Screening for Idiopathic Pulmonary Fibrosis with Comorbid Pattern Recognition in Electronic Health Records

Dmytro Onishchenko1, Robert J. Marlowe2, Gary M. Hunninghake3, Fernando J. Martinez4,5 and Ishanu Chattopadhyay1,6,7

1Department of Medicine, University of Chicago, Chicago, IL USA
6Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL USA
7Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL USA
2Spencer-Fontayne Corporation, Jersey City, NJ, USA
3Director, Interstitial Lung Disease Program, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA USA
4Bruce Webster Professor of Internal Medicine, Medicine, Weill Cornell Medical College, New York, NY USA
5Chief of Division of Pulmonary and Critical Care Medicine at Weill Cornell Medicine and NewYork-Presbyterian Weill Cornell Medical Center, New York, NY USA

⋆To whom correspondence should be addressed: e-mail: ishanu@u.chicago.edu.

Idiopathic pulmonary fibrosis (IPF) is an irreversible, debilitating, and ultimately lethal fibrosing interstitial lung disease (ILD) of unknown cause1–3. The poor prognosis of IPF (mean survival less than 2-5 years post diagnosis4), combined with its worldwide prevalence greater than all but the most common cancers5, makes it a serious health problem. Thus, the need for early diagnosis is paramount. Currently, early screening is hampered by the absence of reliable screening tools, non-specific symptomology, a limited understanding of the phenotypic and genetic markers for early-stage IPF, and the potential need for invasive procedures for confirmatory diagnosis. In this study we introduce a new screening tool that requires no new diagnostic tests, may be universally administered, and does not necessarily require recognition of early symptoms by the patients or care providers. Applying novel pattern discovery algorithms on the detailed history of past medical encounters of individual patients, we leverage subtle comorbidity patterns to compute the Zero-burden Co-Morbidity Risk Score for IPF (ZCoR-IPF), which is expected to be widely, readily, rapidly and inexpensively applicable at points of care. Our algorithm is trained and validated on a large national insurance claims database (n=2,059,559). In out-of-sample validation, we demonstrate that ZCoR-IPF identifies IPF patients with an AUC approaching 84% at 4 years, ≥ 86% at 3 years, ≥ 87% at 2 years, and exceeding ≥ 88% at 1 year before conventional diagnosis. Our out-of-sample accuracy indicates that on average, we correctly predict the eventual IPF status ≥ 94 – 98 out of every 100 patients, with a positive predictive value ≥ 50% irrespective of sex at a specificity of 99%. Maximum PPV and NPV achieved are ≥ 70% and ≥ 98% respectively. We investigate several important sub-populations, including high risk patients with known IPF co-morbidities, low risk patients who lack those indications, in older (> 65) and younger populations, in patients who already have a diagnosis of Chronic obstructive pulmonary disease (COPD), asthma or have cardiac disorders, and for patients who are never recorded to have experienced dyspnea and thus are at a higher risk of a missed diagnosis. High performance of the ZCoR-IPF score across clinically relevant situations suggests that it may aid providers to more selectively flag patients for detailed diagnostic evaluation, substantially improving patient outcomes and quality-of-life, as well as efficiency of healthcare resource utilization. Additionally, earlier diagnosis may improve the odds of disease modifying therapy, optimal timing of a lung transplantation, and accelerate access to interventions and palliative care6,7.
decade, timely, efficient, and confident diagnosis of IPF has been recognized as a major public health challenge worldwide.\textsuperscript{2,5,6,8–15} In this study, we put forth a novel approach to predict future IPF diagnosis by algorithmically identifying subtle patterns in the time, nature and ordering of past medical encounters of individual patients.

Over the past decade, timely, efficient, and confident diagnosis of IPF has been recognized as a major public health challenge worldwide.\textsuperscript{2,5,6,8–15} Identifying IPF cases is a complex, multi-step, and often multi-year process.\textsuperscript{6,11,13–17} A usually necessary but not always sufficient condition is referral for high-resolution chest computed tomography (HRCT), often at an expert center.\textsuperscript{6,11} In most cases, such referral leads to eventual recognition of radiological or histological usual interstitial pneumonia (UIP), the hallmark of IPF, or of other radiological and/or histological signs associated with this disease, and serologic studies to rule-out of other forms of interstitial lung disease (ILD).\textsuperscript{2,6,11} For such referral to happen, one or more of a number of scenarios must unfold: 1) Patients recognize the chronicity, progression, or both of respiratory symptoms and the resultant need for medical attention.\textsuperscript{6,7,15,17,2} 2) Health-care practitioners note the importance of these symptoms, of signs of possible fibrosis on auscultation, or of both.\textsuperscript{5,18} 3) Radiologists, not infrequently non-specialists in chest imaging, note incidental findings of interstitial lung abnormalities (ILA) or ILD on thoracic or abdominal CT.\textsuperscript{10,14,19,20} Alternatively, a pulmonologist or a HRCT referral may only take place after an emergency room visit or hospitalization due to acute exacerbation of IPF.\textsuperscript{15}

Hence, IPF diagnosis often entails multiple physician visits and repeated, sometimes invasive tests, and is often characterized by delay and misdiagnosis.\textsuperscript{6,13,14} Not infrequently, this situation pertains even after ILD is seen on CT or a pulmonologist has been consulted. For example, a retrospective analysis of two US academic medical center cohorts found around a quarter of patients with IPF or other forms of ILD to have waited $\geq$ 1 year after a CT finding of ILD to see a pulmonologist, while a Medicare claims study\textsuperscript{13} found that more than a third of patients saw a pulmonologist $> 3$ years before IPF diagnosis. A medical chart analysis\textsuperscript{10} and a patient survey\textsuperscript{6} suggested misdiagnosis rates approaching or exceeding 40%. Notably, 28\% of respondents in a 600-patient survey\textsuperscript{6} reported that the burden of their journey to an IPF diagnosis contributed to a decision to apply for disability or to retire.

The difficulties in reliable early diagnosing IPF may be ascribed to several causes. First, the most common clinical symptoms of the disease,\textsuperscript{7,17} insidiously progressive chronic exertional dyspnea and/or chronic, often mild cough, are highly non-specific. These symptoms are easily attributed by patients to age or deconditioning,\textsuperscript{6} or by physicians to more common cardiopulmonary diseases.\textsuperscript{5,6,15,16} Important risk factors for IPF, namely, older age, male sex, and cigarette smoking,\textsuperscript{1} are similarly “non-specific”. Second, there is still limited understanding and characterization of phenotypic and genetic findings associated with “early-stage IPF”.\textsuperscript{10} Third, the current diagnostic hallmark of IPF, UIP on HRCT or histology,\textsuperscript{2} is a late-stage finding.\textsuperscript{3} Moreover, UIP may be confirmed via relatively invasive procedures requiring specialized interpretation, e.g., HRCT or surgical lung biopsy.\textsuperscript{6} Lastly, no validated or easily-applicable screening modalities currently exist for IPF.\textsuperscript{9}

In this study we introduce an innovative approach with the potential to ameliorate these key challenges: we introduce a tool that requires no new diagnostic tests, may be universally administered, and does not necessarily require recognition of early symptoms by the patients or care providers. Analyzing large databases of electronic health records via novel pattern discovery algorithms, we identify subtle co-morbidity incidence, timing, and sequence characteristics presaging IPF. Combining these discovered features with standard machine learning then leads to a powerful, accurate, automated screening tool based only on diagnostic codes that exist already in the patient’s past medical record. Here we report on the training and validation of our screening tool, the Pulmonary Fibrosis Co-Morbidity Risk Score (ZCoR-IPF), which is expected to be widely, readily, rapidly and inexpensively applicable at points of care.

**MATERIALS AND METHODS**

**Data Source & Patient Selection**

Our patient data comes from the Truven Health Analytics MarketScan\textsuperscript{®} Commercial Claims and Encounters Database for the years 2003-2018\textsuperscript{21} (referred to as the Truven dataset). This US national database merges data contributed by over 150 insurance carriers and large self-insurance companies, and comprises over seven billion time-stamped diagnosis codes. The entire database tracks over 87 million patients for 1 to 15 years, reflecting a substantial cross-section of the US population. We select our cohort(s) from the Truven dataset in accordance with the inclusion/exclusion criteria described in Table I, ensuring that selected patients have at least three years of medical history recorded in the dataset. The geographical distribution of the patients in our selected cohort(s) is illustrated in Fig. 1a. Fig. 1b illustrates the age distribution at the time of IPF diagnosis (mean $\approx 68$ years), which is consistent with the reported mean onset age for IPF ($\approx 66$ years). We also
**a. Geospatial distribution of positive cases**

**i. Male**

**ii. Female**

**b. Distribution of onset age in the database**

**c. Risk of onset with age**

**d. CONSORT**

![Statistical Characteristics of the Truven Database](image)

**Fig. 1: Statistical characteristics of the Truven database.** Panel a shows where our patients originate from. The geospatial distribution is correlated with population density, as expected. Panel b illustrates the distribution of patient diagnostic ages in the database. Note that the mean age of onset compares favorably with the value reported in the literature (≈ 68 vs 66 years). Panel d shows the probability of diagnosis with age, taking into account the variation of the number of patients of a given age in the database (higher the age beyond 65, smaller the number of patients). Panel d shows patient selection via a CONSORT (Consolidated Standards of Reporting Trials) diagram.

Note that observed risk of onset actually increases with age, which is computed as the number of IPF cases normalized by the total number of patients at the same age, as shown in Fig. 1c.

To view the task of predicting a future IPF diagnosis as a binary classification problem, we aim to categorize time-stamped sequences of diagnostic codes into positive and control categories, where the “positive” category refers to patients diagnosed with IPF 1 year (See Tables IV) from the point of screening. We also consider earlier screening up to 4 years before the actual diagnosis, as described later. The control cohort comprises patients who do not have any target codes in their records, within the next 2 years of the point of screening. For both groups, we base our predictions on the past 2 years of medical history. We consider approximately 42 million diagnostic codes (with over 46K unique codes) for both genders (See Table III for detailed enumeration) in this analysis, and ultimately identified \( n = 2,059,559 \) patients, with 53,407 patients in the positive group and...
TABLE I: Inclusion/Exclusion, Positive/Control Criteria & Cohort Definitions

| Definitions |
|-------------|
| **Inclusion/Exclusion Criteria** |
| Age 48 - 90 |
| Has medical history spanning ≥ 3 yrs before target code\(^2\) (positive, See Tab. IV for list of target codes), or end of record (control) |
| **Positive & Control Cohorts** |
| **Positive Cohort:** Patients with at least one target code (Tab. IV) |
| **Control Cohort:** Patients lacking any target code |

*Medical records denote time stamped history of diagnostic codes
\(^2\)Target codes identify the target condition (IPF or Pulmonary Fibrosis without known cause)

TABLE II: Cohort Sizes

| gender | age group | n       | n\(_{\text{positive}}\) | n\(_{\text{control}}\) |
|--------|-----------|---------|-------------------------|------------------------|
| M      | 45-55     | 226150  | 1870                    | 224280                 |
| M      | 55-65     | 480104  | 9471                    | 470633                 |
| M      | 65-75     | 157017  | 7999                    | 149018                 |
| M      | 75-       | 72469   | 7464                    | 65005                  |
| F      | 45-55     | 291111  | 2554                    | 286557                 |
| F      | 55-65     | 580593  | 10467                   | 570126                 |
| F      | 65-75     | 171022  | 7130                    | 163892                 |
| F      | 75-       | 81093   | 6452                    | 74641                  |
| Total  |           | 2059559 | 53407                   | 2006152                |

TABLE III: Number of diagnostic codes used

| gender | Number of codes | Number of unique codes | codes defining disease categories |
|--------|-----------------|------------------------|----------------------------------|
| M      | 17008378        | 22685                  | 85016                            |
| F      | 25074255        | 23722                  | 85016                            |

TABLE IV: Target Codes: Description of ICD Codes(s) Used To Identify IPF diagnoses

| ICD code | description |
|----------|-------------|
| 516.31   | Idiopathic pulmonary fibrosis |
| J84.112  | Idiopathic pulmonary fibrosis |
| J84.113  | Idiopathic non-specific interstitial pneumonitis |
| J84.111  | Idiopathic interstitial pneumonia not otherwise specified |
| J84.1    | Pulmonary fibrosis unspecified |
| J84.10   | Pulmonary fibrosis unspecified |
| J84.11   | Idiopathic interstitial pneumonia not otherwise specified |

2,006,152 patients in the control group. The cohort sizes are described in Table II.

The diagnostic codes specifically for IPF are 516.31 (ICD9) and J84.112 (ICD10). We use a somewhat broader definition, including five additional ICD10 codes that indicate a diagnosis of pulmonary fibrosis without a known cause (See Table IV). This broadened definition of our target reduces the impact of erroneous coding and diagnostic uncertainties for IPF, and does not violate the key characteristics of a pulmonary fibrosis diagnosis for which no clear causal mechanism (such as toxic exposure) could be established.

Importantly, we do not pre-select or reject any diagnostic or prescription code based on its known or suspected comorbidity with IPF. To investigate if our predictive performance changes substantially for patients who are at “high risk” based on known co-morbidities, we also evaluate our performance within a high risk and a low risk sub-cohort. The high risk sub-cohort comprises patients with one or more of the diagnoses enumerated in Table V, which identify the top known co-morbidities\(^{22}\). The low risk sub-cohort comprises patients who are not at high risk as specified by the previous condition. Our results for the low risk sub-cohort is of particular significance, as these patients are at a higher risk of being not diagnosed early. Additionally, we investigate the applicability of ZCoR-IPF in patients 1) for whom IPF diagnosis might be confounded due a preexisting condition such as COPD, any heart condition, or asthma (See SI-Table III for cohort definition codes), and 2) for whom the absence of any indication of dyspnea reduces the odds of IPF suspicion and might delay diagnosis (See SI-Table II for codes that are absent in this cohort).
| ICD code | Description                                      |
|----------|--------------------------------------------------|
| K21.9    | Gastro-esophageal reflux disease without esophagitis |
| K21      | Gastro-esophageal reflux disease with esophagitis without bleeding |
| I27.20   | Pulmonary hypertension unspecified               |
| I27.0    | Primary pulmonary hypertension                   |
| J44.9    | Chronic obstructive pulmonary disease unspecified |
| G47.33   | Obstructive sleep apnea (adult) (pediatric)      |

*Low-risk patients who lack these diagnoses before IPF diagnosis (positive cohort) or anywhere in their medical history (control cohort).

**TABLE VI:** Predictive performance of ZCoR-IPF for different subpopulations

| sex       | cohort                        | spec. | sens. | PPV  | acc  | NPV  | PPV† | NPV† | AUC  |
|-----------|-------------------------------|-------|-------|------|------|------|------|------|------|
| Female    | All patients                  | 95%   | 0.60  | 0.23 | 0.94 | 0.99 | 0.70 | 0.98 | 0.887±0.006 |
| Female    | All patients                  | 99%   | 0.42  | 0.52 | 0.98 | 0.99 | 0.70 | 0.98 | 0.887±0.006 |
| Male      | All patients                  | 99%   | 0.38  | 0.54 | 0.97 | 0.98 | 0.71 | 0.98 | 0.883±0.005 |
| Male      | All patients                  | 95%   | 0.55  | 0.25 | 0.94 | 0.99 | 0.71 | 0.98 | 0.883±0.005 |
| Female    | COPD/Cardiac/Asthma*          | 95%   | 0.57  | 0.30 | 0.94 | 0.98 | 0.68 | 0.98 | 0.865±0.000 |
| Female    | COPD/Cardiac/Asthma*          | 99%   | 0.41  | 0.62 | 0.97 | 0.98 | 0.68 | 0.98 | 0.865±0.000 |
| Male      | COPD/Cardiac/Asthma*          | 99%   | 0.35  | 0.62 | 0.96 | 0.97 | 0.69 | 0.97 | 0.848±0.000 |
| Male      | COPD/Cardiac/Asthma*          | 95%   | 0.51  | 0.31 | 0.93 | 0.98 | 0.69 | 0.97 | 0.848±0.000 |
| Male      | age 65+                       | 99%   | 0.29  | 0.47 | 0.97 | 0.98 | 0.56 | 0.97 | 0.795±0.010 |
| Male      | age 65+                       | 95%   | 0.44  | 0.21 | 0.94 | 0.98 | 0.56 | 0.97 | 0.795±0.010 |
| Female    | age 65+                       | 99%   | 0.49  | 0.19 | 0.94 | 0.99 | 0.54 | 0.98 | 0.822±0.011 |
| Female    | age 65+                       | 99%   | 0.32  | 0.46 | 0.97 | 0.98 | 0.54 | 0.98 | 0.822±0.011 |
| Male      | age < 65                      | 95%   | 0.54  | 0.24 | 0.94 | 0.99 | 0.80 | 0.98 | 0.842±0.011 |
| Female    | age < 65                      | 99%   | 0.45  | 0.54 | 0.98 | 0.99 | 0.73 | 0.99 | 0.859±0.010 |
| Female    | age < 65                      | 95%   | 0.58  | 0.22 | 0.94 | 0.99 | 0.73 | 0.99 | 0.859±0.010 |
| Male      | age < 65                      | 99%   | 0.41  | 0.57 | 0.97 | 0.98 | 0.80 | 0.98 | 0.842±0.011 |
| Male      | high risk                     | 95%   | 0.58  | 0.40 | 0.93 | 0.98 | 0.69 | 0.97 | 0.884±0.000 |
| Male      | high risk                     | 99%   | 0.39  | 0.70 | 0.96 | 0.97 | 0.69 | 0.98 | 0.884±0.000 |
| Female    | high risk                     | 99%   | 0.38  | 0.65 | 0.96 | 0.97 | 0.64 | 0.98 | 0.894±0.000 |
| Female    | high risk                     | 95%   | 0.61  | 0.36 | 0.93 | 0.98 | 0.64 | 0.98 | 0.894±0.000 |
| Female    | low risk                      | 99%   | 0.32  | 0.22 | 0.99 | 0.99 | 0.70 | 0.99 | 0.852±0.000 |
| Female    | low risk                      | 95%   | 0.48  | 0.07 | 0.95 | 1.00 | 0.70 | 0.99 | 0.852±0.000 |
| Male      | low risk                      | 95%   | 0.50  | 0.09 | 0.95 | 0.99 | 0.72 | 0.99 | 0.853±0.000 |
| Male      | low risk                      | 99%   | 0.31  | 0.24 | 0.98 | 0.99 | 0.72 | 0.99 | 0.853±0.000 |
| Female    | no dyspnea                    | 99%   | 0.29  | 0.23 | 0.98 | 0.99 | 0.71 | 0.99 | 0.842±0.000 |
| Female    | no dyspnea                    | 95%   | 0.49  | 0.09 | 0.95 | 0.99 | 0.71 | 0.99 | 0.842±0.000 |
| Male      | no dyspnea                    | 95%   | 0.48  | 0.10 | 0.94 | 0.99 | 0.87 | 0.99 | 0.823±0.000 |
| Male      | no dyspnea                    | 99%   | 0.31  | 0.29 | 0.98 | 0.99 | 0.87 | 0.99 | 0.823±0.000 |

