further mitigate this potential issue by defining a clear baseline with adjustment specifically up to baseline. Lastly, as is true with all observational studies, residual confounding cannot be completely ruled out even with more complex confounding adjustment approaches (eg, T-BEHRT).

CONCLUSION

We found that patients with COPD in the lowest category of SBP of <120 mm Hg has the lowest risk of cardiovascular events during follow-up. Our findings capture a monotonic relationship between SBP and risk of cardiovascular events in patients with COPD and were in line with the established clinical understanding of the monotonic relationship between SBP and cardiovascular risk in the general population.

ORCID iDs
Shishir Rao http://orcid.org/0000-0001-7331-9416
Mohammad Mamouei http://orcid.org/0000-0002-7757-5238

REFERENCES

1 Ettehad D, Emdin CA, Kiran A, *et al*. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *Lancet* 2016;387:957–67.
2 Blood Pressure Lowering Treatment Trialists' Collaboration. Pharmacological blood pressure lowering for primary and secondary prevention of cardiovascular disease across different levels of blood pressure: an individual participant-level data meta-analysis. *Lancet* 2021;397:1625–36.
3 Whelton SP, McEvoy JW, Shaw L, *et al*. Association of normal systolic blood pressure level with cardiovascular disease in the absence of risk factors. *JAMA Cardiol* 2020;5:1011–8.
4 Byrd JB, Newby DE, Anderson JA, *et al*. Blood pressure, heart rate, and mortality in chronic obstructive pulmonary disease: the Summit trial. *Eur Heart J* 2018;39:3128–34.
5 Adamsson Eryd S, Gudbjörnsdottir S, Manhem K, *et al*. Blood pressure and complications in individuals with type 2 diabetes and no previous cardiovascular disease: national population based cohort study. *BMJ* 2016;354:i4070.
6 Chen W, Thomas J, Sadatsafavi M, *et al*. Risk of cardiovascular comorbidity in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Lancet Respir Med* 2015;3:631–9.
7 Huang Y, Cai X, Mai W, *et al*. Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis. *BMJ* 2016;355:i5953.
8 Rahimi K, Emdin CA, MacMahon S. The epidemiology of blood pressure and its worldwide management. *Circ Res* 2015;116:925–36.
9 Kang Y-Y, Wang J-G. The J-curve phenomenon in hypertension. *Pulse (Basel)* 2016;4:49–60.
10 Rao S, Li Y, Nazarzadeh M, *et al*. Systolic blood pressure and cardiovascular risk in patients with diabetes: a prospective cohort study. *Hypertension* 2023;80:598–607.
11 Li Y, Rao S, Solares JRA, *et al*. BEHRT: transformer for electronic health records. *Sci Rep* 2020;10:7155.
12 Rao S, Mamouei M, Salimi-Khorshidi G, *et al*. Targeted-BEHRT: deep learning for observational causal inference on longitudinal electronic health records. *IEEE Trans Neural Netw Learn Syst* 2022;PP:1–12.
13 Rao S, Li Y, Ramakrishnan R, *et al*. An explainable transformer-based deep learning model for the prediction of incident heart failure. *IEEE J Biomed Health Inform* 2022;26:3362–72.
14 Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
15 Tran J, Norton R, Conrad N, *et al*. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: a population-based cohort study. *PLoS Med* 2018;15:e1002513.
16 Denaxas S, Gonzalez-Izquierdo A, Direk K, *et al*. UK phenomics platform for developing and validating EHR phenotypes: CALIBER. *BioRxiv* 2019.
17 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
18 Payne R, Denholm R. CPRD product code lists used to define long-term preventative, high-risk, and palliative medication. University of Bristol; 2018.
19 Chalmers JD, Poole C, Webster S, *et al*. Assessing the healthcare resource use associated with inappropriate prescribing of inhaled corticosteroids for people with chronic obstructive pulmonary disease (COPD) in gold groups A or B: an observational study using the Clinical Practice Research Datalink (CPRD). *Respir Res* 2018;19:63.
20 Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. *BMJ* 2010;340:bmj.c2289.
21 Hernan M, Robins J. *Causal inference: what if*.
22 Sundström J, Arima H, Woodward M, *et al*. Blood pressure-lowering treatment based on cardiovascular risk: a meta-analysis of individual patient data. *Lancet* 2014;384:591–8.
23 Macmahon S. Blood pressure, stroke, and coronary heart disease *1part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *The Lancet* 1990;335:765–74.
24 Mueller NT, Noya-Alarcon O, Contreras M, *et al*. Association of age with blood pressure across the lifespan in Isolated Yanomami and Yekwana villages. *JAMA Cardiol* 2018;3:1247–9.
25 Vidal-Petiot E, Ford I, Greenlaw N, *et al*. Cardiovascular event rates and mortality according to achieved systolic and diastolic blood pressure in patients with stable coronary artery disease: an international cohort study. *The Lancet* 2016;388:2142–52.
26 Imison C. *Multiple long-term conditions (multimorbidity): making sense of the evidence*. National Institute for Health Research,
27 Cederholm J, Gudbjörnsdottir S, Eliasson B, *et al*. Blood pressure and risk of cardiovascular diseases in type 2 diabetes: further findings from the Swedish national diabetes register (NDR-BP II). *J Hypertens* 2012;30:2020–30.
28 Unger T, Borghi C, Charchar F, *et al*. 2020 international society of hypertension global hypertension practice guidelines. *Hypertension* 2020;75:1334–57.
29 Li Y, Sperrin M, Ashcroft DM, *et al*. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ* 2020;371:m3919.
30 Ding P, Miratrix LW. To adjust or not to adjust? Sensitivity analysis of m-bias and butterfly-bias. *J Causal Inference* 2014;3:41–57.

