BMJ Open

Longitudinal cohorts for harnessing the electronic health record for disease prediction in a US population

Sheila M Manemann,[1] Jennifer L St Sauver ,[1] Hongfang Liu,[2] Nicholas B Larson,[1] Sungrim Moon,[2] Paul Y Takahashi,[3] Janet E Olson,[1] Walter A Rocca ,[1,4,5] Virginia M Miller,[5,6,7,8] Terry M Therneau,[1] Che G Ngufor,[2] Veronique L Roger,[9,10] Yiqing Zhao,[1] Paul A Decker,[1] Jill M Killian,[1] Suzette J Bielinski [1]

To cite: Manemann SM, St Sauver JL, Liu H, *et al*. Longitudinal cohorts for harnessing the electronic health record for disease prediction in a US population. *BMJ Open* 2021;**11**:e044353. doi:10.1136/bmjopen-2020-044353

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-044353).

Received 01 September 2020
Accepted 18 May 2021

Check for updates

© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Suzette J Bielinski;
bielinski.suzette@mayo.edu

ABSTRACT

Purpose The depth and breadth of clinical data within electronic health record (EHR) systems paired with innovative machine learning methods can be leveraged to identify novel risk factors for complex diseases. However, analysing the EHR is challenging due to complexity and quality of the data. Therefore, we developed large electronic population-based cohorts with comprehensive harmonised and processed EHR data.

Participants All individuals 30 years of age or older who resided in Olmsted County, Minnesota on 1 January 2006 were identified for the discovery cohort. Algorithms to define a variety of patient characteristics were developed and validated, thus building a comprehensive risk profile for each patient. Patients are followed for incident diseases and ageing-related outcomes. Using the same methods, an independent validation cohort was assembled by identifying all individuals 30 years of age or older who resided in the largely rural 26-county area of southern Minnesota and western Wisconsin on 1 January 2013.

Findings to date For the discovery cohort, 76 255 individuals (median age 49; 53% women) were identified from which a total of 9 644 221 laboratory results; 9 513 840 diagnosis codes; 10 924 291 procedure codes; 1 277 231 outpatient drug prescriptions; 966 136 heart rate measurements and 1 159 836 blood pressure (BP) measurements were retrieved during the baseline time period. The most prevalent conditions in this cohort were hyperlipidaemia, hypertension and arthritis. For the validation cohort, 333 460 individuals (median age 54; 52% women) were identified and to date, a total of 19 926 750 diagnosis codes, 10 527 444 heart rate measurements and 7 356 344 BP measurements were retrieved during baseline.

Future plans Using advanced machine learning approaches, these electronic cohorts will be used to identify novel sex-specific risk factors for complex diseases. These approaches will allow us to address several challenges with the use of EHR.

Strengths and limitations of this study

- By capitalising on the untapped depth and breadth of clinical data available in modern electronic health record (EHR) systems, we can go beyond traditional risk factors and create comprehensive risk profiles for complex diseases.
- We created an independent validation cohort of patients from a largely rural area in which to assess the generalisability of our findings from our discovery cohort.
- We have biological samples and genomic data in a large subset of patients.
- Using innovative machine learning methods will allow us to address several important and challenging questions associated with the use of EHR data.
- One limitation of this study is that it may be difficult to develop accurate and transportable EHR phenotype algorithms for some female-specific conditions or procedures.

INTRODUCTION

The wide adoption of electronic health records (EHRs) has led to an unprecedented expansion in the availability of comprehensive longitudinal datasets for research.[1] Thus, EHR systems represent an untapped resource for studying life-course biology, multimorbidity and the prediction of complex diseases, such as cardiovascular disease (CVD), dementia, cancers and other ageing-related diseases. However, deriving data from EHRs is challenging and requires extensive harmonisation and processing guided by content experts.

As opposed to research cohort data sources that typically measure a limited set of factors, EHRs capture the full-range of clinical data. However, analysing EHR data can be challenging due to the complex and uneven nature of clinical documentation and data quality.[2] Hallmark challenges for leveraging EHR data in predictive modelling include high degrees of data sparsity, incompleteness, noise and biases.[3 4] Furthermore, changing and evolving EHR systems within

and between institutions add another layer of complexity. Thus data extraction, cleaning, harmonisation, interpretation, management and analyses are major challenges for efficient EHR-based clinical studies.

However, recent advances in data science and machine learning aim to address the uneven nature of clinical documentation intrinsic to EHR data. For example, EHR-based deep learning methods have been proposed for handling missing data imputation,[5] as well as for extracting high-level patient data patterns for prediction algorithms.[6] Furthermore, extensive effort has been dedicated to develop advanced clinical data processing (eg, natural language processing (NLP) technologies) and data management methodologies (eg, ontology-based approaches) to facilitate EHR-based clinical studies.[7 8] Importantly, NLP methods allow for the ascertainment of risk factors recorded in the medical history section of clinical notes that predate EHR systems and/or occurred at another medical centre. Moreover, EHR phenotyping algorithms incorporating multiple data types may be accurate, scalable and transportable.[9 10] Thus, our goal is to capitalise on the depth and breadth of clinical data within the EHR systems to revolutionise risk prediction and to optimise personalised care for every patient.

In order to achieve this goal, we assembled a longitudinal cohort of adult patients in a geographically defined area in southeastern Minnesota, a state in the Upper Midwest region of the USA, to serve as the discovery cohort. Comprehensive EHR data over a 15-year period were ascertained, allowing for complete ascertainment of risk factor profiles. Thus, we have the ability to move beyond traditional risk factors to include reproductive factors, age of risk factor onset and a broader spectrum of clinical tests, diagnoses and patient provided information (PPI). Importantly, we have also created a validation cohort of patients from a largely rural area in which to assess the generalisability of our findings and models. With detailed and rich EHR data, these population-based cohorts can be used for a wide-range of studies, including but not limited to studying novel disease associations (risk factors), clusters of disease or creating sex-specific risk scores for disease prediction.

