A Poisson binomial-based statistical testing framework for comorbidity discovery across electronic health record datasets

Article *in* Nature Computational Science · October 2021

DOI: 10.1038/s43588-021-00141-9

CITATION
1

READS
51

5 authors, including:

Gordon Lemmon
University of Utah
36 PUBLICATIONS **2,102** CITATIONS

SEE PROFILE

ARTICLES
https://doi.org/10.1038/s43588-021-00141-9

nature computational science

Check for updates

A Poisson binomial-based statistical testing framework for comorbidity discovery across electronic health record datasets

Gordon Lemmon[1,2], Sergiusz Wesolowski[1,2], Alex Henrie[1,2], Martin Tristani-Firouzi[3,4] ✉ and Mark Yandell[1,2] ✉

Discovering the concomitant occurrence of distinct medical conditions in a patient, also known as comorbidities, is a prerequisite for creating patient outcome prediction tools. Current comorbidity discovery applications are designed for small datasets and use stratification to control for confounding variables such as age, sex or ancestry. Stratification lowers false positive rates, but reduces power, as the size of the study cohort is decreased. Here we describe a Poisson binomial-based approach to comorbidity discovery (PBC) designed for big-data applications that circumvents the need for stratification. PBC adjusts for confounding demographic variables on a per-patient basis and models temporal relationships. We benchmark PBC using two datasets to compute comorbidity statistics on 4,623,841 pairs of potentially comorbid medical terms. The results of this computation are provided as a searchable web resource. Compared with current methods, the PBC approach reduces false positive associations while retaining statistical power to discover true comorbidities.

Comorbidity refers to the concomitant occurrence of distinct medical conditions in the same patient[1]. Comorbidities can occur together or sequentially across the patient's medical history. Exploring these temporal connections offers further insight into disease progression and disease associations, and promises improved predictive tools for evidence-based medicine[2–5]. Traditionally, comorbidities have been discovered manually, through human chart review, literature search and clinical knowledge[6]. For example, the authors of the Charlson comorbidity index[7] selected comorbid diagnoses based on manual chart review for a 559-patient cohort. Likewise, the well-known Elixhauser Comorbidity index was compiled through review of published studies identifying comorbid conditions[8].

Large collections of electronic health records (EHRs) present promising new opportunities for comorbidity discovery; however, manual review of millions of EHR records in search of comorbidities is infeasible, and ab initio means for discovery and temporal ordering of comorbidities using large collections of EHRs is an area ripe for innovation. Current computational approaches to ab initio comorbidity discovery use statistics such as risk ratio, odds ratio, comorbidity score[9], propensity score or ϕ-correlation to measure effect size. *P*-values are obtained using Fisher's exact test (or hypergeometric), an χ^2 test or binomial test[10–14]. All of these approaches assume that each member of the population has a disease probability equal to the population incidence rate. Confounding variables such as age and sex are controlled for by subsetting the data, a process termed stratification or alternatively through use of matched case-control cohorts[14,15]. Both of these approaches control for confounders, but at the expense of statistical power, as they necessarily reduce sample size. One approach to overcoming this intrinsic limitation is to aggregate massive collections of EHRs, but as we show, even millions of records are too few to explore comorbidities associated with rare diseases when controlling for multiple confounding variables.

Using a Poisson binomial-based approach to comorbidity discovery (PBC), we model the effects of confounding variables, allowing every sample to be personalized, which results in improved statistical power compared with stratification. Briefly, PBC uses logistic regression to model how each patient's demographics impact his or her probability of having a medical term. These personalized probabilities are then used to calculate pairwise expectations and *P*-values under the Poisson binomial distribution, rather than the hypergeometric and binomial distributions that are used for Fisher's exact test and the χ^2 test, respectively. Moreover, with minor modification, this approach can also be used to temporally order comorbidities and to determine the significance of directionalities. As we demonstrate, the Poisson binomial approach provides a considerable advantage as it obviates the need for stratification. The result is increased power for discovery, which we leverage to explore the relationships among diagnoses, medical procedures and medications.

Alongside the need for improved statistical methods, tools are also needed to browse, search and visualize the network of comorbid medical terms discovered in big-data applications. In a manner similar to Siggaard and colleagues[14], we provide a browser-based query engine for navigating these comorbidities and their temporal relationships within the University of Utah Hospitals system (https://pbc.genetics.utah.edu/lemmon2021/pbc-utah).

In what follows, we describe PBC, explore its behavior and benchmark its performance using the contents of the University of Utah Health system and the publicly available MIMIC-IV dataset[16]. For brevity's sake, we will refer to co-occurring medical diagnoses, procedures and medications using the single blanket term, comorbidity. We demonstrate how PBC can be used to transform massive

[1]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [2]Utah Center for Genetic Discovery and Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [3]Division of Pediatric Cardiology, University of Utah School of Medicine, Salt Lake City, UT, USA. [4]Nora Eccles Harrison CVRTI, University of Utah School of Medicine, Salt Lake City, UT, USA. ✉e-mail: Martin.Tristani@utah.edu; myandell@genetics.utah.edu

EHR datasets into a temporal dependency graph for large-scale ab initio discovery of comorbid relationships and investigations of disease progressions.

Results

Modeling effects of confounders using logistic regression. We collected records for 1.6 million patients, encompassing 50 million visits and 150 million diagnosis (DX), procedure (PX) and medication (RX) codes from the University of Utah electronic data warehouse (EDW). For proof of principle analyses presented here, diagnoses were converted from ICD9[17] and ICD10[18] diagnosis codes to clinical classification cystem[19] (CCS) multilevel diagnosis codes. Current Procedural Terminology (CPT)[20] provider billing codes were converted to CCS multilevel procedure codes. We include both leaf nodes and internal nodes so that the researcher can discover more specific comorbidities (for example, CCS 2.1.1: cancer of the colon) as well as more general comorbidities (for example, CCS 2.1: colorectal cancer or CCS 2: neoplasms). Medications were coded using RxNorm concept unique identifiers (CUIs)[21]. This procedure reduced the number of distinct medical terms to 1,007 DX, 259 PX and 1,775 RX codes. We also collected demographic information, including sex, race, ethnicity, insurance class, age and length of medical records.

A logistic regression model (LRM) was determined for each DX, PX and RX term. Each LRM includes demographic information for each patient (age, gender, ancestry, ethnicity, insurance type) and EHR exposure (the length and density of a person's medical record). As medical terms are included or excluded from year to year, and coding practices vary over time, we also include the date of the patient's last visit as a control for this effect. The complete list of features is described in Supplementary Table 1 and Extended Data Fig. 1. The response variable is whether each patient has the term in their medical record. Note that we do not model recurrence—in this analysis we consider only the first instance of a term in a patient's medical history. We use these LRMs to estimate for each medical term, each patient's personalized probability of having that term in their medical record.

We used a regularized regression model under the assumption of collinearity in the confounding features; however, there is no requirement for such an assumption. For example, neural networks could be used in place of LRMs to better capture nonlinear relationships between variables. L1- and L2-penalized logistic regression include a value C that prevents overfitting by penalizing large coefficients. Smaller C-values specify stronger regularization. To determine the optimal C-value for each LRM, it was necessary to choose a score function for LRM evaluation. We experimented with a number of standard and custom score functions as described in Supplementary Section 2. We optimize C for each LRM using stratified threefold cross-validation. A grid search is used to evaluate C-values in the set $\{10^{-14}, 10^{-13}, 10^{-12}, \ldots, 10^{12}, 10^{13}, 10^{14}\}$. Extended Data Fig. 2a shows box plots of the scores reported by each of these score functions. We use entropy as a measure of the ability of a score function to differentiate model quality. We evaluate the distribution of C-values from those score functions achieving high entropy (Extended Data Fig. 2b). We choose J_{cutoff} for all downstream analysis as it includes fewer outliers than the other methods examined; J_{cutoff} is based on Youden's J statistic[22] but, instead of a 50% probability threshold, the classification threshold is determined empirically so that the total number of predicted positives is equal to the actual count of positives.

We train LRMs using stratified threefold cross-validation, evaluated using J_{cutoff}. Under L1-penalized logistic regression, model features can be unselected by setting their coefficients to zero. Figure 1 summarizes how often each demographic feature is included in the trained LRMs. The age at last visit, number of visits, number of terms and length of medical record were grouped together as EHR exposure. Patients may be seen at a non-university clinic, may move in or out of the state over the years, or may have differing proclivities toward visiting the doctor. Electronic health record exposure is an attempt to control for these effects. As Fig. 1 makes clear, EHR exposure is always important in predicting whether a patient has a particular medical term.