1

Supplementary Material
Systolic blood pressure, chronic obstructive pulmonary disease, and cardiovascular risk

Shishir Rao, MSc[1,2], Milad Nazarzadeh, MSc[1,2], Yikuan Li, MSc[1,2], Dexter Canoy, PhD [4], Mohammad

Mamouei, PhD[1,2], Gholamreza Salimi-Khorshidi, DPhil[1,2], Kazem Rahimi, DM, FESC[1,2,3]

Table of Contents

2

Supplementary Methods

Study Design
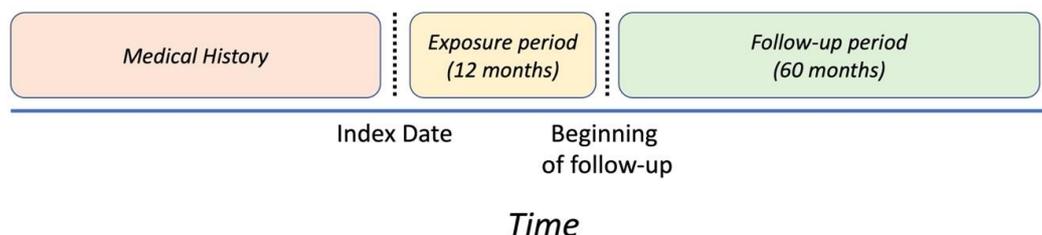
Supplementary Figure S1. Study design of the investigation of the association between systolic blood pressure (SBP) and cardiovascular outcomes in patients with chronic obstructive pulmonary disorder. Index date (baseline) for a given patient is the date of the first SBP measurement recorded between 1990 and 2009 and ages 55 and 90.

A visualisation of the study design can be found in **Supplementary Figure S1.** This visualisation demonstrates the index date, the exposure period where repeat measurements of systolic blood pressure (SBP) are averaged to serve as exposure status, and the follow-up period, which starts 12 months after index date.

Introduction to deep learning and Bidirectional Electronic Health Records Transformer

Deep learning (DL) modelling is a subclass of machine learning (ML), which is in turn a subclass of artificial intelligence (AI) modelling. DL is a more recent paradigm that utilises artificial neural networks to progressively extract more latent and richer features from input data for a given task.