COHORT DESCRIPTION

Setting

Our study uses the resources of the Rochester Epidemiology Project (REP).[11 12] In brief, the REP is a records-linkage system which allows retrieval of nearly all healthcare utilisation and outcomes of residents living in Olmsted County, the home of Mayo Clinic.[12] Thus, the REP captures and updates comprehensive EHR-derived phenotypic data within this population, and is uniquely positioned to characterise longitudinal disease trajectories and outcomes in communities. The electronic indexes of the REP include demographic information, diagnostic and procedure codes, healthcare utilisation data, outpatient drug prescriptions, results of laboratory tests and information about smoking, height, weight and body mass index (BMI).

Starting in 2010, the REP population expanded to include an additional 26-county region in southern Minnesota and western Wisconsin. The REP now includes medical record data from many sources of care across the region including the two largest providers of care in these areas (ie, Mayo Clinic, Mayo Clinic Health System clinics and hospitals, and Olmsted Medical Center and its affiliated clinics).[11] The expansion of the REP from 1 to 27 counties in the Upper Midwest has increased the size of the population fivefold, and its adoption of innovative electronic platforms are important assets to follow our cohorts. The expanded population now offers breadth and depth of data for a large sample size, thereby providing a powerful resource for more precise risk prediction. Importantly, the recent REP expansion markedly increased the proportion of persons living in rural areas to 50%.[11] Additionally, the REP region has similar age, sex and ethnic characteristics as the entire Upper Midwest region of the USA.[11 12]

Our cohorts, updated under the auspices of the expanded REP, will offer a singular opportunity to address the disproportionate burden of disease experienced by rural populations. Rural disparities have been recently underscored by the Centers for Disease Control and Prevention and the American Heart Association, which have called for studies to understand and address these disparities.[13 14]

All individuals 30 years of age or older who resided in Olmsted County, Minnesota on 1 January 2006 were identified for the discovery cohort (figure 1). An age cut-off of 30 was selected because ageing-related diseases are infrequent in children and adults aged 18–29. Additionally, traditional risk factors are not routinely screened in this younger population. The Mayo Clinic EHR began phasing in during the 1990s, and primary care and most specialty departments were added by 2000. Thus an index date of 1 January 2006 was selected to allow a sufficient time period for the electronic ascertainment of patient characteristics and risk factors from the EHR, for the identification of prevalent disease and for more than a decade of follow-up to assess incident or secondary events. Similarly, all individuals 30 years of age or older who were residing in the other 26 counties in southern Minnesota and western Wisconsin on 1 January 2013 were identified for the validation cohort. This region has EHR history beginning in 2010; thus, the index date of 1 January 2013 allows an ample window of time for the collection of data and follow-up.

This study was approved by the Mayo Clinic and Olmsted Medical Center Institutional Review Boards.

Data collection

Baseline data were collected from 2001 to 2005 for the discovery cohort and from 2010 to 2012 for the validation cohort (figure 1). Follow-up for outcomes is ongoing for both cohorts.

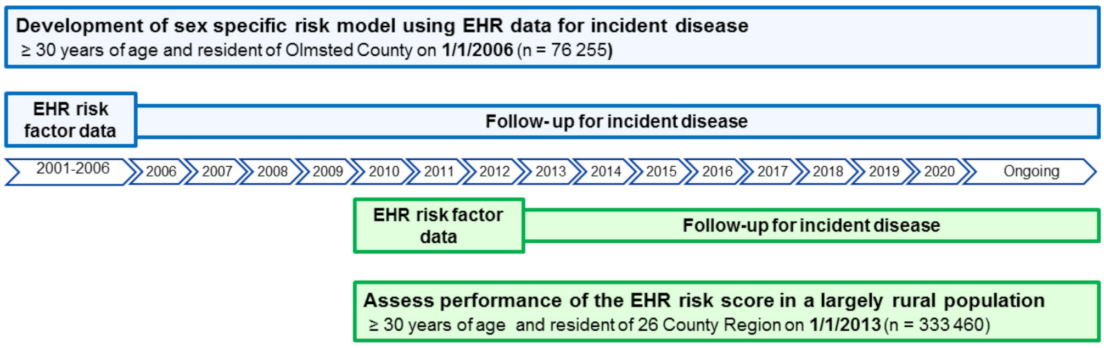

Figure 1 Framework for building electronic health records based risk scores for incident disease. EHR, electronic health record.

Details of our data processing, management and algorithm development are detailed below. All data were collected via the REP, unless indicated otherwise.

Exposures

Demographics

Date of birth, sex, race and ethnicity were obtained. Within the REP, race is classified per the US Census: White, Black, Asian, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander. Categories of 'Other and mixed' and 'Unknown' are also included.[11] Ethnicity is classified per the US Census: Hispanic or non-Hispanic.

Baseline clinical measurements

Heart rate

The median heart rate per calendar day was used for analyses. The most recent daily median heart rate among all heart rate measurements for a person during the baseline data collection period was considered the baseline heart rate. All daily values were retained to assess associations with heart rate variability and outcomes.

Blood pressure

All systolic blood pressure (SBP) and diastolic blood pressure (DBP) measurements with values of 0 and values that were not whole numbers were excluded. For each measurement the following criteria were applied:

1. If the measurement was <1000, it was kept as is.
2. If the measurement was between 1000 and 9999, it was assumed that it was recorded as a two digit SBP and two digit DBP and split apart.
3. If the measurement was ≥10 000, then it was assumed that it was recorded as a three digit SBP and a two digit DBP and split apart.
4. All SBP >300 and all DBP >200 were excluded.

Furthermore, some measurements had time recorded as 00:00 and a real time on the same day. When this occurred, the measurement with time=00:00 was dropped. The median SBP and the median DBP per day were calculated. For each day, any instances where median DBP ≥SBP were deleted. The most recent daily median SBP and DBP among all measurements for a person during the baseline data collection period was considered the baseline blood pressure (BP). All daily values were retained to assess BP variability and outcomes.

