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

Dmytro Onishchenko[1], author2[1], author3[1], author4[1] and Ishanu Chattopadhyay[1,2,3]★

[1]Department of Medicine, University of Chicago, Chicago, IL USA
[2]Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL USA
[3]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL USA
[4]Department of Medicine, University of Chicago, Chicago, IL 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 cause[1–3]. The poor prognosis of IPF (mean survival less than 2-5 years post diagnosis[4]), combined with its worldwide prevalence greater than all but the most common cancers[5], 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, a non-specific symptomology, a limited understanding of the phenotypic and genetic markers for early-stage IPF, and the need for invasive procedures for confirmatory diagnosis. In this study we introduce a new screening tool that requires no new tests and blood-work, 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 Pulmonary Fibrosis Co-Morbidity Risk Score (PCoR), 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=452997). In out-of-sample validation, we demonstrate that PCoR would have identified 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, regardless of the patient's sex. Our out-of-sample accuracy estimates indicate that on average, out of every 100 patients screened with PCoR, we are wrong about their eventual IPF status in less than five cases. The precise and personalized PCoR score can potentially aid care providers to more selectively flag patients for detailed diagnostic referral, substantially improving patient outcomes and quality-of-life, as well as efficiency of healthcare resource utilization. Additionally, earlier diagnosis can improve the odds of getting a lung transplantation, and in general, accelerate access to interventions and palliative care[6,7]. The PCoR score demonstrates the importance of sophisticated pattern discovery in burgeoning patient databases to aid clinical decision-making, in the context of serious illnesses with yet uncharted comorbid patterns and causal mechanisms, to significantly improve patient outcomes with little or no additional burden on patients, care givers and health care resources.

IDIOPATHIC pulmonary fibrosis (IPF) is an irreversible, progressive, debilitating, and ultimately lethal fibrosing interstitial lung disease (ILD) of unknown cause[1–3]. Before the introduction into everyday practice of the anti-fibrotics nintedanib and pirfenidone in 2014, the typical survival of patients with IPF was 2-5 years from the time of diagnosis[4]. This prognosis has been been characterized as "little better than that for inoperable lung cancer"[5], and although the disease is considered rare, as of 2014, IPF had a greater worldwide prevalence than did all but seven most common cancers[5]. Over the past decade, timely, efficient, and confident diagnosis of IPF has been recognized as a major public health challenge worldwide[2,5,6,8–15]. In this study, we put forward a novel approach to predict future IPF diagnoses via algorithmically identifying subtle patterns in the time, nature and ordering of past medical encounters of individual patients.

Identifying IPF cases is a complex, multi-step, and often multi-year process[6,11,13–17]. A usually necessary but not always sufficient condition is referral to a pulmonologist or referral for high-resolution chest computed tomography (HRCT), often at an expert center[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 to anamestic and biochemical rule-out of other forms of interstitial lung disease (ILD)[2,6,11]. For such referral to happen, one or more of a number of scenarios must unfold: 1) Patients must recognize the chronicity, progression, or both of respiratory symptoms and the resultant need for medical attention[6,7,15,17]. 2) Primary care practitioners must appreciate the importance of these symptoms, of signs of possible fibrosis on auscultation, or of both[5,18]. 3) Radiologists, not infrequently non-specialists in chest imaging, must note incidental findings of interstitial lung abnormalities (ILA) or ILD on thoracic or abdominal CT[10,14,19,20]. Alternatively, a pulmonologist or a HRCT referral may only take place[20],after an emergency room visit or hospitalization due to acute exacerbation of IPF[15].