BEHRT, one such DL model, is a Transformer model that has indeed been shown in past works to better represent the complex multimodal EHR than previous DL models such as recurrent and convolutional neural networks in addition to conventional statistical models [3–5]. The flexible BEHRT model allows for including multiple facets of complex EHR data: the encounter itself (e.g., a diagnosis), time information of the encounter (i.e., both age and calendar year), and other attributes such as visit information. While all of these sources of information might provide useful features for utilisation for adjustment in association estimation tasks or risk prediction task, this nuanced data is hard to represent in previous approaches. BEHRT's flexible

architecture allows for encoding this complex arrangement of data, and additionally is able to demonstrate state-of-the-art predictive performance on a host of tasks on EHR data [3–5].

Targeted Bidirectional Electronic Health Records Transformer

We implemented the Targeted Bidirectional Electronic Health Records Transformer (T-BEHRT) for risk ratio (RR) estimation of the association between SBP and cardiovascular outcomes.

In order to include medical history variables in the T-BEHRT model, we conducted some processing of derived CPRD variables. First, the diagnostic records from primary care coded in the Read code format were mapped to the ICD-10 format for consistency with the secondary care coding format (ICD-10). This mapping process yielded a total of 1,497 codes [1]. Second, we mapped the medication codes in the CPRD "product code" format to 386 codes in the BNF coding format [2]. Third, we extracted smoking status (current, former, never a smoker) of a particular patient as the last known status in the 12 months before baseline. Fourth, we extracted patient sex for incorporation as a static variable in the T-BEHRT modelling framework.

A. Targeted BEHRT (T-BEHRT)
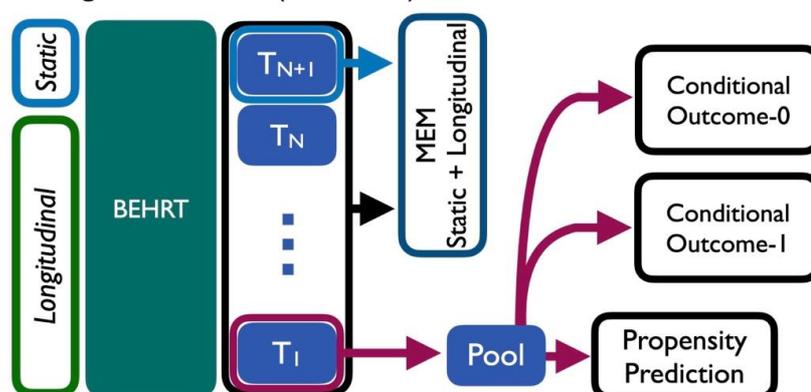
B. Embeddings (*static* and *longitudinal* variables shown above)
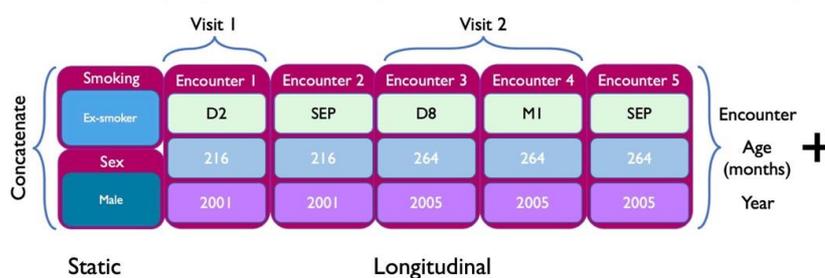
Supplementary Figure S2: T-BEHRT model architecture (A) and embedding design (B). (A) shows the static and longitudinal input, BEHRT feature extractor, the latent outputs for every clinical encounter (outputs T_1 to T_{N+1}) and the tasks for the models: (1) Masked EHR modelling (MEM), (2) propensity score prediction, (3) conditional outcome prediction (given exposure=0 or the reference group), (4) conditional outcome prediction (given exposure=1 or the intervention group). (B) shows the embedding structure. The embeddings include the static and longitudinal embedding structure. The diagnoses (e.g. D2, D8) and medications (e.g. M1) are fed into the model with the appropriate timestamp (age in months and calendar year) of recording. The embeddings for the encounter, age, and year are summed. The SEP element is a separator element used to inform the model that one visit has ended and another has started. The static attributes are similarly represented in high-dimensional embeddings and concatenated to the longitudinal data structure. In, sum the embedding structure incorporates static and longitudinal data inputs. EHR: electronic health records; SEP: Separator; T-BEHRT: Targeted BEHRT; MEM: Masked EHR Modelling

