EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

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

Orla M. Doyle, Roald van der Laan, Marko Obradovic, Peter McMahon, Flora Daniels, Ashley Pitcher, Michael R. Loebinger

Please 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; in press (https://doi.org/10.1183/13993003.00045-2020).

Title

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

Authors

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

Author affiliations

[1] Predictive Analytics, Real World Analytical Solutions, IQVIA, UK
[2] Insmed Utrecht, Netherlands
[3] Real-World Insights, IQVIA, UK
[4] Real-World Insights, IQVIA, Switzerland
[5] Real-World Insights, IQVIA, Denmark
[6] Royal Brompton and Harefield NHS Foundation Trust and Imperial College London, UK

*Authors contributed equally

Corresponding author

Marko Obradovic

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 pre-diagnosis features of NTMLD patients in primary care and to assess the feasibility of using machine learning (ML) to identify undiagnosed NTMLD patients.

IQVIA Medical Research Data (IMRD; 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 pre-diagnosis period, guided by literature and expert clinical opinion. The control population was enriched to have at least one of these features.

A total of 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 chronic obstructive pulmonary disease, asthma, penicillin, macrolides and inhaled corticosteroids. Compared to random testing, ML 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 ML 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.

Keywords

Nontuberculous mycobacterial lung disease (NTMLD); Respiratory Disorders; Electronic Medical Records; United Kingdom (UK); Predictive Modelling; Machine Learning (ML);

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 that 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, chronic obstructive pulmonary disease (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 health care resource utilization [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 (ML) 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]. ML algorithms “learn by example” where a patient’s pre-diagnosis 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 (i) describe the prevalence and incidence rates of diagnosed NTMLD patients, (ii) characterise the pre-diagnosis features of NTMLD patients in primary care and (iii) assess the feasibility of using machine learning to identify undiagnosed NTMLD patients.

Methods

Study design

The IMRD UK EMR primary care database was used (a more detailed description of IMRD is contained in supplementary information). Three criteria were applied to identify positive cases for NTMLD: (i) from IMRD using Read codes A310000 (pulmonary *Mycobacterium avium*-intracellular infection) and A310.00 (pulmonary mycobacterial infection); (ii) by linking to secondary care records in Hospital Episode Statistics (HES) using International Classification of Diseases (ICD)-10 code A31.0 (infection due to other mycobacteria); secondary care records were used only for case identification; and (iii) based on treatment regimens of specific antibiotic combinations (≥180 days; 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 less than 30 days. For patients who were identified by more than one method, the earliest date was chosen as the date of NTMLD diagnosis. This multi-criteria 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 (see Table S2) to ensure that the predictive model would focus on learning to distinguish between different illnesses rather than learn 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 phenomenon. From a random sample of 750,000 patients from the IMRD population, 112,784 patients met the inclusion and matching criteria.

The study period was Sept 2003 to Sept 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 pre-diagnosis 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 three years. Predictors such as age, gender, and risk factors from literature sources including British Thoracic Society and American Thoracic Society guidelines were selected and verified by expert clinical opinion. For a more detailed description of the predictor selection please refer to the supplementary information and specifically Table S2. Metrics were created to quantify the frequency and timing of the predictors; predictors that were not observed were assumed absent rather with missing values (NAs) used for the timing of absent events. 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 pre-diagnosis 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 non-linear associations, 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, non-overlapping partitions of the data using 5-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[1] curves. 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 (5/100,000).

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].

---

[1] 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)*.

[1] Recall (or sensitivity) describes the proportion of true cases that were retrieved by the model (true positives) to the total number of cases in the dataset (sum of true positives and false negatives) that is *true positives/(true positives + false negatives)*.

Results

Participants

A total of 1,082 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 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 "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, have lower BMI and a current or former smoker 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.

Prevalence of diagnosed NTMLD cases

Period prevalence estimated using all cases (1,082 cases) for the entire study period was 9.0 per 100,000. Point prevalence in 2016 was 5.1/100,000; point prevalence considering IMRD/HES cases only was 3.6/ 100,000. Annual prevalence rates of NTMLD in the period from 2006 to 2016 increased from 2.7 to 5.1/100,000. For diagnosis-identified cases, the annual prevalence rates increased from 1.3 to 3.6 while for treatment-identified cases it remained stable at approximately 2/100,000 (Pre-diagnosis primary care patient journey in 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; LFT: Lung function tests; COPD: Chronic obstructive pulmonary diseases; MRC: Medical Research Council breathlessness score

Figure 3).

Machine learning: algorithm performance and interpretation

The AUC was 0.94 indicating high predictive performance. In terms of screening performance, 1,094 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 5 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 ICS) and lung function tests (LFTS) 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 that observed in the full cohort: (i) case identification method: 28% vs 28% diagnosed in IMRD, 7% vs 12% diagnosed in HES and 65% vs 59% diagnosed via treatment regimen, (ii) gender: 56% female vs. 54% female and (iii) age: median ages 64.6 years vs 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 5 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 NTMLD and considering risk thresholds of 0.90 to 0.95, the total prevalence of diagnosed cases and individuals likely to have NTMLD in 2016 was estimated to range from 9 to 16 in 100,000 assuming that our control population was representative of the broader UK population (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 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 at least a 30-fold increase of risk. A single record for diagnosis of cough resulted in relative risk of 0.9 whereas having 5 or more observed records resulted a 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 (Figure S2).; which is 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 or more 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 more than ten prescriptions of macrolides was associated with a 46 fold increase the risk of NTMLD diagnosis compared to patients without an observed record of prescription (Figure 5).