Hence currently, IPF diagnosis often entails multiple physician visits and repeated, sometimes invasive tests, and is often characterized by delay and misdiagnosis[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[14] of two US academic medical center cohorts found around a quarter of patients with IPF or other forms of ILD to have waited ≥ 1 year after a CT finding of ILD to see a pulmonologist, while a Medicare claims study[13] found that more than a third of patients saw a pulmonologist > 3 years before IPF diagnosis. A medical chart analysis10 and a patient survey[6] suggested misdiagnosis rates approaching or exceeding 40%. Notably, 28% of respondents in a 600-patient survey[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 four main causes. First, the most common clinical symptoms of the disease[7,17], insidiously progressive chronic exertional dyspnea and chronic, often mild cough, are highly non-specific. These symptoms are easily attributed by patients to age or deconditoning[6], or by physicians to more common respiratory diseases, *e.g.*, asthma, pneumonia, bronchitis, or chronic obstructive pulmonary disease (COPD), or to heart disease[5,6,15,16]. Important risk factors for IPF, namely, older age, male sex, and cigarette smoking[1], are similarly "non-specific". Second, there is still limited understanding and characterization of phenotypic and genetic findings associated with "early-stage IPF"[10]. Third, the current diagnostic hallmark of IPF, UIP on HRCT or histology[2], is a late-stage finding[3]. Moreover, UIP may be confirmed via relatively invasive procedures requiring specialized interpretation, *e.g.*, HRCT or surgical lung biopsy[6]. Lastly, no validated or easily-applicable screening modalities currently exist for IPF[9].

In this study we introduce a novel approach that has the potential to address these key challenges: we introduce a tool that requires no new tests or blood-work, 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 (PCoR), 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® Commercial Claims and Encounters Database for the years 2003-2018[22] (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, making sure that selected patients have some reasonable length of medical history recorded in the dataset (≥ 3 years). The geographical distribution of the patients in our selected cohort(s) is illustrated in Fig. 1a. Panel b illustrates the age distribution of IPF diagnosis which shows an mean around 68 years. This is consistent with the reported mean onset age for IPF (≈ 66 years[4]). We also 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.

We view the task of predicting a future IPF diagnosis as binary classification problem: time-stamped sequences of diagnostic n codes are to be classified into positive and control categories, where the "positive" category refers to patients diagnosed with IPF 1 year (See Tables V) from the point of screening. We also consider earlier screening upto 4 years before the actual diagnosis, as described later. The control cohort comprises patients

