Identification of potentially undiagnosed patients with nontuberculous mycobacterial lung disease using machine learning applied to primary care data in the UK

Orla M. Doyle[1,7], Roald van der Laan[2,7], Marko Obradovic[2,7], Peter McMahon[3], Flora Daniels[4], Ashley Pitcher[5] and Michael R. Loebinger[6]

Affiliations: [1]Predictive Analytics, Real World Analytical Solutions, IQVIA, London, UK. [2]Insmed Utrecht, Utrecht, The Netherlands. [3]Real-World Insights, IQVIA, London, UK. [4]Real-World Insights, IQVIA, Basel, Switzerland. [5]Real-World Insights, IQVIA, Copenhagen, Denmark. [6]Royal Brompton and Harefield NHS Foundation Trust and Imperial College London, London, UK. [7]These authors contributed equally.

Correspondence: Marko Obradovic, Insmed, The Square 12, Am Flughafen, Frankfurt am Main, Germany. E-mail: marko.obradovic@insmed.com

@ERSpublications
Compared to random testing, machine learning improved detection of undiagnosed patients with NTMLD by almost a thousand-fold with AUC of 0.94 supporting the feasibility of using machine learning applied to primary care data to screen for undiagnosed NTMLD patients https://bit.ly/2WmT5nZ

Cite this article as: Doyle OM, van der Laan R, Obradovic M, *et al.* Identification of potentially undiagnosed patients with nontuberculous mycobacterial lung disease using machine learning applied to primary care data in the UK. *Eur Respir J* 2020; 56: 2000045 [https://doi.org/10.1183/13993003.00045-2020].

ABSTRACT Nontuberculous mycobacterial lung disease (NTMLD) is a rare lung disease often missed due to a low index of suspicion and unspecific clinical presentation. This retrospective study was designed to characterise the prediagnosis features of NTMLD patients in primary care and to assess the feasibility of using machine learning to identify undiagnosed NTMLD patients.

IQVIA Medical Research Data (incorporating THIN, a Cegedim Database), a UK electronic medical records primary care database was used. NTMLD patients were identified between 2003 and 2017 by diagnosis in primary or secondary care or record of NTMLD treatment regimen. Risk factors and treatments were extracted in the prediagnosis period, guided by literature and expert clinical opinion. The control population was enriched to have at least one of these features.

741 NTMLD and 112 784 control patients were selected. Annual prevalence rates of NTMLD from 2006 to 2016 increased from 2.7 to 5.1 per 100 000. The most common pre-existing diagnoses and treatments for NTMLD patients were COPD and asthma and penicillin, macrolides and inhaled corticosteroids. Compared to random testing, machine learning improved detection of patients with NTMLD by almost a thousand-fold with AUC of 0.94. The total prevalence of diagnosed and undiagnosed cases of NTMLD in 2016 was estimated to range between 9 and 16 per 100 000.

This study supports the feasibility of machine learning applied to primary care data to screen for undiagnosed NTMLD patients, with results indicating that there may be a substantial number of undiagnosed cases of NTMLD in the UK.

This article has supplementary material available from erj.ersjournals.com

Received: 9 Jan 2020 | Accepted after revision: 23 April 2020

Introduction

Nontuberculous mycobacterial lung disease (NTMLD) is a rare disease caused by nontuberculous mycobacteria (NTM), which are commonly found in water sources and soil [1–4]. NTMLD is becoming an increasing public health concern with reports of increasing incidence/prevalence worldwide in recent years, which may have been driven by better diagnostic tools, increasing awareness about the disease or a real underlying increase in infection rates [5–7]. The most recent estimates of the annual prevalence of NTMLD in Europe range from 3.3 to 6.0 cases per 100 000 [8–10].

The clinical symptoms of NTMLD include chronic and/or recurring cough, sputum production, fatigue, malaise, dyspnoea, fever, haemoptysis, chest pain and weight loss. However, the diagnosis of NTMLD is challenging as the clinical presentation is similar to common respiratory conditions such as bronchiectasis, COPD and asthma, which frequently co-exist with NTMLD [11–14]. NTMLD often worsens underlying structural lung disease, impairs quality of life and increases mortality and healthcare resource utilisation [15–19]. Given the chronic and progressive nature of NTMLD, a delay in diagnosis could expose patients to the risk of poorer outcomes as lung tissue damage worsens.

Machine learning methods hold considerable potential for finding undiagnosed NTMLD patients, as they can handle large number of clinical predictors and are sensitive to complex relationships. Recent studies provide support for this theory with promising results reported for the prediction of diagnoses and adverse events [20–23]. Machine learning algorithms "learn by example", where a patient's prediagnosis medical history can be mapped to a future outcome of interest (in this case, an NTMLD diagnosis). The algorithm is then tested on independent patients to validate its performance in identifying NTMLD patients who have not yet been diagnosed.

This study was designed to 1) describe the prevalence and incidence rates of diagnosed NTMLD patients; 2) characterise the prediagnosis features of NTMLD patients in primary care; and 3) assess the feasibility of using machine learning to identify undiagnosed NTMLD patients.

Methods

Study design

The IQVIA Medical Research Data (IMRD) UK electronic medical records (EMR) primary care database was used (a more detailed description of IMRD is contained in the supplementary material). Three criteria were applied to identify positive cases for NTMLD: 1) from IMRD using Read codes A310000 (pulmonary *Mycobacterium avium*-intracellular infection) and A310.00 (pulmonary mycobacterial infection); 2) by linking to secondary care records in Hospital Episode Statistics (HES) using International Classification of Diseases-10 code A31.0 (infection due to other mycobacteria) (secondary-care records were used only for case identification); and 3) based on treatment regimens of specific antibiotic combinations (⩾180 days; supplementary table S1), as identified through the British Thoracic Society (BTS) guidelines and clinical expert input [5]. If treatment with a drug appeared to end prior to the next prescription, continuous treatment was assumed if the gap was <30 days. For patients who were identified by more than one method, the earliest date was chosen as the date of NTMLD diagnosis. This multicriteria approach served to maximise NTMLD patient selection.