* The cohort of COPD/Cardiac events/Asthma comprises patients with one of these conditions (See SI-Table III), and is designed to investigate ZCoR-IPF effectiveness under such preexisting diagnoses.
† The cohort of “no dyspnea” comprises patients lacking any record of dyspnea and related conditions (See SI-Table II), and investigates ZCoR-IPF effectiveness under situations aggravating missed diagnosis risk.
‡ Maximum PPV and NPV achieved by ZCoR-IPF

**Modeling & Prediction**

The significant diversity of diagnostic codes, along with the sparsity of codes per patient (0.49 to 0.55 per week diagnostic per patient on average) makes this a difficult learning problem. We proceed by partitioning the disease spectrum into 51 broad categories, *e.g.* infectious diseases, immunologic disorders, and endocrinial disorders (See SI Tab. I for a detailed enumeration of these categories). Some of these categories comprises a relatively large number of diagnostic codes aligning roughly with the categories defined within the ICD framework. The remaining categories represent groups of one or more codes that might have some known or suspected association with pulmonary disorders. Each of the diagnostic categories yield a single time series over weeks...
Fig. 2: Predictive performance of ZCoR-IPF for IPF diagnosis 1 year in future. Panels a and b show the out-of-sample ROC and precision-recall curves for diagnosis 1 year from the point of screening. We achieve AUCs > 88% for both sexes, with sensitivity > 55% at 95% specificity. The PPV achieved is ≈ 50% for females and 45% for males at 99.7% specificity for patients who have at least one respiratory diagnosis with past two years of ZCoR-IPF screening. Panels c and d show the top 20 comorbidity categories sorted in the order of inferred importance in estimating risk. Importantly, the comorbidities modulate risk differentially by sex, although the patterns are broadly similar, e.g., respiratory disorders are the most important co-morbid factor, and other factors such as known IPF high co-morbidities (see Table V), cardiovascular issues, infections, hypertensive abnormalities appear in both males and females, with altered ranking.

(each week being identified as having a value '0' for no code corresponding to the diagnostic category, or '1' if some code is present, and '2' if a diagnostic code from any of the other categories is present). We refer to the individual diagnostic categories as a phenotype in the sequel, since they are observable characteristics of the patients. Once we have defined these diagnosis phenotypes, each patient is represented by 51 sparse stochastic...
Fig. 3: Survival function, hazard rate improvements, and predictive performance for early screening. Panel a shows the estimated lower bounds on the survival function at two specificity levels (90 and 95%), and panel b the estimated upper bounds on hazard rates. The lower (upper) bounds are estimated due to the often missing information on actual deaths. Nevertheless, the last diagnostic record in a patient’s history is a lower bound on the survival time, allowing us to calculate the plots shown above. Panel c shows the variation of the mean survival time as a function of the specificity at which ZCoR-IPF is operated. The baselines shown are calculated from our patient database, and is somewhat lower to what is reported in the literature (2-3 years median post-diagnosis). This is expected since our estimate is a lower bound. We have a clear advantage for even high specificities. Panel d illustrates the variation of estimated raw risk as a function of age for screening four years from actual recorded diagnosis of IPF, showing that risk increases almost linearly with age for the patients eventually diagnosed with IPF. Finally, panel e shows the degradation of out-of-sample AUC as we attempt to screen earlier, stepping back from the time of current diagnosis (in absence of ZCoR-IPF screening). Note that starting from \( \approx 88\% \) AUC at 1 year to diagnosis the performance degrades somewhat to \( \approx 85\% \) for screening at 4 years to diagnosis.

time series of events, which are compressed into specialized Hidden Markov Models known as Probabilistic Finite Automata\(^{24}\). These models are inferred separately for each phenotype, for each sex, and for the control and the positive cohorts, to identify the distinctive average patterns emerging at the population level. Thus, we infer \( 51 \times 2 \times 2 = 204 \) PFSA models in total in this study. Our inference algorithm (See Supplementary Text, Section XI)) for these models do not presuppose a fixed structure, and is able to work with non-synchronized and variable length data streams. Variation in the structure and parameters of these inferred models between the positive and control groups delineate the estimated risk of an IPF diagnosis at the population level. Given these models, and the history of a specific patient, we can then quantify the likelihood of this patient’s particular history being generated by the control PFSA models as opposed to the positive models. We refer to this likelihood difference as the sequence likelihood defect (SLD)\(^{25}\), which is the one of the key informative features in our approach. The SLD is a novel concept, involving the generalization of the notion of KL divergence\(^{26}\) between probability distributions to a generalized divergence between possibly non-iid stochastic processes (See Supplementary text, Section V).

In addition to the phenotype specific Markov models, we use a range of engineered features that reflect various aspects of the patient-specific diagnostic histories, referred to as the “sequence features”, “p-scores”, and “rare scores”. The sequence features include the ratio of number of weeks with the codes of a given phenotype to the total number of weeks in sequence, the ratio of number of weeks with the codes of a given phenotype to the
number of weeks with any diagnosis code recorded and the length of the longest uninterrupted subsequence of weeks with the codes of a given phenotype (See Table VII for complete list of such features). The p-scores and the rare scores encode prevalence characteristics of individual diagnostic codes (see Supplementary text Section ??). Ultimately, we compute a total of 667 features for each patient, which is then used to train a network of standard gradient boosting classifiers27 aiming to map individual patients to a raw risk score. We randomly choose 75% of our patients for training with the rest held-out as a validation set. More details of the predictive pipeline are given in the Supplementray text, Section ??.

**Raw Risk & Relative Risk**

Our predictive pipeline produces a continuous estimate of the raw risk score of an IPF diagnosis in future. Thus, our raw risk estimate is a continuous number, and we must choose a decision threshold to make crisp predictions, i.e., if the raw risk is greater than this calibrated threshold then the individual patient is predicted to be in the positive category. In this study, we select this threshold by maximizing the $F_1$-score, defined as the harmonic mean of sensitivity and specificity, to make a balanced trade-off between Type 1 and Type 2 errors. The relative risk is then defined as the ratio of the raw risk to the decision threshold, and a value $> 1$ indicates a predicted future IPF diagnosis.

**Performance Measurement**

We measure our performance using standard metrics including the Area Under the receiver-operating characteristic curve (AUC), sensitivity, specificity and the Positive Predictive Value (PPV). We also report accuracy (acc, See Table VI), which is the probability of a correct prediction (positive or control).

**Feature Importance & Comorbidity Spectra**

Beyond the demonstrated predictive performance (see Results), calculation of the ZCoR-IPF score offers insights into the comorbid associations of IPF that might actually have predictive value. Estimating the relative importance of the features used is crucial for sanity checks, as well as for insights into the underlying causal mechanisms. We compute the relative importance of the features by estimating the mean change in the raw risk via random perturbation of a particular feature: this is the “feature importance” shown in Fig. 2c for the different diagnostic categories. which illustrates that respiratory disorders are the most important diagnostic category modulating the ZCoR-IPF score.

Importantly, all of our features are based on data already available in the past medical records. We do not demand results from specific tests, or look for specific demographic, bio-molecular, physiological and other parameters; we use what we get in the diagnostic history of patients, which presents un-structured sequence of labels pertaining to the ICD and the prescription codes, and is typically prone to noise, coding errors and sparsity. Our ability to effectively work with uncurated data and achieve high out-of-sample predictive performance showcases the immediate clinical applicability with zero additional burden to patients and providers.

In addition to the patient-specific predictions, we compute the statistically significant log-odds ratio of specific ICD codes occurring in the true positive vs the true negative patient sets. We call these the comorbidity spectra (See Fig. 4). Removing the false positives and the false negatives from consideration in computing the comorbidity spectra allows us to uncover patterns — at the level of individual codes — that are most representative of the patient risk. Importantly, the comorbidity spectra are based on individual codes, as opposed to the feature importances shown in Fig. 2c, which consider aggregated impact of all features that are based on the broad disease categories. Every disorder listed in the co-morbid spectra obviously do not all appear in a single patient, but the idea here is that the codes with high log-odds ratio are significantly more likely in positive cohort. The comorbidity spectra, so named because of disease-category specific color coding, offers unique insight into the predictive co-morbidity burden of IPF.

**RESULTS**

In this study we demonstrate three key results: 1) high out-of-sample predictive performance for identifying a future IPF diagnosis via leveraging subtle comorbidity patterns recorded in the past medical history of individual patients. 2) the ability of our models to maintain high predictive performance for an eventual diagnosis further into future, up to 4 years. And 3) ability to perform effectively under common confounders, such as a preexisting diagnosis of COPD, asthma or heart disease, or the absence of any indication of dyspnea in the past.
Our main prediction results are presented in Fig. 2a-b, which illustrate the ROC and the precision-recall curves respectively (for screening one year before current diagnosis), shown separately for males and females. As noted in the legend of these panels, our out-of-sample predictive performance is > 88% AUC irrespective of the sex of the patient, with > 55% sensitivity at 95% specificity (55% for males and 60% for females). At 99% specificity, we obtain a positive predictive value (PPV) of 52% for females and 54% for males, with sensitivities at 42% and 38% respectively. At these values we obtain an accuracy (denoted as “acc”) of ≈ 94 – 98% (See Table VI) which indicates the overall fraction of correct predictions. The PPV achieved by ZCoR-IPF at maximum accuracy is 70% for females and 71% for males, with a corresponding Negative Predictive Value (NPV) of 98%.

Thus, to summarize: our predictive pipeline detects about 55-60 out of every 100 patients who will be diagnosed with IPF in 1 year, if we operate at 95% specificity. If we wish to operate at the higher specificity of 99%, then out of 100 positive flags we have about 50-52 true positives. The accuracy metric indicates that we are correctly identify the risk status (positive or control) for approximately 94-98 out of 100 patients, irrespective of sex, highlighting the potentially high clinical significance of the ZCoR-IPF score.

From the inferred relative importance of the features (See Fig. 2d-e), we conclude, as expected, that respiratory disorders in the past are the most important modulators of risk, followed by known or suspected IPF comorbidities, metabolic diseases, cardiovascular abnormalities and diseases of the eye. Infections also feature in the top 20 co-morbidities shown in these panels. Importantly while there are sex differences, the overall pattern of the relative importance ranking remains substantially invariant between males and females. With some exceptions, many of these patterns are not particularly surprising; the contribution of this study is to bring them together systematically to realize an accurate risk estimate via the ZCoR-IPF score.

A key metric to evaluate the potential impact of the ZCoR-IPF score is to estimate the expected change in the survival functions via a Kaplan-Meyer (KM) analysis, and the corresponding change in the mean survival time. The standard KM analysis is specifically designed to handle scenarios with incomplete observations, e.g., not knowing the exact time of death for all patients, but merely a lower bound on their survival times. This is particularly useful in our case, since we found that insurance claims often do not record deaths, with expired patients simply dropping off the database. This creates uncertainty on if the patient had actually expired, or if he or she simply dropped insurance, or got dropped from the database for some other reason. Nevertheless, the time over which they are observed in the database is clearly a lower bound on their survival. With this mind, we construct the survival plots shown in Fig. 3a-b, which represents lower bounds on the survival function (panel a) and upper bounds on the hazard rate (panel b), shown along with 95% confidence bounds. Note that since we can operate our predictors at different specificity-sensitivity trade-offs, we get different curves if we vary the specificity. The survival functions are notably similar across the two sexes. Panel c shows the variation of lower bound on the mean survival times, compared with the baseline currently observed in the database; even at 95% specificity we boost the lower bound on mean survival time from around 100 weeks to approach 180-200 weeks.

It is crucial to note that these survival analyses do not take into account the possibility of actually prolonging life via clinical interventions when we push back the time of the diagnosis; thus in actuality we expect the survival times to be markedly better to what is shown in Fig. 3a-c.

Our predictive performance expectedly degrades as we predict earlier, and this is illustrated in Fig. 3e. Importantly however, the degradation is slow enough that we can use ZCoR-IPF with acceptable reliability to predict diagnoses four years into the future. To illustrate how the ZCoR-IPF risk varies over patient age, we estimate the distribution of the scores over the positive and the control cohorts in Fig. 3d. Note that for patients fore who get eventually diagnosed, the risk almost linearly increases with age.

While these results demonstrate the importance of the diverse features used in our approach, understanding the seat of this predictive power is important. The feature importances discussed earlier (Fig. 2d-e) identify the relative impact of broad disease categories. Importantly, to evaluate the feature importance of a specific diagnostic category, we sum the importance of all features based on that category, not just the presence or absence of individual diagnoses. The latter aspect is investigated via the co-morbidity spectra for out-of-sample patients, shown in Figs. 4 separately for males and females. We find that the important co-morbidities are diverse, vary with the sex of the patients, but is clearly dominated by respiratory disorders, followed by diseases of the cardiovascular and circulatory systems. Again, while many of these patterns are expected at the population level, design of the personalized ZCoR-IPF score is not immediately obvious.

Since IPF co-morbidities have been investigated in the literature, a relevant question here is if our performance is dramatically better in sub-cohorts defined by the presence of these high risk diagnoses in the past (defined in Table V). The results are tabulated in Table VI, showing that our performance in the high risk sub-cohort is more or less comparable with full cohort performance. The AUCs achieved for the low risk cohort is somewhat lower (> 85% for males and females), but still acceptably high. Thus, even within the low risk patients, we still
detect 48 out of every 100 patients who are going to have a diagnosis in 1 year.

DISCUSSION

The present study describes the development and performance of ZCoR-IPF in a large US commercial claims database. Our key finding is that in both men and women ages 48-90 years (n=1345, n=1191, respectively), ZCoR-IPF accurately identifies IPF cases 1-4 years sooner than occurred in a variety of everyday community or academic practice settings during 2003-2018: at 208 weeks (4 years) before IPF diagnosis, ZCoR-IPF would have predicted that classification with an AUC approaching 84%, at 156 weeks (3 years) before diagnosis, with an AUC of $\approx 86\%$, at 104 weeks (2 years) with $\approx 87\%$, and at 52 weeks (1 year) before diagnosis, with an AUC approaching 88%, regardless of the patient's sex (Fig. 3e). Importantly, ZCoR-IPF achieves such results non-invasively, inexpensively, and almost instantaneously, since it relies only on diagnostic data already in the patient's electronic medical record, and runs on existing information technology infrastructure. Also of note, ZCoR-IPF expands the data available to both primary care practitioners and expert IPF diagnosticians. The score reflects a sophisticated, highly-detailed automated analysis of comorbidities, considering more than 600 features related to the incidence, timing of individual diagnostic codes. ZCoR-IPF thus supplements information currently used to evaluate and diagnose IPF, which focuses mainly on respiratory signs and symptoms, pulmonary function, and the radiographic and histologic appearance of the lung. Adding comorbidity as a new dimension of assessment, albeit in much simpler fashion, is an approach that has proved fruitful in COPD management.

