BMJ Open

Characteristics and utilisation of the Mayo Clinic Biobank, a clinic-based prospective collection in the USA: cohort profile

Janet E Olson,[1] Euijung Ryu,[2] Matthew A Hathcock,[2] Ruchi Gupta,[2] Joshua T Bublitz,[2] Paul Y Takahashi,[3] Suzette J Bielinski,[1] Jennifer L St Sauver,[1] Karen Meagher,[4] Richard R Sharp,[4] Stephen N Thibodeau,[5] Mine Cicek,[5] James R Cerhan[1]

To cite: Olson JE, Ryu E, Hathcock MA, *et al*. Characteristics and utilisation of the Mayo Clinic Biobank, a clinic-based prospective collection in the USA: cohort profile. *BMJ Open* 2019;**9**:e032707. doi:10.1136/bmjopen-2019-032707

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-032707).

Received 01 July 2019
Revised 14 October 2019
Accepted 15 October 2019

For numbered affiliations see end of article.

Correspondence to
Janet E Olson; olsonj@mayo.edu

ABSTRACT

Purpose The Mayo Clinic Biobank was established to provide a large group of patients from which comparison groups (ie, controls) could be selected for case–control studies, to create a prospective cohort with sufficient power for common outcomes and to support electronic health record (EHR) studies.

Participants A total of 56 862 participants enrolled (21% response rate) into the Mayo Clinic Biobank from Rochester, Minnesota (77%, n=43 836), Jacksonville, Florida (18%, n=10 368) and La Crosse, Wisconsin (5%, n=2658). Participants were all Mayo Clinic patients, 18 years of age or older and US residents.

Findings to date Overall, 43% of participants were 65 years of age or older and female participants were more frequent (59%) than males at all sites. Most participants resided in the Upper Midwest regions of the USA (Minnesota, Iowa, Illinois or Wisconsin), Florida or Georgia. Self-reported race among Biobank participants was 90% white. Here we provide examples of the types of studies that have successfully utilised the resource, including (1) investigations of the population itself, (2) provision of controls for case–control studies, (3) genotype-driven research, (4) EHR-based research and (5) prospective recruitment to other studies. Over 270 projects have been approved to date to access Biobank data and/or samples; over 200 000 sample aliquots have been approved for distribution.

Future plans The data and samples in the Mayo Clinic Biobank can be used for various types of epidemiological and clinical studies, especially in the setting of case–control studies for which the Biobank samples serve as control samples. We are planning cohort studies with additional follow-up and acquisition of genetic information on a large scale.

INTRODUCTION

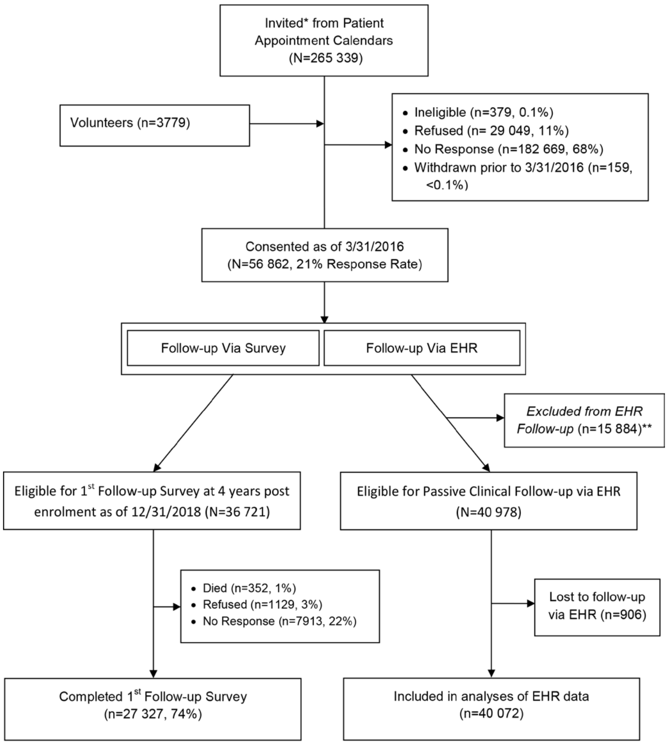
The Mayo Clinic Biobank was established to provide a large group of Mayo Clinic patients from which comparison groups (ie, controls) could be selected for case–control studies, to create a prospective cohort with sufficient power for common outcomes and to support electronic health record (EHR) studies. The Mayo Clinic Biobank was funded by the Mayo Clinic Center for Individualized Medicine, with an initial recruitment goal of 20 000 that was later expanded to 50 000 participants. Recruitment began in 2009 at Mayo Clinic in Rochester, Minnesota. In 2012, recruitment was expanded to the Mayo Clinic Florida in Jacksonville, and in 2013, the Mayo Clinic Health System in La Crosse, Wisconsin.

Strengths and limitations of this study

- ► The Mayo Clinic Biobank is located within the Mayo Clinic healthcare system and tied to the Rochester Epidemiology Project which provides electronic health record data that can be readily mined to passively collect information on our population from the past, present and future.
- ► Participants in the Biobank have provided biological samples which are stored and readily available to researchers.
- ► Participants have consented for additional contacts for future studies including samples and new data collection.
- ► Participants may not fully represent the underlying patient population (21% response rate), with some populations underrepresented.