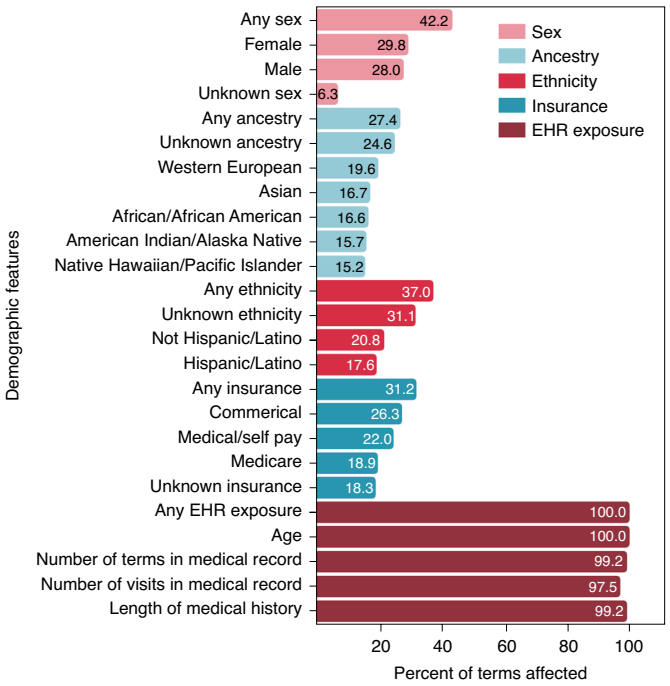


Fig. 1 | Feature selection by L1 regularization. Percent of medical term logistic regression models that include each demographic feature. For example, EHR exposure, that is, the length and density of a person's medical history, is an important predictor for every medical diagnosis, procedure and medication.

Effect of adding features on PBC performance. The binomial distribution models the discrete probability of the number of successes in N independent experiments each with probability P. A naive approach to comorbidity analysis assumes that the probability of seeing terms 1 and 2 ($P_{t1,t2}$) in a medical record is the product of the population incidence rates for terms 1 and 2. Knowing the number of patients with both terms 1 and 2 in their medical record, one can calculate a comorbidity P-value using the binomial test of statistical significance; however, as this method does not adjust for demographic factors, the P-values they generate can be driven by effects such as age and sex. A common approach in comorbidity literature is therefore to stratify the population by age and sex and then calculate the binomial P-value for each stratum[7]. By contrast, the PBC approach uses the Poisson binomial distribution to calculate P-values for each term pair. The Poisson binomial distribution is a generalization of the binomial distribution in which every trial/sample (that is, patient) has a different probability of success (that is, having both terms in their medical record). The probability of an individual patient having both terms in their medical record is calculated as the product of per-patient per-term probabilities generated from corresponding LRMs described above.

Figure 2 explores the relationships between comorbidities discovered by PBC versus stratification as a function of increasing numbers of demographic features. As can be seen, the stratification approach rapidly loses statistical power as more criteria are added to the stratum filter, an effect further exacerbated by reducing the initial cohort size. By contrast, PBC maintains statistical power by modeling demographic features.

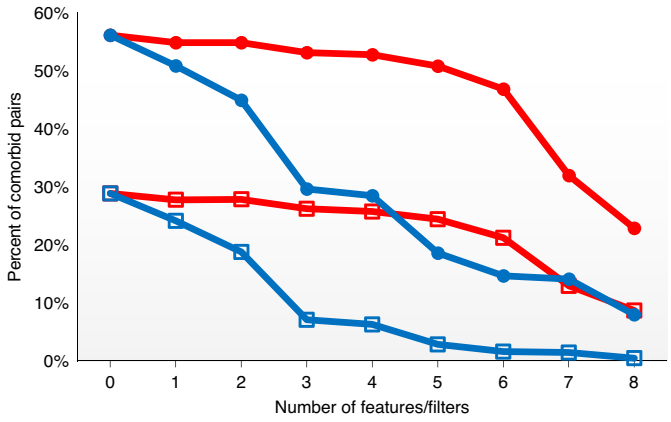


Fig. 2 | Modeling the effects of confounding variables. Percent of significantly co-occurring pairs of medical terms ($P < 1.08 \times 10^{-8}$) using either PBC (red) or stratification (blue) for two different sample sizes (an entire EHR corpus of 1,604,818 (filled circles) and a 78,275 patient sample (open squares)). Moving from left to right we introduce additional features to the PBC approach and additional filters to the stratification approach. The *x*-axis numbering corresponds to the following features/filters: 0, no features; 1, race (African American); 2, sex (female); 3, age (50–59); 4, ethnicity (non-Hispanic); 5, insurance (commercial); 6, span of medical history (at least two years); 7, number of visits (at least three visits); 8, date of last visit (at least as recent as Jan 2018).

Table 1 compares the potential of stratification and PBC to detect three well-known comorbidities[23–25] using our EHR corpus. As in Fig. 2, we compare the *P*-values generated by sequentially adding confounding variables. Notice how stratification dramatically lowers the strength of *P*-values as a function of the size of the stratum. This same behavior is illustrated globally in Fig. 2. Even with millions of EHRs, controlling for more than a few confounding demographic variables leads to strata that are too small to achieve statistical significance. By modeling the effects of multiple confounding variables, PBC retains statistical power to identify comorbidities.

Supplementary Table 2 presents five well-known comorbidities of breast cancer[26–31]. Stratification by age and gender deflates the strength of all *P*-values, and as a result they fall below the Bonferroni-corrected significance threshold (Supplementary Table 2, column 'binomial female 50–59'). By contrast, by explicitly modeling age, gender, race, ethnicity, insurance type and EHR exposure, PBC retains statistical power to capture these true-positive associations.

The complete set of comorbid term pairs discovered by PBC can be visualized using network analysis. In Extended Data Fig. 3, we use the minimum description length algorithm[32] to perform clustering of pairwise comorbidities by *P*-value strength. Terms with similar patterns of comorbidities are closer together in the network. We annotated selected comorbidities discovered within each cluster to illustrate this. A literature search confirmed that these labeled comorbidities represent existing clinical knowledge (see the corresponding citations in Supplementary Table 7). Extended Data Fig. 3 motivated us to produce an interactive tool for querying, exploring and extracting information from the comorbidity network. We developed a browser-based tool for exploration of comorbidities to provide better means to navigate this complex network, as discussed below.

Identification of comorbidities unique to minority groups. PBC can also capture true comorbidities hidden within mixed populations. For instance, consider sickle cell anemia (SCA), a disease that affects 1 in 365 African American newborns and 1 in 100,000 newborn Caucasians in the United States[33]. Malaise and fatigue, while common to many disorders, are among the most common symptoms of SCA[34–36], and usually manifest in an age-dependent manner. Table 2 compares five comorbidity *P*-values. Row 1 presents *P*-values calculated using all data. Although the true comorbidity is discovered, we cannot say with certainty whether the relationship is a true comorbidity or simply driven by a third confounding variable (for example, ancestry). The following rows present χ^2 *P*-values after stratifying by ancestry, ethnicity, gender and age, and PBC *P*-values after including these same features in the regression model. Stratification fails to find a significant comorbidity between SCA and maliase/fatigue once the data is partitioned by ancestry. The rarity of sickle cell disease in Caucasians and the small sample size for African Americans within Utah's EHR corpus make detection of this comorbidity difficult. PBC, by contrast, discovers the comorbidity. In fact, the additions of ancestry, ethnicity and age each increase the strength of the association between SCA and malaise/fatigue.

Application of PBC to a publicly available dataset. The University of Utah data used in this study cannot be shared publicly as it includes protected health information (PHI). We therefore also demonstrate the general applicability of PBC by applying it to the publicly available MIMIC-IV dataset[16], which includes 248,714 patients with associated ICD10 diagnosis or procedure codes. A total of 5,363,338 ICD9 and ICD10 diagnosis and procedure codes were converted to multilevel CCS codes. These include 725 distinct CCS diagnosis codes and 395 distinct CCS procedure codes. We repeated our experiments on this dataset, training a logistic regression model for each diagnosis and procedure code, and calculating comorbidities using the Poisson binomial distribution. Although the MIMIC-IV data are missing much of the detail available in the University of Utah dataset, the PBC results generally mirror those of the Utah dataset (Extended Data Fig. 4). Regression features are shown in Extended Data Fig. 4a. Furthermore, as a random offset is added to each patient's admission dates, it is not possible to control for the changes in use of various billing codes over time. Despite these limitations, the deployment of PBC on the MIMIC-IV public dataset further illustrates how PBC retains statistical power to identify comorbid relationships that are lost by stratification. Tables 1 and 2 are replicated on MIMIC-IV data as Supplementary Tables 3 and 4. Co-occurrence and directional comorbidities discovered within MIMIC-IV data can be queried at the following link: https://pbc.genetics.utah.edu/lemmon2021/pbc-mimic.

Understanding disease progression with temporalized *P*-values. The PBC approach can be extended to provide temporalized (or directional) *P*-values across prespecified time windows (see Methods for details). The inclusion of a direction window is necessary for several reasons. On short time scales, the order of appearance of diagnostic codes is an unreliable indicator of which condition actually preceded the other in the patient. The development of underlying disease, the relevant signs and symptoms, the provider arriving at a given diagnosis and the eventual recording of the said diagnosis might follow staggered paths that have little or no relevance when viewing the data in a time-slice of less than 30 days or so; thus, for many analyses it is probably best to treat these events as contemporaneous, however, in some cases a short window size is optimal for capturing a comorbidity.