Control patients were selected from the IMRD population as those who did not have a record of NTMLD. Controls were also required to have a record of at least one of the selected predictors for NTMLD (supplementary table S2) to ensure that the predictive model would focus on learning to distinguish between different illnesses, rather than learning to distinguish between healthy and ill. Furthermore, controls were matched to cases in terms of the timing of their medical history, to ensure that differences in the distribution of predictors between cases and controls patients reflected "genuine" medical phenomena. From a random sample of 750 000 patients from the IMRD population, 112 784 patients met the inclusion and matching criteria.

Use of IQVIA Medical Research Data is approved by the UK Research Ethics Committee (reference number: 18THIN023-A1). In accordance with this approval, the study protocol was reviewed and approved by an independent Scientific Review Committee (SRC) (reference number: 18THIN023-A1). HES linked data has been reviewed and approved by the Independent Scientific Ethical Advisory Committee (ISEAC) under the same SRC reference number. ISEAC has an independent advisory function to IQVIA. ISEAC's primary purpose is to have oversight of all requests by IQVIA, honorary researchers or sub-licensees to perform analysis on de-identified data held under data sharing agreement with NHS Digital and other healthcare databases. In accordance with the terms and authorisation of data use the following is cited: "Copyright © 2019, re-used with the permission of The Health & Social Care Information Centre. All rights reserved". Reference made to the IMRD database is intended to be descriptive of the data asset licensed by IQVIA. The original data supporting this finding will be available from the corresponding author on request.

The study period was September 2003 to September 2017. The database is updated biannually. Each practice does not necessarily contribute to all updates, so a "last collection date" for each practice is factored in to calculations like incidence and prevalent counts. The index date for cases was defined as the most recent predictor event occurring prior to the first date of NTMLD diagnosis, to ensure only prediagnosis events were included. The index date for controls was temporally matched to index dates of NTMLD cases (±12 weeks). The length of the lookback period was set to a minimum of 3 years. Predictors such as age, sex and risk factors from literature sources, including BTS and American Thoracic Society guidelines were selected and verified by expert clinical opinion. A more detailed description of predictor selection is presented in the supplementary information and specifically supplementary table S2. Metrics were created to quantify the frequency and timing of the predictors; predictors that were not observed were assumed absent rather than missing. These metrics were then used to drive descriptive insights and predictive models. Information related to sputum tests was included in the patient journey description, but were excluded from the predictive model as these tests are most likely to occur when a diagnosis of NTMLD is imminent.

Prevalence calculation methods and assumptions

Yearly incidence rates were calculated as the number of newly diagnosed NTMLD patients with an index date in a given year, divided by the number of patients actively registered in IMRD in the same year (that is, the registration date occurred before the end of a given year and the transfer-out date did not occur before the start of that date). The estimated annual prevalence from primary care linked with IMRD was calculated as follows:

- NTMLD patients selected *via* diagnosis in IMRD or HES in a given year were assumed to be prevalent cases in the subsequent 2 years, while in NTMLD patients selected *via* treatment regimen prevalence was assumed for each year of therapy criterion only
- Patients were censored in years subsequent to transfer-out of the IMRD database, so that patients who had a subsequent record in HES were not counted if they had transferred out of the IMRD database
- The denominator was the number of active patients in a given year of the IMRD database.

Machine-learning methodology

A predictive algorithm was developed on prediagnosis medical history from the diagnosed NTMLD and non-NTMLD cohorts using gradient boosting trees, which is an ensemble of decision trees built successively to correct the errors made by previous trees [24]. This approach is particularly adept at capturing nonlinear associations and interactions and can handle missing data directly, and is reported to be highly performant across use-cases [25]. Specifically for missing data, the algorithm will decide how to handle a missing value for a given observation by learning which is the optimal choice of path in the individual decision trees ensuring that the way missing information is handled is also part of training the algorithm. As a final step, an ensemble of gradient boosting trees was built using bootstrap aggregation (bagging) [26]. A bagged ensemble is a collection of predictive models each trained on different sample of the training data. At testing the predictions are averaged across all models to increase their robustness. These averaged predictions are used as a risk score for NTMLD, that is, a score ranging from zero to one quantifying the risk of NTMLD with higher values associated with higher risk. The algorithm was developed using the R programming language (v3.4.0) and the MLR and XGBoost package [27]. The predictive algorithm was developed and validated on unique, nonoverlapping partitions of the data using five-fold cross-validation and evaluated using the area under the receiver operator characteristic curve (usually referred to as area under curve; AUC), and additionally precision-recall curves. The precision (or positive predictive value) of a model is calculated as the proportion of true cases retrieved by the model (true positives) to the total number of patients predicted to be cases by the model (sum of true positives and false positives), that is true positives/(true positives + false positives). Interpretation of the results focused on the precision-recall curves, which are robust to imbalanced data [28]. The precision-recall curves reported here were scaled to ensure that the performance was representative of what would be expected in the real-world clinical setting, *i.e.* the false positives, and hence precision, were scaled to represent the expected prevalence of NTMLD in the general population (five per 100 000).

Interpretation of predictive models

The feature importance for the bagged ensemble was calculated by averaging the feature importance across all individual models. Risk ratios were calculated for individual predictors whereby the risk score of the NTMLD group was compared across patients with different frequency rates or timings of predictors [29].

Results

Participants

1082 NTMLD cases were identified in the study using three criteria (IMRD-identified, HES-identified and treatment-identified; figure 1). In total, 741 NTMLD cases met the study inclusion/exclusion criteria; these