COHORT DESCRIPTION

Beginning 1 April 2009, patients with medical appointments within 3weeks in selected departments were mailed invitation packets with a cover letter, two copies of the consent form, a baseline questionnaire, a $20 gift selection form and a return envelope. Participants were actively recruited from primary care departments (76%) and specialty clinics


Invited* from Patient Appointment Calendars (N=265 339)
Volunteers (n=3779)
• Ineligible (n=379, 0.1%)
• Refused (n= 29 049, 11%)
• No Response (n=182 669, 68%)
• Withdrawn prior to 3/31/2016 (n=159, <0.1%)
Consented as of 3/31/2016 (N=56 862, 21% Response Rate)
Follow-up Via Survey
Follow-up Via EHR
Excluded from EHR Follow-up (n=15 884)**
Eligible for 1st Follow-up Survey at 4 years post enrolment as of 12/31/2018 (N=36 721)
Eligible for Passive Clinical Follow-up via EHR (N=40 978)
• Died (n=352, 1%)
• Refused (n=1129, 3%)
• No Response (n=7913, 22%)
Lost to follow-up via EHR (n=906)
Completed 1st Follow-up Survey (n=27 327, 74%)
Included in analyses of EHR data (n=40 072)

*Did not include patients who required an interpreter, non-US home address and those with diagnosis codes related to mental capacity (ie, dementia).
**EHR exclusions: No clinical visits in at least 3 out of 5 years prior to date of consent, or did not live in the 26 county region of the Rochester Epi Project.

Figure 1 Modified Consolidated Standards of Reporting Trials flow diagram. EHR, electronic health records.

including Orthopaedics (10%), Executive Health (4%), Obstetrics/Gynaecology (3%), Sports Medicine (1%) and the Breast Clinic (1%). We also allowed volunteers to self-select without a study invitation (5%). The patients were largely selected from departments that provided primary care in order to enrich the Biobank for patients most likely to have comprehensive EHR data. Additionally, these clinical areas were viewed as proving the best population from which to draw research controls, as that was a primary goal of the collection. Eligibility criteria were few and included Mayo Clinic patient, age 18 or older, current US residence and ability to give informed consent. Active recruitment ended March 2016, although those wanting to participate are still accepted. Once informed consent was obtained, the Biobank participants provided a blood sample, completed health-related questionnaires and gave permission to access EHR and link to existing clinical data resources (eg, tissue registry).

Blood samples were collected in a clinical setting that adheres to the requirements set out by the Clinical Laboratory Improvement Amendments. This permits the use of the specimens for clinical grade testing if so desired. A standard blood sample collection included 30 mL in EDTA, 10 mL with no additives and a 4.5 mL with sodium citrate. From these samples, we obtained DNA, EDTA plasma, EDTA platelet poor plasma, citrated plasma and serum. In approximately 10% of patients, blood collected in EDTA was reduced to 20 mL and a 10 mL tube of sodium heparin was substituted to permit processing into slow-frozen white blood cells. Once created, all sample types were stored at −80°C in high-capacity freezers. The aliquots from this collection sum to more than 1.2 million and will be kept indefinitely.

A total of 265 339 patients were invited to the Mayo Clinic Biobank and 3779 participated without mailed study invitation (21% response rate, see figure 1). Of these invited, 379 (0.1%) were ineligible, 29 049 (11%) refused participation and 182 669 (68%) provided no response to our invitations. Prior to the close of the active enrolment, 159 participants withdrew from the Biobank, leaving a total of 56 862 participants enrolled in the Mayo Clinic Biobank as of 31 March 2016. Although the overall response to study invitations was 21%, it varied from 11% at the recruitment site in La Crosse to 16% in Florida and 25% in Rochester. The majority of the participants were recruited at the largest Mayo Clinic site in Rochester, Minnesota (77%, n=43 836), with 18% from Jacksonville, Florida (n=10 368) and 5% (n=2658) from La Crosse, Wisconsin (table 1). Overall, 43% of participants were 65 years of age or older, with the Florida site recruiting 52% of their participants in this age group. Female participants were more frequent (59%) than males at all sites, with the Wisconsin site recruiting 70% females. However, this is due in part to the over-representation of females among Mayo Clinic patient populations; males, in general, are less willing than females to seek medical care. A vast majority of participants resided in the Upper Midwest regions (Minnesota, Iowa, Illinois or Wisconsin), Florida or Georgia (table 1, figure 2). Figure 3 illustrates that participant county-level residency was highest in areas where the recruitment sites were located (Olmsted County in Upper Midwest and Duval County in Florida). Self-reported race among Biobank participants was 90% white overall, ranging from 87% white in Florida to 94% in Rochester Minnesota (table 1).

Differences between participants and the underlying source populations were estimated by comparing participants to data from the Behavioural Risk Factor Surveillance System (BRFSS)[1] for the two recruitment regions (online supplementary table 1). Compared with the closest regional BRFSS data, Florida Biobank participants were most discrepant in age, with greater frequency of older participants (52% age 65+ in Biobank vs 42% in BRFSS). The Wisconsin site was most discrepant in gender, with 70% female participants compared with 56% in the underlying population. The race/ethnicity frequencies reflected the underlying populations for the Upper Midwest; however, the Florida site differed from the underlying population (87% whites in Biobank vs 77% in BRFSS; 3% Hispanic ethnicity in Biobank vs 10% in BRFSS). The Upper Midwest Biobank site participants tended to have higher body mass index (BMI) than in the underlying regional population (40% obese (BMI >30) in Wisconsin, 37% obese in Rochester vs 30% obese in BRFSS). Participants from all sites were more likely to be drinkers of alcohol and less likely to have a history of smoking than the persons in the BRFSS survey.