Consider the following example. PBC reports the following *P*-values for the diagnosis/procedure pair amputation of lower extremity to post-operative infection: within 30 days $= 1 \times 10^{-3{,}663}$, greater than 30 days $= 1 \times 10^{-24}$, greater than 90 days $= 1 \times 10^{-6}$, greater than 365 days $= 0.86$. It is clear that shorter window sizes better capture the increased risk of infection after amputation, as one would logically expect. The window size must therefore be informed by clinical knowledge and research objectives. In this

Table 1 | Change in comorbidity *P*-values as extra confounding variables are considered

Stratification filters	PBC features	*P*-values					
		Concussion and migraine		Multiple myeloma and multiple sclerosis		Cancer of pancreas and hypertension	
		χ^2	PBC	χ^2	PBC	χ^2	PBC
No filters (n=1,538,059)	None	**1×10^{-933}**	**1×10^{-933}**	**1×10^{-121}**	**1×10^{-121}**	**1×10^{-405}**	**1×10^{-405}**
Female (n=794,281)	+Sex	**1×10^{-860}**	**1×10^{-937}**	**1×10^{-70}**	**1×10^{-135}**	**1×10^{-191}**	**1×10^{-400}**
+Age 50–59 (n=69,527)	+Age	**1×10^{-126}**	**$1\times10^{-1,031}$**	1.6×10^{-5}	**2.5×10^{-65}**	7×10^{-7}	**1×10^{-239}**
+Caucasian (n=45,782)	+Ancestry	**1×10^{-92}**	**1×10^{-912}**	1.5×10^{-5}	**2×10^{-87}**	5×10^{-7}	**1×10^{-159}**
+Non-Hispanic (n=39,897)	+Ethnicity	**1×10^{-91}**	**1×10^{-886}**	4.8×10^{-7}	**2×10^{-87}**	**6×10^{-11}**	**1×10^{-61}**
+Commercial (n=21,148)	+Insurance	**1×10^{-4}**	**1×10^{-859}**	0.32	**3×10^{-80}**	8×10^{-4}	**5×10^{-61}**
+3 yr History (n=9,243)	+Span	0.087	**1×10^{-522}**	0.41	**1.6×10^{-60}**	0.03	**5×10^{-78}**

Shown are three established comorbidities from the medical literature: concussion and migraine[23], multiple myeloma and multiple sclerosis[24], and cancer of pancreas and hypertension[25]. Comorbidities below a Bonferroni-corrected alpha threshold of 1.08×10^{-8} are shown in bold.

Table 2 | Identifying comorbidities specific to underrepresented minorities

Stratification filters	PBC features	SCA paired with malaise and fatigue		
		Stratification pair count	χ^2 *P*-value	PBC *P*-value
No filters (n=477,070)	No features	19	**4×10^{-8}**	**4×10^{-8}**
Caucasian (n=276,496)	Ancestry	5	0.002	**2×10^{-8}**
African American (n=7,035)	Ancestry	9	2×10^{-5}	**2×10^{-8}**
+Non-Hispanic (n=6,039)	+Ethnicity	8	2×10^{-4}	**1×10^{-8}**
+Female (n=2,789)	+Gender	7	1×10^{-4}	**9×10^{-8}**
+50–59 (n=302)	+Age	3	0.004	**3×10^{-19}**

Here we compare the stratification-based approach with PBC as regards ability to identify a known comorbidity: SCA paired with malaise and fatigue. *P*-values below a Benjamini–Hockberg-corrected alpha threshold of 1×10^{-6} are shown in bold. Filters applied under stratification produce the sample sizes shown in parentheses in column 1.

manuscript we examined associations based on a 90-day window. Using our website (see below) a user can also query additional window sizes (30 days, 365 days and 730 days).

Supplementary Table 5 presents specific directional associations discovered in an ab initio fashion by PBC and supported by clinical knowledge; for instance, Milrinone is prescribed to patients awaiting heart transplant[37]. The tendency of type-2 diabetics to develop chronic kidney disease is well-known[38,39]. HIV-induced immunocompromization often leads to pneumocystis[40] and obesity is a known risk factor for hypertension[41].

A web-based resource for comorbidity research. Among 4,623,841 pairs of medical terms in our collection of 1.6 million EHRs, we identified 3,311,830 comorbidities co-occurring within a 90-day window, and 1,969,941 temporally directed ones, acting over a time period greater than 90 days. All associations meet a Bonferroni significance threshold of 1.08×10^{-8}. The result is a highly connected network of comorbid diagnoses and associated procedures and medications based on the University of Utah EHR database.

In response to the size and complexity of these big-data, we have created browser-based means to navigate, query and explore them (https://pbc.genetics.utah.edu/lemmon2021/pbc-utah). A screenshot highlighting the functionality of the browser is shown in Fig. 3. The site allows the user to search for relationships between any pairwise diagnosis, procedure or medication. The result is a searchable table of all other DX, PX and RX codes along with statistics about the connection to the query term. These statistics include the counts, expectation and *P*-value of association after adjusting for confounders shown in Fig. 1. We use two-sided *P*-values so that 'less than expected' associations are also discoverable. For comparison, we provide both χ^2 *P*-values and G-test *P*-values.

In addition to patient lifetime co-occurrence *P*-values, we provide within-window and out-of-window directional *P*-values with a selectable window size (30, 90, 365 or 730 days). We also provide statistics on effect size, including relative risk, odds ratio and flow rate, which is the percent of patients coded with term 1 who later are coded with term 2. The flowchart view puts the query term in focus and shows the terms that tend to precede and follow the query term to the left and right, respectively. The user can filter by *P*-value strength and by flow rate.

A slice of that network is shown in Fig. 3. This figure contains a flowchart view for essential hypertension. By switching out the central node, an investigator can step through the temporalized network of diagnoses, procedures and medications. The investigator can further filter by effect size to find co-occurrences that are both significant and prevalent.

Comparisons with other published comorbidity tools. Table 3 compares functionalities provided on our comorbidity website with those found in other published comorbidity discovery tools. To the best of our knowledge, our website is the only available public resource of its kind that considers the individual risk profile for each patient having each medical term. Furthermore, our approach (and website) captures inverse comorbidities (terms occurring together less often than expected) and models temporal relations between pairs of terms.

The R package comoRbidity was published April 2018[10]. We installed the package and reformatted our data to fit the required specifications. Given our input of 150,598,377 CCS and RxNorm codes, the comoRbidity package consumes all available RAM (we used a Linux server with 504 GB of RAM) and fails to complete. We tried to acquire CytoCom[11] and comoR[12], however, the corresponding author has indicated that these projects are no longer maintained nor available for download. The Java package Comorbidity4j was published in January 2019[13]. We installed Comorbidity4j and attempted to compute on our full dataset. The application warned against calculating comorbidities for more than 300,000 pairs of

a

Column visibility | Copy | CSV | Excel | PDF | Print | Search:

Description of T	T Count	Q∧T Count	*P*-value Q∧T	In-window *P*-value	Out-of-window *P*-value T→Q	Out-of-window *P*-value Q→T
Hypertension	181,000	91,203	–26,322	–868,406	–5,208	–3,664
Diseases of the circulatory system	436,766	91,203	–6,948	–273,809	–1,988	–4,174
Disorders of lipid metabolism	100,726	42,746	–4,418	–64,534	–478	–830
Agents acting on the renin-angiote...	96,881	39,272	–4,006	–30,185	–1,266	–3,788
Endocrine; nutritional; and metabolic disease...	376,655	71,850	–3,232	–72,133	–1,087	–71.4
ACE inhibitors, plain	71,170	31,152	–3,037	–23,842	–727	–1890
lisinopril	61,018	26,985	–2,603	–19,716	–650	–1515
Diabetes mellitus without complication	100,643	34,758	–1,974	–28,828	–358	–353
Other and unspecified circulatory disease	83,121	32,375	–1,973	–18,035	–66.0	–257
Obesity	41,659	15,123	–1,913	–11,505	–45.1	–212

b

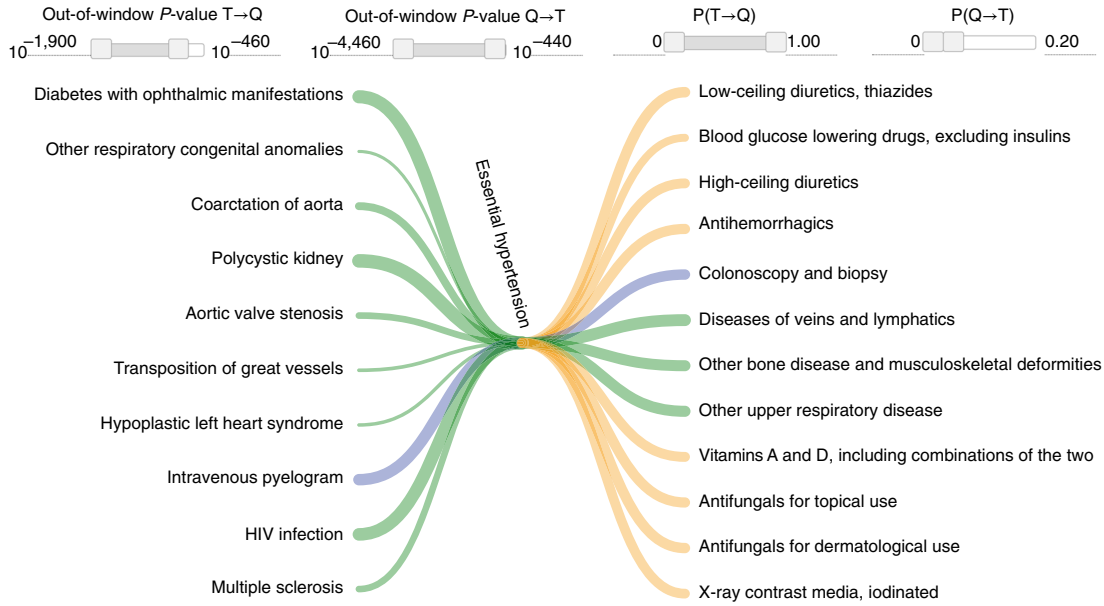


Fig. 3 | Screenshots from PBC-Web. a, Comorbidities of hypertension. Q refers to the query term (in this case hypertension) whereas T refers to the term possibly comorbid with Q. *P*-values that pass the Bonferroni-corrected significance threshold are colored green or red. Green indicates the relationship occurs more often than expected. Red indicates less often than expected. The last two columns represent flow rates, which indicate the actual percent of patients in our database that transit from one term to the next over time. **b**, Flowchart for hypertension. Terms that significantly precede or follow hypertension (separated by at least 90 days) are shown to the left and right of hypertension, respectively. Green connections are diagnoses, blue connections are procedures and orange connections are medications. The thickness of the connection relates to the flow rate—the percent of patients that flow through the given path.