Height, weight and BMI

All heights and weights per person were extracted ±5 years of the index date for the discovery cohort and ±3 years of the index date for the validation cohort. Using a published method as a guide,[15] heights <111.8 cm or >228.6 cm and weights <24.9 kg or >453.6 kg were excluded. For those with more than one height, any height values that met both of the two following conditions were excluded: (1) the absolute difference between that particular height and average height was greater than the SD and (2) the SD was >2.5% of the average height. For those with more than one weight, any weight that met one of the two following conditions was excluded: (1) the range was >22.7 kg and the absolute difference between that specific weight and average weight was >70% of the range or (2) the SD was >20% of the average weight and the absolute difference between that particular weight and average weight was greater than the SD. Heights and weights during the baseline period were retained and all possible BMI combinations were calculated (weight (kg)/height (m^2)). The median BMI was calculated and considered the baseline BMI. BMI values <12 or >70 kg/m^2 were excluded.

Smoking and tobacco use status

All prior smoking responses through the index date per person were ascertained. First, the most recent response per person was identified. If current smoker was indicated then the baseline smoking status was set to current user. Likewise, if the most recent self-report listed former smoker, then the baseline smoking status was set to former smoker. Finally, if self-report indicated never/not currently, then all prior responses were reviewed. If former smoker was indicated, then smoking status was set accordingly. Otherwise, smoking status was listed as never smoker. The same algorithm was used for tobacco use status.

Diagnoses

All International Classification of Diseases, Ninth Revision (ICD-9) and Tenth Revision (ICD-10) diagnosis codes during the baseline period were identified and extracted from the REP electronic indexes. Diagnoses were classified according to the Clinical Classifications Software (CCS), developed at the Agency for Healthcare Research and Quality.[16] CCS is a tool for clustering patient diagnoses and procedures into a manageable number of clinically meaningful categories. Additionally, we used the list of 20 chronic conditions recommended by the US Department of Health and Human Services for studying multimorbidity, as defined by ICD-9 and ICD-10 codes.[17 18]

Procedures

Procedure history was defined by identifying and extracting all CurrentProcedural Terminology (CPT) and ICD-9 and ICD-10 procedure codes during the baseline period. Procedures were classified according to the CCS, as described above.

Gynecological surgeries

Gynecologic surgeries often predate EHR systems or occurred at another medical centre, thus we applied NLP techniques to extract them from the medical history sections of the clinical narratives of the Mayo Clinic EHR. A rule-based algorithm collects these concepts to classify the status of the gynaecological surgery per each patient as six mutually exclusive categories: 'no surgery', 'bilateral oophorectomy only', 'hysterectomy and bilateral oophorectomy', 'unilateral oophorectomy only', 'hysterectomy and unilateral oophorectomy' and 'hysterectomy only'. An expansion of this process to the Olmsted Medical Center EHR is planned.

Female reproductive factors

For the women in the discovery cohort, data were extracted from the following Mayo Clinic Rochester sources: Breast Diagnostic and Cancer Clinic Questionnaire from 2005, Mammography Questionnaire from 2003 to 2005, Mammography database from 2004 to 2005 and the Current Visit Information form from 2001 to 2005. For women in the validation cohort, information in these sources, when available, will be extracted prior to index date and will be augmented with NLP.

Age at menarche

The minimum age of menarche and the most recently reported age of menarche was determined. For the women in whom the minimum does not equal the most recently reported age at menarche, the median of all reports (rounded down to a whole number) was used.

Age at birth of first child

The minimum age at birth of first child and the most recently reported age at birth of first child was determined. For the women in whom the minimum does not equal the most recently reported age at birth of first child, the median of all reports (rounded down to a whole number) was used.

Number of pregnancies and number of live births

The maximum reported number of pregnancies and the most recently reported number of pregnancies were determined. For the women in whom the maximum does not equal the most recently reported number of pregnancies, the median of all reported number of pregnancies (rounded up to a whole number) was used.

Similarly, the maximum reported number of live births and the most recently reported number of live births were determined. For the women in whom the maximum does not equal the most recently reported number of live births, the median of all reported number of live births (rounded up to a whole number) was used.

Ever breastfed

If a woman ever previously reported breastfeeding her child, then breastfeeding status was set to yes. Otherwise, if all prior reports of breastfeeding were no, then breastfeeding status was set to no.

Menopausal status

If a woman ever previously reported menopause, then menopausal status was set to yes. Otherwise, if all prior reports of menopause were no, then menopausal status was set to no.

Between-field checks/corrections were performed. For women who reported an age at birth of their first child, but number of pregnancies=0, both fields were set to missing. For women who reported 0 pregnancies and >0 live births both fields were set to missing. For women who reported breastfeeding, but number of pregnancies=0, both fields were set to missing.

Preterm birth and pregnancy complications

Preterm birth and pregnancy complications including gestational diabetes, gestational hypertension, preeclampsia and eclampsia are identified by diagnoses codes.

ECG

All Mayo Clinic ECG quantitative data and narrative and impressions were extracted during the baseline period. Quantitative variables collected include heart rate, P wave, PR interval, QRS interval, QT interval, QT calculated (Bazett) and QT calculated (Fridericia). In addition, raw wave forms for all ECGs are available.