Table 1 Demographic characteristics of Mayo Clinic Biobank stratified by recruitment locations and overall

	Biobank, Mayo Clinic Florida (n=10 368)	Biobank, MC Health System, Wisconsin (n=2658)	Biobank, Mayo Clinic Rochester (n=43 836)	Entire biobank (n=56 862)
Age at enrolment, N (%)				
18–44	1054 (10%)	489 (18%)	8106 (19%)	9649 (17%)
45–54	1374 (13%)	460 (17%)	7636 (17%)	9470 (17%)
55–64	2511 (24%)	685 (26%)	10 371 (24%)	13 567 (24%)
65 or older	5429 (52%)	1024 (39%)	17 723 (40%)	24 176 (43%)
Female gender, N (%)	5782 (56%)	1867 (70%)	25 738 (59%)	33 387 (59%)
Residence at recruitment, N (%)				
MC catchment areas in Upper Midwest*	5 (0.0%)	2106 (79%)	23 550 (54%)	25 661 (45%)
Other Upper Midwest	41 (0.4%)	544 (21%)	14 125 (32%)	14 710 (26%)
FL/GA	9508 (92%)	2 (0.1%)	515 (1.2%)	10 025 (18%)
Remainder of USA	814 (8%)	6 (0.2%)	5646 (13%)	6466 (11%)
Race (n=55 590), N (%)				
White	8832 (87%)	2450 (93%)	40 166 (94%)	51 448 (90%)
Black/African American	353 (4%)	8 (0.3%)	275 (0.6%)	636 (1%)
Asian	87 (10%)	19 (0.7%)	465 (1.1%)	571 (1%)
Native American/Alaskan Native	23 (0.2%)	4 (0.2%)	69 (0.2%)	96 (0.2%)
Others	825 (8%)	147 (6%)	1869 (4%)	2841 (5%)
Missing	248	30	992	1270
Hispanic ethnicity (n=55 322), N (%)	334 (3%)	22 (0.8%)	486 (1.1%)	842 (1%)
Missing	282	55	1203	1540
BMI (n=54 123), N (%)				
Underweight	188 (2%)	31 (1.2%)	364 (0.9%)	583 (1%)
Normal	3193 (33%)	642 (26%)	11 231 (27%)	15 066 (26%)
Overweight	3550 (37%)	847 (34%)	15 006 (36%)	19 403 (34%)
Obese	2640 (28%)	992 (40%)	15 439 (37%)	19 071 (34%)
Missing	797	146	1796	2739
Education (n=55 022), N (%)				
Less than High School	127 (1.3%)	54 (2%)	833 (2%)	1014 (2%)
High School Graduate	1006 (10%)	509 (19%)	6600 (16%)	8115 (14%)
Associate Degree/Technical School	3147 (31%)	1051 (40%)	13 581 (32%)	17 779 (31%)
College Graduate or Higher	5778 (57%)	1005 (38%)	21 333 (50%)	28 116 (49%)
Missing	310	39	1489	1838
Smoking status (n=54 891), N (%)				
Never	5205 (52%)	1476 (57%)	25 209 (60%)	31 890 (56%)
Former	4418 (44%)	906 (35%)	14 787 (35%)	20 111 (35%)
Current	386 (4%)	216 (8%)	2288 (5%)	2890 (5%)
Missing	359	60	1552	1971
Alcohol (n=55 647), N (%)				
Never	2593 (26%)	563 (21%)	9192 (21%)	12 348 (22%)
Once a month or less	1733 (17%)	584 (22%)	8943 (21%)	11 260 (20%)
2–4 times a month	1719 (17%)	659 (25%)	9807 (23%)	12 185 (21%)
2–5 times a week or more	4076 (40%)	829 (32%)	14 949 (35%)	19 854 (35%)

Continued

Table 1 Continued

	Biobank, Mayo Clinic Florida (n=10 368)	Biobank, MC Health System, Wisconsin (n=2658)	Biobank, Mayo Clinic Rochester (n=43 836)	Entire biobank (n=56 862)
Missing	247	23	945	1215

*27 counties in Minnesota and Wisconsin described by Rocca *et al*.[4]
BMI, body mass index; FL, Florida; GA, Georgia; MC, Mayo Clinic.

Reasons for non-participation

The Mayo Clinic Institutional Review Board does not allow general inquiry into reasons for non-participation. However, we were allowed to conduct a month-long study of all invitees in Rochester during the recruitment phase in August of 2011, and found that refusers were more likely than participants to be older than age 75 and less likely to reside in the region closest to the Mayo Clinic in Rochester.[2] From subsequent interviews of invitees, non-responders were most likely to say they were too busy or did not have time to complete all the requirements for participation. By contrast, the most common concern cited among refusers was related to concerns about privacy and confidentiality.

Cohort follow-up

There are two definitions of follow-up used in the Mayo Clinic Biobank: active follow-up via participant surveys and passive clinical follow-up via linkage to EHR data (figure 1). Biobank participants completed surveys at time of study entry (baseline) and again at 4 years post-enrolment. Survey response rates were 98% at baseline among participants and 74% at 4 years. Additional cross-sectional surveys are planned.

Eligibility for passive clinical follow-up via EHR was restricted to those with adequate EHR data prior to enrolment. Patients were required to have had clinical visits in at least 3 out of 5 years prior to the Biobank consent or live in the 26-county region of the Rochester Epidemiology Project (REP)[3] catchment area at time of consent. The primary reason for the eligibility definition was to exclude patients who were unlikely to receive continuous medical care at Mayo Clinic, including referral patients coming to Mayo Clinic for a single specialty procedure. Of the 56 862 consented patients, 72% (n=40 978) met those requirements and have accumulated 174 087 person-years of follow-up from enrolment to the end of 2017. Among those eligible, EHR data were followed passively until the most recent date that a participant had a contact with Mayo Clinic (clinical visits and/or Biobank study requests), death date (7%) or 31 December 2017, whichever came first. For passive EHR follow-up among eligible female patients, non-deceased women were considered as lost to follow-up if they did not have clinical visits within 2 years prior to 31 December 2017. Male subjects were handled similarly, except the window was extended to 3 years due to their lower rate of medical usage. Those who were lost to follow-up through EHR (n=906) were younger (mean age of 48 vs 60 years), more females (77% vs 60%) and more non-Whites (87% vs 93%) compared with those not lost to follow-up via EHR.