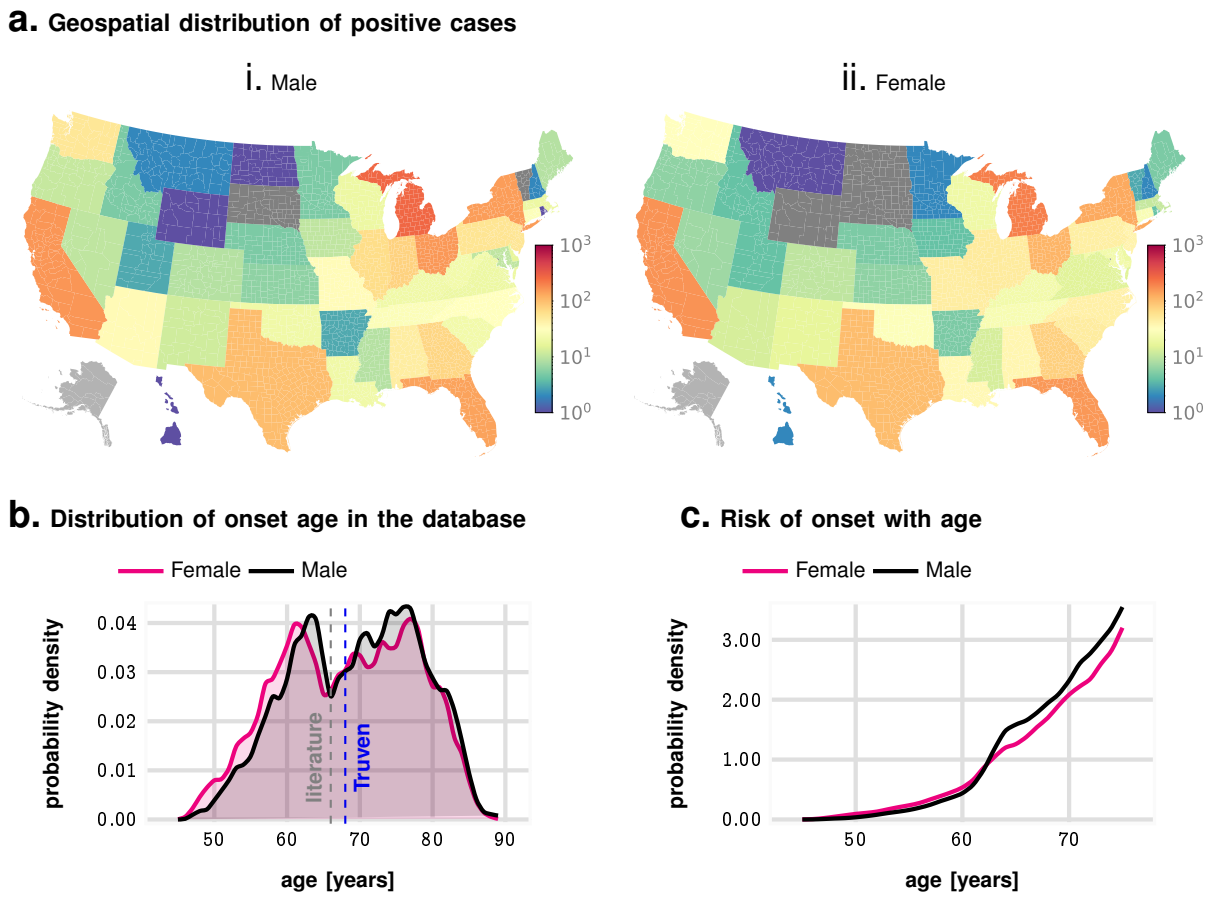

Fig. 1: **Statistical characteristics of the Truven database.** Panel a shows where our patients originate from. The geospatial distribution is correlated with population density, as expected. Panel b illustrates the distribution of patient diagnostic ages in the database. Note that the mean age of onset 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).

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 ultimately end up with $n = 452,997$ patients, with $2,536$ patients in the positive group and $450,461$ patients in the control group. The cohort sizes are described in Table II. In summary, we consider approximately 42 million diagnostic codes (with over 46K unique codes) for both genders (See Table III for detailed enumeration) in this analysis.

How we identify a patient in the positive cohort is important. The diagnostic codes specifically for IPF are 516.31 (ICD9) and J84.112 (ICD10). We use a slightly broader definition in our analysis, including also five additional ICD10 codes that indicate a diagnosis of pulmonary fibrosis without a known cause (See Table V). 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 include results in the Supplementary text with the narrow target definition with only 516.31 and J84.112 codes, where we achieve out-of-sample AUC exceeding 80%, albeit with a reduced number of positive cases.

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 VI, which identify the top known co-morbidities[23]. 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.

Modeling & Prediction

The significant diversity of diagnostic codes, along with the sparsity of codes per patient (XX diagnostic per patient on average) makes this a difficult learning problem. We proceed by partitioning the disease spectrum into 51