The model combines three advances in DL modelling and semiparametric statistics. First, T-BEHRT utilises a modified BEHRT feature extractor architecture to model both static variables, canonically included in standard epidemiological approaches (e.g., sex, smoking status, etc) and longitudinal variables (e.g., diagnoses/medications) in one unified architecture (**Supplementary Figure S2 A**) [3–5]. Each static variable is inputted as a continuous variable or categorical (or binary) variable. If categorical, all possible values of the variable are represented by a two-dimensional embedding matrix, with each value represented as a vector in this matrix[3]. Longitudinal clinical encounters – diagnoses made at primary/secondary care and medications prescribed – are represented by a similar matrix. Age and calendar year attributes of the event date for a particular diagnosis/prescription are also fed to the model via a similar embedding; in this way, the model can adjust for a confounder, for which the effect may vary across time (**Supplementary Figure S2 B**).

Second, the model utilises unsupervised representation learning to better capture confounding elements latent in input EHR, not explicitly adjusted. The unsupervised framework, Masked EHR Modelling (MEM) is used to extract richer latent representations from both static and longitudinal data for propensity score prediction; the model can better capture pre-exposure variables associations with the exposure thereby better capture confounding elements as well [3,6]. The unsupervised learning is conducted in tandem with the causal predictive framework. This unsupervised objective has been consistently shown to improve causal estimation performance – not just with the T-BEHRT architecture but with other architectures as well [3].

5

Third and lastly, semi-parametric "doubly-robust" estimators have found success in mitigating bias and demonstrating more accurate estimates of causal effect. T-BEHRT modelling is powerful when combined with doubly-robust estimation to further reduce bias. To be able to conduct the doubly-robust estimation, the T-BEHRT DL neural first uses a one-layer neural network to predict propensity score (i.e., probability of being treated with a particular exposure) and next, outcome prediction is conducted with two-layer neural networks. After the DL components are used for prediction, propensity score and outcome estimates are utilised in the cross validated targeted maximum likelihood (doubly-robust) estimation (CV-TMLE) algorithm to update the risk estimates utilising the propensity score estimates [7]. Trimming of propensity score greater than 0.97 and less than 0.03 was conducted before pursuing calculation of RR [3].

Risk ratio estimation for T-BEHRT model

The SBP category of 120–129 mm Hg was considered as the reference exposure group in our study; RR was estimated in comparison to this reference category. For a given comparison to the reference group (e.g. 150–159 mm Hg compared to the reference), the T-BEHRT model was first trained to predict exposure category (propensity score) and outcome with k-fold cross-validation (k=10) implemented for training and testing [3]. Risk estimates and propensity score predictions across the 10 test sets were pooled, and by utilising "doubly-robust" post-hoc estimator, Cross Validated Targeted Maximum Likelihood Estimation (CV-TMLE), the risk estimates were further corrected for selection biases, and RR and 95% confidence intervals are derived [7]. The term "T-BEHRT" and associated model in this paper refers to the estimation framework consisting of (1) estimating risk of outcome and propensity score with DL modelling and (2) updating initial estimates with CV-TMLE in order to estimate RR and 95% CI.

Risk ratio estimation for logistic regression model

Logistic regression modelling (LR) was used for the conventional approach in this work. The modelling utilised direct standardisation method for estimation of the RR [8]. As an example, to estimate the effect of 150-159 mm Hg on cardiovascular outcomes with respect to the reference exposure, the trained LR model predicted risk with exposure for all patients set to the categorical variable representing 150-159 mm Hg and predicted risk with exposure similarly set to the reference group. The RR was derived as the ratio of the average of these two sets of predictions. For theoretical guarantees, we implemented k-fold cross-validation (k=10) for causal estimation [9]. RR was calculated as the average of RR estimations on the 10 individual test

6

sets, and the 95% confidence interval (CI) was calculated via bootstrapping [10]. Lastly, the crude RR was calculated as the ratio between the average empirical risk of outcome in a particular exposure group divided by the same in the reference exposure group.