ZCoR-IPF may be expected to contribute in two main ways to increase timeliness and ease of IPF identification. First, ZCoR-IPF might serve as a screening tool to aid primary care physicians, radiologists, or pulmonologists to more selectively flag patients for referral for HRCT, referral to a pulmonologist, or both. Presently, the leading candidates for such flagging would be patients with one or more of chronic dyspnea, chronic cough, and/or chronic “Velcro crackles” on auscultation, restrictive ventilatory patterns on pulmonary function tests, or incidental ILA or ILD on chest or abdominal CT. As these are relatively large groups, ZCoR-IPF might be applied to help distinguish individuals who require immediate referral versus those who require increased surveillance, versus those who require less frequent follow-up. ZCoR-IPF might be an especially useful tool in patients with ILA, since while these findings might reflect an early stage of the disease, only some 0.5%–2% of this group will ever develop IPF. Indeed, given the severity and actionability of the IPF diagnosis, the frequency of smoking history among patients with IPF, and the non-invasive nature, applicability, speed, and low cost of ZCoR-IPF, this tool might be applied in primary care to screen smokers or former smokers for increased surveillance or evaluation for IPF. The feasibility, diagnostic yield, cost-effectiveness, and ultimately, effect on patient outcomes of using ZCoR-IPF in these screening settings merit prospective evaluation.

Second, ZCoR-IPF might serve as a diagnostic aid for pulmonologists, radiologists, pathologists, or multidisciplinary teams in cases showing abnormalities suggestive of, or associated with IPF, but not UIP on HRCT or histopathology. These cases are relatively frequent: roughly half of patients histopathologically diagnosed with IPF lack classic CT findings associated with the disease. Hence ZCoR-IPF might help patients without UIP on HRCT to avoid more invasive tests, especially surgical lung biopsy, and/or may increase clinicians’ diagnostic confidence. The effect of ZCoR-IPF use as a diagnostic aid on healthcare resource utilization and on diagnostic confidence in IPF classification also warrants prospective assessment.

ZCoR-IPF also might speed recruitment and decrease costs of clinical trials of new therapies for IPF and other progressive fibrosing ILD. The tool might do so by allowing more confident inclusion in study samples of patients with provisional IPF diagnoses.

Improving IPF screening and diagnosis may offer considerable opportunity to substantially enhance patient outcomes and quality-of-life, as well as efficiency of healthcare resource utilization. First and foremost, earlier, more efficient and confident diagnosis should enable timelier access to anti-fibrotic treatment, which slows but does not reverse disease. Evidence of similar effect sizes across a spectrum of IPF severity suggests that starting anti-fibrotics earlier in the course of disease may better preserve lung function. Notably, in patients with IPF, each year’s delay following an initial CT scan in referral to a specialized ILD center has been shown to be associated with a 2.94% (95% CI: 0.86–5.03%) increase in pulmonary fibrosis extent on chest CT.

Conversely, more prompt and firm IPF diagnosis should spare patients unneeded or harmful treatments, e.g., corticosteroids, as well as unneeded diagnostic tests and healthcare visits. Earlier identification also should allow more prompt referral for lung transplantation, the only current cure for the disease. Such referral is recommended immediately upon IPF diagnosis, since evaluation for eligibility and waits for graft availability may take months or years; starting this process when younger and less ill may allow patients to avoid disqualification for advanced age or frailty. In the meanwhile, quicker IPF diagnosis will accelerate patients’ access to interven-
ations that may improve lung function and quality-of-life, namely supplemental oxygen, pulmonary rehabilitation, and palliative care, as well as to clinical trials.

Development and use of a risk score for IPF diagnosis that is based on comorbidities aligns well with the longstanding appreciation of multimorbidity as characteristic of patients with IPF. Indeed, a chart review at a German tertiary referral center found 58% of patients with IPF to have 1–3 comorbidities, 30% to have 4–7 comorbidities, and only 12% to have no concomitant illness. The high prevalence of comorbidities in patients with IPF presumably stems from shared risk and pathogenetic factors among IPF and the other conditions, e.g., male gender, cigarette smoking, or aging accelerated by genetic mutations resulting in telomere shortening. Additionally, some comorbid conditions may be hypothesized to be induced by IPF symptoms, e.g., depression, some to contribute to IPF pathogenesis, e.g., gastroesophageal reflux (GER), while others may at least in some instances represent common mis-diagnoses in patients with IPF, e.g., COPD, asthma, or pneumonia.

Strengths and limitations of our work here should be noted. Foremost among the former is our use of detailed “real-world” data from large databases (n=2536 cases, n=450,461 controls) collected over more than a decade throughout the United States or at an academic tertiary referral center in a major US metropolitan area. The huge overall sample size gives ZCoR-IPF robust statistical power to capture subtle patterns of comorbidity that would be indiscernible even in traditional multicenter cohorts. As well, the size and scope of the databases increase likelihood that our findings are generalizable throughout the United States, and, indeed, elsewhere.

Conversely, use of administrative claims databases has certain disadvantages. Absent patient-level case validation, the most important of these is potential lack of accuracy in the ICD-9 and ICD-10 coding for IPF and related diagnoses. A relatively small (n=150 validated cases) patient-level case validation study of another, smaller administrative claims database (Kaiser Permanente) suggested that more than half of IPF codes might be incorrectly applied, including roughly a quarter of cases that appeared to lack even findings of ILD. Arguably, use of data including both patients with IPF and patients with other ILD might minimally affect clinical relevance of our score, given the emerging concept of “progressive chronic fibrosing ILD”. Under this concept, subgroups of patients with IPF (which progresses at heterogenous rates from case-to-case) and subgroups of patients with other ILD, e.g., idiopathic non-specific interstitial pneumonia, hypersensitivity pneumonitis, systemic sclerosis-associated ILD, or rheumatoid arthritis-associated ILD, may have similar phenotypes and disease behavior, and should perhaps be managed similarly, with early initiation of anti-fibrotic therapy. However, inclusion of patients without ILD among our IPF cases would of course be expected to decrease the accuracy and utility of ZCoR-IPF. It would be desirable to perform case validation studies in our databases, which might provide data enabling ZCoR-IPF to be fine-tuned.

Other disadvantages of our use of a commercial insurance administrative database may include excess proportions, relative to those seen in specialist clinical practice, of females, and of less severe cases of both sexes. In other claims database studies, these differences respectively have been attributed to a greater tendency of women than men to seek medical attention, and to a tendency of older and sicker patients to migrate from commercial to government insurance plans. Interestingly, one may speculate that miscoding of soft tissue or rheumatologic diseases with an ILD component as IPF may be more prevalent among females, as are those disorders themselves. Additionally, commercial claims databases contain at best highly limited information on signs and symptoms of disease that are insufficiently severe or impactful to be considered a discrete diagnosis, e.g., dyspnea, cough, or rales. These databases also provide a paucity of data regarding the results of biochemical or imaging investigations. Nor do claims data capture often important lifestyle variables such as smoking, alcohol consumption, exercise, or diet.

Conclusion

We developed and validated ZCoR-IPF, a risk score for IPF capable of identifying with high accuracy this diagnosis 1–4 years before it occurred in everyday practice conditions in community and academic settings across the United States. This score provides clinicians with detailed data on comorbidity patterns to distinguish patients with IPF from those without the disease, a strategy of particular interest given the well-acknowledged predominance of multimorbidity in patients with IPF. Thus ZCoR-IPF supplements the functional and morphological data regarding the lung that have been the mainstay in evaluating potential IPF cases. ZCoR-IPF has the advantages of considering only diagnostic data already in the patient’s electronic medical record, of running on current information technology infrastructure, and of operating inexpensively and almost instantaneously at the point of care. Hence ZCoR-IPF holds promise as a screening tool for primary care practitioners, pulmonologists, and radiologists to use in patients with early signs and/or symptoms of IPF or those at elevated risk for that disease. As well, ZCoR-IPF may offer utility as a diagnostic aid to pulmonologists, radiologists, pathologists,
and multidisciplinary teams in cases with radiological and/or histopathological suggestions, but not the classical radiological and/or histopathological picture of IPF. The effect of ZCoR-IPF on speed, accuracy, and confidence of IPF diagnosis, on health care resource utilization, and ultimately, on patient outcomes, should be assessed prospectively.

REFERENCES

[1] Lederer, D. & Martinez, F. Idiopathic pulmonary fibrosis. *N Engl J Med* 378, 1811–23.
[2] Raghu, G., Remy-Jardin, M. & Myers, J. Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline. *Am J Respir Crit Care Med* 198, 44– 68.
[3] Raghu, G. Idiopathic pulmonary fibrosis: shifting the concept to irreversible pulmonary fibrosis of many entities. *Lancet Respir Med* 7, 926–9.
[4] Ley, B., Collard, H. & King, T., Jr. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 183, 431–40.
[5] Antoniou, K., Symvoulakis, E., Margaritopoulos, G., Lionis, C. & Wells, A. Early diagnosis of ipf: time for a primary-care case-finding initiative? *Lancet Respir Med* 2, 1.
[6] Cosgrove, G. P., Bianchi, P., Danese, S. & Lederer, D. J. Barriers to timely diagnosis of interstitial lung disease in the real world: the intensity survey. *BMC pulmonary medicine* 18, 9 (2018).
[7] Hewson, T. et al. Timing of onset of symptoms in people with idiopathic pulmonary fibrosis. *Thorax*.
[8] Adegunsoye, A. Diagnostic delay in idiopathic pulmonary fibrosis: Where the rubber meets the road. *Ann Am Thorac Soc* 16, 310–2.
[9] Cottin, V. & Richeldi, L. Neglected evidence in idiopathic pulmonary fibrosis and the importance of early diagnosis and treatment. *Eur Respir Rev* 23, 106–10.
[10] Putman, R., Rosas, I. & Hunninghake, G. Genetics and early detection in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 189, 770–8.
[11] Lamas, D. et al. Delayed access and survival in idiopathic pulmonary fibrosis: a cohort study. *Am J Respir Crit Care Med* 184, 842–7.
[12] Hoyer, N., Prior, T., Bendstrup, E., Wilcke, T. & Shaker, S. Risk factors for diagnostic delay in idiopathic pulmonary fibrosis. *Respir Res* 20, 103.
[13] Mooney, J., Chang, E. & Lalla, D. Potential delays in diagnosis of idiopathic pulmonary fibrosis in medicare beneficiaries. *Ann Am Thorac Soc* 16, 393–6.
[14] Pritchard, D., Adegunsoye, A. & Lafond, E. Diagnostic test interpretation and referral delay in patients with interstitial lung disease. *Respir Res* 20, 253.
[15] Schoenheit, G., Becattelli, I. & Cohen, A. Living with idiopathic pulmonary fibrosis: an in-depth qualitative survey of european patients. *Chron Respir Dis* 8, 225–31.
[16] Collard, H., Tino, G. & Noble, P. Patient experiences with pulmonary fibrosis. *Respir Med* 101, 1350–4.
[17] Thickett, D., Voorham, J. & Ryan, R. Historical database cohort study addressing the clinical patterns prior to idiopathic pulmonary fibrosis (ipf) diagnosis in uk primary care. *BMJ Open* 10, 034428.
[18] Cottin, V. & Cordier, J. Velcro crackles: the key for early diagnosis of idiopathic pulmonary fibrosis? *Eur Respir J* 40, 519–21.
[19] Hart, S. Machine learning molecular classification in ipf: Uip or not uip, that is the question. *Lancet Respir Med* 7, 466–7.
[20] Oldham, J. & Noth, I. Idiopathic pulmonary fibrosis: early detection and referral. *Respir Med* 108, 819–29.
[21] Hansen, L. The truven health marketscan databases for life sciences researchers. *Truven Health Analytics IBM Watson Health* (2017).
[22] Raghu, G., Amatto, V., Behr, J. & Stowasser, S. Comorbidities in idiopathic pulmonary fibrosis patients: a systematic literature review. *Eur Respir J* 46, 1113–30.
[23] Organization, W. H. *et al.* International classification of diseases—ninth revision (icd-9). *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire* 63, 343–344 (1988).
[24] Chalttopadyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* 371, 20110543 (2013).
[25] Huang, Y. & Chalttopadyay, I. Data smashing 2.0: Sequence likelihood (sl) divergence for fast time series comparison. arXiv preprint arXiv:1909.12243 (2019).
[26] Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing) (Wiley-Interscience, New York, NY, USA, 2006).
[27] Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, 3146–3154 (2017).
[28] Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American
statistical association 53, 457–481 (1958).

[29] Hyldgaard, C., Hilberg, O. & Bendstrup, E. How does comorbidity influence survival in idiopathic pulmonary fibrosis? *Respir Med* **108**, 647–53.

[30] Divo, M., Cote, C. & Torres, J. Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* **186**, 155–61.

[31] JP, T., C, C. & JM, M. Prognostic evaluation of copd patients: Gold 2011 versus bode and the copd comorbidity index cote. *Thorax* **69**, 799–804.

[32] Oldham, J., Adegunsoye, A. & Khera, S. Underreporting of interstitial lung abnormalities on lung cancer screening computed tomography. *Ann Am Thorac Soc* **15**, 764–6.

[33] Walsh, S., Humphries, S., Wells, A. & Brown, K. Imaging research in fibrotic lung disease; applying deep learning to unsolved problems. *Lancet Respir Med* **8**, 1144–53.

[34] Raghu, G., Flaherty, K. & Lederer, D. Use of a molecular classifier to identify usual interstitial pneumonia in conventional transbronchial lung biopsy samples: a prospective validation study. *Lancet Respir Med* **7**, 487–96.

[35] Kropski, J. Biomarkers and early treatment of idiopathic pulmonary fibrosis. *Lancet Respir Med* **7**, 725–7.

[36] Farrand, E., Iribarren, C. & Vittinghoff, E. Impact of idiopathic pulmonary fibrosis on longitudinal health-care utilization in a community-based cohort of patients. *Chest*.

[37] Kreuter, M., Ehlers-Tenenbaum, S. & Palmowski, K. Impact of comorbidities on mortality in patients with idiopathic pulmonary fibrosis. *PLoS One* **11**, 0151425.

[38] Esposito, D., Lanes, S. & Donneyong, M. Idiopathic pulmonary fibrosis in united states automated claims. incidence, prevalence, and algorithm validation. *Am J Respir Crit Care Med* **192**, 1200–7.

[39] Ley, B., Urbania, T. & Husson, G. Code-based diagnostic algorithms for idiopathic pulmonary fibrosis. *Case Validation and Improvement. Ann Am Thorac Soc* **14**, 880–7.

[40] Inoue, Y., Kaner, R. & Guiot, J. Diagnostic and prognostic biomarkers for chronic fibrosing interstitial lung diseases with a progressive phenotype. *Chest* **158**, 646–59.

[41] George, P., Spagnolo, P. & Kreuter, M. Progressive fibrosing interstitial lung disease: clinical uncertainties, consensus recommendations, and research priorities. *Lancet Respir Med* **8**, 925–34.

[42] Mortimer, K. *et al.* Characterizing idiopathic pulmonary fibrosis patients using us medicare-advantage health plan claims data. *BMC Pulm Med* **19**, 11.