Echocardiography

Echocardiography data were retrieved through the Mayo Clinic Echocardiography database during the baseline period. Methods from prior work were used.[19] Ejection fraction, interventricular septum thickness end diastole, left atrial (LA) volume end systole, LA volume index end systole, left ventricular (LV) internal dimension end diastole, LV internal dimension end systole, mitral valve systolic effective regurgitant orifice, LV mass and LV mass index values were averaged when multiple measurements

were performed. E/A and E/e′ were calculated using the corresponding values. The most severe descriptor word (severe, moderate–severe, moderate, mild–moderate, mild, trivial or none) was used to define aortic regurgitation, aortic stenosis, mitral regurgitation, mitral stenosis, pulmonary regurgitation, pulmonary stenosis, tricuspid regurgitation and tricuspid stenosis. ECG rhythms including atrial fibrillation, atrial flutter and sinus rhythm were ascertained from the echocardiogram. LV size descriptor other than normal (ie, borderline, left, mild, mild–moderate, moderate, moderate–severe or severe) was classified as enlarged. Non-missing values for LV filling pressure were considered increased. LV diastolic dysfunction category (normal, grade 1, grade 1A, grade 2, grade 3, grade 3–4 and grade 4) was collected. Finally, LV wall motion score index was ascertained and when 'no' was indicated the score was set to 1 (normal, ie, no regional wall motion abnormalities).

Prescription medications

All prescriptions during the baseline period were electronically ascertained. Medications were organised according to the National Drug File Reference Terminology (NDF-RT) classifications. For each NDF-RT class, a variable was created to indicate whether each person had received a prescription for that class in the 1 year prior to index.

Laboratory values

All laboratory values were extracted from the electronic laboratory system that started in 1992. Laboratory tests were mapped to Logical Observation Identifiers Names and Codes (LOINC), which is the most widely used classification system for laboratory tests. Tests are often reported in more than 1 unit of measure and LOINC provides a unique code for each.[20]

Qualitative test results were harmonised such that they conformed to a uniform set of unique outcomes. For example, there are 51 unique tests for ABO blood type and Rh factor available in REP within the time period with different textural representations of the same result (eg, B POS, B POSTIVE, B, POS, POSITIVE). During harmonisation two variables were created, ABO Type (possible values of A, B, AB and O) and Rh Type (possible values of negative or positive).

Results such as 'not performed', 'invalid results', 'unable to calculate', etc. were dropped. The midpoint value was retained for all results reported as a range (eg, 0–2=1).

Patient provided information

PPI from Current Visit Information forms, which patients are asked to complete annually at Mayo Clinic, was extracted for the discovery cohort. Sociodemographic data were retrieved including: educational attainment, employment status, relationship status and with whom the patient currently lives. Functional status data were also retrieved including: does the patient have difficulty eating, dressing, using the toilet, bathing or getting in and out of bed; does the patient have difficulty climbing two flights of stairs, does he/she have home care assistance available if needed, is he/she breathing device dependent, is he/she mobility device dependent and does the patient use dentures or hearing aids. The most recent response during the baseline period for each item was retained for baseline.

A modified Katz Index[21] was calculated with the following activities of daily living (ADLs): eating, dressing, using the toilet, bathing or getting in and out of bed. Patients received one point for each ADL that they could perform without difficulty; thus scores could range from 0 (low independence) to 5 (high independence).

Family history of disease

All family history content was retrieved from the 'family history' section of unstructured clinical notes. An NLP pipeline (MedTagger) was used to extract mentions of family members.[22] Disease mentions were extracted using MetaMap API which used Unified Medical Language System (UMLS) dictionary 2018AA.[23 24] UMLS concepts were further mapped to CCS codes. Relationships between family member and disease were extracted using combined semantical rules and distance-based rules.

Biologic specimens

There are two sources of stored biological specimens on a subset of the discovery and validation cohorts. First, the Mayo Clinic Biobank is an institutional resource comprised of over 56 000 volunteers who donated biological specimens, and provided risk factor data, access to EHR data, and consent to participate in additional studies.[25] Biological samples collected on each participant include DNA (median 183 µg), 4 mL serum, 12 mL plasma and an aliquot of frozen white blood cells. The second source of biological samples is the Cardiovascular Disease Repository (CaDRe). CaDRe is a collection of samples (ie, serum, plasma, DNA, buffy coat) collected historically and prospectively from patients with myocardial infarction (MI), coronary artery bypass graft (CABG) surgery, percutaneous coronary interventions (PCI), heart failure and atrial fibrillation in the Olmsted County population.[26–30] Currently, approximately 13 000 persons in the discovery cohort and approximately 9000 participants from the validation cohort are participants in at least one of the above mentioned studies.

Genomic data

In 2019, Mayo Clinic formalised a partnership with Regeneron Pharmaceuticals called Project Generation. As part of this collaboration, exome sequencing and genome-wide association data are being generated for all participants of the Mayo Clinic Biobank and CaDRe, which includes approximately 13 000 persons from the discovery cohort and 9000 participants from the validation cohort. Although we do not have these data for everyone in the cohorts, the genomic data available can be used for ancillary studies.