Implementation details

The code for this work was implemented in the python coding language. The DL models was implemented using Pytorch – a DL framework validated on many past works in DL and EHR specifically [11]. Two graphical processing units (NVIDIA Titan Xp) were used for DL model training and evaluation. Hyperparameters of the model (manually selected, non-trainable parameters of the model) are shown in **Supplementary Table 1**. More details on the DL modelling can be found in the original methods paper [3].

Supplementary Table 1. T-BEHRT model hyperparameters

Hyperparameter	Attribute
Hidden BEHRT size	150
Intermediate BEHRT Layer size	108
Hidden dropout probability	0.3
Attention dropout probability	0.4
Number of hidden layers (BEHRT)	5
Hidden activation functions	Exponential Linear Unit
Initialiser range of parameters	0.02
N (number of tokens/clinical encounters)	300
Mini-batch size	128

Supplementary Results
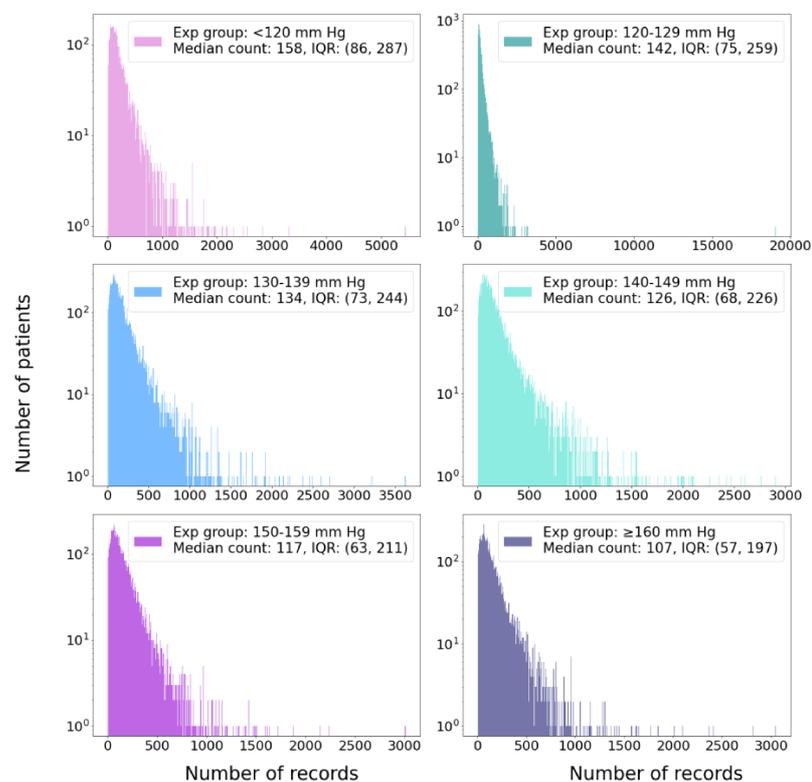
Supplementary Figure S3. Distributions of number of encounters for patients stratified by exposure status. The x axis is counts of the records; y axis the number of patients. Number of bins used for the distribution analyses is 500 for each of the six analyses. Median number of records with interquartile range (IQR) is presented in the legend for each of the six analyses.

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

Heart

8

Supplementary Figure S4. Forest plot of risk ratio estimates of the crude and adjusted logistic regression (LR) models with 95% confidence intervals (CI) for association of systolic blood pressure and the primary composite outcome. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, 120-129 mm Hg. The effect size is plotted on a logarithmic scale. For the reference category, there is no confidence interval.