[43] Mortimer, K., Bartels, D. & Hartmann, N. Characterizing health outcomes in idiopathic pulmonary fibrosis using us health claims data. *Respiration* **99**, 108–18.
Fig. 4: Co-morbidity Spectrum (males). Disorders that increase the odds of the patient being a “true positive” vs a “true negative”. Such disorders (ranked according to the log-odds ratio) are more likely to be found in patients who are in the positive cohort. Comparing panel a with panel b, we note that these odds change from males to females, but as expected the patterns are broadly similar, with over-representation of respiratory disorders.
| Feature-Phenotype | Explanation | n Features |
|-------------------|-------------|------------|
| aggregate score   | Mean p-score of feature-phenotype codes within sequence divided by mean p-score of all codes in the record | 9 |
| high scores proportion | proportion of codes with very high p-scores among all codes in the record | 1 |
| low scores proportion | proportion of codes with very low p-scores among all codes in the record | 1 |
| standard deviation of p-scores of second half of the record divided by standard deviation of p-scores of first half of the record | 1 |
| range of p-scores of second half of the record divided by range of p-scores of first half of the record | 1 |
| skew of p-scores of second half of the record divided by skew of p-scores of first half of the record | 1 |
| aggregation relative to p-score of feature-phenotype | aggregation of all feature-phenotype's mean scores divided by mean p-score of all codes in the record | 6 |
| aggregation relative to p-score of feature-phenotype | aggregation of all feature-phenotype's mean scores divided by corresponding general p-score of feature-phenotype | 6 |
| Max/Min/Std/Range of the lengths of subsequences of consequent weeks with any diagnosis code recorded | Time interval from observation date to the first phenotype code, normalized by record length | 51 |
| Max/Min/Std/Range of the lengths of subsequences of consequent weeks with codes | Time interval from observation date to the first phenotype code, normalized by record length | 51 |
| Max/Min/Std/Range of the lengths of subsequences of consequent weeks without codes | Time interval from observation date to the first phenotype code, normalized by record length | 51 |
| Proportion of weeks in a record observed where at least one DX code was recorded | Proportion of weeks in a record observed where at least one DX code was recorded | 1 |
| Sequence Likelihood Defect for a given phenotype | Sequence Likelihood Defect for a given phenotype | 51 |
| Negative LogLikelihood score for a given phenotype | Negative LogLikelihood score for a given phenotype | 51 |
| Ratio of Positive to Negative LogLikelihood score for a given phenotype | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Mean Sequence Likelihood Defect | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Mean Negative LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Mean Range of LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Standard Deviation of Negative LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Standard Deviation of Negative LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Range of Negative LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Range of Negative LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Range of Positive LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Range of Positive LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Standard Deviation of Positive LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Standard Deviation of Positive LogLikelihood score | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Range of LogLikelihood score ratio | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| Standard Deviation of LogLikelihood score ratio | Time interval from observation date to the last phenotype code, normalized by record length | 51 |
| predicted risk from pfsa model | predicted risk from pfsa model | 1 |
| predicted risk from seq model | predicted risk from seq model | 1 |
| predicted risk from pscore model | predicted risk from pscore model | 1 |
| predicted risk from rare model | predicted risk from rare model | 1 |
| patient age at the moment of the screening | Patient age at the moment of the screening | 1 |

*feature: Corresponds to the ICD disease categories, or sets of diagnostic codes tracked, or medications tracked either as individual active ingredients or as sets e.g. antidepressants

$\Delta$: Sequence Likelihood Defect (See SI text Sec V)

neg loglikelihood: loglikelihood of observed sequence being generated by the model inferred from control (See Methods)

pos loglikelihood: loglikelihood of observed sequence being generated by the model inferred from positive (See Methods)
Supplementary Text: Screening for Idiopathic Pulmonary Fibrosis with Comorbid Pattern Recognition in Electronic Health Records

Dmytro Onishchenko¹, Robert J. Marlowe², Gary M. Hunninghake³, Fernando J. Martinez⁴,⁵ and Ishanu Chattopadhyay¹,⁶,⁷⋆

¹Department of Medicine, University of Chicago, Chicago, IL USA
²Spencer-Fontayne Corporation, Jersey City, NJ, USA
³Director, Interstitial Lung Disease Program, Brigham and Women's Hospital, Harvard Medical School, Boston, MA USA
⁴Bruce Webster Professor of Internal Medicine, Medicine, Weill Cornell Medical College, New York, NY USA
⁵Chief of Division of Pulmonary and Critical Care Medicine at Weill Cornell Medicine and NewYork-Presbyterian Weill Cornell Medical Center, New York, NY USA

⋆To whom correspondence should be addressed: e-mail: ishanu@u.chicago.edu.

CONTENTS

I Detailed Mathematical Approach 3
II Time-series Modeling of Diagnostic History 3
III Inference & Event Periods 3
IV Step 1: Partitioning The Human Disease Spectrum 3
V Step 2: Model Inference & The Sequence Likelihood Defect △ 3
VI Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules 4
VII Threshold Selection on ROC Curve 4
VIII Note on Receiver Operating Characteristics (ROC) and Precision-recall Curves 5
IX Effect of Class Imbalance 5
X Generating PFSA Models From Set of Input Streams with Variable Input Lengths 6
XI Probabilistic Finite State Automata Inference 6
XII Probabilistic Finite-State Automaton 6
XIII Theoretical Development of Sequence Likelihood Defect 8
XIV Pipeline Optimization: Hyper-training, Training, & Validation 10
LIST OF TABLES

I  Disease Categories With Detailed Set of ICD Codes Used .................................................. 12
II Diagnostic codes for dyspnea for investigating ZCoR-IPF effectiveness in identifying IPF without such indications* ................................................................. 24
III Diagnostic codes for COPD, heart disease and asthma for investigating ZCoR-IPF effectiveness under other pre-existing diagnoses. The COPD/Heart disease/Asthma cohort comprises patients with one or more codes from this list ......................... 25

LIST OF FIGURES

1 Schematic of prediction pipeline. Panels A, B and C show the sequential training steps, namely the hyper-training of the PFSA models and the p-score dictionary, the pre-aggregation training of the four LGBM models, and the aggregation training of the final LGBM model respectively. Panel D shows the configuration of the trained pipeline in operation. In the training steps, the filled box represents the component being trained in that step ................................................................. 11

LIST OF ALGORITHMS

1 GenESeSS .............................................................. 8
2 Log-likelihood .................................................... 9
I. Detailed Mathematical Approach

II. Time-series Modeling of Diagnostic History

Individual diagnostic histories can have long-term memory\(^1\), implying that the order, frequency, and comorbid interactions between diseases are important for assessing the future risk of our target phenotype. We analyze patient-specific diagnostic code sequences by first representing the medical history of each patient as a set of stochastic categorical time-series — one each for a specific group of related disorders — followed by the inference of stochastic models for these individual data streams. These inferred generators are a special class of Hidden Markov Models (HMMs), referred to as Probabilistic Finite State Automata (PFSA)\(^2\). The inference algorithm we use is distinct from classical HMM learning, and has important advantages related to its ability to infer structure, and its sample complexity (See Supplementary text, Section XI). We infer a separate class of models for the positive and control cohorts, and then the problem reduces to determining the probability that the short diagnostic history from a new patient arises from the positive as opposed to the control category of the inferred models.

III. Inference & Event Periods

We train our predictive pipeline with all diagnostic codes that are recorded in the past 2 years from the point at which a prediction is made. This period from which we use data to train our pipeline is called the “inference window”. Our aim is to make predictions on the occurrence of the target diagnostic codes at 1 year from the end of the inference window. For patients in the control cohort, we make sure that no target code appears for 2 years after the end of the inference window. Additionally, when making predictions further into the future (upto 4 years, as described in the main text), we always make sure that the control group has no target codes for 1 year after the predicted time of diagnosis, i.e., if we are making a prediction of a diagnosis 4 years in future, then control group patients are chosen to have no diagnosis in at least next 5 years.

IV. Step 1: Partitioning The Human Disease Spectrum

We begin by partitioning the human disease spectrum into 51 non-overlapping categories. Each category is defined by a set of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD9) (See Table SI-I for description of the categories used in this study). For this study, we ended up using 17008378 and 25074255 diagnostic codes for males and females respectively (22685 and 23722 unique codes) spanning both ICD9 and ICD10 protocols (using ICD10 General Equivalence Mappings (GEMS)\(^3\) equivalents where necessary), from a total 2,059,559 patients. Transforming the diagnostic histories to report only the broad categories reduces the number of distinct codes that the pipeline needs to handle, thus improving statistical power.

Our categories largely align with the top-level ICD9 categories, with small adjustments, e.g. bringing all infections under one category irrespective of the pathogen or the target organ. We do not pre-select the phenotypes; we want our algorithm to seek out the important patterns without any manual curation of the input data.

For each patient, the past medical history is a sequence \((t_1, x_1), \ldots, (t_m, x_m)\), where \(t_i\) are timestamps and \(x_i\) are ICD9 codes diagnosed at time \(t_i\). We map individual patient history to a three-alphabet categorical time series \(z^k\) corresponding to the disease category \(k\), as follows. For each week \(i\), we have:

\[
z^k_i = \begin{cases} 
  0 & \text{if no diagnosis codes in week } i \\
  1 & \text{if there exists a diagnosis of category } k \text{ in week } i \\
  2 & \text{otherwise}
\end{cases}
\]

The time-series \(z^k\) is observed in the inference period. Thus, each patient is represented by 43 mapped trinary series.

V. Step 2: Model Inference & The Sequence Likelihood Defect \(\Delta\)

The mapped series, disease-category, and IPF diagnosis-status are considered to be independent sample paths, and we want to explicitly model these systems as specialized HMMs (PFSA). We model the positive and the control cohorts and each disease category separately, ending up with a total of 86 HMMs at the population level (43 categories, 2 IPF status categories: positive and control). Each of these inferred models is a PFSA; a directed graph with probability-weighted edges, and acts as an optimal generator of the stochastic process driving the sequential appearance of the three letters (as defined by Eq. (1)) corresponding to disease category, and IPF status-type (See Section XI in the Supplementary text for background on PFSA inference).

To reliably infer the IPF status-type of a new patient, i.e., the likelihood of a diagnostic sequence being generated by the corresponding IPF status-type model, we generalize the notion of Kullbeck-Leibler (KL) divergence\(^4,5\).
between probability distributions to a divergence $D_{KL}(G||H)$ between ergodic stationary categorical stochastic processes $G, H$ as:

$$D_{KL}(G||H) = \lim_{n \to \infty} \frac{1}{n} \sum_{x=|x|-n} p_G(x) \log \frac{p_G(x)}{p_H(x)}$$  \hspace{1cm} (2)

where $|x|$ is the sequence length, and $p_G(x), p_H(x)$ are the probabilities of sequence $x$ being generated by the processes $G, H$ respectively. Defining the log-likelihood of $x$ being generated by a process $G$ as:

$$L(x, G) = -\frac{1}{|x|} \log p_G(x)$$  \hspace{1cm} (3)

The cohort-type for an observed sequence $x$ — which is actually generated by the hidden process $G$ — can be formally inferred from observations based on the following provable relationships (See Supplementary Text Section XI, Theorem 6 and 7):

$$\lim_{|x| \to \infty} L(x, G) = \mathcal{H}(G)$$  \hspace{1cm} (4a)

$$\lim_{|x| \to \infty} L(x, H) = \mathcal{H}(G) + D_{KL}(G||H)$$  \hspace{1cm} (4b)

where $\mathcal{H}(. )$ is the entropy rate of a process $^4$. Importantly, Eq. (4) shows that the computed likelihood has an additional non-negative contribution from the divergence term when we choose the incorrect generative process. Thus, if a patient is eventually going to be diagnosed with IPF, then we expect that the disease-specific mapped series corresponding to her diagnostic history be modeled by the PFSA in the positive cohort. Denoting the PFSA corresponding to disease category $j$ for positive and control cohorts as $G^j_+, G^j_0$ respectively, we can compute the sequence likelihood defect (SLD, $\Delta^j$) as:

$$\Delta^j \triangleq L(G^j_+, x) - L(G^j_0, x) \to D_{KL}(G^j_0||G^j_+)$$  \hspace{1cm} (5)

With the inferred PFSA models and the individual diagnostic history, we estimate the SLD measure on the right-hand side of Eqn. (5). The higher this likelihood defect, the higher the similarity of diagnosis history to that of women with IPF.

VI. Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules

The risk estimation pipeline operates on patient specific information limited to the available diagnostic history in the inference period, and produces an estimate of the relative risk of IPF, with an associated confidence value. To learn the parameters and associated model structures of this pipeline, we transform the patient specific data to a set of engineered features, and the feature vectors realized on the positive and control sets are used to train a gradient-boosting classifier $^7$. The complete list of 667 features used is provided in Tab. VII in the main text.

We need two training sets: one to infer the models, and one to train the classifier with features derived from the inferred models. Thus, we do a random 3-way split of the set of unique patients into feature-engineering (25%), training (25%) and test (50%) sets. We use the feature-engineering set of ids first to infer our PFSA models (unsupervised model inference in each category), which then allows us to train the gradient-boosting classifier using the training set and PFSA models (classical supervised learning), and we finally execute out-of-sample validation on the test set. Fig. 2c in the main text shows the top 20 features ranked in order of their relative importance (relative loss in performance when dropped out of the analysis).

VII. THRESHOLD SELECTION ON ROC CURVE

Once the ROC curve has been computed, we must choose a decision threshold to trade-off true positive rate and false positive rate. In situations where the number of negatives vastly outnumber the number of positives (which is the case in our problem), it is better to base this trade-off on a measure that is independent of the number of true negatives. The two popular measures considered in the literature are accuracy and the F1-score:

$$\text{accuracy} = \frac{t_p + t_n}{t_p + f_p + f_n + t_n}$$  \hspace{1cm} (6)

$$\text{F1} = \frac{2t_p}{2t_p + f_p + f_n}$$  \hspace{1cm} (7)

The F1-score is the same as accuracy where the number of true negatives is the same as the number of true positives, thus partially correcting for the class imbalance.

The selection of the threshold may also be dictated by the current practice of ensuring high specificities in screening tests. Thus, the most relevant clinically operating point is either the one corresponding to 95%
specificity, which is highlighted in Fig. 2a in the main text, or even higher specificities due to the low prevalence of IPF (See Table VI in the main text).

VIII. NOTE ON RECEIVER OPERATING CHARACTERISTICS (ROC) AND PRECISION-RECALL CURVES

The ROC curve is a plot between the False Positive rate (FPR) and the True Positive Rate (TPR), and the area under the ROC curve (AUC) is often used as a measure of classifier performance. For the same of completeness, we introduce the relevant definitions:

In the following $P$ denotes the total number of positive samples (number of patients who are eventually diagnosed), and $N$ denotes the total number of negative samples (number of patients in the control group).

**Definition 1.** True positive rate, true negative rate, false positive rate, positive predictive value (PPV), and prevalence ($\rho$) are defined as:

$$TPR = \frac{t_p}{P} = \frac{t_p}{t_p + f_n} \quad (8)$$
$$TNR = \frac{t_n}{N} = \frac{t_n}{t_n + f_p} \quad (9)$$
$$FPR = 1 - TNR \quad (10)$$
$$PPV = \frac{t_p}{t_p + f_p} \quad (11)$$
$$\rho = \frac{P}{N + P} \quad (12)$$

where as before $t_p, t_n, f_p, f_n$ are true positives, true negatives, false positives, and false negatives respectively.

Note that TPR is also referred to as recall or sensitivity, and PPV is also referred to as precision. True negative rate is also known as specificity.

A precision-recall curve, or a PPV-sensitivity curve is a plot between PPV and TPR.

Denoting sensitivity by $s$, and specificity by $c$, it follows that:

$$PPV = \frac{t_p}{t_p / P + (f_p / N)(N / P)} = \frac{TPR}{TPR + (N - t_n)(N / P)} \quad (13)$$
$$\Rightarrow PPV = \frac{s}{s + (1 - c)(\frac{1}{\rho} - 1)} \quad (14)$$

Thus, we note that for a fixed specificity and sensitivity, the PPV depends on prevalence. Indeed, it is clear from the above argument that PPV decreases with decreasing prevalence, and vice versa.

IX. EFFECT OF CLASS IMBALANCE

ROC curves are generally assumed to be robust to class imbalance. Note that if we assume that patient outcomes are independent (which is well-justified in the case of a non-communicable condition, particularly in large databases), then $t_p$ should scale linearly with the total number of positives $P$, implying:

$$TPR = \frac{t_p}{P} = \frac{t'_p}{P'} \quad (15)$$

implying that with different sizes of the set of positive samples (or negative samples), the ROC curve remains unchanged. In particular, note that even if the prevalence is very small (say 0.01%), we cannot cheat to boost the AUC by labeling all predictions as negative, or stating that risk is always zero: in that case, our $P$ is very small, but our $t_p = 0$ strictly, implying that our TPR $= 0$, thus leading to a zero AUC. We can cheat to boost the accuracy (See the previous section), but not the AUC.

Note that while relative class sizes or imbalance does not affect the ROC (under the assumption that true positives and true negatives scale with the number of positives and negatives), very small absolute sample sizes might still result in poor performance of the model.

The precision-recall curves do get affected by class imbalance, or the prevalence, as shown by Eq (14). However, in diagnostic analysis, they are important since we are generally less interested in the number of true negatives; the ratio of false positives to the total number of positive recommendations by the algorithm is much more relevant, i.e., the PPV or the precision.
X. Generating PFSA Models From Set of Input Streams with Variable Input Lengths

Our PFSA reconstruction algorithm\(^2\) is distinct from standard HMM learning. We do not need to pre-specify structures, or the number of states in the algorithm, and all model parameters are inferred directly from data. Additionally, we can operate either with 1) a single input stream, or 2) a set of input streams of possibly varying lengths which are assumed to be different and independent sample paths from the unknown stochastic generator we are trying to infer. At an intuitive level, we use the input data to infer the length of histories one must remember to estimate the current state, and predict futures for the process being modeled. Thus, we do not step through the symbol streams with a pre-specified model structure, and avoid the need to have equal-length inputs. More details of the algorithm are provided in the next section.

The ability to model a set of input streams of varying lengths is particularly important, since medical histories of different patients are typically of different lengths.

XI. Probabilistic Finite State Automata Inference

XII. Probabilistic Finite-State Automaton

Let \( \Sigma \) be a finite alphabet of symbols with size \(|\Sigma|\). The set of sequences of length \( d \) over \( \Sigma \) is denoted by \( \Sigma^d \). The set of finite but unbounded sequences over \( \Sigma \) is denoted by \( \Sigma^* \), the Kleene star operation, i.e., \( \Sigma^* = \bigcup_{d=0}^{\infty} \Sigma^d \).

We use lower case Greek, for example \( \sigma \) or \( \tau \), for symbols in \( \Sigma \), and lower case Latin, for example \( x \) or \( y \), for sequences of symbols, i.e., \( x = \sigma_1 \sigma_2 \cdots \sigma_n \). We use \(|x|\) to denote the length of \( x \). The empty sequence is denoted by \( \lambda \).