Table 1 Baseline characteristics for the discovery cohort

	Discovery cohort: Olmsted County			
	Overall n=76 255	**Female n=40 463**	**Male n=35 792**	**P value**
Demographics				
Female	40 463 (53)			
Age on index date*, median (IQR)	49 (40, 61)	50 (40, 62)	49 (40, 60)	<0.001
Race				<0.001
American Indian	191 (0.3)	99 (0.2)	92 (0.3)	
Asian	2914 (3.8)	1572 (3.9)	1342 (3.7)	
Black	2257 (3.0)	1113 (2.8)	1144 (3.2)	
White	67 372 (88)	36 033 (89)	31 339 (88)	
Hawaiian/Pacific Islander	117 (0.2)	60 (0.1)	57 (0.2)	
Other/multiracial	2153 (2.8)	1091 (2.7)	1062 (3.0)	
Unknown	1251 (1.6)	495 (1.2)	756 (2.1)	
Hispanic ethnicity	2860 (3.8)	1364 (3.4)	1496 (4.2)	<0.001
Clinical characteristics				
BMI, median (IQR)	28 (24, 32)	27 (23, 32)	28 (26, 32)	<0.001
Unknown	17 172 (23)	6594 (16)	10 578 (30)	
Smoking status				<0.001
Unknown	24 630 (32)	11 316 (28)	13 314 (37)	
Never	31 690 (42)	19 088 (47)	12 602 (35)	
Ever	19 935 (26)	10 059 (25)	9876 (28)	
Systolic BP†, median (IQR)	122 (110, 133)	120 (110, 132)	124 (114, 135)	<0.001
Unknown	7658 (10)	3446 (9)	4212 (12)	
Diastolic BP†, median (IQR)	72 (66, 80)	70 (64, 80)	76 (68, 82)	<0.001
Unknown	7658 (10)	3446 (9)	4212 (12)	
Heart rate†, median (IQR)	72 (66, 80)	74 (68, 80)	72 (64, 80)	<0.001
Unknown	8946 (12)	4094 (10)	4852 (14)	
Clinical conditions‡				
Hypertension	21 766 (29)	11 623 (29)	10 143 (28)	0.239
Hyperlipidaemia	25 307 (33)	12 580 (31)	12 727 (36)	<0.001
Coronary artery disease	7969 (11)	3274 (8)	4695 (13)	<0.001
Cardiac arrhythmias	11 926 (16)	6343 (16)	5583 (16)	0.769
Heart failure	2308 (3.0)	1233 (3.1)	1075 (3.0)	0.725
Diabetes	9230 (12)	4616 (11)	4614 (13)	<0.001
Stroke	3004 (3.9)	1577 (3.9)	1427 (4.0)	0.526
COPD	8235 (11)	4731 (12)	3504 (9.8)	<0.001
Chronic kidney disease	3174 (4.2)	1556 (3.8)	1618 (4.5)	<0.001
Arthritis	15 285 (20)	9269 (23)	6016 (17)	<0.001
Osteoporosis	4030 (5.3)	3512 (8.7)	518 (1.5)	<0.001
Asthma	5838 (7.7)	3838 (9.5)	2000 (5.6)	<0.001
Cancer	10 497 (14)	6245 (15)	4252 (12)	<0.001
Depression	13 299 (17)	9009 (22)	4290 (12)	<0.001
Anxiety	7771 (10)	5084 (13)	2687 (7.5)	<0.001
Dementia	1958 (2.6)	1181 (2.9)	777 (2.2)	<0.001
Substance abuse	3137 (4.1)	1317 (3.3)	1820 (5.1)	<0.001

Continued

Table 1 Continued

	Discovery cohort: Olmsted County			
	Overall n=76 255	Female n=40 463	Male n=35 792	P value
Schizophrenia	1293 (1.7)	748 (1.8)	545 (1.5)	<0.001

Results are presented as n (%) unless otherwise noted.
*Index date is 1 January 2006.
†Closest measurement within 5 years prior to index date.
‡Conditions were ascertained using ICD codes recommended by the US Department of Health and Human Services, with the exception of anxiety which was defined by CCS category, using electronic medical history from 2001 to index date.
BMI, body mass index; BP, blood pressure; COPD, chronic obstructive pulmonary disease; ICD, International Classification of Diseases.

Follow-up and outcomes

Patients are followed after their index date to assess disease and ageing-related outcomes. Below are details of specific outcomes that we have collected thus far.

Myocardial infarction

MIs collected for a long-standing surveillance study were used for this project.[27] Residents admitted to Olmsted County hospitals with a troponin T level of 0.03 ng/mL or higher were identified.[27] MIs were validated using standard epidemiologic criteria which integrate cardiac pain, ECG changes and elevated biomarkers.[31] The presence or absence of a change (rise or fall) between any two troponin T measurements was defined by a difference of at least 0.05 ng/mL, which is greater than the level of imprecision of the assay at all concentrations.[32] Circumstances that might invalidate biomarker values were recorded.[33]

Up to three ECGs per episode were coded using the Minnesota Code Modular ECG Analysis System.[34] According to the algorithm, MIs were classified as definite, probable, suspect or no infarction.[31 35] Only incident (first-ever) cases were included in the cohort.

PCI and CABG surgery

Data were extracted from the Mayo Clinic Coronary Artery Percutaneous Intervention (PCI) registry. Because Mayo Clinic is the sole provider of coronary angiography in Olmsted County, a complete retrieval is possible via the database. By contrast, CPT codes were used to identify PCI in the validation cohort. For both cohorts, CABG surgery was identified using CPT codes.

Cardiovascular death

Minnesota death certificate and National Death Index Plus data were ascertained. CVD death is defined as underlying cause of death code ICD-9 390–459 and ICD-10 I00–I99.[36]

Stroke

A stroke algorithm was trained on an atrial fibrillation (AF) cohort.[26] First occurrence of ischaemic strokes, transient ischaemic attack and haemorrhagic strokes after incident AF from 1 January 2000 through 31 March 2015 were identified using diagnostic codes and were validated by trained nurse abstractors who manually reviewed the clinical notes. The algorithm includes diagnosis and procedure codes electronically extracted via the REP indexes and stroke-related keywords. The algorithm was trained using random forest models, and the resulting algorithm involved different weight (importance) on different features (ICD, CPT and keywords). The algorithm identifies stroke incidence dates with a precision of 0.900, recall of 0.918 and F-score of 0.909 in the general population.[37]

Patient and public involvement

Patients or the public were not involved in the design, conduct, reporting or dissemination plans of this study.

FINDINGS TO DATE

We identified 76 255 individuals (median age 49; 53% women) 30 years of age or older, residing in Olmsted County on 1 January 2006 (table 1) for the discovery cohort. A total of 9 644 221 laboratory results; 9 513 840 diagnosis codes; 10 924 291 service/procedure codes; 1 277 231 outpatient prescriptions; 966 136 heart rate measurements and 1 159 836 BP measurements were retrieved during the baseline time period. Seventy-one thousand two hundred and twenty-two (93%) patients had at least one clinical contact during the baseline period. The five most prevalent conditions in this cohort overall were hyperlipidaemia, hypertension, arthritis, depression and cardiac arrhythmias (table 1).