A subset of participants who have ever lived in a 27-county region of southern Minnesota and western Wisconsin have clinical data available through their inclusion in the REP.[3 4] Using the REP infrastructure, we can identify medical care received outside of Mayo Clinic affiliates from key healthcare providers in the region (ie, Olmsted Medical Center (OMC) and its satellites, Olmsted County Public Health Services and others).[3] In 2016, the Mayo Clinic Biobank received approval from these key healthcare providers to receive their data to aid investigators in determining whether participants had meaningful quantities of data at these sites to aid in disease definitions, either for inclusion (cases) or exclusion (controls). Figure 4 depicts the density of various data elements (eg, diagnosis codes, vital data, image data, prescribed medication and laboratory test results) available in participants' EHR between 1994 (the time when EHR was available at Mayo Clinic) and 2017, implying that the Biobank participants have excellent coverage of longitudinal EHR.

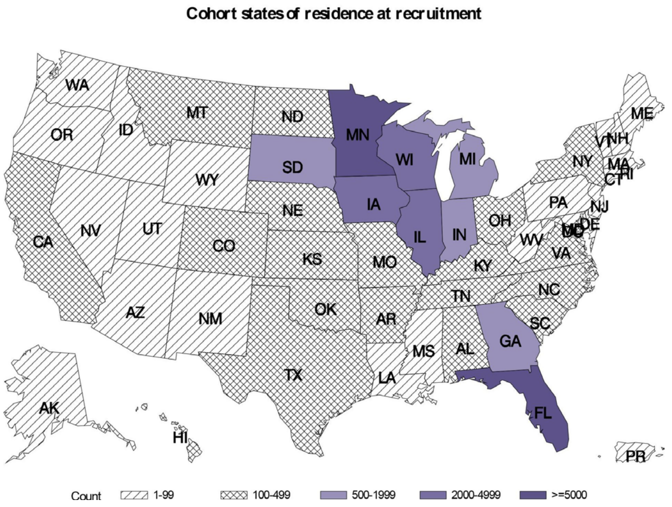
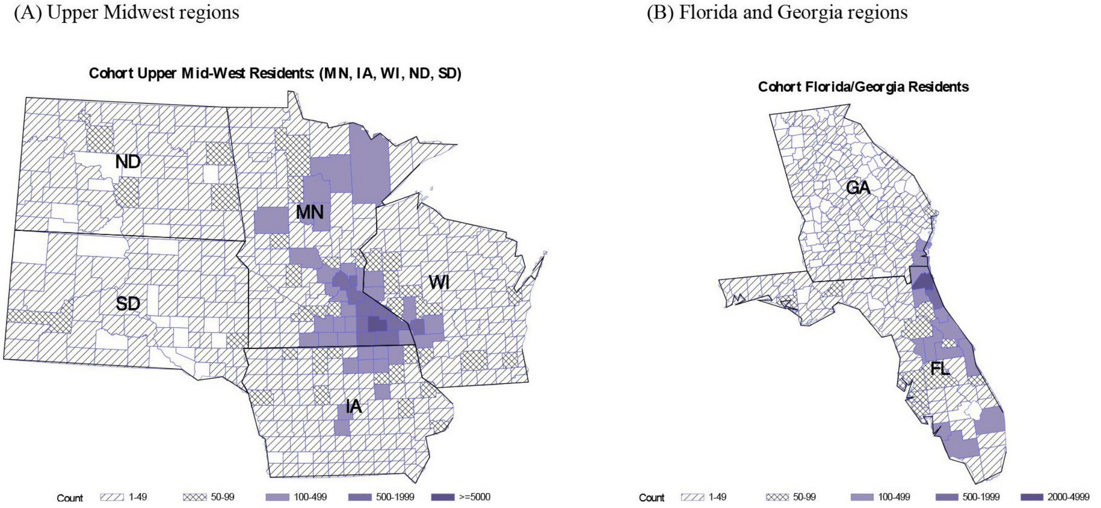


Figure 2 Geographical distribution (state-level) of home residences among participants at the time of enrolment into the Mayo Clinic Biobank.



(A) Upper Midwest regions

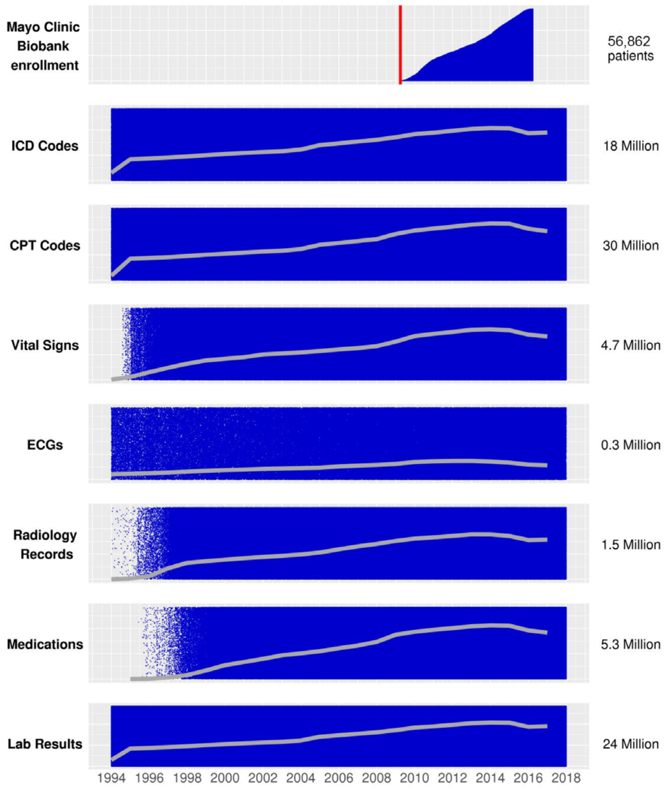
(B) Florida and Georgia regions

Cohort Upper Mid-West Residents: (MN, IA, WI, ND, SD)
ND
MN
WI
SD
IA
Count ▨ 1-49 ▦ 50-99 100-499 500-1999 >=5000
Cohort Florida/Georgia Residents
GA
FL
Count ▨ 1-49 ▦ 50-99 100-499 500-1999 2000-4999

Figure 3 Geographical distribution (county-level) of home residences among Mayo Clinic Biobank participants from Upper Midwest regions (A) and Florida/Georgia (B).