terms (774 distinct terms). After several hours the calculation timed out, having allocated 50.2 GB of RAM. By contrast, our method uses a maximum of 174 MB of RAM and has no upper limit on the number of patients, visits or unique terms.

The disease trajectory browser (DTB) allows navigation of temporal relationships among medical records of 7.2 million Danish patients[14]. Direct comparison between our results is not possible as our groups have access to different EHR datasets, however, we can consider differences in methodology. The DTB measures effect size using relative risk, significance is measured using the binomial test and confounders are controlled for using case/control matching. For comparison, on PBC-Web we provide relative risk estimates and χ^2 *P*-values (which are a close approximation to the binomial test *P*-values). As described above, our approach models patient features rather than controlling for them through stratification or case/control matching.

Discussion

Although the term comorbidity is often used to denote significant associations between medical outcomes (for example, hypertension and heart attack), the concept is easily extended to include associated variables such as medical procedures and medications. For brevity's sake, in what follows, we refer to statistically significant associations among these collective variables as comorbidities.

Big data offer many opportunities and challenges for comorbidity discovery. One limitation imposed by data size is that morbidity discovery is necessarily pairwise; hence the term comorbidity, as opposed to multimorbidity discovery. Tools for multimorbidity-based

Table 3 | Comparison of comorbidity discovery tools

Package	CytoCom	comoR	comoRbidity	Comorbidity4j	DTB	PBC-Web
Release date	2014	2014	2018	2019	2020	2021
Relative risk	✓	✓	✓	✓	✓	✓
φ-Correlation	✓	✓	✓	✓		✓
Comorbidity score			✓	✓		✓
Odds ratio			✓	✓		✓
Fisher's exact test		✓	✓	✓		✓
χ^2 or binomial test					✓	✓
Poisson binomial						✓
Inverse comorbidities						✓
Temporal directionality			✓	✓	✓	✓
Interactive network of comorbidities	✓				✓	✓
Arbitrarily large datasets					See caption	✓
Dealing with confounders	Stratify	No support	Stratify	Stratify	Matching	Logistic regression
Platform	Cytoscape	R package	R package	Java/website	Publicly available website	Publicly available website

Features available in our online query tool, PBC-Web, are compared against the following comorbidity discovery tools: CytoCom[11], comoR[12], comoRbidity[10], Comorbidity4j[13], DTB[14]. Relative risk, φ-correlation, comorbidity score, odds ratio and Fisher's Exact test rely on the population incidence rate; Poisson binomial rely on per-person per-term probabilities. For DTB for arbitrarily large datasets, authors apply a prefiltering step as calculating all pairs of comorbidities is too computationally demanding.

discovery[42] are necessarily limited in scale due to computational constraints, considering for instance, 34 disease clusters[43]. By contrast, we have calculated pairwise comorbidities among 37,997 ICD10 diagnosis codes.

Commonly used statistical approaches to ab initio comorbidity discovery are hindered by the assumption that every member of the population (or stratum) has a disease probability equal to the population (or stratum) incidence rate. Stratification is commonly used to subset EHR collections to meet this requirement; however, stratification necessarily reduces sample size and statistical power (see Fig. 2). In practice, stratified data quickly become limiting even for very large datasets, as inclusion criteria grow more complex. This problem is exacerbated for ab initio approaches aimed at simultaneous discovery of comorbid relationships among thousands of diagnoses, procedures and medications. This process necessitates many millions of statistical tests; the requirement for multiple testing corrections mean that statistical power is of paramount importance.

Our motivation in developing PBC was to overcome the need for stratification while still achieving high accuracy and statistical power, which would allow discovery of comorbidities of rare diseases using small datasets, and ab initio discovery using very large EHR collections. Our results document the efficacy of PBC for achieving these ends. It is still important to bear in mind that the comorbidities it discovers do not necessarily indicate mechanistic relationships. For instance, two diagnoses may both be driven by smoking, but as smoking was not included in the logistic regression model, we cannot say anything about smoking as a potential driver (cause) of the relationship; thus PBC—like all existing methods in this domain—cannot assign comorbidities to one of the four etiological models described by Valderas and colleagues with with certainty[1], it can only say that the relationship is not due to factors included in the logistic regression model.

PBC provides an effective solution for a foundational step for outcomes research. The curse of dimensionality is a well-known phenomenon in which training a predictor with too many features can lead to higher error rates[44,45]. Considering there are about 69,000 ICD10 diagnosis codes, 70,000 ICD10 procedure codes and 350,000 RxNorm CUI codes, dimensionality reduction is necessary for effective machine learning on EHRs. By discovering which variables influence which outcome, PBC can reduce dimensionality and facilitate the creation of downstream tools for outcome predictions; thus, PBC's role in feature selection becomes clear. For a given clinical outcome, PBC can produce a manageable set of pairwise associations that become the inputs for training predictive models of disease.

It is important to note the limitations inherent in the use of data from a single EHR for comorbidity discovery. Data from a single EHR represents a non-random sampling of the general population and PBC does not model this sampling bias. University of Utah Healthcare sees many patients on a referral basis. For these patients, our view of their medical history is incomplete. This bias could result in PBC failing to discover true directionalities. Billing practices can vary within hospital systems—for instance, between inpatient and outpatient services and between provider billing and hospital billing. Hospital billing is performed by medical billing specialists, whereas provider billing is performed by clinicians. In this paper, we have restricted our analysis to provider billing terms. Clinical notes provide a still richer, more nuanced source of data and may more accurately describe a patient's medical condition. Application of PBC to the outputs of natural language processing tools will clearly be a fruitful path for future research.

Capobianco and Lio[46] present a vision for comorbidity discovery and analysis that is multidisciplinary and enabled by dynamic networks, with time as a key component in explaining disease relationships. PBC directly addresses these challenges, creating a scalable approach that (1) adjusts for confounding demographic variables and (2) models temporal relationships in large EHR datasets. We offer the PBC web-browser as a first-generation navigation tool for exploring these relationships. The PBC website provides a community resource for outcomes research, laying the foundation for improving current comorbidity-based outcomes tools, creation of new ones and, more generally, fueling healthcare discovery for improved care.

Methods

University of Utah medical records. The University of Utah maintains an EDW—a central storage and search facility for all data collected from all university hospitals and clinics, and all departments and specialties. SQL queries were composed to the following information: (1) medical record number, sex, race, ethnicity and age of each patient; (2) a list of patient visits, along with the visit date

and medical terms associated with each visit, such as diagnostic codes, procedure codes and medications ordered. Data were deidentified.

We collect ICD9[17] and ICD10[18] diagnosis codes, CPT procedural codes[20] and RXNorm[21] medication codes (concept unique identifiers) from University of Utah electronic medical records. ICD and CPT diagnosis and procedure codes were mapped to the hierarchical CCS[19]. CCS codes allow for more powerful statistics at the expense of concept resolution. After mapping to CCS, we retain 1,007 distinct diagnosis codes and 259 distinct procedure codes. These codes include both internal nodes and leaf nodes in the hierarchical CCS tree. In all, we collected records for 1.6 million patients, 50 million visits and 150 million diagnosis (DX), procedure (PX) and medication (RX) codes.

Counts of EDW patient demographics are displayed in Supplementary Table 1. Extended Data Fig. 1 displays how our data are distributed by gender and age decade. Panel B of Extended Data Fig. 1 shows how the length of patient medical records (in years) is distributed. Note that these lengths are limited by the history of electronic data collection at the University of Utah, which began in the early 2000s but started to ramp up around 2009 and has since increased rapidly. We thus see the 95th percentile for medical history length is around 12–15 years for most age bins. Panels C and D of Extended Data Fig. 1 show how the number of visits and the number of terms in a medical record trend with age. In almost all decades, women have more medical visits and medical terms than men, though this effect is most pronounced between 20 and 50.

Logistic regression for person-term probabilities. The initial step in comorbidity analysis is to ascertain the probability of a given term being found in a given person's medical record. A naive approach could assign everyone the same probability based on the term's frequency in the database or within each age–gender strata as seen in other methods.