Discussion

Under-diagnosis 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, non-specific 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 pre-diagnosis history and applying an ML 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 (3.3-6/100,000) [8-10]. The prevalence of NTMLD in the treatment-identified cases was relatively stable which, considering only GP prescriptions are available in this data, may indicate that more testing is carried out by GPs in more recent years 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 9-16/100,000. This is comparable to other published results from the UK which looked at diagnosed NTMLD: Shah et al reported that the incidence of NTM isolates rose from 5.6/100,000 in 2007 to 7.6/100,000 in 2012 [36]. Axson et al. showed an average annual prevalence of NTM disease in the UK primary care over the period from 2006 to 2016 of 6.38 per 100,000 [37].

Limited data are available on physicians' awareness of NTMLD: both the European Respiratory Society 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.

An ML algorithm was used to model complex associations in terms of frequency and timing within the rich pre-diagnosis digital footprints. Based on an assumed prevalence of 5 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. algorithm was not biased for or against patients according to characteristics such as age, gender and method of case identification [29].

There are several limitations of 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 General Practitioners 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 was 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 ML 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 ML 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 ML-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 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 an UK primary care database enabled the development of a predictive ML algorithm to identify individuals likely to suffer from NTMLD. The algorithm exhibited almost a thousand-fold better detection of cases with NTMLD vs. 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.

List of abbreviations

BMI: Body Mass Index; BTS: British Thoracic Society; COPD: Chronic Obstructive Pulmonary Disease; CTS: Corticosteroids; CPRD: Clinical Practice Research DataLink; Dx: Diagnosis; GP: General Practitioner; HES: Hospital Episode Statistics; HIV: Human Immunodeficiency Virus; ICD: International Classification of Diseases; IMRD: IQVIA Medical Research Data; K: Thousand; LFT: Lung Function Test; MAC: *Mycobacterium avium* Complex; ML: Machine Learning; MRC: Medical Research Council; NTM: Nontuberculous Mycobacterial; NTMLD: Nontuberculous Mycobacterial Lung Disease; PRC: Precision-Recall Curve; RA: Rheumatoid Arthritis; Rx: Drug-Identified; TB: Tuberculosis;; UK: United Kingdom; UK: United States

Declarations

Ethics approval and consent to participate

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 to the terms the 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”.

*IQVIA Medical Research Data UK incorporating THIN, a registered trademark of Cegedim SA in the United Kingdom and other countries. Reference made to the IMRD database is intended to be descriptive of the data asset licensed by IQVIA.

Consent for publication

Not applicable

Availability of data and material

All relevant data are available in the manuscript and there are no supplementary files. The original data supporting this finding will be available from the corresponding author on request.

Competing interests

MO and RvdL are employees of Insmed. OD, FD, PM and AP are employees of IQVIA, a company that has provided consultancy services to Insmed.

Funding

This study was financially supported by Insmed.

Authors' contributions

OD, PM, FD, AP, MO, RvdL and ML 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.

Acknowledgements

Medical writing support was provided by Lorna Barclay, Chandresh Kumar, Paranjoy Saharia and Ishneet Kaur on behalf of IQVIA 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.