We denote the set of strictly infinite sequences over \( \Sigma \) by \( \Sigma^\infty \), and the set of strictly infinite sequences having \( x \) as prefix by \( x\Sigma^\infty \). Let \( S = \{ x\Sigma^\infty : x \in \Sigma^* \} \cup \{ \emptyset \} \), we can verify that \( S \) is a semiring\(^9\) over \( \Sigma^\infty \). We use \( \mathcal{F} \) to denote the sigma algebra generated by \( S \).

Definition 2 (Stochastic Process over \( \Sigma \)). A stochastic process over a finite alphabet \( \Sigma \) is a collection of \( \Sigma \)-valued random variables \( \{ X_t \}_{t \in \mathbb{N}} \) indexed by positive integers\(^10\).

We are specifically interested in processes in which the \( X_t \)s are not necessarily independently distributed.

Definition 3 (Sequence-Induced Measure and Derivative). For a process \( \mathcal{P} \), let \( Pr_\mathcal{F}(x) \) or simply \( Pr(x) \) denote the probability \( \mathcal{P} \) producing a sample path prefixed by \( x \). The measure \( \mu_x \) induced by a sequence \( x \in \Sigma^* \) is the extension\(^8\) to \( \mathcal{F} \) of the premeasure defined on the semiring \( S \) given by

\[
\forall x, y \in \Sigma^*, \mu_x(y \Sigma^\infty) = \frac{Pr_\mathcal{F}(xy)}{Pr_\mathcal{F}(x)}, \text{ if } Pr_\mathcal{F}(x) > 0
\]

(16)

For any \( d \in \mathbb{N} \), the \( d \)-th order derivative of a sequence \( x \), written as \( d_x^d \), is defined to be the marginal distribution of \( \mu_x \) on \( \Sigma^d \), with the entry indexed by \( y \) denoted by \( d_x^d(y) \). The first-order derivative is called the symbolic derivative and is denoted by \( \phi_x \) for short.

Definition 4 (Probabilistic Nerode Equivalence and Causal States\(^11\)). For any pair of sequences \( x, y \in \Sigma^* \), \( x \) is equivalent to \( y \), written as \( x \sim y \), if and only if either \( Pr(x) = Pr(y) = 0 \), or \( \mu_x = \mu_y \). The equivalence class of a sequence \( x \) is denoted by \(|x| \) and is called a causal state\(^12\). The cardinality of the set of causal states is called the probabilistic Nerode index, or the Nerode index for simplicity.

We can see from the definition that causal states captures how the history of a process influences its future. Since the probabilistic Nerode equivalence is right invariant, it gives rise naturally to a automaton structure introduced below.

Definition 5 (Probabilistic Finite-State Automaton (PFSA)). A PFSA \( G \) is defined by a quadruple \( (Q, \Sigma, \delta, \pi) \), where \( Q \) is a finite set, \( \Sigma \) is a finite alphabet, \( \delta : Q \times \Sigma \to \Sigma \) is called the transition map, and \( \pi : Q \to P_{\Sigma^*} \), where \( P_{\Sigma^*} \) is the space of probability distributions over \( \Sigma \), is called the transition probability. The entry of \( \pi(q) \) indexed by \( \sigma \) is denoted by \( \pi(q, \sigma) \).

Definition 6 (Transition and Observation Matrices). The transition matrix \( \Pi \) is the \(|Q| \times |Q| \) matrix with the entry indexed by \( q, q' \), written as \( \pi_{q, q'} \), satisfying

\[
\pi_{q, q'} = \sum_{\sigma \in \Sigma} \pi(q, \sigma)_{(q', \sigma)} - q' \]

(17)
and the observation matrix $\hat{\Pi}$ is a $|Q| \times |\Sigma|$ matrix with the entry indexed by $q, \sigma$ equaling $\hat{\pi}(q, \sigma)$.

We note that both $\Pi$ and $\hat{\Pi}$ are stochastic, i.e. non-negative with rows summing up to 1.

**Definition 7** (Extension of $\delta$ and $\pi$ to $\Sigma^*$). For any $x = \sigma_1 \ldots \sigma_k$, $\delta(q, x)$ is defined recursively by

$$
\delta(q, x) \triangleq \delta(\delta(q, \sigma_1 \ldots \sigma_{k-1}), \sigma_k)
$$

with $\delta(q, \lambda) = q$, and $\pi(q, x)$ is defined recursively by

$$
\pi(q, x) \triangleq \prod_{i=1}^{k} \pi(\delta(q, \sigma_1 \ldots \sigma_{i-1}), \sigma_i)
$$

with $\pi(q, \lambda) = 1$.

**Definition 8** (Strongly Connected PFSA). We say a PFSA is strongly connected if the underlying directed graph is strongly connected\(^{13}\). More precisely, a PFSA $G = (Q, \Sigma, \delta, \pi)$ is strongly connected if for any pair of distinct states $q$ and $q'$ in $Q$, there is an $x \in \Sigma^*$ such that $\delta(q, x) = q'$.

We assume all PFSA in the discussions in the sequel are strongly connected if not specified otherwise. For strongly connected PFSA $G$, there is a unique probability distribution over $Q$ that satisfies $\nu^T \Pi = \nu^T$. This is the **stationary distribution**\(^{14,15}\) of $G$ and is denoted as $\nu_G$, or $\nu$ if $G$ is understood.

**Definition 9** ($\Gamma$-Expression). We can encode the information contained in $\delta$ and $\pi$ by a set of $|Q| \times |Q|$ matrices $\Gamma = \{\Gamma_\sigma|\sigma \in \Sigma\}$, where

$$
\Gamma_\sigma|_{q,q'} \triangleq \begin{cases} 
\pi(q, \sigma) & \text{if } \delta(q, \sigma) = q', \\
0 & \text{if otherwise.}
\end{cases}
$$

$\Gamma_\sigma$ is called the **event-specific transition matrix**, with the event being that $\sigma$ is current the output. $\Gamma_\sigma$ can also be extended to arbitrary $x \in \Sigma^*$ by defining $\Gamma_x = \prod_{i=1}^{k} \Gamma_{\sigma_i}$, with $\Gamma_\lambda = I$.

**Definition 10** (Sequence-Induced Distribution on States). For a PFSA $G = (Q, \Sigma, \delta, \pi)$ and a distribution $p_0$ on $Q$, the **distribution on $Q$ induced by a sequence** $x$ is given by $p_{G,p_0}(x) = p_0^T \Gamma_x$ with $p_{G,p_0}(\lambda) = p_0$. The entry indexed by $q \in Q$ of the vector $p_{G,p_0}(x)$ is written as $p_{G,p_0}(x, q)$. When $p_0 = \nu_G$, the stationary distribution of $G$, we write $p_{G,\nu_G}(x)$ as $p_G(x)$, or simply as $p(x)$, if $G$ is understood.

**Definition 11** (Stochastic Process Generated by a PFSA). Let $G = (Q, \Sigma, \delta, \pi)$ be a PFSA and let $p_0$ be a distribution on $Q$, the **$\Sigma$-valued stochastic process** $\{X_t\}_{t \in \mathbb{N}}$ generated by $G$ and $p_0$ satisfies that $X_1$ follows the distribution $p_0$ and $X_{t+1}$ follows the distribution $p_{G,\nu_G}(X_1 \cdots X_t)$ for $t \in \mathbb{N}$.

For the rest of this paper, we will assume $p_0 = \nu_G$ if not specified otherwise. We can show that, when initialized with $p_G$, the process generated by a PFSA $G$ is stationary and ergodic. We also note that, for the process generate by $G$, we have $\phi_\pi = p_G(x)^T \hat{\Pi}$. Since $p_G(\lambda) = p_G$, the symbolic derivative of the empty sequence $\phi_\lambda$ is the stationary distribution on the symbols.

**Definition 12** (Synchronizable PFSA and Synchronizing Sequence). A **synchronizing sequence** is a finite sequence that sends an arbitrary state of the PFSA to a fixed state\(^{16}\). To be more precise, let $G = (Q, \Sigma, \delta, \pi)$ be a PFSA, we say a sequence $x \in \Sigma^*$ is a synchronizing sequence to a state $q \in Q$ if $\delta(q', x) = q$ for all $q' \in Q$. A PFSA is **synchronizable** if it has at least one synchronizing sequence. Given a sample path generated by a PFSA, we say the PFSA is **synchronized** if a synchronizing sequence transpires in the sample path.

**Definition 13** (Equivalence and Irreducibility). Two PFSA $G$ and $H$ are equivalent if they generate the same stochastic process. A PFSA $G$ is said to be irreducible, if there is not another PFSA with smaller state set that is equivalent to $G$.

**Definition 14**. Consider a PFSA $G$ over state set $Q$. For a give $\epsilon > 0$, we say a sequence $x$ is a $\epsilon$-synchronizing sequence to a state $q \in Q$ if

$$
\|p_G(x) - \nu_q\|_{\infty} \leq \epsilon.
$$

While there exists PFSA that is not synchronizable, we can show that an irreducible PFSA always has an $\epsilon$-synchronizing sequence for some state $q$ for arbitrarily small $\epsilon > 0$. Moreover, we can show that as length increases, sequences produced by PFSA become uniformly $\epsilon$-synchronizing. These two are the underpinning properties for the inference algorithm of PFSA (See Alg. 1), because they imply that $\phi_\pi$ can be used to approximate $\pi(q)$ if $x$ is properly prefixed and long enough.

**Definition 15** (Joint $\epsilon$-Synchronizing Sequence). Let $G$ and $H$ be two PFSA over state sets $Q_G$ and $Q_H$, respectively. For a fixed $\epsilon$, a sequence $x$ is said to be **jointly $\epsilon$-synchronizing** to $(q, \tau) \in Q_G \times Q_H$ if $x$ is
\(\epsilon\)-synchronizing to \(q\) and to \(r\) simultaneously. We define
\[
\Sigma^d_{\epsilon,(q,r)} = \{ \{ x \in \Sigma^d : x \text{ jointly } \epsilon\text{-synchronizing to (}q,r\} \}
\]

**Definition 16 (Joint Pair of States).** Let \(G\) and \(H\) be two PFSA over state sets \(Q_G\) and \(Q_H\), respectively. Define
\[
p_G(q,r) = \lim_{d \to \infty} p_G \left( \Sigma^d_{\epsilon,(q,r)} \right)
\]
A pair of states \((q,r)\) in \(Q_G \times Q_H\) is called a \(G\)-joint pair of states if \(p_G(q,r) > 0\). We also define
\[
Q_c = \{ (q,r) \in Q_G \times Q_H : (q,r) \text{ is a } G\text{-joint pair} \}
\]

The inference algorithm for PFSA is called GenESeSS for Generator Extraction Using Self-similar Semantics. With an input sequence \(x\) and a hyperparameter \(\epsilon\), GenESeSS outputs a PFSA in the following three steps: 1) approximate an almost synchronizing sequence; 2) identify the transition structure of the PFSA; 3) calculate the transition probabilities of the PFSA. See Alg. 1 for details.

**Algorithm 1: GenESeSS**

**Data:** A sequence \(x\) over alphabet \(\Sigma\), \(0 < \epsilon < 1\)

**Result:** State set \(Q\), transition map \(\delta\), and transition probability \(\pi\)

*/ Step One: Approximate \(\epsilon\)-synchronizing sequence */

1. Let \(L = \left\lfloor \log_{|\Sigma|} 1/\epsilon \right\rfloor \);
2. Calculate the derivative heap \(D^\pi_x\) equaling \(\left\{ \hat{\varphi}^x_y : y \text{ is a sub-sequence of } x \text{ with } |y| \leq L \right\}\);
3. Let \(C\) be the convex hull of \(D^\pi_x\);
4. Select \(x_0\) with \(\hat{\varphi}^x_{x_0}\) being a vertex of \(C\) and has the highest frequency in \(x\);

*/ Step Two: Identify transition structure */

5. Initialize \(Q = \{ q_0 \}\);
6. Associate to \(q_0\) the sequence identifier \(x_{q_0} = x_0\) and the probability vector \(d_{q_0} = \hat{\varphi}^x_{x_0}\);
7. Let \(Q\) be the set of states that are just added and initialize it to be \(Q\);
8. while \(Q \neq \emptyset\) do
9.  let \(Q_{\text{new}} = \emptyset\) be the set of new states;
10. for \((q, \sigma) \in \tilde{Q} \times \Sigma\) do
11.  let \(x = x_{q_\sigma}\) and \(d = \hat{\varphi}^x_{x_\sigma}\);
12.  if \(\|d - d_{q_\sigma}\|_\infty < \epsilon\) for some \(q' \in Q\) then
13.    let \(\delta(q, \sigma) = q'\);
14.  else
15.    let \(Q_{\text{new}} = Q_{\text{new}} \cup \{ q_{\text{new}} \}\) and \(Q = Q \cup \{ q_{\text{new}} \}\);
16.    associate to \(q_{\text{new}}\) the sequence identifier \(x_{q_{\text{new}}} = x_{q_\sigma}\) and the probability vector \(d_{q_{\text{new}}} = d\);
17.    let \(\delta(q, \sigma) = q_{\text{new}}\);
18.  end
19. end
20. take a strongly connected subgraph of the labeled directed graph defined by \(Q\) and \(\delta\), and denote the vertex set of the subgraph again by \(Q\)

*/ Step Three: Identify transition probability */

21. Initialize counter \(N[ q, \sigma ]\) for each pair \((q, \sigma) \in Q \times \Sigma\);
22. Choose a random starting state \(q \in Q\);
23. for \(\sigma \in \Sigma\) do
24.  let \(N[ q, \sigma] = N[ q, \sigma] + 1\);
25.  let \( q = \delta(q, \sigma)\);
26.  let \( p(q) = \left\{ (N[ q, \sigma])_{\sigma \in \Sigma} \right\} \)
27. return \(Q, \delta, \pi\);

**XIII. THEORETICAL DEVELOPMENT OF SEQUENCE LIKELIHOOD DEFECT**

**Definition 17 (Entropy Rate and KL Divergence).** By entropy rate of a PFSA, we mean the entropy rate of the stochastic process generated by the PFSA\(^{17}\). Similarly, by KL divergence of two PFSA, we mean the KL divergence between the two processes generated by them\(^{18}\). More precisely, we have
\[
H(G) = -\lim_{d \to \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p(x) \log p(x)
\]
and the KL divergence
\[
\mathcal{D}_{KL}(G || H) = \lim_{d \to \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)}
\]
whenever the limits exist.

**Theorem 1** (Closed-form Formula for Entropy Rate and KL Divergence). The entropy rate of a PFSA \( G = (\Sigma, Q, \delta, \bar{\pi}) \) is given by
\[
\mathcal{H}(G) = \sum_{q \in Q} \nu_G(q) \cdot h(\bar{\pi}(q))
\]
where \( h(\nu) \) is the based-2 entropy of the probability vector \( \nu \).

Consider two PFSA \( G = (Q_G, \Sigma, \delta_G, \bar{\pi}_G) \) and \( H = (Q_H, \Sigma, \delta_H, \bar{\pi}_H) \) with \( \mu_G \) being absolutely continuous with respect to \( \mu_H \). Let \( Q_c \) be the set of \( G \)-joint pairs of states, we have
\[
\mathcal{D}_{KL}(G || H) = \sum_{(q, r) \in Q_c} p_G(q, r) \mathcal{D}_{KL}(\bar{\pi}_G(q) || \bar{\pi}_H(r))
\]
**Definition 18** (Log-likelihood). Let \( x \in \Sigma^d \), the log-likelihood\(^17\) of a PFSA \( G \) generating \( x \) is given by
\[
L(x, G) = -\frac{1}{d} \log p_G(x)
\]
The calculation of log-likelihood is detailed in Alg. 2.

**Algorithm 2: Log-likelihood**

**Data:** A PFSA \( G = (\Sigma, Q, \delta, \bar{\pi}) \) and a sequence \( x \) over alphabet \( \Sigma \)

**Result:** Log-likelihood \( L(x, G) \) of \( G \) generating \( x \)

1. Calculate the state transition matrix \( \Gamma \) and observation \( \bar{\Gamma} \);
2. Calculate the stationary distribution over states \( p_G \) of \( G \) from \( \Pi \);
3. Calculate the stationary distribution of alphabet \( \delta_G^x = p_G^x \bar{\Gamma} \);
4. Initialize \( p \) by \( p_G \) and \( q \) by \( \phi_\lambda \);
5. Let \( L = 0 \);
6. for \( i \) from 1 to \( |x| \) do
   7. Let \( \sigma \) be the \( i \)-th entry of \( x \);
   8. Let \( L = L - \log q_\sigma \);
   9. Let \( p^x = [p^T \Gamma_\sigma] \) where \( \Gamma_\sigma \) is defined in 9;
   10. Let \( q^x = p^T \bar{\Gamma} \);
7. return \( L/|x| \);

**Theorem 2** (Convergence of log-likelihood). Let \( G \) and \( H \) be two reduced PFSA, and let \( x \in \Sigma^d \) be a sequence generated by \( G \). Then we have
\[
L(x, H) \to \mathcal{H}(G) + \mathcal{D}_{KL}(G || H)
\]
in probability as \( d \to \infty \).