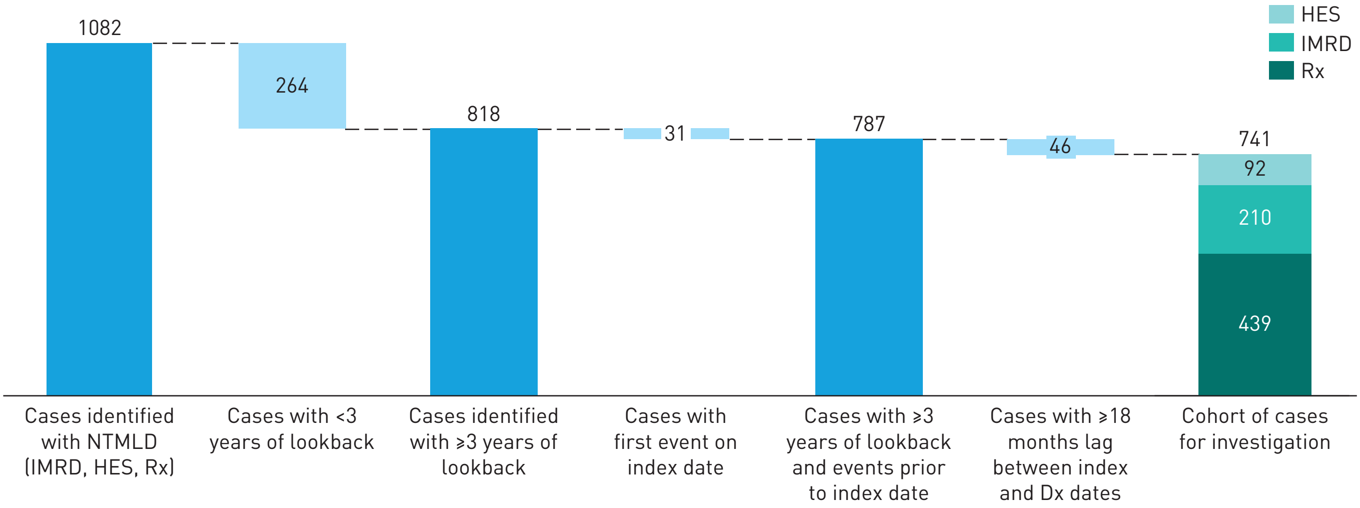

FIGURE 1 Nontuberculous mycobacterial lung disease (NTMLD) cases meeting inclusion/exclusion criteria. IMRD: IQVIA Medical Research Data; HES: Hospital Episode Statistics; Rx: treatment-identified; Dx: diagnosis.

included 210 cases identified from IMRD (31.9% of whom also met the treatment-identified criteria), 92 from HES (10.9% of which also met the treatment-identified criteria) and 439 from treatment criterion (9.8% of which also met the IMRD or HES-identified criteria). The control cohort comprised 112 874 patients.

Patient journey of NTMLD cases

The patient journey of the NTMLD cohort illustrates that risk factors and symptoms related to NTMLD are experienced in the years leading up to diagnosis of NTMLD (figure 2). A "tuberculosis (TB) diagnosis" was observed in 18.2% of IMRD-/HES-identified cases and 34.9% of treatment-identified cases. TB occurred on average within weeks of the first NTMLD diagnosis, which could reflect a suspicion of TB later to be confirmed as NTMLD, and therefore TB predictors were excluded from the model.

Comparison of NTMLD cases versus *controls*

Table 1 presents the patient demographics for the case and control cohorts. NTMLD cases were more likely to be older, female, a current or former smoker and have lower body mass index than controls.

The top 10 most frequent predictors are summarised in table 2. Seven of these are shared across both cohorts, indicating that the selected non-NTMLD patients are those with healthcare-seeking behaviour that is relevant to respiratory disorders. A higher proportion of NTMLD patients were exposed to treatments and the time since first observed prescription was longer.

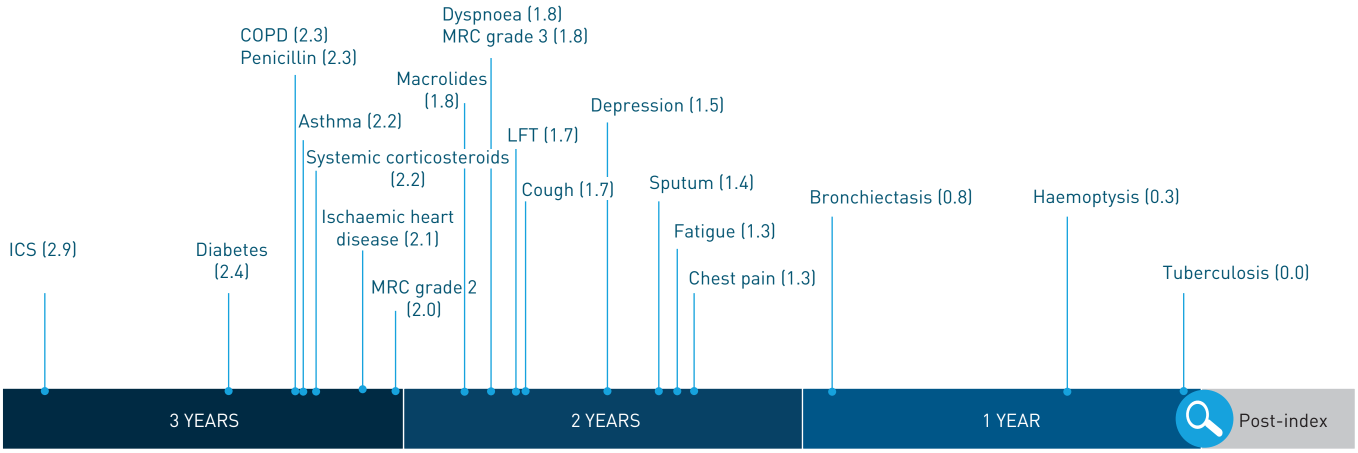

FIGURE 2 Prediagnosis primary care patient journey in nontuberculous mycobacterial lung disease (NTMLD). Numbers in brackets indicate the median time since index date in years for the top 20 most frequent predictors in the NTMLD case cohort. ICS: inhaled corticosteroids; MRC: Medical Research Council breathlessness score; LFT: lung function tests.

TABLE 1 Proportion of nontuberculous mycobacterial lung disease (NTMLD) cases and controls by demographic segments of interest

	NTMLD cases	Controls
Subjects n	741	112 874
Age at diagnosis years	59.8±19.2	48.5±24.7
<18	5.8	14.8
18–20	1.2	1.9
21–30	4.6	9.1
31–40	3.2	11.2
41–50	7.2	13.9
51–60	17.1	13.3
61–70	28.9	12.9
71–80	22.7	11.1
>80	9.3	11.8
Sex		
Female	53.8	46.6
Male	46.2	53.3
BMI		
Underweight	16.9	2.5
Normal	45.9	24.4
Overweight	17.4	15.6
Obese	19.8	22.8
Unknown	0.0	34.7
Smoking status		
Never	1.2	1.6
Former	38.5	24.7
Current	24.3	17.6
Unknown	36.0	56.2
Alcohol status		
Harmful	1.6	1.3
Hazardous	5.5	4.0
Unknown	92.8	94.7
GP location		
England	58.7	72.8
Northern Ireland	4.0	3.9
Wales	11.2	10.0
Scotland	26.0	13.3

Data are presented as mean±SD or %, unless otherwise stated. BMI: body mass index; GP: general practitioner.