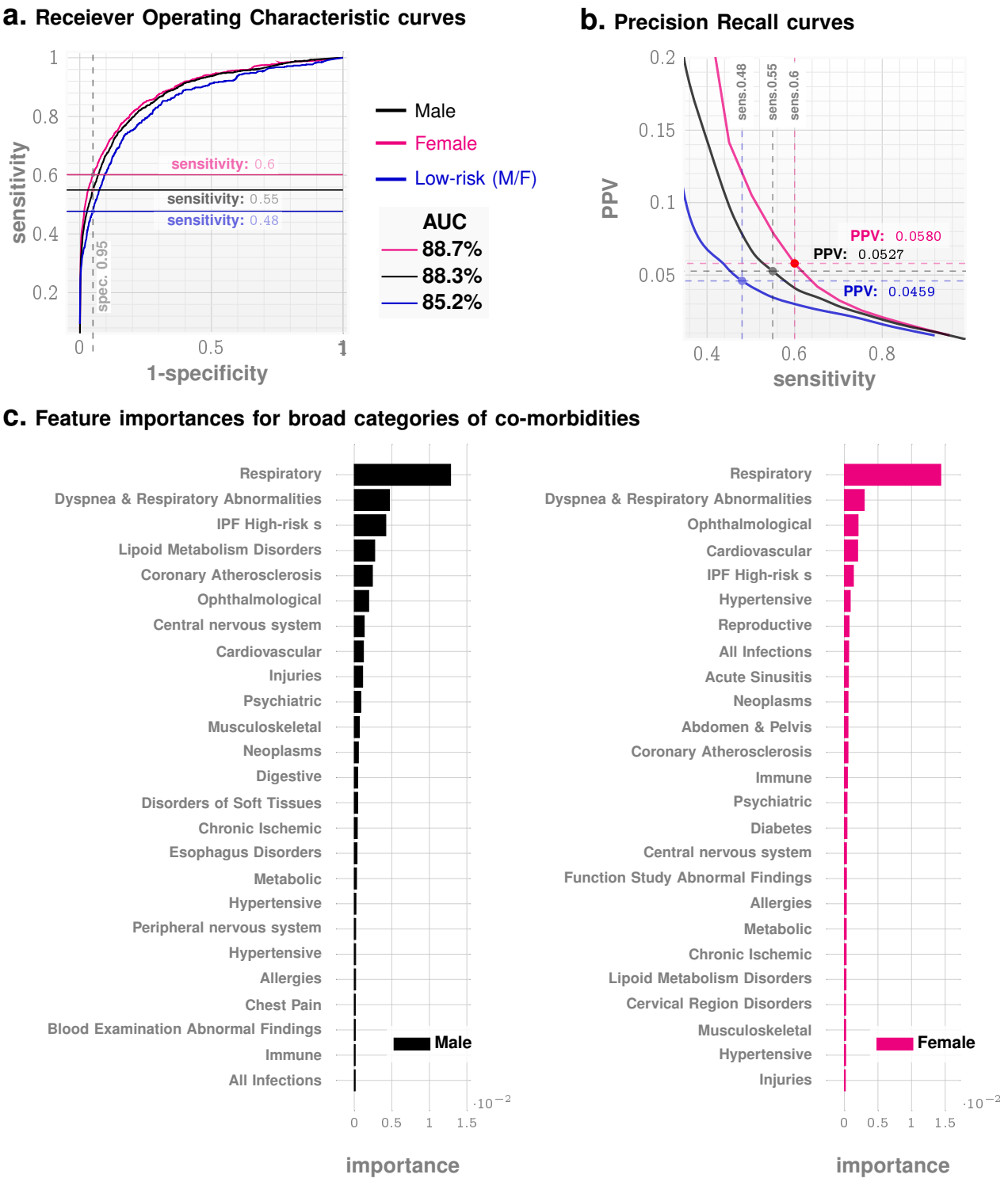

Fig. 2: **Predictive performance of PCoR for IPF diagnosis 1 year in future.** Panels a and b show the out-of-sample ROC and precision-recall curves for diagnosis 1 year from the point of screening. We achieve AUCs $> 88\%$ for both sexes, with sensitivity $> 55\%$ at 95% specificity. The PPV achieved is $> 5\%$ for both males and close to 6% for females at 95% speciificity. Note that the theoretical maximum PPV possible at this specificity for IPF population prevalence of 0.004945[21] (2011, US) is $\approx 9\%$ (achieved at 100% sensitivity). 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 Tab;e VI), cardiovascular issues, infections, hypertensive abnormalities appear in both males and females, with altered ranking.

broad categories, *e.g.* infectious diseases, immunologic disorders, and endocrinal disorders (See SI Tab. **??** for a detailed enumeration of these categories). Some of these categories comprise of a relatively large number of diagnostic codes aligning roughly with the categories defined within the ICD framework[24]. The remaining categories represent groups of one or more codes that might have some known or suspected association with