FINDINGS TO DATE

The most common self-reported prevalent conditions at enrolment and incident conditions at the time of 4-year follow-up are presented in table 2. The most common prevalent conditions at enrolment included hyperlipidaemia (41%), hypertension (40%) and osteoarthritis (35%). Among the subjects who completed both the baseline and follow-up surveys by the end of 2017 (n=24 016), those who first reported a particular condition on the 4-year follow-up survey were considered as having an incident condition (table 2). The most commonly reported conditions in this group during 4 years post-enrolment included osteoarthritis (19%), cataracts (17%), hyperlipidaemia (14%) and atrial fibrillation (14%). Twenty-six percent (26%) of those eligible for follow-up survey at 4 years post-enrolment did not complete one after three mailed requests. Compared with those who completed the 4-year follow-up survey, non-respondents were younger (mean age of 54 vs 60 years) and more likely to be non-Whites (89% vs 94%). Gender distribution was similar between the two groups (59% female in both groups).



Red vertical line indicates the initiation of enrolment, each blue dot represents unique data point on a given day for each patient and grey curves depict the proportion of the participants having at least one record for a given year for each data element.

ICD, International Classification of Diseases; CPT, Current Procedural Terminology.

Figure 4 Depth of electronic health records of Mayo Clinic Biobank participants since 1994.

Table 3 displays a summary of 20 chronic conditions defined by the United States Department of Health and Human Services (US DHHS) (https://www.cdc.gov/pcd/issues/2013/12_0239.htm) on the 40 978 participants eligible for passive EHR follow-up as described above. It also contains data on the entire cohort. The prevalence of each chronic condition was defined as the proportion of participants with at least two ICD nine or 10 codes that were at least 31 days apart and the date of the first code occurred (index date) prior to the consent date. After excluding prevalent cases, the cumulative incidence of each condition until the last follow-up date defined above (up to approximately 8.5 years after enrolment) was calculated using the Kaplan Meier method.[5] Based on the definition, the most common prevalent condition was hyperlipidaemia (53%) followed by arthritis and hypertension (44% for both), and the most common incident condition was arthritis (16%) followed by cancer (13%). Similar trends were observed in the whole cohort (table 3). Mortality endpoints are identified through linkage to the Mayo Clinic EHR, State of Minnesota electronic death certificates and the National Death Index-Plus (https://www.cdc.gov/nchs/ndi/index.htm).

Table 2 Top 15 self-reported prevalent conditions at enrolment into Mayo Clinic Biobank among all 56 862 participants. Incident conditions are those reported by the 24 016 subjects with data available on both the enrolment (baseline) and at 4-year follow-up surveys through the end of 2017

Self-reported disease	Prevalent cases at enrolment (%)	Incident cases at follow-up (%)
Hyperlipidaemia	23 121 (41%)	1885 (14%)
Hypertension	22 373 (40%)	1615 (11%)
Osteoarthritis	19 393 (35%)	3066 (19 %)
Cancer*	18 497 (33%)	2007 (12%)
Non-melanoma skin cancer	8825 (16%)	1583 (8%)
Prostate cancer (men only)	2655 (12 %)	318 (4%)
Breast cancer	3003 (9%)	293 (2%)
Melanoma	2118 (4%)	510 (2%)
Other cancer	1262 (2%)	1044 (4%)
Gastro-oesophageal reflux disorder	16 846 (30%)	2098 (12%)
Cataracts	15 845 (28%)	2966 (17%)
Depression	13 648 (24%)	1175 (6%)
Anxiety	11 500 (21%)	1412 (7%)
Migraine headaches	10 787 (19%)	765 (4%)
Sleep apnoea	9486 (17%)	1481 (7%)
Hyper/hypothyroidism	8344 (15%)	928 (5%)
Asthma	7568 (14%)	501 (2%)
Diabetes	6516 (12%)	626 (3%)
Irritable bowel disease	6108 (11%)	783 (4%)
Atrial fibrillation/ arrhythmia	23 121 (41%)	1885 (14%)

*Incident cases were defined as those who reported to have a condition at follow-up survey, but either reported 'no' at the baseline or did not answer the particular question. Among those who completed both questionnaires, 58.7% (n=14 091) were females. For prostate cancer, number of males was used as a denominator.

Key findings from the Mayo Clinic Biobank exemplify the types of research that can utilise a biobank of this type, including (1) investigations of the population itself, (2) provision of controls for case–control studies, (3) genotype-driven research, (4) EHR-based research and (5) prospective recruitment to other studies. Over 270 projects have been approved to date to access Biobank data and/or samples; over 200 000 sample aliquots have been approved for distribution. Despite this high usage, many aliquots remain and are available for additional future research projects.