**Proof.** We first notice that
\[
\sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)} = \sum_{x \in \Sigma^{d-1}} \sum_{\sigma \in \Sigma} p_G(x) p_G(x) \bar{\pi}_G |_{\sigma} \log \frac{p_G(x) p_G(x)}{p_H(x) p_H(x)} \bar{\pi}_H |_{\sigma} \left( p_G(x) \bar{\pi}_G \right) + \sum_{x \in \Sigma^{d-1}} p_G(x) \sum_{\sigma \in \Sigma} p_G(x) \bar{\pi}_G |_{\sigma} \log \frac{p_G(x) \bar{\pi}_G}{p_H(x) \bar{\pi}_H} |_{\sigma} \left( p_G(x) \bar{\pi}_G \right)
\]
By induction, we have \( \mathcal{D}_{KL}(G || H) = \lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{d} D_i \), and hence by Cesàro summation theorem\(^19\), we have \( \mathcal{D}_{KL}(G || H) = \lim_{d \to \infty} D_d \). Let \( x = \sigma_1 \sigma_2 \ldots \sigma_n \) be a sequence generated by \( G \). Let \( x^{i-1} \) is the truncation of \( x \) at the \((i - 1)\)-th symbols, we have
\[
-\frac{1}{n} \sum_{i=1}^{n} \log p_H(x) \left( x^{i-1} \right) \bar{\pi}_H |_{\sigma_i} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{p_G(x^{i-1}) \bar{\pi}_G}{p_H(x^{i-1}) \bar{\pi}_H} |_{\sigma_i} - \frac{1}{n} \sum_{i=1}^{n} \log p_G \left( x^{i-1} \right) \bar{\pi}_G |_{\sigma_i}
\]
Since the stochastic process $G$ generates is ergodic, we have

$$\lim_{n \to \infty} A_{x,n} = \lim_{d \to \infty} D_d = D_{KL}(G \| H)$$

and $\lim_{n \to \infty} B_{x,n} = \mathcal{H}(G)$. □

### XIV. PIPELINE OPTIMIZATION: HYPER-TRAINING, TRAINING, & VALIDATION

Our pipeline comprises a network of individually trained light gradient boosting machine (LGBM)\(^2\) classifiers that focus on complementary aspects of the problem, and operate on different categories of input features as described next. Importantly, some of these features need to be generated non-trivially from the raw data, and these feature generators have parameters that need to be trained as well (or comprise models that need to be inferred). We call this inference of the feature generators as hyper-training. Importantly, this is different from the more common notion of hyper-parameters. Hyper-parameters are one or more variables whose scalar values are commonly tuned by grid-search or via some meta-heuristics to optimize classifiers, whereas hyper-training produces generators of features, not simply a set of numbers.

#### Hyper-training & Training

**Trinary Quantization of Medical Histories:** The medical histories are mapped into trinary disease-phenotype-specific data-streams to enable generation of some of the features described below, as outlined in Section IV (Step 1).

**Feature Categories:** The features used in the pipeline maybe categorized as follows:

- **PFSA scores:** The PFSA scores are computed on the basis of the inferred PFSA models as described in the previous sections. The generation of the PFSA models from the trinary data-streams is the first hyper-training step. These scores consist of the negative and positive log-likelihood of a phenotype-specific quantized medical history being generated by the PFSA models for the positive cohort and the control cohort of sex-stratified patients, and the corresponding sequence likelihood defects (See Section ??). Recall that PFSAs are specialized HMMs, and these measures encode the dynamics of the underlying processes, and are sensitive to the ordering, and frequency of the codes at the resolution of the disease phenotypes. Also, recall that diseases phenotypes are broad categories of diagnostic codes, and that we generate PFSA models for each category, and separately for the sexes and the positive cohort and the control cohorts).

- **Prevalence scores (p-scores):** The p-scores focus on individual diagnostic codes, and we create a dictionary of the ratio of relative prevalence of each code (relative to the set of all codes present) in the positive category (for each sex) to the control category. This is the second hyper-training step. In the later steps of the pipeline, we use dictionary look ups to map codes to their p-scores, and also their aggregate measures such as mean, median, and variance to train a downstream LGBM.

- **Rare scores:** These scores consist of a subset of p-scores which correspond to codes with particularly high and low relative prevalences (p-score $> 2$ or $< .5$). Thus, this feature category depends on the p-score dictionary generated in the second hyper-training step.

- **Sequence scores:** Sequence scores are relatively straight-forward statistical measures such as mean, median, variance, time since last occurrence etc., on the trinary phenotype-specific sex-stratified histories. No hyper-training is required for the generation of the sequence features.

Thus we require three splits of the training dataset. The first split is used to carry out hyper-training of the PFSA models and the p-score dictionary. The second split is used to train the score-category specific LGBMs, one for each feature category. And the third split is used to train the final LGBM that takes inputs from the outputs of the four LGBMs in the previous layer. The network layout is shown in Fig. 1.

#### Validation

In validation, or actual prediction of patient fate, we use the trinary mapping, generate the features using the PFSA models and the p-score dictionary, and calculate the raw-risk via the trained LGBM network. The relative score is then obtained by a choice of the operating point reflecting the specificity/sensitivity trade-off discussed before.
SI-Fig. 1: Schematic of prediction pipeline. Panels A, B and C show the sequential training steps, namely the hyper-training of the PFSA models and the p-score dictionary, the pre-aggregation training of the four LGBM models, and the aggregation training of the final LGBM model respectively. Panel D shows the configuration of the trained pipeline in operation. In the training steps, the filled box represents the component being trained in that step.

Data Splits: Training and Validation Hold-out

All eligible patients are randomly split into the Training set (≈ 75% of data) and the Test set (≈ 25% of data). The Training set is then split into 3 subsets: 1) The hyper-training set (See Fig. 1A) is used to train PFSA models p-score dictionary, 2) the second split (referred to as the pre-aggregation split, see Fig. 1B) is used to train the four feature-category specific LGBMs, and 3) the final split (referred to as the aggregation split, see Fig. 1C) is used to train the aggregate LGBM which uses outputs from the trained LGBMs in the previous layer. This trained pipeline is then validated on the held out validation split (≈ 25% of data).
SI Table I: Disease Categories With Detailed Set of ICD Codes Used

| Constituent ICD9 Codes |
|------------------------|
| **Disease Categories** | **Constituent ICD9 Codes** |
| 250.0 | 250.20 | 250.61 | 250.81 | E11.29 | E10.92 | E10.81 | E10.69 | E10.59 | 250.23 | 250.33 | 250.22 | 250.83 | 250.62 | 250.40 | 250.32 | 250.63 | 250.43 | 250.02 | 250.72 | 250.80 | 250.31 |
| 250.13 | 250.41 | 250.51 | E11.10 | E10.62 | E11.88 | E11.60 | E11.01 | E11.31 | 250.50 | 250.51 | E11.65 | E11.53 | E11.25 | E11.50 | 250.35 | E11.20 | E11.59 | 250.30 | 250.52 | 250.53 | 250.39 |
| 250.70 | 250.72 | 250.65 | 250.80 | 250.51 | 250.81 | 250.61 | 250.81 | 250.61 | 250.81 | 250.61 | 250.81 | 250.61 | 250.81 | 250.61 | 250.81 | 250.61 | 250.81 | 250.61 | 250.81 | 250.61 | 250.81 |

**Continued on next page**
| Code | Description |
|------|-------------|
| J04.0 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J10.6 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J12.2 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.0 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.1 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.2 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.3 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.4 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.5 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.6 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.7 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.8 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J15.9 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.0 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.1 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.2 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.3 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.4 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.5 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.6 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.7 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.8 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J16.9 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.0 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.1 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.2 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.3 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.4 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.5 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.6 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.7 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.8 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J17.9 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.0 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.1 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.2 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.3 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.4 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.5 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.6 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.7 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.8 | Respiratory Infections - Infections of the Respiratory Tract and Ear |

| Code | Description |
|------|-------------|
| J18.9 | Respiratory Infections - Infections of the Respiratory Tract and Ear |
### SI-Table II: Diagnostic codes for dyspnea for investigating ZCoR-IPF effectiveness in identifying IPF without such indications

| ICD code | Description                                    |
|----------|-----------------------------------------------|
| R06.9    | Unspecified abnormalities of breathing        |
| 784.99   | Head & neck symptoms NEC                      |
| 786.0    | Respiratory abnorm NOS                        |
| 786.9    | Resp sys/chest symp NEC                       |
| R06.02   | Shortness of breath                           |
| R06.01   | Orthopnea                                     |
| R06.89   | Other abnormalities of breathing               |
| 786.8    | Hiccough                                      |
| 786.05   | Shortness of breath                           |
| R06.7    | Sneezing                                      |
| 786.09   | Respiratory abnorm NEC                        |
| R06.00   | Dyspnea unspecified                           |
| 786.06   | Tachypnea                                     |
| 786.07   | Wheezing                                      |
| R06.82   | Tachypnea not elsewhere classified            |
| R06     | Dyspnea unspecified                           |
| 786.04   | Cheyne-stokes respiratn                       |
| R06.6    | Hiccough                                      |
| R06.1    | Stridor                                       |
| R06.83   | Snoring                                       |
| R06.2    | Wheezing                                      |
| R06.09   | Other forms of dyspnea                        |
| 786.03   | Apnea                                         |
| R06.5    | Mouth breathing                               |
| R06.81   | Apnea not elsewhere classified                |
| 786.01   | Hyperventilation                              |
| R06.03   | Acute respiratory distress                    |
| R06.4    | Hyperventilation                              |
| R06.3    | Periodic breathing                            |
| 786.1    | Stridor                                       |

*The No-dyspnea cohort comprises patients who lack codes from this list. Thus these patients are at higher risk of mis-diagnosis*
SI-Table III: Diagnostic codes for COPD, heart disease and asthma for investigating ZCoR-IPF effectiveness under other pre-existing diagnoses. The COPD/Heart disease/Asthma cohort comprises patients with one or more codes from this list.

| ICD code | Description |
|----------|-------------|
| 394.0    | Mitral stenosis                         |
| J45.998  | Other asthma                            |
| 422.9    | Acute myocarditis NOS                   |
| 415      | Acute cor pulmonale                     |
| I25.111  | Atherosclerotic heart disease of native coronary artery with angina pectoris with documented spasm |
| 426      | Atriovent block complete                |
| I27.21   | Secondary pulmonary arterial hypertension |
| J42      | Unspecified chronic bronchitis          |
| 395.9    | Rheum aortic dis NEC/NOS                |
| I20.1    | Angina pectoris with documented spasm   |
| 425.0    | Endomyocardial fibrosis                 |
| I51.9    | Disease of pericardium unspecified      |
| 398.91   | Rheumatic heart failure                 |
| 423.3    | Cardiac tamponade                       |
| J45.21   | Mild intermittent asthma with (acute) exacerbation |
| 427.6    | Premature beats NOS                     |
| 394.2    | Mitral stenosis w insuff                |
| J47      | Bronchiectasis with acute lower respiratory infection |
| I51.5    | Myocardial degeneration                 |
| I50.810  | Right heart failure unspecified         |
| 395.2    | Rheum aortic sten/insuff                |
| 428.1    | Left heart failure                      |
| 410.72   | Subendo infarct subseq                  |
| 491.2    | Obst chr bronc w/o exc                 |
| 493.9    | Asthma NOS                              |
| I20      | Unstable angina                         |
| I13.10   | Hypertensive heart and chronic kidney disease without heart failure with stage 1 through stage 4 chronic kidney disease or unspecified chronic kidney disease |
| I49.2    | Junctional premature depolarization     |
| I50.42   | Chronic combined systolic (congestive) and diastolic (congestive) heart failure |
| 425.4    | Prim cardiomyopathy NEC                 |
| J41.0    | Simple chronic bronchitis               |
| 426.1    | Atriovent block NOS                     |
| 429      | Myocarditis NOS                         |
| 491      | Simple chr bronchitis                   |
| I25.701  | Atherosclerosis of coronary artery bypass graft(s) unspecified with angina pectoris with documented spasm |
| I25.84   | Coronary atherosclerosis due to calcified coronary lesion |
| I50.84   | End stage heart failure                 |
| I51.0    | Cardiac septal defect acquired          |
| 414.01   | Crnry athrscl natve vssl               |
| I37.1    | Nonrheumatic pulmonary valve insufficiency |
| 115.03   | Histoplasmosis caps pericard           |
| I21.01   | ST elevation (STEMI) myocardial infarction involving left main coronary artery |
| I06      | Rheumatic aortic stenosis               |
| I45.9    | Conduction disorder unspecified        |
| 410.02   | AMI anterolateral subseq                |
| 426.89   | Conduction disorder NEC                 |
| 422.92   | Septic myocarditis                      |
| 494.1    | Bronchiectasis w ac excac              |
| I23.1    | Atrial septal defect as current complication following acute myocardial infarction |
| I31.8    | Other specified diseases of pericardium |
| J45.41   | Moderate persistent asthma with (acute) exacerbation |
| I25.769  | Atherosclerosis of bypass graft of coronary artery of transplanted heart with unspecified angina pectoris |
| I34.1    | Nonrheumatic mitral (valve) prolapse    |
| I31      | Chronic adhesive pericarditis          |
| 402.01   | Mal hypert hrt dis w hf                 |
| 410.52   | AMI lateral NEC subseq                  |
| 493.0    | Extrinsic asthma NOS                    |
| I09.2    | Chronic rheumatic pericarditis          |
| J45.901  | Unspecified asthma with (acute) exacerbation |
| B39.5    | Histoplasmosis duboisii                 |
| 425.8    | Cardiomyopathy in oth dis              |
| I05.9    | Rheumatic mitral valve disease unspecified |
| 411      | Post MI syndrome                       |