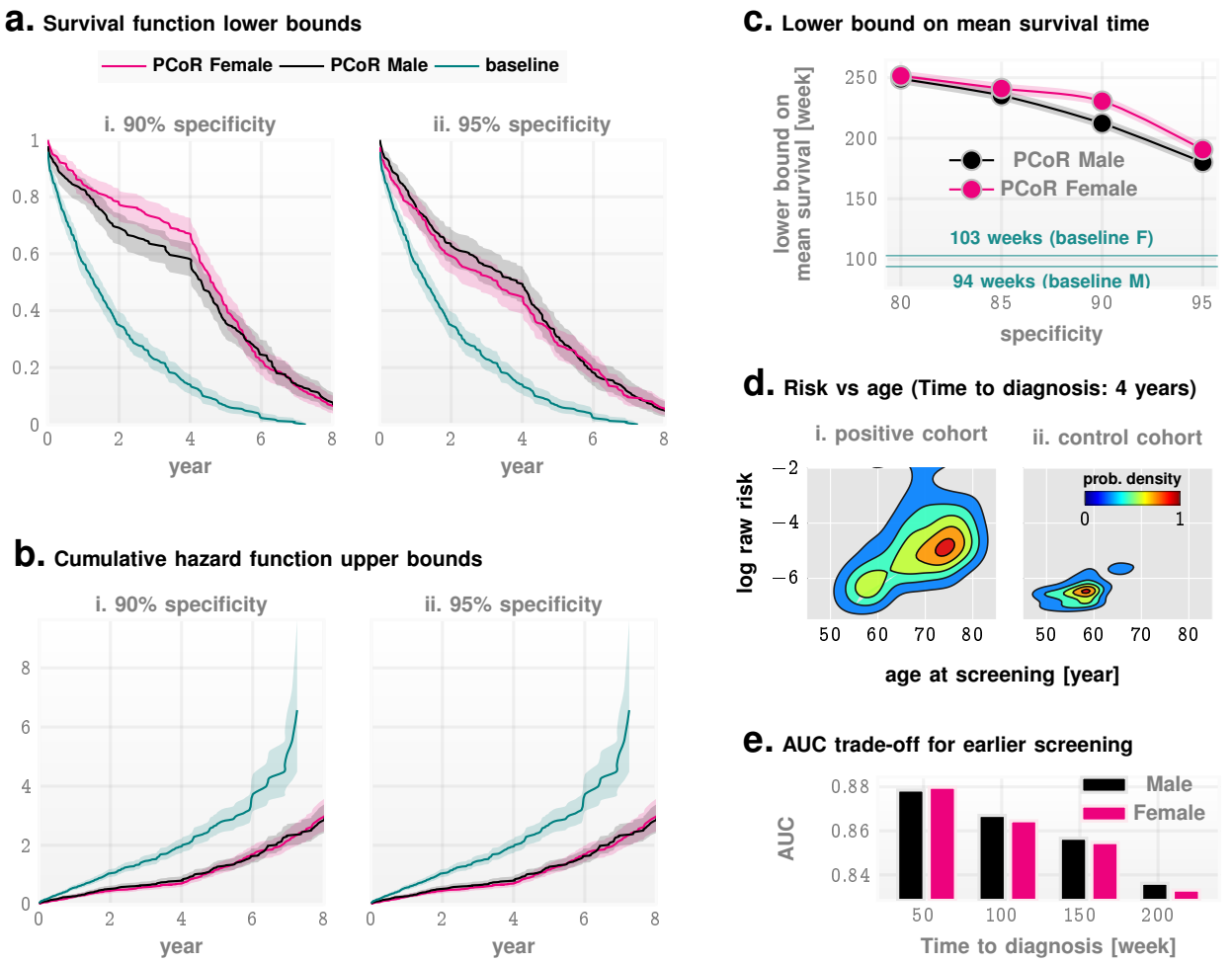

Fig. 3: **Survival function, hazard rate improvements, and predictive performance for early screening. Panel a** shows the estimated lower bounds on the survival function at two specificity levels (90 and 95%), and **panel b** the estimated upper bounds on hazard rates. The lower (upper) bounds are estimated 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 specificty at which PCoR is operated. The baselines shown are calculated from our patient databse, 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 PCoR 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.

pulmonary disorders. Each of the diagnostic categories yield a single time series over weeks (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 time series of events, which are compressed into specialized Hidden Markov Models known as Probabilistic Finite Automata[25]. 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 **??**)) 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)[26], 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[27] between probability distributions to a generalized divergence between possibly non-iid stochastic processes (See Supplementary

TABLE I: Inclusion/Exclusion, Positive/Control Criteria & Cohort Definitions

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

⋆Medical records denote time stamped history of diagnostic codes

‡Target codes identify the target condition (IPF or Pulmonary Fibrosis without known cause)