Investigations of the Biobank populations

Various aspects of the Mayo Clinic Biobank population have been described earlier.[6–10] Takahashi *et al*[11] helped us understand our participants by comparing participants in the Mayo Clinic Biobank who were assigned a primary care provider in Rochester, Minnesota (39% at time of publication) to the entire Rochester primary care (Employee and Community Health, ECH) population. We found that Biobank patients were older and had more chronic conditions than the underlying ECH population overall. However, we found that the associations of chronic condition burden with risk of hospitalisation and emergency department visits were similar for both populations. We also found that self-perceived health status and alcohol use had a strong association with risk of hospitalisation in our population.[12]

The Biobank as a source of unaffected subjects

Numerous studies have used samples (DNA, serum and/or plasma) from the Biobank for controls. For example, Kleinstern *et al*[13] used 1267 Biobank subjects as controls in their validation study of a polygenic risk score (PRS) for chronic lymphocytic lymphoma. The PRS association was replicated in other populations and provides a method of identifying those at increased risk of chronic lymphocytic leukaemia as well as its precursor, monoclonal B-cell lymphocytosis. Numerous studies of bipolar disorder have benefitted from the Mayo Clinic Biobank. For example, Winham *et al*[14] used Biobank controls to identify possible sex-specific effects of P2RX7 variants in female bipolar patients, but not male patients. Another frequent use of the Biobank includes studies of breast cancer,[15 16] glioma[17] and prostate cancer.[18]

Genotype-driven research studies

Genotype-driven research is an especially good example of science that is enhanced by the existence of a biobank. Prior to the existence of the Biobank, identifying individuals with rare variants was quite challenging. Shah *et al*[19] utilised the stored DNA in the Biobank to identify subjects with a relatively rare genotype in TCF7L2. These investigators first identified the subjects with the genetic variant of interest via genotyping of 5000 participants. Subjects with the rare genotype were then invited to participate in clinical studies of diabetes to better understand the biological mechanism underlying the diabetes/gene association. Although the underlying mechanism has not yet been elucidated, results suggest that TCF7L2 impairs glucose tolerance through its effects on glucagon secretion.[19]

EHR-based research studies

Numerous users of the Biobank have conducted EHR-based studies involving all or a portion of the Biobank participants. For example, Bielinski and colleagues[20] have made use of the extensive EHR data available to develop and validate an EHR-driven algorithm for heart failure using EHR data for the Biobank population. The final algorithm utilised various EHR components from both structured and unstructured data to classify each patient into categories of definite, probable or possible heart failure, and controls, depending on the confidence of

Table 3 Prevalence (at enrolment) and cumulative incidence through the last follow-up period (up to 8.5 years after enrolment) of the 20 US DHHS chronic conditions among all consented subjects and the subset eligible for clinical follow-up via electronic health records

	Eligible participants for follow-up (n=40 978)		All consented (n=56 862)	
US DHHS 20 chronic conditions	**Prevalent cases (%)**	**Incident case (%*)**	**Prevalent cases (%)**	**Incident case (%*)**
Hypertension	17 976 (44%)	2034 (12%)	20 876 (37%)	2835 (11%)
Congestive heart failure	1394 (3%)	1121 (4%)	1592 (3%)	1259 (4%)
Coronary artery disease	7388 (18%)	1373 (6%)	8492 (15%)	1742 (5%)
Cardiac arrhythmias	13 508 (33%)	2158 (10%)	15 349 (27%)	2762 (9%)
Hyperlipidaemia	21 817 (53%)	1773 (12%)	25 038 (44%)	2596 (11%)
Stroke	2689 (7%)	975 (4%)	2969 (5%)	1116 (3%)
Arthritis	17 944 (44%)	2861 (16%)	21 173 (37%)	3847 (14%)
Asthma	3874 (9%)	489 (2%)	4331 (8%)	637 (2%)
Autism	3 (<1%)	6 (<1%)	3 (<1%)	6 (<1%)
Cancer (selected†)	13 648 (33%)	2681 (13%)	15 946 (28%)	3379 (12%)
Chronic kidney disease	3604 (9%)	2094 (8%)	4153 (7%)	2413 (7%)
Chronic obstructive pulmonary disease	3600 (9%)	728 (3%)	3932 (7%)	897 (2%)
Dementia	714 (2%)	813 (3%)	893 (2%)	906 (3%)
Depression	7368 (18%)	1452 (6%)	8040 (14%)	1804 (5%)
Diabetes	11 779 (29%)	2473 (11%)	13 351 (23%)	3051 (9%)
Hepatitis	574 (1%)	126 (<1%)	706 (1%)	174 (<1%)
HIV	27 (<1%)	3 (<1%)	35 (<1%)	8 (<1%)
Osteoporosis	4063 (10%)	926 (3%)	4405 (8%)	1102 (3%)
Schizophrenia	153 (<1%)	116 (<1%)	168 (<1%)	127 (<1%)
Substance Abuse	1305 (3%)	662 (3%)	1461 (3%)	752 (2%)

*Cumulative incidence was calculated based on the Kaplan-Meier (KM) method, which is a sum of the KM estimate of incidence of a given condition at a specified time point over time. The KM estimate at a specified time point is the number of participants having a given condition at that time, divided by the number of patients at risk (alive, disease-free and not lost to follow-up), multiplied by the probability of disease-free survival just prior to that time. Note that this estimate cannot be calculated as a simple proportion of the number of participants with a given condition, divided by the number of participants at risk at the biobank entry.
†Cancer category includes breast, colorectal, lung and prostate cancers.
US DHHS, United States Department of Health and Human Services.

classification and the depth of EHR data reflecting a real-world patient population.

The Biobank as a source of study participants

Finally, an example of use of the Biobank as a population from which we drew a prospective cohort is provided by the 'Right Drug, Right Dose, Right Time: Using Genomic Data to Individualize Treatment Protocol' by Bielinski *et al.*[21] For this study, we invited 2000 Biobank participants to enrol in a pre-emptive pharmacogenomics project. Over 50% responded, provided a new blood sample and received pharmacogenomic results. This initial project has now been expanded to invite 20 000 participants to enrol an additional 10 000 subjects.