Continued on next page
| ICD code | Description |
|----------|-------------|
| J43.8    | Other emphysema |
| I27      | Primary pulmonary hypertension |
| 428.31   | Ac diastolic hrt failure |
| 428.43   | Ac/chr syst/dia hrt fail |
| 295.5    | Chordae tendinae rupture |
| I46      | Cardiac arrest due to underlying cardiac condition |
| I48.92   | Unspecified atrial flutter |
| 410.01   | AMI anterolateral init |
| 411.1    | Intermed coronary synd |
| I37.8    | Other nonrheumatic pulmonary valve disorders |
| J43.1    | Panlobular emphysema |
| 402.91   | Hyp ht dis NOS w ht fail |
| I13.2    | Hypertensive heart and chronic kidney disease with heart failure and with stage 5 chronic kidney disease or end stage renal disease |
| I11      | Hypertensive heart disease with heart failure |
| J41.1    | Mucopurulent chronic bronchitis |
| I25.710  | Atherosclerosis of autologous vein coronary artery bypass graft(s) with unstable angina pectoris |
| 428.4    | Syst/diast hrt fail NOS |
| I31.0    | Chronic adhesive pericarditis |
| 394      | Mitral stenosis |
| I44.4    | Left anterior fascicular block |
| 496      | Chr airway obstruct NEC |
| I24.9    | Acute ischemic heart disease unspecified |
| I25.110  | Atherosclerotic heart disease of native coronary artery with unstable angina pectoris |
| I20.8    | Other forms of angina pectoris |
| I47.1    | Supraventricular tachycardia |
| I11.0    | Hypertensive heart disease with heart failure |
| 420.99   | Acute pericarditis NEC |
| 417.8    | Pulmon circulat dis NEC |
| 426.12   | Atrioven block-mobitz ii |
| I28.1    | Aneurysm of pulmonary artery |
| I50.83   | High output heart failure |
| 410.3    | AMI interopost unspec |
| I42.1    | Obstructive hypertrophic cardiomyopathy |
| I49      | Ventricular fibrillation |
| 426.52   | Rt bbb/lft ant fasc blk |
| I22.3    | Subsequent ST elevation (STEMI) myocardial infarction of other sites |
| 410.51   | AMI lateral NEC initial |
| I25.739  | Atherosclerosis of nonautologous biological coronary artery bypass graft(s) with unspecified angina pectoris |
| 428.23   | Ac on chr syst hrt fail |
| 429.1    | Myocardial degeneration |
| 423.8    | Pericardial disease NEC |
| I08      | Rheumatic disorders of both mitral and aortic valves |
| I48.91   | Unspecified atrial fibrillation |
| I37      | Nonrheumatic pulmonary valve stenosis |
| I26.09   | Other pulmonary embolism with acute cor pulmonale |
| I50.20   | Unspecified systolic (congestive) heart failure |
| 410.91   | AMI NOS initial |
| 402.0    | Mal hyp ht dis w/o hf |
| 426.53   | Bilat bb block NEC |
| 492.8    | Emphysema NEC |
| J47.1    | Bronchiectasis with (acute) exacerbation |
| I25.700  | Atherosclerosis of coronary artery bypass graft(s) unspecified with unstable angina pectoris |
| I07.2    | Rheumatic tricuspid stenosis and insufficiency |
| I08.8    | Other rheumatic multiple valve diseases |
| 396.2    | Mitral insuf/aort stenos |
| 414.11   | Aneurysm coronary vessel |
| I45.81   | Long QT syndrome |
| 495.4    | Malt workers’ lung |
| I39      | Endocarditis and heart valve disorders in diseases classified elsewhere |
| I24.1    | Dressler’s syndrome |
| 427.9    | Cardiac dysrhythmia NOS |
| 491.0    | Simple chr bronchitis |
| 414.2    | Chr tot occlus cor artry |
| I36      | Nonrheumatic tricuspid (valve) stenosis |
| ICD code | Description |
|----------|-------------|
| I44.39   | Other atrioventricular block |
| I21.A1   | Myocardial infarction type 2 |
| I49.02   | Ventricular flutter |
| 404.0    | Mal hy ht/kd I-IV w/o hf |
| I27.82   | Chronic pulmonary embolism |
| 493.82   | Cough variant asthma |
| 404.12   | Ben hy ht/kd st V w/o hf |
| I06.2    | Rheumatic aortic stenosis with insufficiency |
| J45.32   | Mild persistent asthma with status asthmaticus |
| 427.41   | Ventricular fibrillation |
| I34      | Nonrheumatic mitral (valve) insufficiency |
| I07.1    | Rheumatic tricuspid insufficiency |
| 410.11   | AMI anterior wall init |
| 414      | Cor ath unsp vsl ntv/gft |
| 421      | Ac/subac bact endocard |
| 410.81   | AMI NEC initial |
| I05.8    | Other rheumatic mitral valve diseases |
| J45.902  | Unspecified asthma with status asthmaticus |
| I28.0    | Arteriovenous fistula of pulmonary vessels |
| 427.0    | Parox atrial tachycardia |
| 397.1    | Rheum pulmon valve dis |
| I09      | Rheumatic myocarditis |
| I09.1    | Rheumatic diseases of endocardium valve unspecified |
| 428.9    | Heart failure NOS |
| 495.1    | Bagassosis |
| 398.0    | Rheumatic myocarditis |
| T80.0xxA | Air embolism following infusion transfusion and therapeutic injection initial encounter |
| I49.01   | Ventricular fibrillation |
| I26.92   | Saddle embolus of pulmonary artery without acute cor pulmonale |
| I44.2    | Atriointerventricular block complete |
| I25.718  | Atherosclerosis of autologous vein coronary artery bypass graft(s) with other forms of angina pectoris |
| 414.06   | Cor ath natv art tp hrt |
| I31.3    | Pericardial effusion (noninflammatory) |
| J67.5    | Mushroom-worker’s lung |
| J67.9    | Hypersensitivity pneumonitis due to unspecified organic dust |
| 410.2    | AMI inferolateral unspec |
| 427.2    | Parox tachycardia NOS |
| I25.708  | Atherosclerosis of coronary artery bypass graft(s) unspecified with other forms of angina pectoris |
| 413.1    | Pirizmetal angina |
| 410.7    | Subendo infarct unspec |
| I21.09   | ST elevation (STEMI) myocardial infarction involving other coronary artery of anterior wall |
| 493.91   | Asthma w status asthmatic |
| J67.4    | Maltworker’s lung |
| I50.43   | Acute on chronic combined systolic (congestive) and diastolic (congestive) heart failure |
| 426.11   | Atrioint block-1st degr |
| I13      | Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease or unspecified chronic kidney disease |
| 410.71   | Subendo infarct initial |
| 410.1    | AMI anterior wall unspec |
| I50.812  | Chronic right heart failure |
| 115.94   | Histoplasmosis endocard |
| J45.909  | Unspecified asthma uncomplicated |
| I25.751  | Atherosclerosis of native coronary artery of transplanted heart with angina pectoris with documented spasm |
| 424.2    | Nonheum tricusp val dis |
| I25.798  | Atherosclerosis of other coronary artery bypass graft(s) with other forms of angina pectoris |
| I37.9    | Nonrheumatic pulmonary valve disorder unspecified |
| J43.0    | Unilateral pulmonary emphysema [MacLeod’s syndrome] |
| I28.8    | Other diseases of pulmonary vessels |
| 411.0    | Post MI syndrome |
| 422.91   | Idiopathic myocarditis |
| 424.3    | Pulmonary valve disorder |
| I08.3    | Combined rheumatic disorders of mitral and tricuspid valves |
| I42.7    | Cardiomyopathy due to drug and external agent |
| I50.9    | Heart failure unspecified |
| 495.8    | Allerg alveol/pneum NEC |
| ICD code | Description |
|----------|-------------|
| 414.12   | Dissection cor artery |
| 424.9    | Endocarditis NOS |
| 404.03   | Mal hyp ht/kd stg V w hf |
| I21.21   | ST elevation (STEMI) myocardial infarction involving left circumflex coronary artery |
| I48.4    | Atypical atrial flutter |
| I45.3    | Trifascicular block |
| I05.1    | Rheumatic mitral insufficiency |
| 395      | Rheumat aortic stenosis |
| I31.2    | Hemopericardium not elsewhere classified |
| I42.8    | Other cardiomyopathies |
| I45.19   | Other right bundle-branch block |
| I25.10   | Atherosclerotic heart disease of native coronary artery without angina pectoris |
| I44.7    | Left bundle-branch block unspecified |
| I50.1    | Left ventricular failure unspecified |
| 397.0    | Tricuspid valve disease |
| I50.23   | Acute on chronic systolic (congestive) heart failure |
| 493.11   | Int asthma w status asth |
| I21.11   | ST elevation (STEMI) myocardial infarction involving right coronary artery |
| J45.51   | Severe persistent asthma with (acute) exacerbation |
| I45.10   | Unspecified right bundle-branch block |
| 420.0    | AC pericardit in oth dis |
| J67.6    | Maple-bark-stripper's lung |
| I25.83   | Coronary atherosclerosis due to lipid rich plaque |
| I46.9    | Cardiac arrest cause unspecified |
| I06.8    | Other rheumatic aortic valve diseases |
| I27.29   | Other secondary pulmonary hypertension |
| J41      | Simple chronic bronchitis |
| I07.9    | Rheumatic tricuspid valve disease unspecified |
| J41.8    | Mixed simple and mucopurulent chronic bronchitis |
| 425      | Endomyocardial fibrosis |
| I51      | Cardiac septal defect acquired |
| I25.729  | Atherosclerosis of autologous artery coronary artery bypass graft(s) with unspecified angina pectoris |
| 410.82   | AMI NEC subsequent |
| I42.6    | Alcoholic cardiomyopathy |
| I50.41   | Acute combined systolic (congestive) and diastolic (congestive) heart failure |
| 402      | Mal hyp ht dis w/o hf |
| 426.54   | Trifascicular block |
| J45.40   | Moderate persistent asthma uncomplicated |
| 495.3    | Suberosis |
| I44.60   | Unspecified fascicular block |
| 494      | Bronchiectas w/o ac excac |
| 414.9    | Chr ischemic trt dis NOS |
| 414.0    | Cor ath uns psvl ntv/gft |
| I23.3    | Rupture of cardiac wall without hemopericardium as current complication following acute myocardial infarction |
| I44      | Atrioventricular block first degree |
| 493.92   | Asthma NOS w (ac) excac |
| 426.81   | Lown-ghanong-levine synd |
| I51.7    | Cardiomegaly |
| 417.9    | Pulmon circular dis NOS |
| I09.89   | Other specified rheumatic heart diseases |
| I21.29   | ST elevation (STEMI) myocardial infarction involving other sites |
| I44.5    | Left posterior fascicular block |
| I47.2    | Ventricular tachycardia |
| I24      | Acute coronary thrombosis not resulting in myocardial infarction |
| I22      | Subsequent ST elevation (STEMI) myocardial infarction of anterior wall |
| I25.760  | Atherosclerosis of bypass graft of coronary artery of transplanted heart with unstable angina |
| 414.05   | Cor ath bypass graft NOS |
| 404.02   | Mal hyt/kd st V w/o hf |
| I51.2    | Rupture of papillary muscle not elsewhere classified |
| I50.31   | Acute diastolic (congestive) heart failure |
| I42.5    | Other restrictive cardiomyopathy |
| J67.7    | Air conditioner and humidifier lung |
| 425.18   | Oth hyptrophic cardiomy |
| 394.1    | Rheumatic mitral insuff |

Continued on next page
| ICD code  | Description                                                                 |
|-----------|-----------------------------------------------------------------------------|
| I22.2     | Subsequent non-ST elevation (NSTEMI) myocardial infarction                   |
| I25.811   | Atherosclerosis of native coronary artery of transplanted heart without angina pectoris |
| I30.9     | Acute pericarditis unspecified                                               |
| I27.1     | Kyphoscoliotic heart disease                                                 |
| I27.81    | Cor pulmonale (chronic)                                                     |
| 427.31    | Atrial fibrillation                                                          |
| 414.4     | Cor ath d/t calc cor lsn                                                     |
| J40       | Bronchitis not specified as acute or chronic                                |
| I25.9     | Chronic ischemic heart disease                                               |
| 429.71    | Acq cardiac sept defect                                                      |
| 410.41    | AMI inferior wall init                                                       |
| 416.0     | Prim pulm hypertension                                                       |
| 428.3     | Diastolic hrt failure NOS                                                     |
| 429.2     | Ascvd                                                                       |
| 410.22    | AMI inferolateral subseq                                                     |
| 416.2     | Chr pulmonary embolism                                                       |
| I08.1     | Rheumatic disorders of both mitral and tricuspid valves                      |
| I33       | Acute and subacute infective endocarditis                                   |
| 410.61    | True post infarct init                                                       |
| J45.52    | Severe persistent asthma with status asthmatic                               |
| I50.813   | Acute on chronic right heart failure                                         |
| 402.9     | Hyp hrt dis NOS w/o hf                                                       |
| 396.3     | Mitral/aortic val insuff                                                     |
| 428.42    | Chr syst/diast hrt fail                                                      |
| I36.9     | Nonrheumatic tricuspid valve disorder unspecified                            |
| I05.0     | Rheumatic mitral stenosis                                                    |
| I40.1     | Isolated myocarditis                                                        |
| I07       | Rheumatic tricuspid stenosis                                                 |
| 422.0     | Ac myocardit in oth dis                                                      |
| I45       | Right fascicular block                                                       |
| 429.81    | Papillary muscle dis NEC                                                     |
| I21.3     | ST elevation (STEMI) myocardial infarction of unspecified site               |
| 428.32    | Chr diastolic hrt fail                                                       |
| 398       | Rheumatic myocarditis                                                       |
| 410.31    | AMI inferopost initial                                                       |
| 491.22    | Obs chr bronc w ac bronc                                                     |
| 410.62    | True post infarct subseq                                                     |
| I40.0     | Infective myocarditis                                                       |
| I45.2     | Bilascicular block                                                           |
| 428.22    | Chr systolic hrt failure                                                     |
| I25.3     | Aneurysm of heart                                                            |
| I28       | Arteriovenous fistula of pulmonary vessels                                   |
| 414.04    | Cor ath arty bypass grft                                                     |
| 410       | AMI anterolateral unspec                                                     |
| 417.0     | Arterioven fistu pul ves                                                     |
| I26       | Septic pulmonary embolism with acute cor pulmonale                           |
| 429.6     | Papillary muscle rupture                                                     |
| I21.49    | Other myocardial infarction type                                             |
| I22.9     | Subsequent ST elevation (STEMI) myocardial infarction of unspecified site    |
| I25.790   | Atherosclerosis of other coronary artery bypass graft(s) with unstable angina pectoris |
| J67.0     | Farmer’s lung                                                                |
| 424.1     | Aortic valve disorder                                                        |
| I50.21    | Acute systolic (congestive) heart failure                                    |
| 427.5     | Cardiac arrest                                                               |
| 495       | Farmers’ lung                                                                |
| 424.0     | Mitral valve disorder                                                        |
| I08.2     | Rheumatic disorders of both aortic and tricuspid valves                      |
| I25.82    | Chronic total occlusion of coronary artery                                   |
| 416.1     | Kyphoscoliotic heart dis                                                     |
| I21.02    | ST elevation (STEMI) myocardial infarction involving left anterior descending coronary artery |
| 410.42    | AMI inferior wall subseq                                                     |
| 493.01    | Ext asthma w status asth                                                     |
| J45.30    | Mild persistent asthma uncomplicated                                         |
| I50.814   | Right heart failure due to left heart failure                                 |

Continued on next page
| ICD code | Description |
|----------|-------------|
| 414.19   | Aneurysm of heart NEC |
| 414.3    | Cor ath d/t lpd rch plaq |
| I43      | Cardiomyopathy in diseases classified elsewhere |
| 422.99   | Acute myocarditis NEC |
| 493.21   | Ch ob asthma w stat asth |
| I48      | Paroxysmal atrial fibrillation |
| I97.0    | Postcardiomy syndrome |
| 495.2    | Bird-fanciers’ lung |
| 426.3    | Left bb block NEC |
| 427      | Parox atrial tachycardia |
| I36.1    | Nonrheumatic tricuspid (valve) insufficiency |
| I42.3    | Endomyocardial (eosinophilic) disease |
| I25.711  | Atherosclerosis of autologous vein coronary artery bypass graft(s) with angina pectoris with documented spasm |
| I51.4    | Myocarditis unspecified |
| 414.07   | Cor ath bps graft tp hrt |
| I33.9    | Acute and subacute endocarditis unspecified |
| 428.33   | Ac on chr diast hrt fail |
| I26.01   | Septic pulmonary embolism with acute cor pulmonale |
| 414.8    | Chr ischemic hrt dis NEC |
| 423.2    | Constrictiv pericarditis |
| 491.9    | Chronic bronchitis NOS |
| I41      | Myocarditis in diseases classified elsewhere |
| 414.02   | Crn ath atlg vn bps grft |
| 404.92   | Hy ht/kd NOS st V w/o hf |
| 426.2    | Left bb hemiblock |
| 493.12   | Int asthma w (ac) exec |
| 425.11   | Hypertrophic obst cardiomy |
| I35.1    | Nonrheumatic aortic (valve) insufficiency |
| I09.9    | Rheumatic heart disease unspecified |
| 413      | Angina decubitus |
| I33.0    | Acute and subacute infective endocarditis |
| 115.04   | Histoplasm caps endocard |
| 425.2    | Obsc afric cardiomyopath |
| I24.8    | Other forms of acute ischemic heart disease |
| 410.21   | AMI inferolateral init |
| I37.0    | Nonrheumatic pulmonary valve stenosis |
| 429.82   | Hyperkinetic heart dis |
| 429.79   | Other sequelae of MI NEC |
| I05      | Rheumatic mitral stenosis |
| 491.1    | Mucopurul chr bronchitis |
| I25.738  | Atherosclerosis of nonautologous biological coronary artery bypass graft(s) with other forms of angina pectoris |
| 411.81   | Acute cor occlsn w/o Mi |
| J47.9    | Bronchiectasis uncomplicated |
| I30.8    | Other forms of acute pericarditis |
| I23.5    | Rupture of papillary muscle as current complication following acute myocardial infarction |
| I25.759  | Atherosclerosis of native coronary artery of transplanted heart with unspecified angina pectoris |
| 495.5    | Mushroom workers’ lung |
| I34.9    | Nonrheumatic mitral valve disorder unspecified |
| 397.9    | Rheum endocarditis NOS |
| 414.1    | Aneurysm of heart |
| I25.721  | Atherosclerosis of autologous artery coronary artery bypass graft(s) with angina pectoris with documented spasm |
| I52      | Other heart disorders in diseases classified elsewhere |
| 396.8    | Mitr/aortic multi involv |
| 410.32   | AMI interpost subseq |
| I25.731  | Atherosclerosis of nonautologous biological coronary artery bypass graft(s) with angina pectoris with documented spasm |
| 429.3    | Cardiomegaly |
| J44.9    | Chronic obstructive pulmonary disease unspecified |
| I25.761  | Atherosclerosis of bypass graft of coronary artery of transplanted heart with angina pectoris with documented spasm |
| I27.23   | Pulmonary hypertension due to lung diseases and hypoxia |
| I48.2    | Chronic atrial fibrillation unspecified |
| 396.9    | Mitral/aortic v dis NOS |
| I47.0    | Re-entry ventricular arrhythmia |
| I44.1    | Atrioventricular block second degree |
| I21.19   | ST elevation (STEMI) myocardial infarction involving other coronary artery of inferior wall |