TABLE II: Cohort Sizes

gender	age group	n	n_{positive}	n_{control}
M	45-55	57160	66	57094
M	55-65	101105	402	100703
M	65-75	32392	459	31933
M	75-	13055	418	12637
F	45-55	75370	98	75272
F	55-65	122903	373	122530
F	65-75	35936	371	35565
F	75-	15076	349	14727
Total		452997	2536	450461

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: Summarized MCoR Performance for Different Cohorts & Prediction Problems

gender	AUC	AUC low-risk	AUC high-risk	specificity	sensitivity	sensitivity low-risk	sensitivity high-risk	accuracy
M	0.883	0.853	0.884	0.95	0.56	0.50	0.58	0.95
F	0.887	0.852	0.894	0.95	0.60	0.48	0.61	0.95

†Calculated at 95% specificity and prevalence of 0.004945. AUC estimates have uncertainty of ± 0.01 at 95% confidence

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

ICD code	description
516.31	Idiopath pulmon 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

Tex, Section **??**).

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"). These 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). Ultimately, we compute a total of 667 features for each patient, which is then used to train a standard gradient boosting classifier[28] aiming to map individual patients to a raw risk score. We randomly choose 50% of our patients for training with the rest held-out as a validation set. 50% of the training data is used for PFSA inference, and the rest for training the gradient boosting classifier.

TABLE VI: High risk comorbidities which define our hig-risk 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).

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 (See Table IV), 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 PCoR 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 PCoR 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 dirty 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 two 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. And 2) the ability of our models to maintain high predictive performance for an eventual diagnosis further into future, upto 4 years.

Our main prediction results are presented in Fig. 2a-b, which illustrate the ROC and the precision-recall curves respectively, 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). We achieve accuracies $> 95\%$ (See Table IV), which indicates the overall fraction of correct predictions. We achieve a PPV of $5-6\%$, which compares favorably with the maximum theoretical value of $\approx 9\%$ at 95% specificity, given the low population prevalence of IPF at 0.004945 in the US[21].

Thus, to summarize: our predictive pipeline detects about 55-60 out of every 100 patients who are going to have a diagnosis in 1 year, and out of 100 positive flags we have about 5-6 true positives. The high false positive rate here partly arises from the low prevalence, since even with 100% sensitivity at 95% specificity, we would be able to get only 9 true positives out of every 100 flags on average. The accuracy metric indicates that out of every 100 patients, we are wrong in less than 5 cases about their risk status (positive or control), irrespective of the sex of the patients, highlighting the potentially high clinical significance of the PCoR 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 PCoR score.

A key metric to evaluate the potential impact of the PCoR score is to estimate the expected change in the survival functions via a Kaplan-Meyer (KM) analysis[29], 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 int 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 PCoR with acceptable reliability to predict diagnoses four years into the future. To illustrate how the PCoR 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 PCoR 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 VI). The results are tabulated in Table IV, 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. The ROC and the precision-recall curves for the

low risk sub-cohort is shown in Fig. 2a-b, where we note that our sensitivity drops to 48% at 95% specificity, with the PPV dropping slightly to 4.6%. 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, and out of 100 positive flags we have about 5 true positives on average. And our accuracy in this cohort is still about 95%, so as before, our decision is wrong in the case of less than 5 out of every 100 patients.

DISCUSSION

The present study describes the development and performance of PCoR 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), PCoR 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, PCoR 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, PCoR 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, PCoR expands the data available to both primary care practitioners and expert IPF diagnosticians. The score reflects a sophisticated, highly-detailed automated analysis of cormorbidities, considering more than 600 features related to the incidence, timing of individual diagnostic codes. PCoR 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[1,2,30]. Adding comorbidity as a new dimension of assessment, albeit in much simpler fashion[31], is an approach that has proved fruitful in COPD management[32].

PCoR may be expected to contribute in two main ways to increase timeliness and ease of IPF identification. First, PCoR 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[5,7,10,14,17,18,20,33]. As these are relatively large groups, PCoR 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. PCoR 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[10,34]. Indeed, given the severity and actionability of the IPF diagnosis, the frequency of smoking history among patients with IPF1 , and the non-invasive nature, applicabilty, speed, and low cost of PCoR, 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 PCoR in these screening settings merit prospective evaluation.