Women were slightly older than men (50 vs 49 years old) and were less likely to have a diagnosis of hyperlipidaemia, coronary artery disease, diabetes, chronic kidney disease and substance abuse (table 1). Conversely, women were more likely to be diagnosed with chronic obstructive pulmonary disease (COPD), arthritis, osteoporosis, asthma, cancer, depression, anxiety, dementia and schizophrenia.

In preliminary analyses, individuals in the discovery cohort without CVD (n=70 826) were followed from index date through 30 September 2017 for CVD-related outcomes: 1353 MIs, 1476 PCIs, 602 CABG, 912 strokes and 1770 CVD-related deaths occurred.

Table 2 Baseline characteristics for the validation cohort

	Validation cohort: 26-county region			
	Overall n=333 460	Female n=173 840	Male n=159 620	P value
Demographics				
Female	173 840 (52)			
Age on index date*, median (IQR)	54 (43, 67)	55 (43, 68)	54 (43, 66)	<0.001
Race				<0.001
American Indian	636 (0.2)	372 (0.2)	264 (0.2)	
Asian	2846 (0.9)	1711 (1.0)	1135 (0.7)	
Black	3421 (1.0)	1402 (0.8)	2019 (1.3)	
White	313 873 (94)	164 452 (95)	149 421 (94)	
Hawaiian/Pacific Islander	303 (0.1)	144 (0.1)	159 (0.1)	
Other/multiracial	4917 (1.5)	2427 (1.4)	2490 (1.6)	
Unknown	7464 (2.2)	3332 (1.9)	4132 (2.6)	
Hispanic ethnicity	9359 (2.8)	4623 (2.7)	4736 (3.0)	<0.001
Clinical characteristics				
BMI, median (IQR)	29 (25, 34)	29 (25, 34)	30 (27, 33)	<0.001
Unknown	104 805 (31)	46 712 (27)	58 093 (36)	
Smoking status				<0.001
Unknown	58 312 (18)	25 612 (15)	32 700 (21)	
Never	141 207 (42)	84 155 (48)	57 052 (36)	
Ever	133 941 (40)	64 073 (37)	69 868 (44)	
Systolic BP†, median (IQR)	123 (112, 134)	122 (110, 132)	125 (116, 136)	<0.001
Unknown	56 419 (17)	26 362 (15)	30 057 (19)	
Diastolic BP†, median (IQR)	74 (67, 80)	72 (65, 80)	76 (70, 82)	<0.001
Unknown	56 419 (17)	26 362 (15)	30 057 (19)	
Heart rate†, median (IQR)	72 (65, 80)	73 (66, 81)	72 (64, 80)	<0.001
Unknown	59 643 (18)	28 223 (16)	31 420 (20)	
Clinical conditions‡				
Hypertension	110 847 (33)	57 060 (33)	53 787 (34)	<0.001
Hyperlipidaemia	116 189 (35)	58 574 (34)	57 615 (36)	<0.001
Coronary artery disease	31 054 (9.3)	12 023 (6.9)	19 031 (12)	<0.001
Cardiac arrhythmias	51 065 (15)	25 947 (15)	25 118 (16)	<0.001
Heart failure	12 726 (3.8)	6378 (3.7)	6348 (4.0)	<0.001
Diabetes	59 706 (18)	29 343 (17)	30 363 (19)	<0.001
Stroke	13 766 (4.1)	6956 (4.0)	6810 (4.3)	<0.001
COPD	32 862 (9.9)	18 331 (11)	14 531 (9)	<0.001
Chronic kidney disease	21 429 (6.4)	10 455 (6.0)	10 974 (6.9)	<0.001
Arthritis	62 727 (19)	36 606 (21)	26 121 (16)	<0.001
Osteoporosis	14 216 (4.3)	12 459 (7.2)	1757 (1.1)	<0.001
Asthma	19 591 (5.9)	12 723 (7.3)	6868 (4.3)	<0.001
Cancer	39 254 (12)	21 784 (13)	17 470 (11)	<0.001
Depression	53 327 (16)	35 849 (21)	17 478 (11)	<0.001
Anxiety	38 396 (12)	25 567 (15)	12 829 (8.0)	<0.001
Dementia	9440 (2.8)	5675 (3.3)	3765 (2.4)	<0.001
Substance abuse	13 127 (3.9)	4985 (2.9)	8142 (5.1)	<0.001

Continued

Table 2 Continued

	Validation cohort: 26-county region			
	Overall n=333 460	Female n=173 840	Male n=159 620	P value
Schizophrenia	6432 (1.9)	3439 (2.0)	2993 (1.9)	0.031

Results are presented as n (%) unless otherwise noted.
*Index date is 1 January 2013.
†Closest measurement within 3 years prior to index date.
‡Conditions were ascertained using ICD codes recommended by the US Department of Health and Human services, with the exception of anxiety which was defined by CCS category, using electronic medical history from 2010 to index date.
BMI, body mass index; BP, blood pressure; COPD, chronic obstructive pulmonary disease; ICD, International Classification of Diseases.

We identified 333 460 individuals 30 years of age or older residing in the 26-county region of southern Minnesota and western Wisconsin (median age 54; 52% women; table 2) on 1 January 2013. To date, this validation cohort includes a total of 48 587 189 laboratory results; 19 926 750 diagnosis codes; 24 843 462 services/procedures; 7 083 721 outpatient prescriptions; 10 527 444 heart rate measurements and 7 356 344 BPs during the baseline time period. A total of 303 479 (91%) patients had at least one clinical contact during the baseline period. Overall, the five most prevalent conditions were hyperlipidaemia, hypertension, arthritis, diabetes and depression (table 2).