Continued on next page
| ICD code | Description |
|----------|-------------|
| I06.1    | Rheumatic aortic insufficiency |
| 402.11   | Benign hypertrophy of heart with heart failure |
| I25.791  | Atherosclerosis of other coronary artery bypass graft(s) with angina pectoris with documented spasm |
| 402.1    | Benign hypertrophy of heart without heart failure |
| I46.2    | Cardiac arrest due to underlying cardiac condition |
| 410.8    | Acute myocardial infarction unspecified |
| 420.91   | Atrial idiopathic pericarditis |
| I36.2    | Nonrheumatic tricuspid (valve) stenosis with insufficiency |
| 427.81   | Sinoatrial node dysfunction |
| I48.0    | Paroxysmal atrial fibrillation |
| I35.2    | Nonrheumatic aortic (valve) stenosis with insufficiency |
| 395.0    | Rheumatoid aortic stenosis |
| 425.7    | Metabolic cardiomyopathy |
| 398.99   | Rheumatic heart disease NEC |
| 394.9    | Mitral valve disease NEC/NOS |
| I45.4    | Nonspecific intraventricular block |
| 425.82   | Long QT syndrome |
| I37.2    | Nonrheumatic pulmonary valve stenosis with insufficiency |
| 425.5    | Alcoholic cardiomyopathy |
| I25.42   | Coronary artery dissection |
| I32.0    | Pericarditis in diseases classified elsewhere |
| I34.8    | Nonspecific intraventricular block |
| 423.9    | Pericardial disease NOS |
| 410.5    | Acute cor pulmonale |
| I08.9    | Rheumatic multiple valve disease unspecified |
| 410.92   | Acute myocardial infarction unspecified |
| I50.0    | Acute nonspecific idiopathic pericarditis |
| J43.2    | Centrilobular emphysema |
| 427.1    | Paroxysmal ventricular tachycardia |
| I21.9    | Acute myocardial infarction unspecified |
| I25.41   | Coronary artery aneurysm |
| I44.0    | Atrioventricular block first degree |
| I25.750  | Atherosclerosis of native coronary artery of transplanted heart with unstable angina |
| 423.9    | Pericardial disease NOS |
| 410.5    | Acute cor pulmonale |
| I08.9    | Rheumatic multiple valve disease unspecified |
| 410.92   | Acute myocardial infarction unspecified |
| I50.0    | Left ventricular failure unspecified |
| 115.14   | Histoplasmosis endocarditis |
| 421.0    | Acute subacute bacterial endocarditis |
| I05.2    | Rheumatic mitral stenosis with insufficiency |
| I05.32   | Chronic diastolic (congestive) heart failure |
| 410.4    | Acute myocardial infarction unspecified |
| I09.0    | Rheumatic myocarditis |
| 495.0    | Farmers' lung |
| I20.0    | Unstable angina |
| I25.2    | Old myocardial infarction |
| J45.0    | Mild intermittent asthma uncomplicated |
| 426.0    | Atrioventricular block complete |
| 426.5    | Bundle branch block NOS |
| J67.2    | Bird fancier's lung |
| I09.81   | Rheumatic heart failure |
| J45.31   | Mild persistent asthma with (acute) exacerbation |
| 115.13   | Histoplasmosis endocarditis |
| 404.11   | Benign hypertrophy of heart and/or vessels |
| 411.89   | Acute ischemic heart disease NEC |
| I51.1    | Rupture of chordae tendineae not elsewhere classified |
| 495.7    | "Ventilation" pneumonitis |
| J45.22   | Mild intermittent asthma with status asthmaticus |
| I24.0    | Acute coronary thrombosis not resulting in myocardial infarction |
| J43      | Unilateral pulmonary emphysema [MacLeod's syndrome] |
| I49.1    | Atrial premature depolarization |
| I26.90   | Septic pulmonary embolism without acute cor pulmonale |
| J47.0    | Bronchiectasis with acute lower respiratory infection |
| J45.990  | Exercise induced bronchospasm |
| 429.4    | Heart disease postcardiac surgery |

Continued on next page
| ICD code | Description |
|----------|-------------|
| 410.0    | AMI anterolateral unspec |
| 424      | Mitral valve disorder |
| I27.0    | Primary pulmonary hypertension |
| 427.42   | Ventricular flutter |
| 404.93   | Hyp ht/kd NOS st V w hf |
| 491.8    | Chronic bronchitis NEC |
| 415.19   | Pulm embol/infarct NEC |
| I25.719  | Atherosclerosis of autologous vein coronary artery bypass graft(s) with unspecified angina pectoris |
| 426.13   | Av block-2nd degree NEC |
| I50.89   | Other heart failure |
| 410.9    | AMI NOS unspecified |
| I51.9    | Heart disease unspecified |
| 427.69   | Premature beats NEC |
| I49.49   | Other premature depolarization |
| I23.2    | Ventricular septal defect as current complication following acute myocardial infarction |
| I25.799  | Atherosclerosis of other coronary artery bypass graft(s) with unspecified angina pectoris |
| 427.61   | Atrial premature beats |
| 495.9    | Allerg alveol/pneum NOS |
| 412      | Old myocardial infarct |
| 426.51   | Rt bbb/lft post fasc blk |
| 404.13   | Ben hyp ht/kd stg V w hf |
| I49.8    | Other specified cardiac arrhythmias |
| 395.1    | Rheumatic aortic insuff |
| 426.9    | Conduction disorder NOS |
| 493      | Extrinsic asthma NOS |
| I30.1    | Infective pericarditis |
| 404.91   | Hyp ht/kd NOS I-IV w hf |
| I45.6    | Pre-excitation syndrome |
| I25.89   | Other forms of chronic ischemic heart disease |
| I49.5    | Sick sinus syndrome |
| I08.0    | Rheumatic disorders of both mitral and aortic valves |
| J44      | Chronic obstructive pulmonary disease with (acute) lower respiratory infection |
| I47      | Re-entry ventricular arrhythmia |
| J44.1    | Chronic obstructive pulmonary disease with (acute) exacerbation |
| I23.4    | Rupture of chordae tendineae as current complication following acute myocardial infarction |
| 429.9    | Heart disease NOS |
| 429.83   | Takotsubo syndrome |
| I13.11   | Hypertensive heart and chronic kidney disease without heart failure with stage 5 chronic kidney disease or end stage renal disease |
| I44.69   | Other fascicular block |
| I51.81   | Takotsubo syndrome |
| I13.0    | Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease or unspecified chronic kidney disease |
| B39.4    | Histoplasmosis capsulati unspecified |
| I50.30   | Unspecified diastolic (congestive) heart failure |
| 493.2    | Chronic obst asthma NOS |
| I31.4    | Cardiac tamponade |
| 414.03   | Crn ath nonatlg blg grft |
| 417.1    | Pulmon artery aneurysm |
| I50.82   | Biventricular heart failure |
| 492      | Emphysematous bleb |
| I35.9    | Nonrheumatic aortic valve disorder unspecified |
| I20.9    | Angina pectoris unspecified |
| 422      | Ac myocarditis in oth dis |
| I45.89   | Other specified conduction disorders |
| I50.33   | Acute on chronic diastolic (congestive) heart failure |
| I25.768  | Atherosclerosis of bypass graft of coronary artery of transplanted heart with other forms of angina pectoris |
| I42.4    | Endocardial fibroelastosis |
| 425.3    | Endocard fibroelastosis |
| I23.8    | Other current complications following acute myocardial infarction |
| I25.730  | Atherosclerosis of nonautologous biological coronary artery bypass graft(s) with unstable angina pectoris |
| J45.42   | Moderate persistent asthma with status asthmaticus |
| I49.40   | Unspecified premature depolarization |
| 421.1    | Ac endocarditis in oth dis |
| J45.991  | Cough variant asthma |
| ICD code | Description |
|----------|-------------|
| 398.9 | Rheumatic heart dis NOS |
| 429.89 | Ill-defined hrt dis NEC |
| J45.20 | Mild intermittent asthma uncomplicated |
| J44.0 | Chronic obstructive pulmonary disease with (acute) lower respiratory infection |
| I38 | Endocarditis valve unspecified |
| I07.0 | Rheumatic tricuspid stenosis |
| I22.1 | Subsequent ST elevation (STEMI) myocardial infarction of inferior wall |
| 423.1 | Adhesive pericarditis |
| 426.6 | Other heart block |
| 427.32 | Atrial flutter |
| 495.6 | Mapl bark-stripprs' lung |
| 428.41 | Ac syst/diastol hrt fail |
| 427.89 | Cardiomyopathy unspecified |
| I35 | Nonrheumatic aortic (valve) stenosis |
| I51.3 | Intracardiac thrombosis not elsewhere classified |
| 410.6 | True post infarct unspec |
| I46.8 | Cardiac arrest due to other underlying condition |
| J46 | |
| 415.13 | Saddle embol pulmon art |
| 'J43.9' | |
| B39.9 | Histoplasmosis unspecified |
| 424.99 | Endocarditis NEC |
| 492.0 | Emphysematous bleb |
| I11.9 | Hypertensive heart disease without heart failure |
| I27.24 | Chronic thromboembolic pulmonary hypertension |
| I27.83 | Eisenmenger’s syndrome |
| I42.9 | Cardiomyopathy unspecified |
| 491.21 | Obs chr bronc w(ac) excal |
| 421.9 | Ac/subac endocardit NOS |
| J45.50 | Severe persistent asthma uncomplicated |
| 423.0 | Hemopericardium |
| I06.0 | Rheumatic aortic stenosis |
| I45.0 | Right fascicular block |
| 397 | Tricuspid valve disease |
| 420 | Ac pericardit in oth dis |
| I25.810 | Atherosclerosis of coronary artery bypass graft(s) without angina pectoris |
| I44.30 | Unspecified atrioventricular block |
| I22.0 | Subsequent ST elevation (STEMI) myocardial infarction of anterior wall |
| 426.4 | Rt bundle branch block |
| J67.8 | Hypersensitivity pneumonitis due to other organic dusts |
| I27.9 | Pulmonary heart disease unspecified |
| 416 | Prim pulm hypertension |
| 413.9 | Angina pectoris NEC/NOS |
| 428.0 | CHF NOS |
| I40.8 | Other acute myocarditis |
| 428.2 | Systolic hrt failure NOS |
| 428 | CHF NOS |
| J67.1 | Bagassosis |
| I50.22 | Chronic systolic (congestive) heart failure |
| I48.3 | Typical atrial flutter |
| I36.8 | Other nonrheumatic tricuspid valve disorders |
| 423 | Hemopericardium |
| 415.12 | Septic pulmonary embolism |
| I23.7 | Postinfarction angina |
| I25.118 | Atherosclerotic heart disease of native coronary artery with other forms of angina pectoris |
| J43.9 | Emphysema unspecified |
| I06.9 | Rheumatic aortic valve disease unspecified |
| 404.9 | Hy hist kd NOS I-IV w/o hf |
| I47.9 | Paroxysmal tachycardia unspecified |
| I45.5 | Other specified heart block |
| 429.0 | Myocarditis NOS |
| I25.720 | Atherosclerosis of autologous artery coronary artery bypass graft(s) with unstable angina pectoris |
| I25.812 | Atherosclerosis of bypass graft of coronary artery of transplanted heart without angina pectoris |
| I25 | Atherosclerotic heart disease of native coronary artery without angina pectoris |

Continued on next page
| ICD code  | Description                                                                 |
|----------|-----------------------------------------------------------------------------|
| I50.40   | Unspecified combined systolic (congestive) and diastolic (congestive) heart failure |
| I49.3    | Ventricular premature depolarization                                         |
| 413.0    | Angina decubitus                                                            |
| 425.9    | Second cardiomyopathy NOS                                                   |
| I28.9    | Disease of pulmonary vessels unspecified                                     |
| I48.1    | Longstanding persistent atrial fibrillation                                  |
| 396      | Mitral/aortic stenosis                                                      |
| I42      | Dilated cardiomyopathy                                                      |
| I21.4    | Non-ST elevation (NSTEMI) myocardial infarction                             |
| I34.2    | Nonrheumatic mitral (valve) stenosis                                        |
| I36.0    | Nonrheumatic tricuspid (valve) stenosis                                      |
| I25.119  | Atherosclerotic heart disease of native coronary artery with unspecified angina pectoris |
| I50.811  | Acute right heart failure                                                   |
| 404.01   | Mal hyp ht/kd I-IV w hf                                                      |
| 493.22   | Ch obst asth w (ac) exac                                                    |
| 420.9    | Acute pericarditis NOS                                                      |
| 424.91   | Endocarditis in oth dis                                                     |
| I26.02   | Saddle embolus of pulmonary artery with acute cor pulmonale                 |
| 422.93   | Toxic myocarditis                                                           |
| 490      | Bronchitis NOS                                                              |
| I27.89   | Other specified pulmonary heart diseases                                     |
| I25.6    | Silent myocardial ischemia                                                  |
| 426.7    | Anomalous av excitation                                                     |
| I21      | ST elevation (STEMI) myocardial infarction involving left main coronary artery |
| I26.99   | Other pulmonary embolism without acute cor pulmonale                        |
| I42.0    | Dilated cardiomyopathy                                                      |
| I42.2    | Other hypertrophic cardiomyopathy                                           |
| I07.8    | Other rheumatic tricuspid valve diseases                                     |
| I49.9    | Cardiac arrhythmia unspecified                                               |
| I34.0    | Nonrheumatic mitral (valve) insufficiency                                    |
| 404.1    | Ben hy ht/kd I-IV w/o hf                                                    |
| 415.11   | Iatrogen pulm emb/infarc                                                    |
| I23      | Hemopericardium as current complication following acute myocardial infarction |
| 416.9    | Chr pulmon heart dis NOS                                                    |
| 410.12   | AMI anterior wall subseq                                                     |
| I27.20   | Pulmonary hypertension unspecified                                           |
| I23.6    | Thrombosis of atrium auricular appendage and ventricle as current complications following acute myocardial infarction |
| I40.9    | Acute myocarditis unspecified                                               |
| I35.0    | Nonrheumatic aortic (valve) stenosis                                        |
| I31.1    | Chronic constrictive pericarditis                                           |
| 396.1    | Mitral stenos/aort insuf                                                    |
| I25.758  | Atherosclerosis of native coronary artery of transplanted heart with other forms of angina pectoris |
| I30.0    | Acute nonspecific idiopathic pericarditis                                    |
| I25.709  | Atherosclerosis of coronary artery bypass graft(s) unspecified with unspecified angina pectoris |
| I27.22   | Pulmonary hypertension due to left heart disease                             |
| 493.02   | Ext asthma w(acute) exac                                                    |
| 493.1    | Intrinsic asthma NOS                                                        |
| 115.93   | Histoplasmosis pericard                                                     |
| I25.5    | Ischemic cardiomyopathy                                                     |
| 393      | Chr rheumatic pericard                                                      |
| I40      | Infective myocarditis                                                      |
| 494.0    | Bronchiectas w/o ac exac                                                    |
| 396.0    | Mitral/aortic stenosis                                                      |
| 493.81   | Exercise ind bronchospasm                                                   |
| 416.8    | Chr pulmon heart dis NEC                                                     |
| I35.8    | Other nonrheumatic aortic valve disorders                                    |
| I51.89   | Other ill-defined heart diseases                                            |
| I25.728  | Atherosclerosis of autologous artery coronary artery bypass graft(s) with other forms of angina pectoris |
| J67.3    | Suberosis                                                                   |
| 428.21   | Ac systolic hrt failure                                                      |
REFERENCES

[1] Granger, C. W. J. & Joyeux, R. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–29.

[2] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).

[3] General equivalence mappings. URL https://www.cms.gov/Medicare/Coding/ICD10/downloads/ICD-10_GEM_fact_sheet.pdf.

[4] Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley-Interscience, New York, NY, USA, 1991).

[5] Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86 (1951). URL https://doi.org/10.1214/aoms/1177729694.

[6] Doob, J. *Stochastic Processes*. Wiley Publications in Statistics (John Wiley & Sons, 1953). URL https://books.google.com/books?id=KvJQAAAAMAAJ.

[7] Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002). URL http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

[8] Hopcroft, J. E. *Introduction to automata theory, languages, and computation* (Pearson Education India, 2008).

[9] Klenke, A. *Probability theory: a comprehensive course* (Springer Science & Business Media, 2013).

[10] Doob, J. *Stochastic processes*. Wiley publications in statistics (Wiley, 1990). URL https://books.google.com/books?id=7Bu8jgEACAAJ.

[11] Chattopadhyay, I. & Ray, A. Structural transformations of probabilistic finite state machines. *International Journal of Control** **81**, 820–835 (2008).

[12] Chattopadhyay, I. & Lipson, H. Data smashing: uncovering lurking order in data. *Journal of The Royal Society Interface** **11**, 20140826 (2014).

[13] Bondy, J. & Murty, U. *Graph theory* (2008). *Grad. Texts in Math* (2008).

[14] Vidyasagar, M. *Hidden markov processes: Theory and applications to biology*, vol. 44 (Princeton University Press, 2014).

[15] Kai, L. C. *Markov Chains: With Stationary Transition Probabilities* (Springer-Verlag, 1967).

[16] Trahtman, A. N. The road coloring and Černý conjecture. In *Proc. of Prague stringology conference*, vol. 1, 12 (Citeseer, 2008).

[17] Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

[18] Matthews, A. G. d. G., Hensman, J., Turner, R. & Ghahramani, Z. On sparse variational methods and the kullback-leibler divergence between stochastic processes. *Journal of Machine Learning Research** **51**, 231–239 (2016).

[19] Hardy, G. Divergent series, with a preface by je littlewood and a note by ls bosanquet, reprint of the revised (1963) edition. *Éditions Jacques Gabay, Sceaux* (1992).

[20] Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, 3146–3154 (2017).