A. Primary outcome

	Events/Total		Risk Ratio (95% CI)
Cardiovascular death, ischemic heart disease, stroke, heart failure			
<120 mm Hg	1077 / 3943		1.06 (1.04, 1.08)
120-129 mm Hg	1512 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	2104 / 8097		0.99 (0.98, 1.01)
140-149 mm Hg	2427 / 9050		1.02 (1.00, 1.03)
150-159 mm Hg	1758 / 6101		1.07 (1.05, 1.10)
≥160 mm Hg	2109 / 6541		1.19 (1.17, 1.20)

0.6   1   1.6
Lower risk   Higher risk

B. Secondary outcomes

	Events/Total		Risk Ratio (95% CI)
Heart failure			
<120 mm Hg	453 / 3943		1.10 (1.05, 1.15)
120-129 mm Hg	610 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	855 / 8097		1.00 (0.97, 1.02)
140-149 mm Hg	1016 / 9050		1.03 (1.01, 1.06)
150-159 mm Hg	759 / 6101		1.12 (1.08, 1.16)
≥160 mm Hg	977 / 6541		1.30 (1.27, 1.33)
Ischaemic heart disease (myocardial infarction and coronary heart disease)			
<120 mm Hg	693 / 3943		1.07 (1.03, 1.10)
120-129 mm Hg	966 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	1304 / 8097		0.97 (0.95, 1.00)
140-149 mm Hg	1460 / 9050		0.98 (0.96, 0.99)
150-159 mm Hg	1075 / 6101		1.06 (1.02, 1.10)
≥160 mm Hg	1182 / 6541		1.11 (1.09, 1.13)
Stroke			
<120 mm Hg	217 / 3943		0.89 (0.85, 0.93)
120-129 mm Hg	358 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	507 / 8097		1.02 (0.97, 1.07)
140-149 mm Hg	645 / 9050		1.15 (1.10, 1.20)
150-159 mm Hg	444 / 6101		1.15 (1.10, 1.19)
≥160 mm Hg	592 / 6541		1.36 (1.31, 1.41)

0.6   1   1.6
Lower risk   Higher risk

Supplementary Figure S5. Forest plot of risk ratio estimates of the adjusted logistic regression (LR) model with extended predictor set with 95% confidence intervals (CI) for association of systolic blood pressure and (A) the primary outcome and (B) the secondary outcomes. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, 120-129 mm Hg. The effect size is plotted on a logarithmic scale. For the reference category, there is no confidence interval.

10

	Events/Total		Risk Ratio (95% CI)

Heart failure

Crude
<120 mm Hg	453 / 3943		1.11 (1.11, 1.11)
120-129 mm Hg	610 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	855 / 8097		1.02 (1.02, 1.02)
140-149 mm Hg	1016 / 9050		1.08 (1.08, 1.08)
150-159 mm Hg	759 / 6101		1.20 (1.20, 1.20)
≥160 mm Hg	977 / 6541		1.44 (1.44, 1.44)

LR
<120 mm Hg	453 / 3943		1.11 (1.06, 1.16)
120-129 mm Hg	610 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	855 / 8097		0.99 (0.97, 1.01)
140-149 mm Hg	1016 / 9050		1.01 (0.98, 1.04)
150-159 mm Hg	759 / 6101		1.07 (1.03, 1.12)
≥160 mm Hg	977 / 6541		1.23 (1.20, 1.26)

Ischaemic heart disease (myocardial infarction and coronary heart disease)

Crude
<120 mm Hg	693 / 3943		1.07 (1.07, 1.07)
120-129 mm Hg	966 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	1304 / 8097		0.98 (0.98, 0.98)
140-149 mm Hg	1460 / 9050		0.98 (0.98, 0.98)
150-159 mm Hg	1075 / 6101		1.07 (1.07, 1.07)
≥160 mm Hg	1182 / 6541		1.10 (1.10, 1.10)

LR
<120 mm Hg	693 / 3943		1.08 (1.05, 1.11)
120-129 mm Hg	966 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	1304 / 8097		0.97 (0.94, 0.99)
140-149 mm Hg	1460 / 9050		0.94 (0.93, 0.96)
150-159 mm Hg	1075 / 6101		1.02 (0.99, 1.06)
≥160 mm Hg	1182 / 6541		1.03 (1.01, 1.06)