Prevalence of diagnosed NTMLD cases

Period prevalence estimated using all cases (1082 cases) for the entire study period was 9.0 per 100 000. Point prevalence in 2016 was 5.1 per 100 000; point prevalence considering IMRD/HES cases only was 3.6 per 100 000. Annual prevalence rates of NTMLD in the period from 2006 to 2016 increased from 2.7 to 5.1 per 100 000. For diagnosis-identified cases, the annual prevalence rates increased from 1.3 to 3.6 per 100 000, while for treatment-identified cases it remained stable at ∼2 per 100 000 (figure 3).

Machine learning: algorithm performance and interpretation

The AUC was 0.94, indicating high predictive performance. In terms of screening performance, 1094 individuals (with at least one risk factor for NTMLD) would need to be screened to detect 100 out of 741 (*i.e.* precision of 9.1% at a sensitivity of 13.5%) identified NTMLD patients based on projection to a prevalence of five in 100 000 (figure 4). To detect the same number of patients with NTMLD by random testing for NTMLD within this enriched population would require 1 million individuals to be screened (*i.e.* precision of 0.01%). The algorithm thus improves detection of patients with NTMLD by almost a thousand-fold. Age, and the timing of symptoms (cough), treatments (macrolides and inhaled corticosteroids) and lung function tests in the pre-index period were the highest contributors to algorithm performance (figure 4b).

To assess potential bias, characteristics of the top 100 true positives were compared to the full NTMLD case cohort. For all characteristics assessed, the distribution observed in the true positives was similar to

TABLE 2 Proportion and event frequency of top 10 most frequent predictors of nontuberculous mycobacterial lung disease

	Cases					Controls				
	Predictor	Type	Proportion of cases	Events pre-index n	Time of first exposure prior to index years	Predictor	Type	Proportion of cases	Events pre-index n	Time of first exposure prior to index years
Subjects n			741					112 874		
1	Penicillin	Rx	77.3	4	2.3	Penicillin	Rx	54.9	2	1.9
2	Macrolides	Rx	55.6	3	1.8	Cough	Dx	40.7	1	0.8
3	ICS	Rx	55.5	20	2.9	LFT	Test	19.0	3	1.7
4	LFT	Test	52.4	6	2.3	Macrolides	Rx	18.0	1	1.6
5	Systemic corticosteroids	Rx	48.0	5	2.2	Systemic corticosteroids	Rx	17.0	2	1.2
6	Cough	Dx	47.6	2	1.7	Imaging	Test	15.6	1	2.6
7	Imaging	Test	47.1	2	1.2	ICS	Rx	15.6	6	0.7
8	COPD	Dx	33.9	5	2.3	Asthma	Dx	13.5	4	2.1
9	Dyspnoea	Dx	31.8	2	1.8	Fatigue	Dx	13.5	1	0.5
10	Tuberculosis	Dx	28.1	1	0	Depression	Dx	12.9	2	0.7

Data are presented as % or median, unless otherwise stated. Rx: drug-identified; Dx: diagnosis; ICS: inhaled corticosteroids; LFT: lung function test.

that observed in the full cohort: 1) case identification method: 28% *versus* 28% diagnosed in IMRD, 7% *versus* 12% diagnosed in HES and 65% *versus* 59% diagnosed *via* treatment regimen; 2) sex: 56% female *versus* 54% female; and 3) age: median age 64.6 years *versus* 61.5 years; the lower median age for the true positives is largely driven by cystic fibrosis cases (12 patients) in the true-positive group; when removing these patients, the median age is 63.8 years.

Estimated rates of undiagnosed NTMLD cases

The predictive algorithm metrics were projected to assume a diagnosed prevalence rate of five per 100 000 in the UK general population, which is in line with the study results and previously published literature from UK and European studies [9, 10, 30]. Using the algorithm to identify individuals likely to have

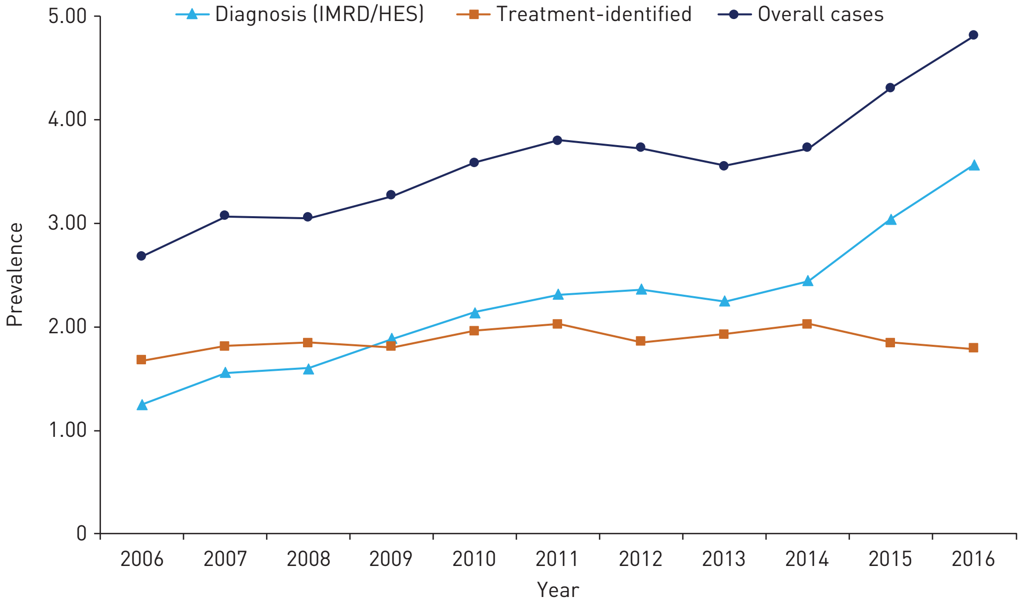

FIGURE 3 Annual prevalence of diagnosed nontuberculous mycobacterial lung disease (NTMLD). IMRD: IQVIA Medical Research Data; HES: Hospital Episode Statistics.