Similar to the discovery cohort, women were slightly older than men (55 vs 54 years old). Women were less likely to have a diagnosis of hypertension, hyperlipidaemia, coronary artery disease, cardiac arrhythmias, heart failure, diabetes, stroke, chronic kidney disease and substance abuse, and were more likely to be diagnosed with COPD, arthritis, osteoporosis, asthma, cancer, depression, anxiety, dementia and schizophrenia (table 2).

STRENGTHS AND LIMITATIONS

By leveraging harmonised and processed EHR data for clinical and translational research, our methods have several strengths. We are capitalising on the untapped depth and breadth of clinical data available in modern EHR systems in order to comprehensively identify risk factors of diseases, thus overcoming the inherent limitation of relying on a relatively small number of risk factors as is common in prospective research cohorts. We are using a foundational model that goes beyond traditional risk factors to include reproductive factors, age at onset of risk factors and a broad spectrum of clinical tests and diagnoses. Importantly, we have also created an independent validation cohort of patients from a largely rural area in which to assess the generalisability and transportability of our findings and models from the discovery cohort. Furthermore, in a large subset of patients, we have biological samples and genomic data. Thus, by developing and extending EHR algorithms for population research, these cohorts include a wide-range of sex-specific and other important risk factors or phenotypes occurring throughout the lifespan. Furthermore, we are identifying barriers and determining best practices for implementing study results from one type of medical practice to another.

Future use of innovative machine learning methods, such as gradient boosting machine and deep learning, will allow us to address several important and challenging questions associated with the use of EHR data such as how to efficiently (1) deal with missing values, (2) assess and use a large number of variables without over-fitting, (3) learn from non-linear relationships in the data and (4) design time-to-event models. In a community EHR environment, missing values will be frequent and will, in many cases, be informative. For example, the fact that a particular test was not ordered can itself be predictive. Traditional modelling approaches, such as linear or Cox regression, do not explicitly handle missing data, and this is one reason that risk modelling has traditionally been confined to prospective research cohorts.

The biggest limitation in utilisation of these techniques is the ability to develop accurate and transportable EHR phenotype algorithms for female-specific variables that are difficult to phenotype (eg, adverse pregnancy outcomes and gynaecological surgeries). Likewise, there can be challenges with determining the correct combination of gynaecological surgeries (eg, unilateral/bilateral oophorectomy with/without hysterectomy) and timing in regards to hormone therapy. By contrast, we do not foresee issues related to identifying male-specific factors, because these conditions are diagnosis based and thus available in the EHR. Finally, we did not collect information regarding usage of over-the-counter medications or supplements and multi-vitamins.

There are some additional limitations for the validation cohort in the 26 counties of southern Minnesota and western Wisconsin. Preliminary information indicates that EHR data will be more limited for this population. In particular, PPI, including family history of disease and difficulty climbing stairs is not routinely electronically available. In addition, because the EHR data are only available from 2010 forward, historic data on reproductive and gynecologic factors are more limited. However, this real world validation step will assess the performance of the phenotype algorithms to determine risk factor status as well as the prediction models including such

information when available. If inclusion of historic health information significantly improves the models, we will have evidence that such information should be routinely collected during healthcare visits to adequately assess disease risk. In the future, collection of historic health information may then be incorporated as part of clinical practice to improve disease risk assessment. Furthermore, some healthcare encounters were not captured that occurred outside of the REP. Although coverage varies by county, the REP captures approximately 100% of the Olmsted County population compared to the US Census, whereas coverage of the 26-county population is approximately 60%.[11]

Finally, the availability of biological samples and genome-wide and exome sequence information on a large sample of cohort participants is a strength. However, those with biologic samples and genomic information were not selected from the population randomly; therefore, they are not representative of the discovery or validation cohort.

CONCLUSION

With detailed and rich EHR data and using innovative machine learning methods, the population-based cohorts described herein can be used for a wide-range of studies, including but not limited to studies of novel disease associations, defining clusters of disease or creating risk scores for disease prediction.

Author affiliations
[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, USA
[2]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, Minnesota, USA
[3]Division of Community Internal Medicine, Department of Medicine, Mayo Clinic, Rochester, Minnesota, USA
[4]Department of Neurology, Mayo Clinic, Rochester, Minnesota, USA, Mayo Clinic, Rochester, Minnesota, USA
[5]Mayo Clinic Women's Health Research Center, Mayo Clinic, Rochester, Minnesota, USA
[6]Department of Physiology and Biomedical Engineering, Mayo Clinic, Rochester, Minnesota, USA
[7]Department of Surgery, Mayo Clinic, Rochester, Minnesota, USA
[8]Mayo Clinic Specialized Center of Research Excellence, Mayo Clinic Rochester, Minnesota, USA, Mayo Clinic, Rochester, Minnesota, USA
[9]Division of Cardiovascular Diseases, Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA
[10]Epidemiology and Community Health Branch National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland, USA

Acknowledgements We thank Ellen Koepsell, RN and Mary Roberts for their study support.

Contributors SJB and NBL jointly conceived the study. SJB, NBL, YZ, SM, HL, PAD and JMK handled the data management and analyses. SMM and SJB drafted the manuscript. SMM, JLS, HL, NBL, SM, PYT, JEO, WAR, VMM, TMT, CGN, VLR, YZ, PAD, JMK and SJB critically revised the manuscript for important intellectual content and approved the manuscript.