Second, PCoR 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[10**?**] . Hence PCoR might help patients without UIP on HRCT to avoid more invasive tests, especially surgical lung biopsy[2,35], and/or may increase clinicians' diagnostic confidence. The effect of PCoR use as a diagnostic aid on healthcare resource utilization and on diagnostic confidence in IPF classification also warrants prospective assessment.

PCoR 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[36]. 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[14].

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[1,6,11–16,37]. 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[1,6]. In the meanwhile, quicker IPF diagnosis will accelerate patients' access to interventions that may improve lung function and quality-of-life, namely supplemental oxygen, pulmonary rehabilitation, and palliative care[6,7], 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[23]. Indeed, a chart review at a German tertiary referral center[38] (N=272) 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[38]. 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)[2?] , while others may at least in some instances represent common mis-diagnoses in patients with IPF, *e.g.*, COPD, asthma, or pneumonia[6,15,16].

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 PCoR robust statistical power to capture subtle patterns of comorbidity that would be indiscernible even in traditional multicenter cohorts[34]. 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[39,40]. 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[40]. 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[3,41,42]. However, inclusion of patients without ILD among our IPF cases would of course be expected to decrease the accuracy and utility of PCoR. It would be desirable to perform case validation studies in our databases, which might provide data enabling PCoR 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[43], and to a tendency of older and sicker patients to migrate from commercial to government insurance plans[44]. 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.36 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[44].

Conclusion

We developed and validated PCoR, 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 patents with IPF. Thus PCoR supplements the functional and morphological data regarding the lung that have been the mainstay in evaluating potential IPF cases. PCoR 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 PCoR 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, PCoR may offer utility as a diagnostic aid to pulmonolgists, 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 PCoR 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] Sauleda, J., Núñez, B., Sala, E. & Soriano, J. B. Idiopathic pulmonary fibrosis: epidemiology, natural history, phenotypes. *Medical Sciences* **6**, 110 (2018).

[22] Hansen, L. The truven health marketscan databases for life sciences researchers. *Truven Health Ananlytics IBM Watson Health* (2017).

[23] Raghu, G., Amatto, V., Behr, J. & Stowasser, S. Comorbidities in idiopathic pulmonary fibrosis patients: a systematic literature review. *Eur Respir J* **46**, 1113–30.

[24] Organization, W. H. *et al.* International classification of diseases—ninth revision (icd-9). *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire* **63**, 343–344 (1988).

[25] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013). URL http://rsta.royalsocietypublishing.org/content/371/1984/20110543.full.pdf.

[26] Huang, Y. & Chattopadhyay, I. Data smashing 2.0: Sequence likelihood (sl) divergence for fast time series

comparison. *arXiv preprint arXiv:1909.12243* (2019).
[27] Cover, T. M. & Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).
[28] Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, 3146–3154 (2017).
[29] Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**, 457–481 (1958).
[30] Hyldgaard, C., Hilberg, O. & Bendstrup, E. How does comorbidity influence survival in idiopathic pulmonary fibrosis? *Respir Med* **108**, 647–53.
[31] 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.
[32] 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.
[33] Oldham, J., Adegunsoye, A. & Khera, S. Underreporting of interstitial lung abnormalities on lung cancer screening computed tomography. *Ann Am Thorac Soc* **15**, 764–6.
[34] 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.
[35] 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.
[36] Kropski, J. Biomarkers and early treatment of idiopathic pulmonary fibrosis. *Lancet Respir Med* **7**, 725–7.
[37] Farrand, E., Iribarren, C. & Vittinghoff, E. Impact of idiopathic pulmonary fibrosis on longitudinal health-care utilization in a community-based cohort of patients. *Chest* .
[38] Kreuter, M., Ehlers-Tenenbaum, S. & Palmowski, K. Impact of comorbidities on mortality in patients with idiopathic pulmonary fibrosis. *PLoS One* **11**, 0151425.
[39] 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.
[40] 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.
[41] Inoue, Y., Kaner, R. & Guiot, J. Diagnostic and prognostic biomarkers for chronic fibrosing interstitial lung diseases with a progressive phenotype. *Chest* **158**, 646–59.
[42] George, P., Spagnolo, P. & Kreuter, M. Progressive fibrosing interstitial lung disease: clinical uncertainties, consensus recommendations, and research priorities. *Lancet Respir Med* **8**, 925–34.
[43] Mortimer, K. *et al.* Characterizing idiopathic pulmonary fibrosis patients using us medicare-advantage health plan claims data. *BMC Pulm Med* **19**, 11.
[44] 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. Comapring **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.