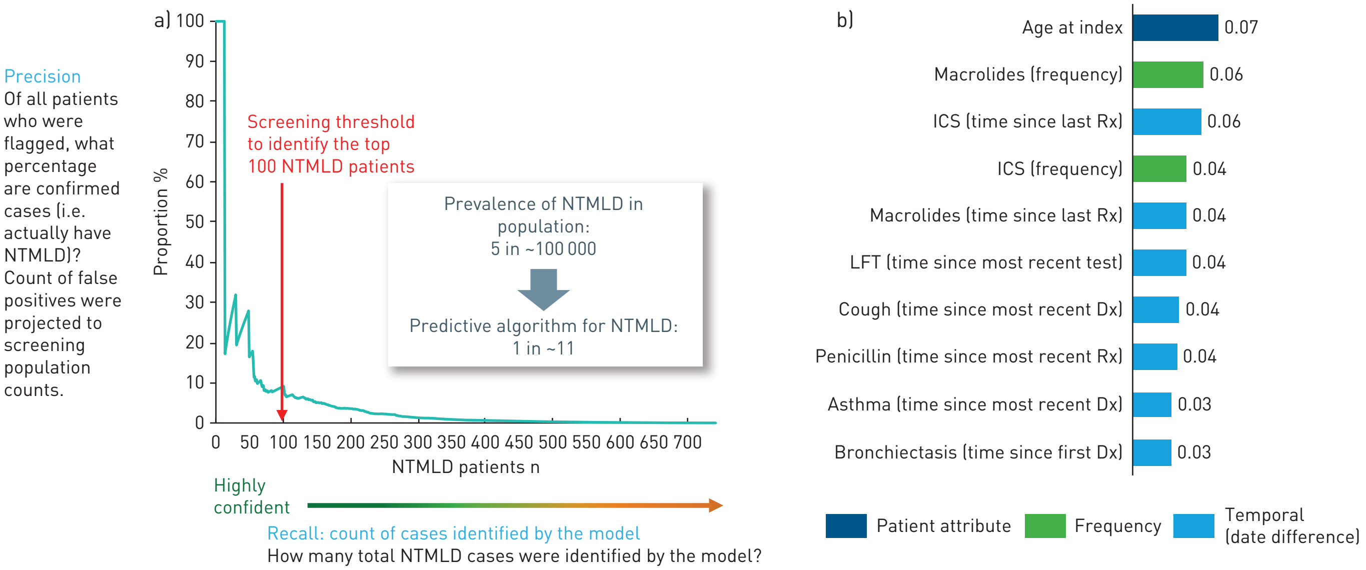

FIGURE 4 Projected precision-recall curve and variable importance for bagged modelling cohort. a) Projected precision-recall curve and b) variable importance for the predictive algorithm. NTMLD: nontuberculous mycobacterial lung disease; ICS: inhaled corticosteroids; Rx: prescription; LFT: lung function test; Dx: diagnosis.

NTMLD and considering risk thresholds of 0.90–0.95, the total prevalence of diagnosed cases and individuals likely to have NTMLD in 2016 was estimated to range from nine to 16 in 100 000, assuming that our control population was representative of the broader UK population (supplementary figure S1).

Relative risks of NTMLD diagnosis-based diagnoses and treatments

The relationship between the number of records observed for diagnoses and treatments as well as their timing was investigated for those considered to be key drivers for NTMLD (figure 5 and supplementary figure S2). For diagnoses of asthma, COPD and bronchiectasis, having at least one observed record was associated with a substantial increase in relative risk with diagnosis of bronchiectasis being associated with ⩾30-fold increase of risk. A single record for diagnosis of cough resulted in relative risk of 0.9, whereas having five or more observed records resulted in at least a four-fold increase in risk, highlighting that considering the extent of the diagnosis is an important consideration. In terms of the timing of diagnoses, risk of NTMLD increased as the time since first exposure to COPD and asthma increased, whereas risk declined as time since first occurrence of bronchiectasis increased (supplementary figure S2). This is

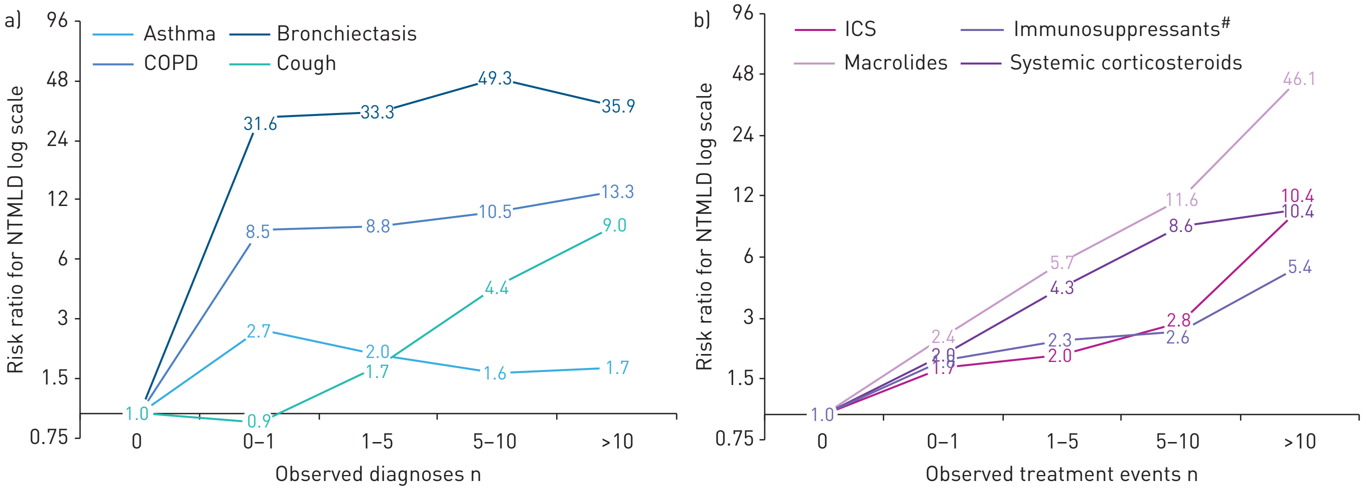

FIGURE 5 Relative risk ratio of nontuberculous mycobacterial lung disease (NTMLD) for selected a) diagnoses and b) treatments. The change in risk ratio is illustrated as the number of observed records increases with patients with an absence of the event as the comparator group. ICS: inhaled corticosteroids. #: including, but not limited to systemic and inhaled corticosteroids, tumour necrosis factor-α inhibitors, calcineurin inhibitors, interleukin inhibitors.