Funding This work was supported by grants from the National Heart, Lung and Blood Institute (R01 HL136659, R01 HL59205 and R01 HL72435) and the American Heart Association (11SDG7260039) and was made possible by the Rochester Epidemiology Project, Rochester, Minnesota (R01 AG034676) from the National Institute on Aging. The funding sources played no role in the design, conduct, or reporting of this study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Jennifer L St Sauver http://orcid.org/0000-0002-9789-8544
Walter A Rocca http://orcid.org/0000-0002-1832-7664
Suzette J Bielinski http://orcid.org/0000-0002-2905-5430

REFERENCES

1 Hripcsak G, Bloomrosen M, FlatelyBrennan P, *et al*. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 health policy meeting. *J Am Med Inform Assoc* 2014;21:204–11.
2 Duan R, Cao M, Wu Y, *et al*. An empirical study for impacts of measurement errors on EHR based association studies. *AMIA Annu Symp Proc* 2016;2016:1764–73.
3 Weiskopf NG, Hripcsak G, Swaminathan S, *et al*. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform* 2013;46:830–6.
4 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144–51.
5 Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. *Pac Symp Biocomput* 2017;22:207–18.
6 Miotto R, Li L, Kidd BA, *et al*. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016;6:26094.
7 Sung NS, Crowley WF Jr, Genel M, *et al*. Central challenges facing the National clinical research enterprise. *JAMA* 2003;289:1278–87.
8 Payne PRO, Johnson SB, Starren JB, *et al*. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. *J Investig Med* 2005;53:192–201.
9 Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE Consortium. *Sci Transl Med* 2011;3:79re1.
10 Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* 2013;20:e147–54.
11 Rocca WA, Grossardt BR, Brue SM, *et al*. Data resource profile: expansion of the Rochester Epidemiology Project medical records-linkage system (E-REP). *Int J Epidemiol* 2018;47:368–368j.
12 St Sauver JL, Grossardt BR, Yawn BP, *et al*. Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system. *Int J Epidemiol* 2012;41:1614–24.
13 Moy E, Garcia MC, Bastian B, *et al*. Leading causes of death in nonmetropolitan and metropolitan areas- United States, 1999-2014. *MMWR Surveill Summ* 2017;66:1–8.
14 Harrington RA, Califf RM, Balamurugan A, *et al*. Call to Action: Rural Health: A Presidential Advisory from the American Heart Association and American Stroke Association. *Circulation* 2020;141:e615–44.
15 Cheng FW, Gao X, Mitchell DC, *et al*. Body mass index and all-cause mortality among older adults. *Obesity* 2016;24:2232–9.
16 Cohen JW, Cohen SB, Banthin JS. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. *Med Care* 2009;47:S44–50.
17 Goodman RA, Posner SF, Huang ES, *et al*. Defining and measuring chronic conditions: imperatives for research, policy, program, and practice. *Prev Chronic Dis* 2013;10:E66.
18 U.S. Department of Health and Human Services. Multiple Chronic Conditions - A Strategic Framework: Optimum Health and Quality

of Life for Individuals with Multiple Chronic Conditions. Washington, DC. December 2010.
19 Bursi F, McNallan SM, Redfield MM, *et al*. Pulmonary pressures and death in heart failure: a community study. *J Am Coll Cardiol* 2012;59:222–31.
20 Forrey AW, McDonald CJ, DeMoor G, *et al*. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem* 1996;42:81–90.
21 Katz S, Ford AB, Moskowitz RW, *et al*. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. *JAMA* 1963;185:914–9.
22 Liu H, Bielinski SJ, Sohn S, *et al*. An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc* 2013;2013:149–53.
23 Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17:229–36.
24 U.S. National Library of Medicine. Unified medical language system (UMLS). Available: https://www.nlm.nih.gov/research/umls/ [Accessed 06 Sep 2020].
25 Olson JE, Ryu E, Johnson KJ, *et al*. The Mayo Clinic Biobank: a building block for individualized medicine. *Mayo Clin Proc* 2013;88:952–62.
26 Chamberlain AM, Gersh BJ, Alonso A, *et al*. Decade-long trends in atrial fibrillation incidence and survival: a community study. *Am J Med* 2015;128:260–7.
27 Roger VL, Weston SA, Gerber Y, *et al*. Trends in incidence, severity, and outcome of hospitalized myocardial infarction. *Circulation* 2010;121:863–9.
28 Gerber Y, Weston SA, Redfield MM, *et al*. A contemporary appraisal of the heart failure epidemic in Olmsted County, Minnesota, 2000 to 2010. *JAMA Intern Med* 2015;175:996–1004.
29 Gerber Y, Jaffe AS, Weston SA, *et al*. Prognostic value of cardiac troponin T after myocardial infarction: a contemporary community experience. *Mayo Clin Proc* 2012;87:247–54.
30 Gerber Y, Dunlay SM, Jaffe AS, *et al*. Plasma lipoprotein-associated phospholipase A2 levels in heart failure: association with mortality in the community. *Atherosclerosis* 2009;203:593–8.
31 White AD, Folsom AR, Chambless LE, *et al*. Community surveillance of coronary heart disease in the Atherosclerosis Risk in Communities (ARIC) Study: methods and initial two years' experience. *J Clin Epidemiol* 1996;49:223–33.
32 Alpert JS, Thygesen K, Antman E, *et al*. Myocardial infarction redefined--a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *J Am Coll Cardiol* 2000;36:959–69.
33 Jaffe AS. Elevations in cardiac troponin measurements: false false-positives: the real truth. *Cardiovasc Toxicol* 2001;1:87–92.
34 Kors JA, Crow RS, Hannan PJ, *et al*. Comparison of computer-assigned Minnesota Codes with the visual standard method for new coronary heart disease events. *Am J Epidemiol* 2000;151:790–7.
35 Roger VL, Jacobsen SJ, Weston SA, *et al*. Trends in the incidence and survival of patients with hospitalized myocardial infarction, Olmsted County, Minnesota, 1979 to 1994. *Ann Intern Med* 2002;136:341–8.
36 Virani SS, Alonso A, Benjamin EJ, *et al*. Heart Disease and Stroke Statistics-2020 Update: A Report from the American Heart Association. *Circulation* 2020;141:e139–596.
37 Zhao Y, Fu S, Bielinski SJ, Decker PA, Chamberlain AM, Roger VL, Liu H, Larson NB. Using natural language processing and machine learning to identify incident stroke from electronic health records. AHA EPI Lifestyle Scientific Sessions, Phoenix, AZ Circulation. 141(Suppl 1):AP259; 2020.