Our approach involves developing a logistic regression model for each term. The independent features, X, are the list of persons in the EHR along with their gender, race, ethnicity, financial class and risk exposure. Risk exposure includes the age of a person at the time of their last visit as well as the length and density of their medical record. Length is defined as the number of days between first visit and last visit, whereas density is approximated by the number of visits within a medical record. The dependent outcome $\mathbf{y}$ is a binary vector indicating whether each person has the term in their medical record. The value C is the inverse of regularization strength in L2-penalized logistic regression. Smaller values of C indicate stronger regularization. The coefficients are determined by minimizing the following loss function, where β represents the coefficients and c is a constant (see Scikit-learn documentation[47] for a more complete discussion of logistic regression):

$$\min_{\beta,c} \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n} \log\left(\exp\left(-\mathbf{y}_i\left(X_i^T\beta + \mathrm{c}\right)\right) + 1\right) \quad (1)$$

For each term, stratified threefold cross-validation is used to determine the optimal value for C within the set $\{10^{-14}, 10^{-13}, 10^{-12}, \ldots, 10^{12}, 10^{13}, 10^{14}\}$. Cross-validation relies on a scoring function to assess the accuracy of logistic regression, given differing values for C. We evaluated standard and custom score functions based on their ability to differentiate logistic regression results with differing C values (see Extended Data Fig. 2).

The above approach resulted in logistic regression models for each term, capable of predicting the probability that a given person has a given term. For rare terms, we find that the probabilities output by logistic regression may not sum up to the actual number of patients with the term. To account for this we adjust each probability by a bias correction factor such that the sum of the adjusted patient probabilities is equal to the actual number of patients with the term. The exact form of this correction factor is given in Supplementary Section 2.

A limitation of logistic regression is the lack of a confidence metric on each predicted probability output by the regression model. For instance, predicted probabilities might be more accurate for patients of western European than African ancestry, as the corpus data are skewed for this demographic variable (see Supplementary Table 1). To overcome this limitation, we divide our data into six randomized partitions that were balanced so that the number of affected individuals in each partition is approximately equal. This partitioning is accomplished using StratifiedKFold from the python package sklearn[48]. Next, each partition is used to fit a logistic regression model, using the previously determined regularization strength. Each model is used to predict the probability of each person having the term. These six probabilities are used to calculate a sample variance for each person-term probability, $s^2_{P_{t\in m}}$, where m represents a patient's medical record, t represents a medical term, and $P_{t\in m}$ is the probability of term t in m. To be clear these six LRMs are only used to calculate sample variance, whereas the LRM trained on the full dataset is used per-patient per-term probability predictions.

Poisson binomial for term pair *P*-values. Our null hypothesis is that pairs of medical terms are independently distributed in University of Utah medical records, $H_0 : t1 \perp t2 | \mathcal{M}$. A significant P-value would indicate that a pair of terms co-occur more often than would occur by chance. Given two independent terms, $t1 \perp t2$ the probability the two would occur by chance in a given person's medical record m, is the product of their individual probabilities:

$$P_{\{t1,t2\}\subseteq m} = P_{t1\in m} * P_{t2\in m} \quad (2)$$

naive approach assumes person-term probabilities are equal to population incidence rates:

$$P_{t1\in m} = \frac{|\mathcal{M}_{t1}|}{|\mathcal{M}|}; P_{t2\in m} = \frac{|\mathcal{M}_{t2}|}{|\mathcal{M}|} \quad (3)$$

Using the naive approach, person-term-pair probabilities follow the binomial distribution with

$$P_{\{t1,t2\}\subseteq m} = \frac{|\mathcal{M}_{t1}|}{|\mathcal{M}|} * \frac{|\mathcal{M}_{t2}|}{|\mathcal{M}|} \quad (4)$$

However, using logistic regression, we have different probabilities for each person/term pair. The Poisson binomial distribution is the discrete distribution of a sum of Bernoulli trials where the probability of each trial differs; thus, using logistic regression, our data follow a Poisson binomial distribution.

As the cumulative distribution function of a Poisson binomial is computationally tractable only for a small number of values, numerous approximations have been developed[49]. We use the normal approximation as it is fast and accurate for large datasets. To determine a P-value using the normal approximation, the mean and variance for the Poisson binomial are calculated and used as parameters for a normal distribution. The mean of a Poisson binomial represents the expected number of University of Utah patients who will have in their medical record both terms in the pair and is calculated as the sum of probabilities for each person m:

$$\mu_{t1,t2} = \sum_{m}^{\mathcal{M}} P_{t1,t2|m} \quad (5)$$

The variance of a Poisson binomial is likewise similar in form to a binomial distribution:

$$\sigma^2_{t1,t2} = \sum_{m}^{\mathcal{M}} P_{\{t1,t2\}\subseteq m}\left(1 - P_{\{t1,t2\}\subseteq m}\right) \quad (6)$$

The Poisson binomial variance is augmented with the logistic regression variances described in the previous section $\left(s^2_{P_{t1\in m}}, s^2_{P_{t2\in m}}\right)$ using the product rule and the law of total variance. One can think of these variances as measurement error for $P_{t1\subseteq m}$ and $P_{t2\subseteq m}$ and they are larger for rare terms.

The probability that a person has a pair of terms, $P_{\{t1,t2\}\subseteq m}$, can be so rare it exceeds the limits of floating-point arithmetic. We thus implement our methods in log-space. Our logistic regression models report per-patient per-term log probabilities. We calculate $\ln(\mu_{t1,t2})$ and $\ln(\sigma^2_{t1,t2})$ rather than summing in normal space. We use a numerical approximation to calculate the $\ln(P\text{-value})$ of a normal distribution[50] (implemented as gsl_sf_log_erfc in the Gnu scientific library[51]). To report significance, we use throughout this paper an alpha threshold of 0.05. As we calculate P-values for 4,622,320 pairs of medical terms, our Bonferroni-corrected alpha is set to 1.08×10^{-8}.

Direction *P*-values. Imagine two terms that occur together in medical records more often than would occur by chance: which term tends to occur first in the medical record, or do they tend to occur in the same time frame? We calculate P-values for the temporal nature of each association. For each patient with medical record m, the date of the first occurrence of each term in m is recorded. Pairs of terms occurring within a window of size W are labeled as in-window, whereas pairs of terms occurring outside of W contribute to the $t1 \rightarrow t2$ count or the $t2 \rightarrow t1$ count. For the analyses presented here, we chose a 90-day window, as this duration decreases noise associated with the date of information capture within the medical record, but the approach is valid over any interval. Note that some term relationships—such as a surgery procedure followed by an infection diagnosis—will only show a significant direction with a window smaller than 90 days. PBC-Web includes four window sizes: 30, 90, 365, 730.

For a person with medical record m that contains terms 1 and 2 ($\{t1, t2\} \subseteq m$), the probability that term 1 occurs before term 2 is a function of the ratio of the probabilities of the two terms:

$$P^{m_{t1,t2}}_{t1\rightarrow t2} = P^{m}_{t1\rightarrow t2|\{t1,t2\}\subseteq m} = \frac{P^m_{t1}}{P^m_{t1} + P^m_{t2}} \quad (7)$$

Given $\{t1, t2\} \subseteq m$, the probability $t1$ and $t2$ occur within a time window of size W, is a function of the span or length of a person's medical history, Span_m:

$$P^{m_{t1,t2}}_{t1 \overset{W}{\sim} t2} = P^{m}_{t1 \overset{W}{\sim} t2|\{t1,t2\}\subseteq m} = \begin{cases} 1, & \mathrm{Span}_m \le W \\ \frac{W}{\mathrm{Span}_m}\left(2 - \frac{W}{\mathrm{Span}_m}\right), & \text{otherwise} \end{cases} \quad (8)$$

The above formula represents the percent of timepoints $t1$ and $t2$ that fall within W days of each other. The derivation of the formula is given in Supplementary Section 2.

The product of equations (7) and (8) gives the probability term 1 would precede term 2 within a window of size W:

$$P^{m_{t1,t2}}_{t1 \overset{in\ W}{\rightarrow} t2} = P^{m}_{t1 \overset{in\ W}{\rightarrow} t2 | \{t1,t2\} \subseteq m} = P^{m_{t1,t2}}_{t1 \rightarrow t2} \cdot P^{m_{t1,t2}}_{t1 \overset{W}{\sim} t2} \quad (9)$$

Similar logic can be applied to derive the out-of-window probability of term 1 occurring at least W days before term 2. A more complete explanation is found in Supplementary Section 2. Direction P-values are calculated using a normal approximation of the Poisson Binomial cumulative distribution function as in the previous section.

Statistics. Each P-value computed using the methods described above is compared against an adjusted alpha threshold to determine significance. Our browser-based query tool allows selection of either Bonferroni or Benjamini–Hochberg correction. Throughout this manuscript, unless otherwise noted, we report significance on the basis of a Bonferroni adjusted significance threshold.

Code, web-development and calculations. We implement our statistical analysis using Python, Cython and C. Cython is a static compiler for Python and the extended Cython programming language[52]. Scikit-learn[48] was used for logistic regression studies. All of the figures accompanying this article were generated using Matplotlib[53]. Our website is built using the Flask web framework[54]. The backend is pure python and the front end is JavaScript and D3[55]. Logistic regression modeling and pairwise calculation of Poisson Binomial P-values were performed at the University of Utah Center for High Performance Computing PHI protected environment. Training 3,041 logistic regression models while tuning the regularization strength with cross-validation took 1,959 CPU hours. The maximum memory usage was 174 MB. Calculating comorbidity statistics for 4,623,841 term pairs took 7,455 CPU hours while maximum memory usage was 87 MB.