fitting, given that a diagnosis for bronchiectasis may trigger testing for NTMLD, which is less likely in other diagnoses such as asthma and COPD. The risk of NTMLD diagnosis declined when first date of TB occurred ⩾18 months prior to index date. While history of TB was associated with NTMLD diagnosis [31], diagnosis of TB in quick succession with a diagnosis of NTMLD is likely driven by the initial suspicion of TB.

For treatments, a higher number of prescriptions and longer time since first exposure increased the risk of NTMLD substantially. Exposure to >10 prescriptions of macrolides was associated with a 46-fold increase of risk of NTMLD diagnosis compared to patients without an observed record of prescription (figure 5).

Discussion

Underdiagnosis and delayed diagnosis is a key challenge in the management of NTMLD and may lead to worsening of underlying disease and increased mortality [32–34]. Lack of suspicion, nonspecific symptoms and co-existing pulmonary conditions that are frequent in patients with NTMLD may further complicate the timely and accurate diagnosis of NTMLD. This study aims to provide a better understanding of the epidemiology of NTMLD in the UK by profiling prediagnosis history and applying a machine learning algorithm to screen for likely undiagnosed cases of NTMLD in a primary-care population.

This study found that prevalence rates of NTMLD from 2006 to 2016 increased from 2.7 to 5.1 per 100 000, in line with the prevalence data reported for Europe (3.3–6 per 100 000) [8–10]. The prevalence of NTMLD in the treatment-identified cases was relatively stable, which, considering that only general practitioner (GP) prescriptions are available in these data, may indicate that in recent years more testing is carried out by GPs and then patients are referred to secondary care for diagnosis confirmation, treatment initiation and monitoring. This hypothesis is supported by findings observing steep increases in NTM isolation in UK secondary care [35, 36].

The results of this study suggest that the total estimated prevalence of diagnosed cases combined with individuals likely to have NTMLD in 2016 ranged between nine and 16 per 100 000. This is comparable to other published results from the UK which looked at diagnosed NTMLD: Shah *et al.* [36] reported that the incidence of NTM isolates rose from 5.6 per 100 000 in 2007 to 7.6 per 100 000 in 2012. Axson *et al.* [37] showed an average annual prevalence of NTM disease of 6.38 per 100 000 in UK primary care over the period from 2006 to 2016.

Limited data are available on physicians' awareness of NTMLD: both the European Respiratory Society and BTS bronchiectasis guidelines recommend testing for NTM in patients with bronchiectasis, but testing was only performed in 17.2% of the UK patients enrolled in the EMBARC study [38]. Moreover, a recent survey of pulmonologists in several European countries confirmed that recommendations for testing for NTM in patients with bronchiectasis are only partly followed, with physicians not testing for NTM managing significantly fewer NTMLD patients [39]. This suggests that greater awareness of NTM testing recommendations is needed, and this is likely to lead to earlier diagnosis and an increased number of NTMLD cases in the future.

In this study, a set of predictors (risk factors) relevant to NTMLD were identified from the published literature and guidelines [5, 40] and subsequently confirmed using clinical expert guidance. Known risk factors such as bronchiectasis, COPD, inhaled corticosteroid use, asthma and exposure to immunosuppressant medications were highly ranked by model. In addition, the model identified prescriptions to antibiotic medicines as key predictors with multiple (>10) prescriptions for macrolides associated with an elevated risk of NTMLD. The first observed prescription for macrolides was, on average, 1.8 years prior to diagnosis of NTMLD. This may reflect that patients who are prescribed macrolides are more likely to have a chronic respiratory condition (*e.g.* COPD) which is a risk factor for NTMLD; alternatively, it may reflect that this population are, in general, doing less well clinically in the period prior to diagnosis and are therefore more likely to be screened for NTMLD.

A machine learning algorithm was used to model complex associations in terms of frequency and timing within the rich prediagnosis digital footprints. Based on an assumed prevalence of five per 100 000, 1094 patients would need to be screened to identify 100 true positive NTMLD patients representing a precision of 9.1% and a sensitivity of 13.5%, which when compared to random screening (precision of 0.01%) within the enriched control cohort leads to almost a thousand-fold improvement. Moreover, the true positive rates were largely consistent across protected patient characteristics, *i.e.* the algorithm was not biased for or against patients according to characteristics such as age, sex and method of case identification [29].

There are several limitations to this study. The number of practices contributing data to IMRD varied over time, with the most recent years having fewer practices than earlier years, as data are contributed in a batched process spanning time periods ranging from months to years. Therefore, the end of the data period was chosen conservatively as September 2017 to help alleviate this, but nonetheless we note than patient numbers are less in recent years than earlier years.

The treatment-identification criterion was based on BTS guidelines and validated by clinical expert opinion; however, it is conceivable that these patients may have been treated for another non-NTMLD infection, for example, TB. This was deemed unlikely since TB is typically treated in secondary care rather than primary care, and treatment of both active and latent TB is in most cases only up to 6 months. Additionally, NTMLD is normally defined, recorded and treated when a species is documented. Conversely, TB is often diagnosed empirically without a positive culture. Consequently, the suspected diagnosis of TB is often made based on either the radiological findings or a mycobacterial growth or smear prior to the identification and this may then be changed to NTMLD when the species is identified, whereas the reverse is much less likely to happen. Nonetheless, treatment-identified patients did have a higher prevalence of diagnosed or suspected TB (34.9%) than the diagnosis-identified patients (18.2%). For future studies, it may be possible to validate the inferred diagnosis of NTMLD by applying the selection criteria used in the study to a database which offers a service whereby GPs are contacted by letter to answer a bespoke questionnaire around diagnoses: confirmation of diagnosis including tests done, diagnosis dates, resolution dates, family members testing, reticulation, *etc*.