In summary, these studies demonstrate the utility of the Mayo Clinic Biobank for a wide range of research questions. The Biobank infrastructure enhances the abilities of investigators to conduct novel studies in a timely and cost-effective manner.

Patient and public involvement

The Mayo Clinic Biobank was supported by patient and community input from the very beginning. Prior to the initiation of the Mayo Clinic Biobank, we sought advice on development, management and operations of the Biobank via a Deliberative Community Engagement in the fall of 2007.[6] This group developed into a Rochester area Community Advisory Board (CAB) that has met quarterly since its inception in 2008. Some, but not all, of CAB members are also Biobank participants. In May 2013, the CAB expanded into a multisite network, creating an additional CAB in Jacksonville, Florida.[22] The CAB members regularly meet to review Biobank projects, provide comment on participant materials and advise on new initiatives. In addition to their role in Biobank stewardship, the CABs provide researchers with important resources to gain understanding of local attitudes relevant to their specific research projects, for example, in their consideration of a proposal to place research results into the EHR to facilitate translation of the results from a pharmacogenomics study.[23] A unique collaboration between the Mayo Clinic and Mountain Park Health Center Biobanks and CABs has also helped to raise and address questions about under-representation of Latino patients in precision medicine research.[24 25]

Strengths and limitations

The strengths of the Mayo Clinic Biobank are numerous, including its location within the Mayo Clinic healthcare system which provides EHR data that can be readily mined to passively collect follow-up information on our population. The Biobank's existence within Mayo Clinic in Rochester also provides ties to the REP, which provides an infrastructure to collect EHR data from multiple institutions in the region.[3 4] A unique aspect of the Mayo Clinic Biobank is the connection to the tissue resources at Mayo Clinic. Tissue slides and formalin fixed paraffin-embedded (FFPE) blocks have been created and retained on all patients who ever had a surgical procedure at Mayo Clinic since 1907.[26] Mayo Clinic policy is to retain slides and FFPE blocks indefinitely. The consent form specifically mentions this resource and allows the use of these tissues for research. Diagnosis and procedure codes are electronically available beginning in 1935; full electronic healthcare records began in 2000. The existence of stored samples makes future studies with biological specimens highly cost-effective as investigators can withdraw samples for a fraction of the cost required to recruit patients and obtain new samples. In addition, enrolled patients have consented to multiple contacts, which may include requests for additional biological samples, thus making a broad number of longitudinal studies more feasible. Patients were selected mostly from general care clinics; thus, the population is not purposely enriched for any particular health condition. The resource is especially good for serving as a source of controls for clinic-based case–control studies or to provide samples and data on patients with common health conditions.

Another major strength of the Mayo Clinic Biobank is its network of CABs that provide ongoing advice on management and operation of the Biobank. As mentioned above in the Patient and Public Involvement section, the community has been significantly involved in the governance of the Mayo Clinic Biobank since its inception.

Despite its strengths, the Mayo Clinic Biobank has limitations. First, participants may not fully represent the underlying population. As described above, only patients in the Mayo Clinic system were invited to participate. This is less impactful in the Minnesota and Wisconsin sites as the proportion of the population served by the Mayo Clinic system is much greater than is seen at the Florida site. In addition, among the patients invited, the overall response rate to study invitations was only 21% and certain groups, such as current smokers and those who do not utilise the healthcare systems, are under-represented. This resource also under-represents some racial groups, especially at the Florida site (online supplementary table 1). This is due in part to under-representation of the same racial groups among patients at Mayo Clinic. To address this disparity, Mayo Clinic is developing two sister biobanks: The *Sangre Por Salud* Biobank[24] among Hispanic populations in Phoenix, Arizona, and Biobank Mississippi at the University of Mississippi Medical Center which is enriched for persons of African descent. Another limitation of the Mayo Clinic Biobank is tied to the original intent of the Biobank, which was to provide a source of controls for primarily DNA-based, case–control studies. Consequentially, the Mayo Clinic Biobank is unlikely to have pretreatment samples on patients with rare conditions.

Conclusion

The Mayo Clinic Biobank is an important resource for clinical research. It is embedded in a clinical practice with access to both EHR data and cohort-specific patient surveys. It has provided subjects for numerous research projects in both a time and cost-effective way, thus attaining its primary goals of providing comparison groups for case–control studies, establishing a prospective cohort for future research and supporting EHR studies. We encourage collaborations with researchers from other institutions, both academic and industrial. The data and samples in the Mayo Clinic Biobank can be used for various types of epidemiological and clinical studies, especially in the setting of case–control studies for which the Biobank samples serve as control samples. In the future, we will conduct additional follow-up of the cohort and are in the process of acquiring genetic information on a large scale.

Author affiliations
[1]Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[2]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[3]Division of Primary Care Internal Medicine, Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA
[4]Biomedical Ethics Research Program, Mayo Clinic, Rochester, Minnesota, USA
[5]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA

Acknowledgements The authors wish to acknowledge the institutional leadership at Mayo Clinic for their foresight and support of this project, including Gianrico Farrugia, M.D. and Keith Stewart, M.B., Ch.B. We also thank the many efforts of the study coordinators who did the bulk of the work to successfully recruit all the participants. We especially thank Jody Morrisette for her leadership of the coordinator team.

Collaborators All data from the Mayo Clinic Biobank presented in this article are stored by the research group of the authors on safe servers at Mayo Clinic, USA and handled confidentially. To apply for access to data, contact Biobank@mayo.edu and request the current application for access to samples and/or data. Due to the specifics of the informed consent language, a Mayo Clinic researcher must be included as a collaborator on all projects. If you cannot identify a staff member of Mayo Clinic, one can be identified for you. Contact Dr Janet E Olson (olsonj@mayo.

edu) for enquires. More information is available at our website at: https://www.mayo.edu/research/centers-programs/mayo-clinic-biobank/overview.