Data availability

In this paper we calculate comorbidity statistics for all pairs of medical billing codes, including diagnoses, procedures and medications. All of these P-values are available to query and download from the following link: https://pbc.genetics.utah.edu/lemmon2021. Furthermore, a 2.3 GB file containing comorbidity statistics for all 4,623,841 pairs of medical terms can be downloaded from the Open Science Framework[56]. Source Data for Figs. 1 and 2 and Extended Data Fig. 1–4 are available with this manuscript. The original input data includes detailed medical records from University of Utah Health. As this data include PHI (patient demographics, birth dates and dated medical diagnosis, procedure, and medication codes): we cannot make the data available with this publication.

Code availability

We provide a CodeOcean capsule[57] including code and sample input data.

Received: 13 December 2020; Accepted: 16 September 2021;
Published online: 21 October 2021

References

1. Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C. & Roland, M. Defining comorbidity: implications for understanding health and health services. *Ann. Fam. Med.* **7**, 357–363 (2009).
2. Lone, N. I. et al. Predicting risk of unplanned hospital readmission in survivors of critical illness: a population-level cohort study. *Thorax* **74**, 1046–1054 (2019).
3. Wang, H. et al. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**, 1968–1978 (2018).
4. Facchinetti, G. et al. Continuity of care interventions for preventing hospital readmission of older people with chronic diseases: a meta-analysis. *Int. J. Nurs. Stud.* **101**, 103396 (2020).
5. Atashi, A., Sarbaz, M., Marashi, S., Hajialiasgari, F. & Eslami, S. Intensive care decision making: using prognostic models for resource allocation. *Stud. Health Technol. Inform.* **251**, 145–148 (2018).
6. Yurkovich, M., Avina-Zubieta, J. A., Thomas, J., Gorenchtein, M. & Lacaille, D. A systematic review identifies valid comorbidity indices derived from administrative health data. *J. Clin. Epidemiol.* **68**, 3–14 (2015).
7. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
8. Elixhauser, A., Steiner, C., Harris, D. R. & Coffey, R. M. Comorbidity measures for use with administrative data. *Med. Care* **36**, 8–27 (1998).
9. Roque, F. S. et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.* **7**, e1002141 (2011).
10. Gutiérrez-Sacristán, A. et al. comoRbidity: an R package for the systematic analysis of disease comorbidities. *Bioinformatics* **34**, 3228–3230 (2018).
11. Moni, M. A., Xu, H. & Liò, P. CytoCom: a Cytoscape app to visualize, query and analyse disease comorbidity networks. *Bioinforma. Oxf. Engl.* **31**, 969–971 (2015).
12. Moni, M. A. & Liò, P. comoR: a software for disease comorbidity risk assessment. *J. Clin. Bioinforma.* **4**, 8 (2014).
13. Ronzano, F., Gutiérrez-Sacristán, A. & Furlong, L. I. Comorbidity4j: a tool for interactive analysis of disease comorbidities over large patient datasets. *Bioinforma. Oxf. Engl.* **35**, 3530–3532 (2019).
14. Siggaard, T. et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nat. Commun.* **11**, 4952 (2020).
15. Winter, A. C., Rist, P. M., Buring, J. E. & Kurth, T. Prospective comorbidity-matched study of Parkinson's disease and risk of mortality among women. *BMJ Open* **6**, e011888 (2016).
16. Johnson, A. et al. *MIMIC-IV (Version 1.0)* (PhysioNet, 2021); https://doi.org/10.13026/S6N6-XD98
17. *ICD-9-CM—International Classification of Diseases, Ninth Revision, Clinical Modification* (CDC, 2019); https://www.cdc.gov/nchs/icd/icd9cm.htm
18. *ICD-10-CM—International Classification of Diseases, Tenth Revision, Clinical Modification* (CDC, 2020); https://www.cdc.gov/nchs/icd/icd10cm.htm
19. *Clinical Classifications Software Refined (CCSR)* (AHRQ, 2021); https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp
20. *CPT Codes* (AAPC, 2021); https://www.aapc.com/resources/medical-coding/cpt.aspx
21. Liu, S. et al. RxNorm: prescription for electronic drug information exchange. *IT Prof.* **7**, 17–23 (2005).
22. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
23. Seifert, T. The relationship of migraine and other headache disorders to concussion. *Handb. Clin. Neurol.* **158**, 119–126 (2018).
24. Shimanovsky, A. et al. Autoimmune manifestations in patients with multiple myeloma and monoclonal gammopathy of undetermined significance. *BBA Clin.* **6**, 12–18 (2016).
25. Saif, M. W., Kaley, K. & Lamb, L. Pancreatic adenocarcinoma complicated by sinistral portal hypertension. *Cureus* **8**, e689 (2016).
26. Han, H. et al. Hypertension and breast cancer risk: a systematic review and meta-analysis. *Sci. Rep.* **7**, 44877 (2017).
27. Li, X. et al. Comorbidities among patients with cancer who do and do not develop febrile neutropenia during the first chemotherapy cycle. *J. Oncol. Pharm. Pract.* **22**, 679–689 (2016).
28. Chia, V. M. et al. Chronic comorbid conditions associated with risk of febrile neutropenia in breast cancer patients treated with chemotherapy. *Breast Cancer Res. Treat.* **138**, 621–631 (2013).
29. Toma-Dasu, I., Wojcik, A. & Kjellsson Lindblom, E. Risk of second cancer following radiotherapy. *Phys. Med.* **42**, 211–212 (2017).
30. Donin, N. et al. Risk of second primary malignancies among cancer survivors in the United States, 1992 through 2008. *Cancer* **122**, 3075–3086 (2016).
31. Grantzau, T. & Overgaard, J. Risk of second non-breast cancer among patients treated with and without postoperative radiotherapy for primary breast cancer: a systematic review and meta-analysis of population-based studies including 522,739 patients. *Radiother. Oncol.* **121**, 402–413 (2016).
32. Rissanen, J. Modeling by shortest data description. *Automatica* **14**, 465–471 (1978).
33. Hassell, K. L. Population estimates of sickle cell disease in the U.S. *Am. J. Prev. Med.* **38**, S512–S521 (2010).
34. Ahmadi, M., Poormansouri, S., Beiranvand, S. & Sedighie, L. Predictors and correlates of fatigue in sickle cell disease patients. *Int. J. Hematol.-Oncol. Stem Cell Res.* **12**, 69–76 (2018).
35. Herson, J., Sharma, S., Crocker, C. L. & Jones, D. Physical complaints of patients with sickle cell trait. *J. Reprod. Med.* **14**, 129–132 (1975).
36. Aich, A., Jones, M. K. & Gupta, K. Pain and sickle cell disease. *Curr. Opin. Hematol.* **26**, 131–138 (2019).
37. Tariq, S. & Aronow, W. S. Use of inotropic agents in treatment of systolic heart failure. *Int. J. Mol. Sci.* **16**, 29060–29068 (2015).
38. Anders, H.-J., Huber, T. B., Isermann, B. & Schiffer, M. CKD in diabetes: diabetic kidney disease versus nondiabetic kidney disease. *Nat. Rev. Nephrol.* **14**, 361–377 (2018).
39. Koye, D. N., Magliano, D. J., Nelson, R. G. & Pavkov, M. E. The global epidemiology of diabetes and kidney disease. *Adv. Chronic Kidney Dis.* **25**, 121–132 (2018).
40. El Fane, M. et al. Pneumocystosis during HIV infection. *Rev. Pneumol. Clin.* **72**, 248–254 (2016).
41. Seravalle, G. & Grassi, G. Obesity and hypertension. *Pharmacol. Res.* **122**, 1–7 (2017).
42. Hassaine, A., Salimi-Khorshidi, G., Canoy, D. & Rahimi, K. Untangling the complexity of multimorbidity with machine learning. *Mech. Ageing Dev.* **190**, 111325 (2020).
43. Hassaine, A. et al. Learning multimorbidity patterns from electronic health records using non-negative matrix factorisation. *J. Biomed. Inform.* **112**, 103606 (2020).
44. Chandrasekaran, B. & Jain, A. K. Quantization complexity and independent measurements. *IEEE Trans. Comput.* **C–23**, 102–106 (1974).