Patient data were extracted from primary care and were not inclusive of other care settings, which impacts predictors such as exposure to immunosuppressive/immunomodulating medications. These medications are more often prescribed in secondary care and therefore underrepresented in primary-care records. Associated diagnoses which require immunosuppressive/immunomodulatory treatment were captured in the algorithm acting as a proxy, albeit in the absence of granularity such as duration of exposure.

For machine learning in respiratory medicine, the greatest progress has been observed in algorithms for medical images [41] with tools in development for detection of pulmonary nodules of lung disorders and infections [42] and screening for TB [43]. Beyond imaging, the potential of machine learning for structured healthcare data such as EMR and medical claims data has been reported in COPD for predicting subsequent diagnosis in asthma patients [44] and predicting hospital re-admission [45]. The algorithm described here paves the way for machine learning-based screening of rare respiratory diseases in using primary-care data with the predictive performance for this NTMLD-specific algorithm supporting further development and pilot studies. A natural immediate next step would be conducting external validation to provide a more wide-ranging assessment of performance. External validation involves applying the algorithm to a completely distinct dataset (*e.g.* geographically and/or temporally) in order to assess its performance in a new setting. Given that the model developed here was based on the IMRD EMR dataset, the Clinical Practice Research Datalink dataset is a promising candidate for an external validation study given similarities in data capture (*e.g.* diagnoses and prescriptions) and underlying data model to IMRD (*e.g.* use of Read codes), care setting (primary care across the UK) and ability to link to secondary care for a subset of patients. Application of the NTMLD screening model to this dataset would provide robust insight into the ability of the model to generalise in an external setting with sufficiently similar properties.

Conclusions

The data captured in a UK primary-care database enabled the development of a predictive machine learning algorithm to identify individuals likely to suffer from NTMLD. The algorithm exhibited almost a thousand-fold better detection of cases with NTMLD *versus* random testing in a cohort with at least one risk factor for NTMLD. Moreover, the predictive algorithm indicates that there may be a substantial number of undiagnosed cases of NTMLD in the UK.

Acknowledgements: Medical writing support was provided by Lorna Barclay, Chandresh Kumar, Paranjoy Saharia and Ishneet Kaur on behalf of IQVIA (UK and India) and was funded by Insmed. The authors were fully responsible for all content and editorial decisions and were involved at all stages of manuscript development and have approved the final version.

Author contributions: O.M. Doyle, P. McMahon, F. Daniels, A. Pitcher, M. Obradovic, R. van der Laan and M.R. Loebinger were involved in the study conception and design, data analysis and drafting of the manuscript. All authors have read and approved the final version of the manuscript.

Conflict of interest: O.M. Doyle reports funding from Insmed, during the conduct of the study; and is employed by IQVIA. R. van der Laan reports funding from Insmed, during the conduct of the study; and is employed by Insmed. M. Obradovic reports funding from Insmed, during the conduct of the study; and is employed by Insmed. P. McMahon reports funding from Insmed, during the conduct of the study; and is employed by IQVIA. F. Daniels reports funding from Insmed, during the conduct of the study; and is employed by IQVIA. A. Pitcher reports funding from Insmed, during the conduct of the study; and is employed by IQVIA. M.R. Loebinger reports personal fees and funding from

Insmed, during the conduct of the study; personal fees from Savara, Grifols, Bayer, AstraZeneca, Polyphor and Meiji, outside the submitted work.

Support statement: This study was financially supported by Insmed. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Falkinham JO 3rd. Environmental sources of nontuberculous mycobacteria. *Clin Chest Med* 2015; 36: 35–41.

2 Falkinham JO 3rd. Current epidemiologic trends of the nontuberculous mycobacteria (NTM). *Curr Environ Health Rep* 2016; 3: 161–167.

3 Sood G, Parrish N. Outbreaks of nontuberculous mycobacteria. *Curr Opin Infect Dis* 2017; 30: 404–409.

4 Lyman MM, Grigg C, Kinsey CB, *et al.* Invasive nontuberculous mycobacterial infections among cardiothoracic surgical patients exposed to heater-cooler devices. *Emerg Infect Dis* 2017; 23: 796–805.

5 Haworth CS, Banks J, Capstick T, *et al.* British Thoracic Society guidelines for the management of non-tuberculous mycobacterial pulmonary disease (NTM-PD). *Thorax* 2017; 72: Suppl. 2, ii1–ii64.

6 Sarro YD, Kone B, Diarra B, *et al.* Simultaneous diagnosis of tuberculous and non-tuberculous mycobacterial diseases: time for a better patient management. *Clin Microbiol Infect Dis* 2018; 3: 10.15761/CMID.1000144.

7 Shibata Y, Horita N, Yamamoto M, *et al.* Diagnostic test accuracy of anti-glycopeptidolipid-core IgA antibodies for *Mycobacterium avium* complex pulmonary disease: systematic review and meta-analysis. *Sci Rep* 2016; 6: 29325.

8 Prevots DR, Loddenkemper R, Sotgiu G, *et al.* Nontuberculous mycobacterial pulmonary disease: an increasing burden with substantial costs. *Eur Respir J* 2017; 49: 1700374.

9 Wagner D, van Ingen J, Adjemian J, *et al.* Annual prevalence and treatment estimates of nontuberculous mycobacterial pulmonary disease in Europe: a NTM-NET collaborative study. *Eur Respir J* 2014; 44: Suppl. 58, P1067.

10 Ringshausen FC, Wagner D, de Roux A, *et al.* Prevalence of nontuberculous mycobacterial pulmonary disease, Germany, 2009–2014. *Emerg Infect Dis* 2016; 22: 1102–1105.