TABLE VII: Feature Definitions (Total number of features used: 667)

feature name	explanation	n_{features}
feature-phenotype scores relative to phenotype score	Mean p-score of **feature-phenotype** codes within sequence divided by general p-score of **feature-phenotype**	51
feature-phenotype scores relative to whole score	Mean p-score of **feature-phenotype** codes within sequence divided by mean p-score of all codes in the record	51
aggregation score	aggregation of the p-scores 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
dynamics of mean score	mean p-score of second half of the record divided by mean p-score of first half of the record	1
dynamics of st.dev score	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
dynamics of score range	range of p-scores of second half of the record divided by range of p-scores of first half of the record	1
dynamics of score skew	skew of p-scores of seconf half of the record divided by skew of p-scores of first half of the record	1
aggregation relative to phn score	aggregation of all **feature-phenotype**'s mean scores divided by corresponding general p-score of **feature-phenotype**	6
aggregation relative to whole score	aggregation of all **feature-phenotype**'s mean scores divided by mean p-score of all codes in the record	6
feature-phenotype proportion	Ratio of number of weeks with the codes of a given phenotype to the total number of weeks in sequence	51
feature-phenotype prevalence	Ratio of number of weeks with the codes of a given phenotype to the number of weeks with any diagnosis code recorded	51
feature-phenotype first incident	Time interval from observation date to the first phenotype code, normalized by record length	51
feature-phenotype last incident	Time interval from observation date to the last phenotype code, normalized by record length	51
feature-phenotype mean position	Mean time position of phenotype codes in the record, normalized by record length	51
feature-phenotype streak	Length of the longest uninterrupted subsequence of weeks with the codes of a given phenotype recorded	51
Max/Mean/Std/Range intermission	Maximum/Mean/Standard Deviation/Range of the lengths of subsequences of consequent weeks with codes	4
Max/Mean/Std cluster	Maximum/Mean/Standard Deviation of the lengths of subsequences of consequent weeks without codes	3
Max/Std/Range prevalence	Maximum/Standard Deviation/Range of the phenotype prevalences	3
Density of DX Record	Proportion of weeks in a record observed where at least one DX code was recorded	1
feature-phenotype	Sequence Likelihood Defect for a given phenotype	51
feature-phenotype neg llk	Negative LogLikelihood score for a given phenotype	51
feature-phenotype pos llk	Positive LogLikelihood score for a given phenotype	51
feature-phenotype llk ratio	Ratio of Positive to Negative LogLikelihood score for a given phenotype	51
Max Δ	Mean Sequence Likelihood Defect	1
Std Δ	Standard Deviation of Sequence Likelihood Defects	1
Range Δ	Range of Sequence Likelihood Defects	1
Mean neg llk	Mean Negative LogLikelihood score	1
Range neg llk	Range of Negative LogLikelihood score	1
Std. deviation neg llk	Standard Deviation of Negative LogLikelihood score	1
Mean pos llk	Mean Positive LogLikelihood score	1
Range pos llk	Range of Positive LogLikelihood score	1
Std. deviation pos llk	Standard Deviation of Positive LogLikelihood score	1
Mean llk ratio	Mean LogLikelihood score ratio	1
Range llk ratio	Range of LogLikelihood score ratio	1
Std. deviation llk ratio	Standard Deviation of LogLikelihood score ratio	1
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
age at 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

†Δ: Sequence Likelihood Defect (See Methods)

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

pos loglikelihood: loglikelhood of observed sequence being generated by the model inferred from positive (See Methods)