Contributors JEO, ER, PYT, SNT and JRC jointly conceived the study. ER, MAH, RG and JTB handled the data management and analyses. JEO and ER drafted the manuscript. SJB, JSS, KM, RRS, MC and JRC critically revised the manuscript for important intellectual content. JEO and ER were responsible for the final version of the manuscript.

Funding The creation and operation of the Mayo Clinic Biobank is supported by funding provided by the Mayo Clinic Center for Individualized Medicine.

Map disclaimer The depiction of boundaries on the map(s) in this article do not imply the expression of any opinion whatsoever on the part of BMJ (or any member of its group) concerning the legal status of any country, territory, jurisdiction or area or of its authorities. The map(s) are provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Centers for Disease Control and Prevention. Behavioral risk factor surveillance system. Available: https://www.cdc.gov/brfss/index.html [Accessed 9 Jan 2019].

2 Ridgeway JL, Han LC, Olson JE, *et al*. Potential bias in the bank: what distinguishes refusers, nonresponders and participants in a clinic-based biobank?. *Public Health Genomics* 2013;16:118–26.

3 Rocca WA, Yawn BP, St. Sauver JL, *et al*. History of the Rochester Epidemiology Project: half a century of medical records linkage in a US population. *Mayo Clin Proc* 2012;87:1202–13.

4 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.

5 Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. *Clin Cancer Res* 2007;13:559–65.

6 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.

7 Ryu E, Takahashi PY, Olson JE, *et al*. Quantifying the importance of disease burden on perceived general health and depressive symptoms in patients within the Mayo Clinic Biobank. *Health Qual Life Outcomes* 2015;13:95.

8 Ryu E, Olson JE, Juhn YJ, *et al*. Association between an individual housing-based socioeconomic index and inconsistent self-reporting of health conditions: a prospective cohort study in the Mayo Clinic Biobank. *BMJ Open* 2018;8:e020054.

9 Olson JE, Ryu E, Lyke KJ, *et al*. Acceptability of electronic visits for return of research results in the Mayo Clinic Biobank. *Mayo Clin Proc Innov Qual Outcomes* 2018;2:352–8.

10 Ryu E, Juhn YJ, Wheeler PH, *et al*. Individual housing-based socioeconomic status predicts risk of accidental falls among adults. *Ann Epidemiol* 2017;27:415–20.

11 Takahashi PY, Ryu E, Olson JE, *et al*. Hospitalizations and emergency department use in Mayo Clinic Biobank participants within the employee and community health medical home. *Mayo Clin Proc* 2013;88:963–9.

12 Takahashi PY, Ryu E, Olson JE, *et al*. Health behaviors and quality of life predictors for risk of hospitalization in an electronic health record-linked biobank. *Int J Gen Med* 2015;8:247–54.

13 Kleinstern G, Camp NJ, Goldin LR, *et al*. Association of polygenic risk score with the risk of chronic lymphocytic leukemia and monoclonal B-cell lymphocytosis. *Blood* 2018;131:2541–51.

14 Winham SJ, Bobo WV, Liu J, *et al*. Sex-Specific effects of gain-of-function P2RX7 variation on bipolar disorder. *J Affect Disord* 2019;245:597–601.

15 Garcia-Closas M, Couch FJ, Lindstrom S, *et al*. Genome-Wide association studies identify four ER negative–specific breast cancer risk loci. *Nat Genet* 2013;45:392–8.

16 Michailidou K, Hall P, Gonzalez-Neira A, *et al*. Large-Scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 2013;45:353–61.

17 Melin BS, Barnholtz-Sloan JS, Wrensch MR, *et al*. Genome-Wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet* 2017;49:789–94.

18 Teerlink CC, Leongamornlert D, Dadaev T, *et al*. Genome-Wide association of familial prostate cancer cases identifies evidence for a rare segregating haplotype at 8q24.21. *Hum Genet* 2016;135:923–38.

19 Shah M, Varghese RT, Miles JM, *et al*. TCF7L2 genotype and α-Cell function in humans without diabetes. *Diabetes* 2016;65:371–80.

20 Bielinski SJ, Pathak J, Carrell DS, *et al*. A robust e-epidemiology tool in phenotyping heart failure with differentiation for preserved and reduced ejection fraction: the Electronic Medical Records and Genomics (eMERGE) Network. *J Cardiovasc Transl Res* 2015;8:475–83.

21 Bielinski SJ, Olson JE, Pathak J, *et al*. Preemptive genotyping for personalized medicine: design of the Right Drug, Right Dose, Right Time—Using Genomic Data to Individualize Treatment protocol. *Mayo Clin Proc* 2014;89:25–33.

22 Allyse MA, McCormick JB, Sharp RR. Prudentia populo: involving the community in biobank governance. *Am J Bioeth* 2015;15:1–3.

23 Kimball BC, Nowakowski KE, Maschke KJ, *et al*. Genomic data in the electronic medical record: perspectives from a biobank community advisory board. *J Empir Res Hum Res Ethics* 2014;9:16–24.

24 Shaibi G, Singh D, De Filippis E, *et al*. The Sangre POR Salud Biobank: facilitating genetic research in an underrepresented Latino community. *Public Health Genomics* 2016;19:229–38.

25 Shaibi GQ, Kullo IJ, Singh DP, *et al*. Developing a process for returning medically actionable genomic variants to Latino patients in a federally qualified health center. *Public Health Genomics* 2019;21:77–84.

26 Giannini C, Oelkers MM, Edwards WD, *et al*. Maintaining clinical tissue archives and supporting human research: challenges and solutions. *Arch Pathol Lab Med* 2011;135:347–53.