11 Park TY, Chong S, Jung JW, *et al.* Natural course of the nodular bronchiectatic form of *Mycobacterium avium* complex lung disease: long-term radiologic change without treatment. *PLoS One* 2017; 12: e0185774.

12 Khan Z, Miller A, Bachan M, *et al.* *Mycobacterium avium* complex (MAC) lung disease in two inner city community hospitals: recognition, prevalence, co-infection with *Mycobacterium tuberculosis* (MTB) and pulmonary function (PF) improvements after treatment. *Open Respir Med J* 2010; 4: 76–81.

13 Kotilainen H, Valtonen V, Tukiainen p, *et al.* Clinical findings in relation to mortality in non-tuberculous mycobacterial infections: patients with *Mycobacterium avium* complex have better survival than patients with other mycobacteria. *Eur J Clin Microbiol Infect Dis* 2015; 34: 1909–1918.

14 Mirsaeidi M, Hadid W, Ericsoussi B, *et al.* Non-tuberculous mycobacterial disease is common in patients with non-cystic fibrosis bronchiectasis. *Int J Infect Dis* 2013; 17: e1000-4.

15 Mehta M, Marras TK. Impaired health-related quality of life in pulmonary nontuberculous mycobacterial disease. *Respir Med* 2011; 105: 1718–1725.

16 Huang CT, Tsai YJ, Wu HD, *et al.* Impact of non-tuberculous mycobacteria on pulmonary function decline in chronic obstructive pulmonary disease. *Int J Tuberc Lung Dis* 2012; 16: 539–545.

17 Diel R, Jacob J, Lampenius N, *et al.* Burden of non-tuberculous mycobacterial pulmonary disease in Germany. *Eur Respir J* 2017; 49: 1602109.

18 Marras TK, Mirsaeidi M, Chou E, *et al.* Health care utilization and expenditures following diagnosis of nontuberculous mycobacterial lung disease in the United States. *J Manag Care Spec Pharm* 2018; 24: 964–974.

19 Marras TK, Vinnard C, Zhang Q, *et al.* Relative risk of all-cause mortality in patients with nontuberculous mycobacterial lung disease in a US managed care population. *Respir Med* 2018; 145: 80–88.

20 Annavarapu S, Goldfarb, S, Gelb M, *et al.* Development and validation of a predictive model to identify patients at risk of severe COPD exacerbations using administrative claims data. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 2121–2130.

21 Ross EG, Shah NH, Dalman RL, *et al.* The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J Vasc Surg* 2016; 64: 1515–1522.

22 Uspenskaya-Cadoz O, Alamuri C, Wang L, *et al.* Machine learning algorithm helps identify non-diagnosed prodromal Alzheimer's disease patients in the general population. *J Prev Alzheimers Dis* 2019; 6: 185–191.

23 Kiely D, Doyle O, Drage E, *et al.* Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension. *Pulm Circ* 2019; 9: 2045894019890549.

24 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232.

25 Olson RS, Cava W, Mustahsan Z, *et al.* Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput* 2018; 23: 192–203.

26 Breiman L. Bagging predictors. *Mach Learn* 1996; 24: 123–140.

27 Chen T, Guestrin C. Xgboost: a scalable tree boosting system. *In*: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Fransicso, Association for Computing Machinery, 2016.

28 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10: e0118432.

29 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. *In:* Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Curran Associates Inc., 2016; pp. 3323–3331.

30 Prevots DR, Marras TK. Epidemiology of human pulmonary infection with nontuberculous mycobacteria: a review. *Clin Chest Med* 2015; 36: 13–34.

31 Stout JE, Koh W-J, Yew WW. Update on pulmonary disease due to non-tuberculous mycobacteria. *Int J Infect Dis* 2016; 45: 123–134.

32 Park HY, Jeong BH, Chon HR, *et al.* Lung function decline according to clinical course in nontuberculous mycobacterial lung disease. *Chest* 2016; 150: 1222–1232.

33 Hwang JA, Kim S, Jo KW, *et al.* Natural history of *Mycobacterium avium* complex lung disease in untreated patients with stable course. *Eur Respir J* 2017; 49: 1600537.

34 O'Connell ML, Birkenkamp KE, Kleiner DE, *et al.* Lung manifestations in an autopsy-based series of pulmonary or disseminated nontuberculous mycobacterial disease. *Chest* 2012; 141: 1203–1209.

35 Cowman S, Burns K, Benson S, *et al.* The antimicrobial susceptibility of non-tuberculous mycobacteria. *J Infect* 2016; 72: 324–331.

36 Shah NM, Davidson JA, Anderson LF, *et al.* Pulmonary *Mycobacterium avium*-intracellulare is the main driver of the rise in non-tuberculous mycobacteria incidence in England, Wales and Northern Ireland, 2007–2012. *BMC Infect Dis* 2016; 16: 195.

37 Axson EL, Bloom CI, Quint JK. Nontuberculous mycobacterial disease managed within UK primary care, 2006–2016. *Eur J Clin Microbiol Infect Dis* 2018; 37: 1795–1803.

38 Finch S, van der Laan R, Crichton M, *et al.* Non-tuberculous mycobacteria testing in bronchiectasis in the UK: data from the EMBARC registry. *Thorax* 2019; 74: A238–A239.

39 Wagner D, van Ingen J, van der Laan R, *et al.* Screening for NTM lung disease in adult non-CF adult bronchiectasis patients – physician survey in Germany, UK, Italy, France and the Netherlands. *Thorax* 2018; 73: A102–A102.

40 Griffith DE, Aksamit T, Brown-Elliott BA, *et al.* An official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. *Am J Respir Crit Care Med* 2007; 175: 367–416.

41 Angelini E, Dahan S, Shah A. Unravelling machine learning: insights in respiratory medicine. *Eur Respir J* 2019; 54: 1901216.

42 Qin C, Yao D, Shi Y, *et al.* Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomed Eng Online* 2018; 17: 113.

43 Qin ZZ, Sander MS, Rai B, *et al.* Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep* 2019; 9: 15000.

44 Himes BE, Dai Y, Kohane IS, *et al.* Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *J Am Med Inform Assoc* 2009; 16: 371–379.

45 Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. *Sci Rep* 2019; 9: 2362.