45. Trunk, G. V. A problem of dimensionality: a simple example. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 306–307 (1979).
46. Capobianco, E. & Lio', P. Comorbidity: a multidimensional approach. *Trends Mol. Med.* **19**, 515–521 (2013).
47. *Linear Models* Section 1.1, Scikit-learn 0.24.1 Documentation (ScikitLearn, 2021); https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
48. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
49. Hong, Y. On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.* **59**, 41–51 (2013).
50. Hart, J. F. *Computer Approximations* (Wiley, 1968).
51. *GNU Scientific Library: Reference Manual* (Network Theory, 2009).
52. Behnel, S. et al. Cython: the best of both worlds. *Comput. Sci. Eng.* **13**, 31–39 (2011).
53. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
54. Grinberg, M. *Flask Web Development: Developing Web Applications with Python* (O'Reilly, 2018).
55. Bostock, M., Ogievetsky, V. & Heer, J. D^3: data-driven documents. *IEEE Trans. Vis. Comput. Graph.* **17**, 2301–2309 (2011).
56. Lemmon, G., Wesolowski, S., Henrie, A., Tristani-Firouzi, M., & Yandell, M. *PBC Comorbidities* (OSF, 2021); https://doi.org/10.17605/OSF.IO/TH239
57. Lemmon, G., Wesolowski, S., Henrie, A., Tristani-Firouzi, M., Yandell, M. *A Poisson Binomial Based Statistical Testing Framework for Comprehensive Comorbidity Discovery Across Massive Electronic Health Record Datasets* (CodeOcean, 2021); https://doi.org/10.24433/CO.2251918.v1

Acknowledgements

The following collaborators have provided valuable discussion, feedback, and insight which has guided development of PBC: B. Bray, V. Deshmukh, K. Eilbeck, E. J. Hernandez and R. Shah. We thank members of the University of Utah EDW for facilitating access to medical records. The computational resources used were partially funded by the NIH Shared Instrumentation Grant 1S10OD021644-01A1. This research was supported by the AHA Children's Strategically Focused Research Network grant (17SFRN33630041) and the Nora Eccles Treadwell Foundation. G. Lemmon was supported by NRSA training grant T32H757632. S. Wesolowski was supported by NRSA training grant T32DK110966-04 and the AHA Children's Strategically Focused Research Network Fellowship award (17SFRN33630041).

Author contributions

G.L. was the senior research associate leading PBC development and validation. S.W. is an applied mathematician who has helped formalize our approach to statistical testing. A.H. was a software engineer on the project. M.T.-F. and M.Y. conceived of the project and secured research funding and played a key role in scientific discussions regarding development of PBC. All authors edited the manuscript.

Competing interests

G.L. and M.Y. own shares in Backdrop Health, a University of Utah effort to commercialize Bayesian inference on health records; however, there are no financial ties regarding this research. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43588-021-00141-9.

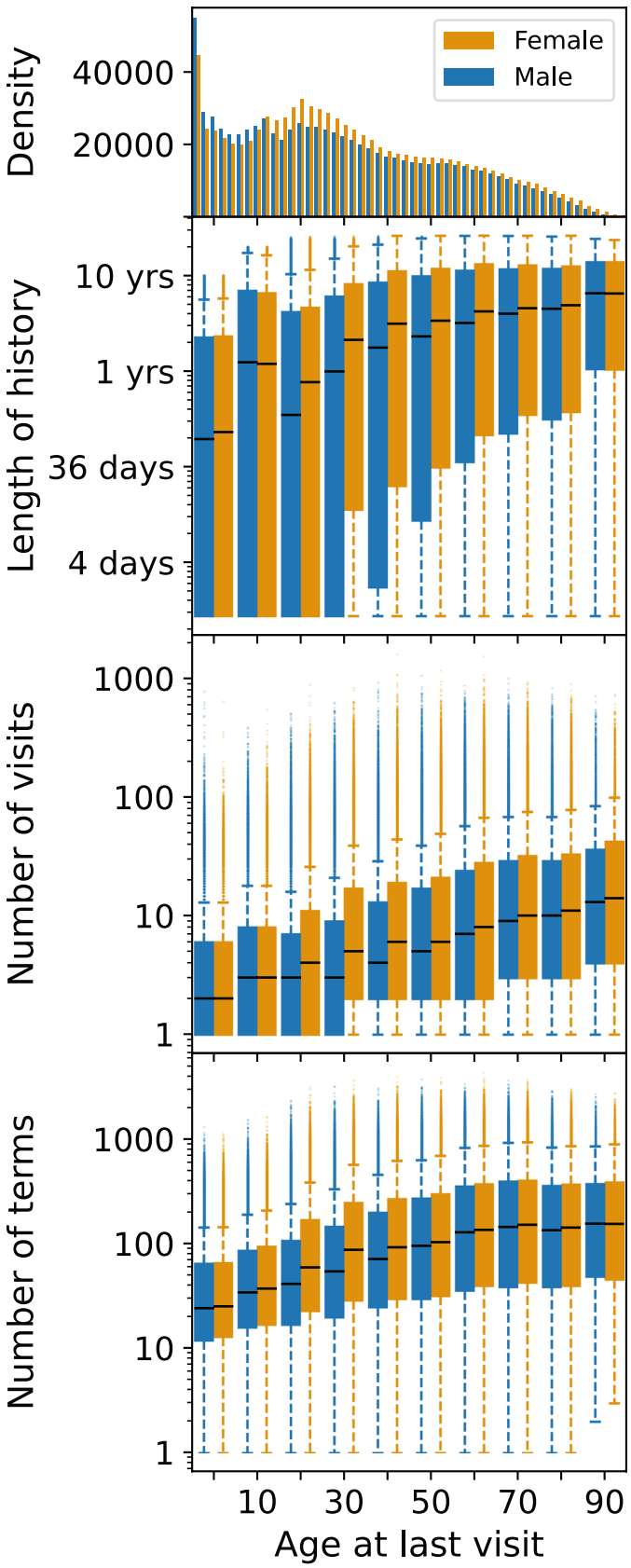
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-021-00141-9.

Correspondence and requests for materials should be addressed to Martin Tristani-Firouzi or Mark Yandell.

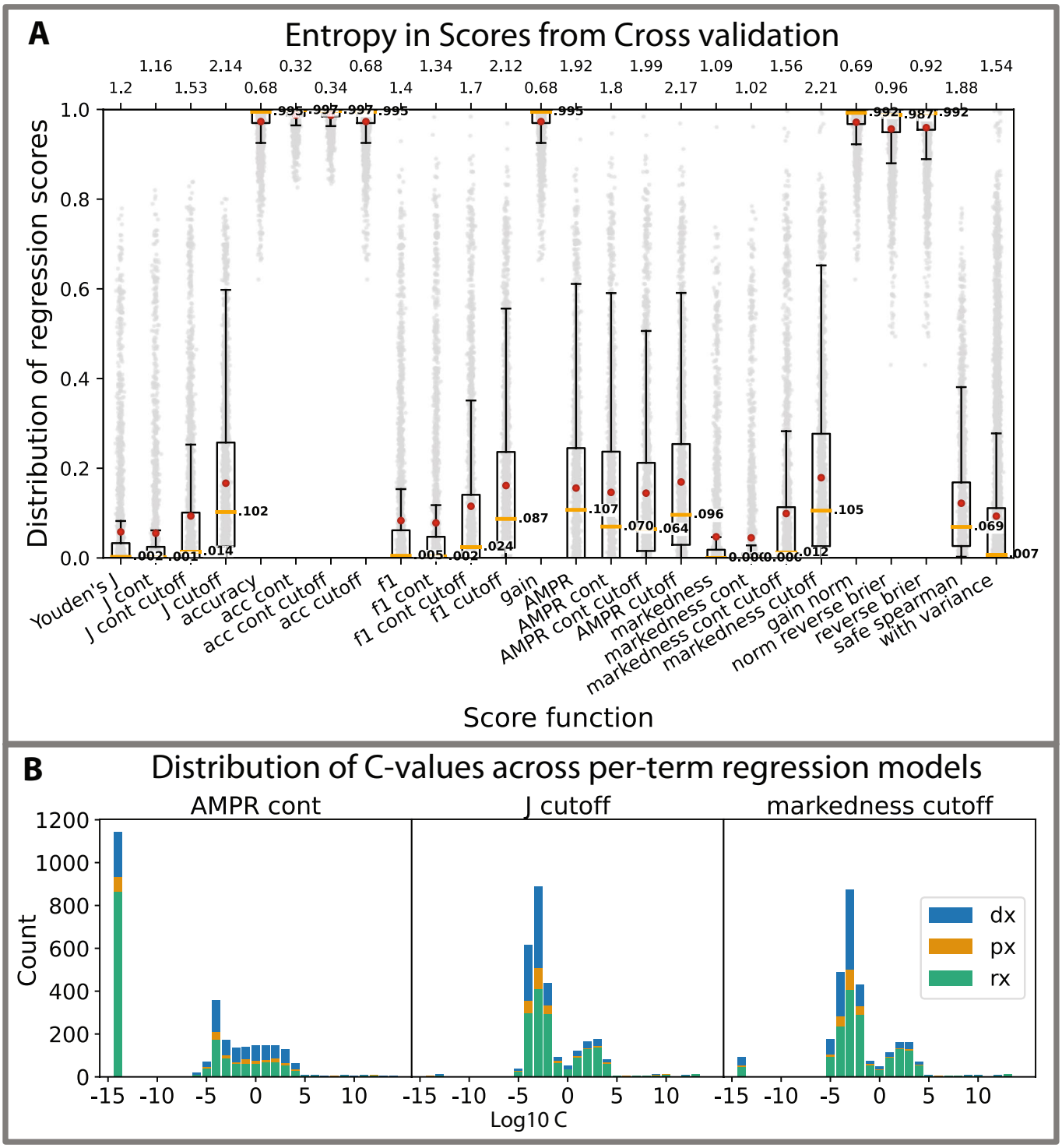
Peer review information *Nature Computational Science* thanks Jeffrey P. Rewley and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Handling editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team.