Stroke

Crude
<120 mm Hg	217 / 3943		0.90 (0.90, 0.90)
120-129 mm Hg	358 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	507 / 8097		1.03 (1.03, 1.03)
140-149 mm Hg	645 / 9050		1.17 (1.17, 1.17)
150-159 mm Hg	444 / 6101		1.19 (1.19, 1.19)
≥160 mm Hg	592 / 6541		1.48 (1.48, 1.48)

LR
<120 mm Hg	217 / 3943		0.91 (0.88, 0.94)
120-129 mm Hg	358 / 5870		1.00 (1.00, 1.00)
130-139 mm Hg	507 / 8097		1.01 (0.96, 1.06)
140-149 mm Hg	645 / 9050		1.10 (1.06, 1.15)
150-159 mm Hg	444 / 6101		1.10 (1.06, 1.15)
≥160 mm Hg	592 / 6541		1.29 (1.23, 1.34)

0.6   1   1.6

← Lower risk   Higher risk →

Supplementary Figure S6. Forest plot of risk ratio estimates of the crude and adjusted logistic regression (LR) models with 95% confidence intervals (CI) for association of systolic blood pressure and the secondary outcomes. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio

estimates are shown in the right-most column relative to reference class, 120-129 mm Hg. The effect size is

plotted on a logarithmic scale. For the reference category, there is no confidence interval.

Supplementary References

1. NHS Digital. Read Code Map, https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9 (2020).

2. Trowell WJ. "British National Formulary". *British Medical Journal (Clinical research ed.)*. Epub ahead of print 1981. DOI: 10.1136/bmj.282.6269.1078.

3. Rao S, Mamouei M, Salimi-Khorshidi G, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Rahimi K. Targeted-BEHRT: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records. *IEEE Trans Neural Netw Learn Syst* 2022; 1–12.

4. Rao S, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Cleland JG, Lukasiewicz T, Salimi-Khorshidi G, Rahimi K. An explainable Transformer-based deep learning model for the prediction of incident heart failure. *IEEE J Biomed Health Inform* 2022; 1–1.

5. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: Transformer for Electronic Health Records. *Sci Rep* 2020; 10: 7155.

6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. Epub ahead of print 1983. DOI: 10.1093/biomet/70.1.41.

7. Levy J. An Easy Implementation of CV-TMLE. *arXiv*.

8. Hernan M, Robins J. Causal inference: what if.

9. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 2018; 21: C1–C68.

10. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. 1982. Epub ahead of print 1982. DOI: 10.1137/1.9781611970319.

11. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library, http://arxiv.org/abs/1912.01703 (2019).

12. Tran J, Norton R, Conrad N, Rahimian F, Canoy D, Nazarzadeh M, Rahimi K. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: A population-based cohort study. *PLoS Med* 2018; 15: e1002513.

13. Rao S, Li Y, Nazarzadeh M, Canoy D, Mamouei M, Hassaine A, Salimi-Khorshidi G, Rahimi K. Systolic Blood Pressure and Cardiovascular Risk in Patients With Diabetes: A Prospective Cohort Study. *Hypertension*. Epub ahead of print 30 December 2022. DOI: 10.1161/HYPERTENSIONAHA.122.20489.

14. Nazarzadeh M, Bidel Z, Canoy D, Copland E, Bennett DA, Dehghan A, Davey Smith G, Holman RR, Woodward M, Gupta A, Adler AI, Wamil M, Sattar N, Cushman WC, McManus RJ, Teo K, Davis BR, Chalmers J, Pepine CJ, et al. Blood pressure-lowering treatment for prevention of major cardiovascular diseases in people with and without type 2 diabetes: an individual participant-level data meta-analysis. *Lancet Diabetes Endocrinol* 2022; 10: 645–654.

15. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: Longitudinal cohort study using cardiovascular disease as exemplar. *The BMJ*; 371. Epub ahead of print 2020. DOI: 10.1136/bmj.m3919.