Reprints and permissions information is available at www.nature.com/reprints.

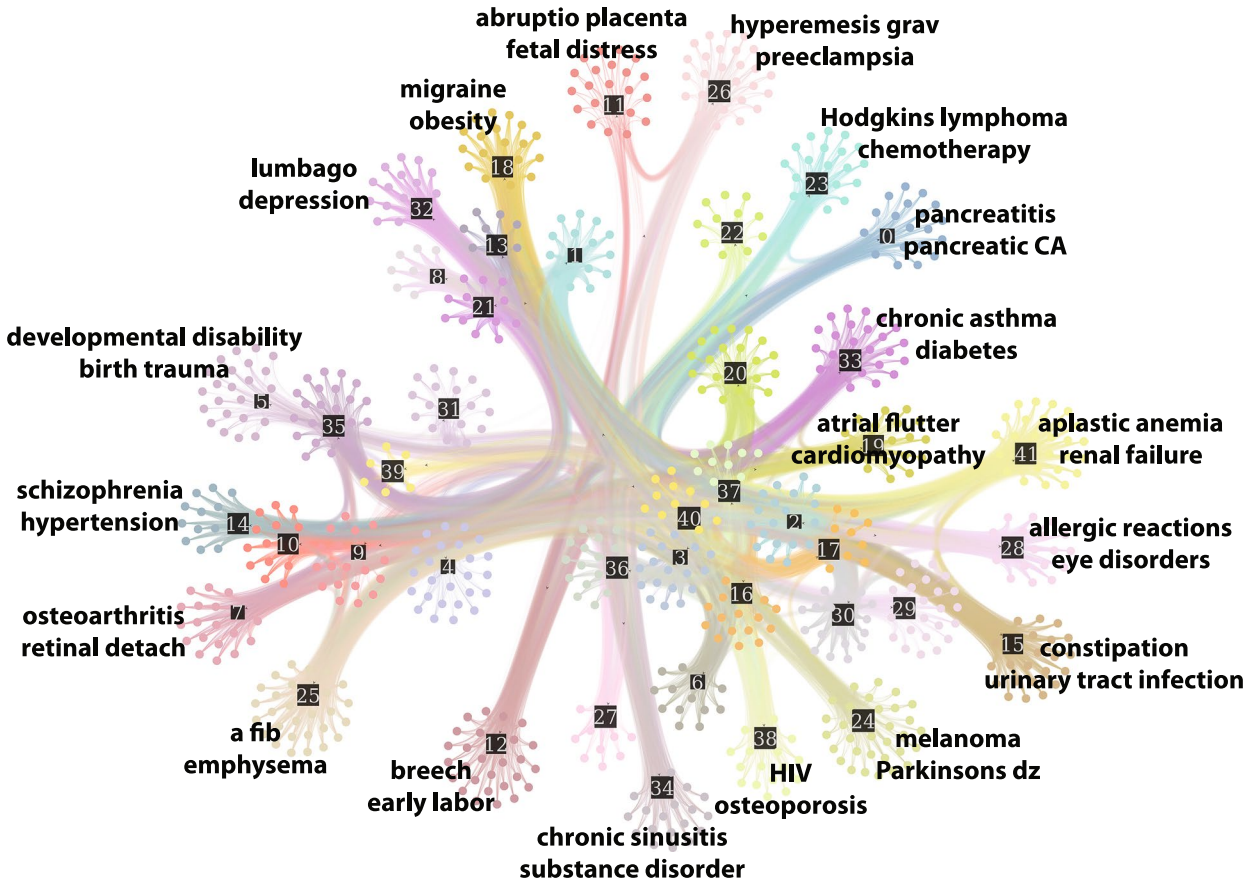
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



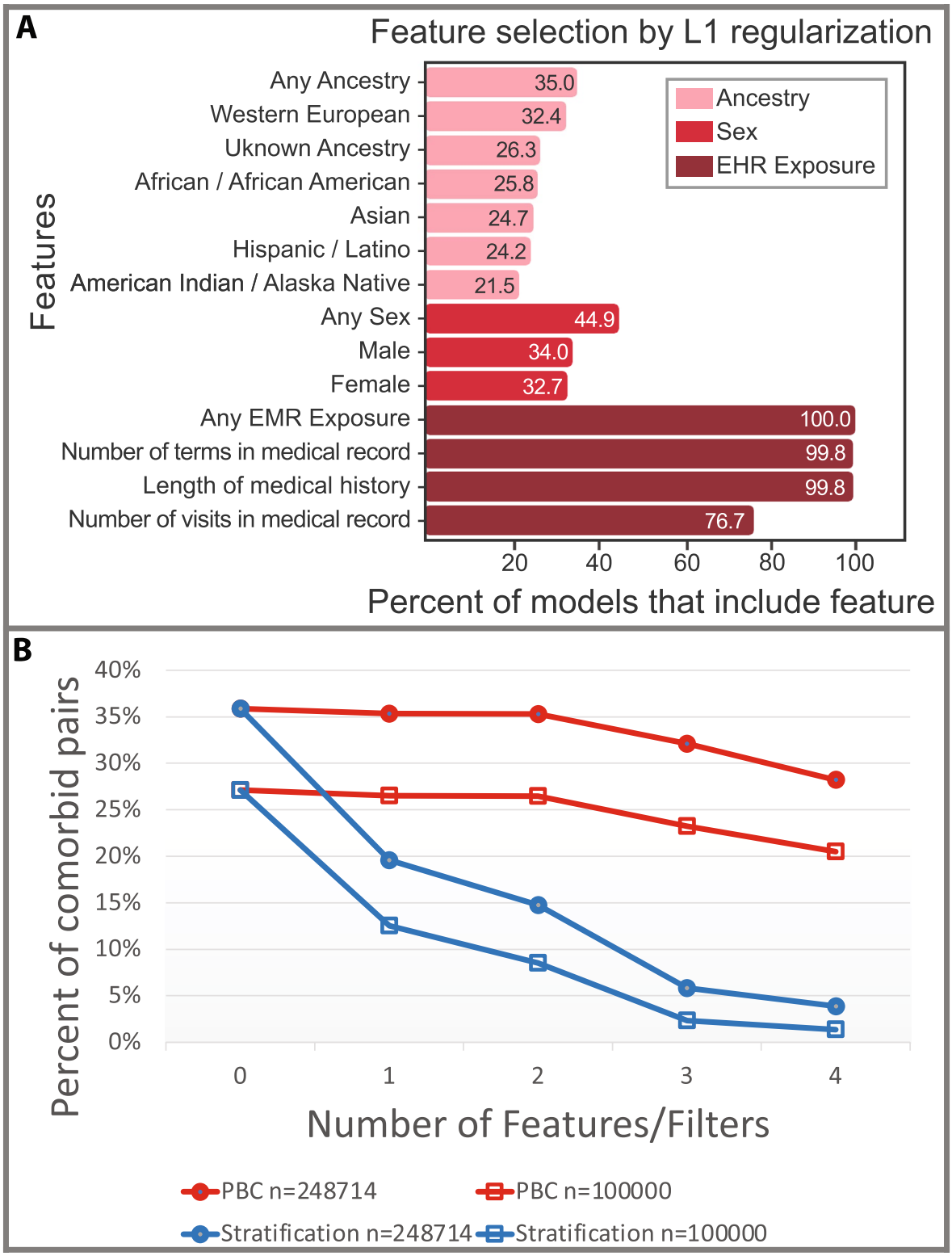
Extended Data Fig. 1 | University of Utah medical records binned by age-decade. Boxplots show median (black line), 25th and 75th percentile (box ends), 95th and 5th percentile (whisker caps) and outliers. Number of terms (bottom panel) is a count of distinct diagnoses, procedures and medications found in each patient's medical history.



Extended Data Fig. 2 | Comparison of score functions for logistic regression C-value optimization. For each score function, we evaluated C-values ranging from 10^{-14} to 10^{14}. **(a)** For each of 3041 diagnosis (DX), procedure (PX), and medication (RX) terms, we use cross validation to select the C-value that achieves the best score. Each boxplot contains these 3041 best scores as evaluated with different score functions. **(b)** Distribution of C-values for 3 score functions with high entropy. J_{cutoff} was chosen for downstream analysis because it has high entropy and has a smooth C-value distribution without the large outlier at C = −14.



Extended Data Fig. 3 | Minimum description length of the comorbidity network discovered by PBC for diagnoses in the University of Utah EDW. Examples of significantly associated medical conditions within each cluster are displayed. Citations supporting these associations are listed in Supplementary Table 6.



Extended Data Fig. 4 | Deployment of PBC on MIMIC-IV EHR data. See Fig. 1 legend for description of **(a)** and Fig. 2 legend for description of **(b)**. In **(b)**, the X-axis ticks correspond to the addition of regression features (PBC) or stratification criteria from left to right: 0 - no features, no stratification, 1- gender/female, 2 - ancestry/African American, 3 - length of medical history/at least 2 years, 4 - number of visits/at least 3 visits. The MIMIC-IV results are very similar to the University of Utah results, reinforcing a key message of this paper - that PBC retains the power to identify comorbid relationships that are lost by stratification.