References

1. Falkinham, J.O., 3rd, *Environmental sources of nontuberculous mycobacteria.* Clin Chest Med, 2015. **36**(1): p. 35-41.
2. Falkinham, J.O., 3rd, *Current Epidemiologic Trends of the Nontuberculous Mycobacteria (NTM).* Curr Environ Health Rep, 2016. **3**(2): p. 161-7.
3. Sood, G. and N. Parrish, *Outbreaks of nontuberculous mycobacteria.* Curr Opin Infect Dis, 2017. **30**(4): p. 404-409.
4. Lyman, M.M., et al., *Invasive Nontuberculous Mycobacterial Infections among Cardiothoracic Surgical Patients Exposed to Heater-Cooler Devices(1).* Emerg Infect Dis, 2017. **23**(5): p. 796-805.
5. Haworth, C.S., et al., *British Thoracic Society guidelines for the management of non-tuberculous mycobacterial pulmonary disease (NTM-PD).* Thorax, 2017. **72**(Suppl 2): p. ii1-ii64.
6. Sarro, Y.D., et al., *Simultaneous diagnosis of tuberculous and non-tuberculous mycobacterial diseases: Time for a better patient management.* Clinical microbiology infectious diseases 2018. **3**(3).
7. Shibata, Y., et al., *Diagnostic test accuracy of anti-glycopeptidolipid-core IgA antibodies for Mycobacterium avium complex pulmonary disease: systematic review and meta-analysis.* Scientific reports, 2016. **6**: p. 29325.
8. Prevots, D.R., et al., *Nontuberculous mycobacterial pulmonary disease: an increasing burden with substantial costs.* European Respiratory Journal, 2017. **49: 1700374**.
9. Wagner, D., et al., *Annual prevalence and treatment estimates of nontuberculous mycobacterial pulmonary disease in Europe: A NTM-NET collaborative study.* European Respiratory Journal, 2014. **44**(Suppl 58): p. P1067.
10. Ringshausen, F.C., et al., *Prevalence of Nontuberculous Mycobacterial Pulmonary Disease, Germany, 2009-2014.* Emerg Infect Dis, 2016. **22**(6): p. 1102-5.
11. Park, T.Y., 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**(10): p. e0185774.
12. Khan, Z., 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**: p. 76-81.
13. Kotilainen, H., 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**(9): p. 1909-18.
14. Mirsaeidi, M., et al., *Non-tuberculous mycobacterial disease is common in patients with non-cystic fibrosis bronchiectasis.* Int J Infect Dis, 2013. **17**(11): p. e1000-4.
15. Mehta, M. and T.K. Marras, *Impaired health-related quality of life in pulmonary nontuberculous mycobacterial disease.* Respir Med, 2011. **105**(11): p. 1718-25.
16. Huang, C.T., et al., *Impact of non-tuberculous mycobacteria on pulmonary function decline in chronic obstructive pulmonary disease.* Int J Tuberc Lung Dis, 2012. **16**(4): p. 539-45.
17. Diel, R., et al., *Burden of non-tuberculous mycobacterial pulmonary disease in Germany.* Eur Respir J, 2017. **49**(4).
18. Marras, T.K., 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**(10): p. 964-974.
19. Marras, T.K., 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**: p. 80-88.

20. Annavarapu, S., 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**: p. 2121-2130.
21. Ross, E.G., et al., *The use of machine learning for the identification of peripheral artery disease and future mortality risk.* J Vasc Surg, 2016. **64**(5): p. 1515-1522.e3.
22. Uspenskaya-Cadoz, O., et al., *Machine Learning Algorithm Helps Identify Non-Diagnosed Prodromal Alzheimer's Disease Patients in the General Population.* J Prev Alzheimers Dis, 2019. **6**(3): p. 185-191.
23. Kiely, D., et al., *EXPRESS: Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension.* Pulmonary Circulation, 2019. **0**(ja): p. 2045894019890549.
24. Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.
25. Olson, R.S., et al., *Data-driven advice for applying machine learning to bioinformatics problems.* Pac Symp Biocomput., 2018. **23**: p. 192-203.
26. Breiman, L., *Bagging predictors.* Machine learning, 1996. **24**(2): p. 123-140.
27. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system.* in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016. ACM.
28. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PLoS One, 2015. **10**(3): p. e0118432.
29. Hardt, M., E. Price, and N. Srebro, *Equality of opportunity in supervised learning*, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, Curran Associates Inc.: Barcelona, Spain. p. 3323-3331.
30. Prevots, D.R. and T.K. Marras, *Epidemiology of human pulmonary infection with nontuberculous mycobacteria: a review.* Clinics in chest medicine, 2015. **36**(1): p. 13-34.
31. Stout, J.E., W.-J. Koh, and W.W. Yew, *Update on pulmonary disease due to non-tuberculous mycobacteria.* International Journal of Infectious Diseases, 2016. **45**: p. 123-134.
32. Park, H.Y., et al., *Lung function decline according to clinical course in nontuberculous mycobacterial lung disease.* Chest, 2016. **150**(6): p. 1222-1232.
33. Hwang, J.A., et al., *Natural history of Mycobacterium avium complex lung disease in untreated patients with stable course.* European Respiratory Journal, 2017. **49**(3): p. 1600537.
34. O'Connell, M.L., et al., *Lung manifestations in an autopsy-based series of pulmonary or disseminated nontuberculous mycobacterial disease.* Chest, 2012. **141**(5): p. 1203-1209.
35. Cowman, S., et al., *The antimicrobial susceptibility of non-tuberculous mycobacteria.* J Infect, 2016. **72**(3): p. 324-31.
36. Shah, N.M., 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**: p. 195.
37. Axson, E.L., C.I. Bloom, and J.K. Quint, *Nontuberculous mycobacterial disease managed within UK primary care, 2006-2016.* Eur J Clin Microbiol Infect Dis, 2018. **37**(9): p. 1795-1803.
38. Finch, S., et al., *M8 Non-tuberculous mycobacteria testing in bronchiectasis in the UK: data from the EMBARC registry.* Thorax, 2019. **74**: p. A238-A239.
39. Wagner, D., et al., *P11 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**: p. A102-A102.
40. Griffith, D.E., et al., *An Official ATS/IDSA Statement: Diagnosis, Treatment, and Prevention of Nontuberculous Mycobacterial Diseases.* American Journal of Respiratory and Critical Care Medicine, 2007. **175**(4): p. 367-416.
41. Angelini, E., S. Dahan, and A. Shah, *Unravelling machine learning: insights in respiratory medicine.* Eur Respir J, 2019. **54**(6).

42. Qin, C., et al., *Computer-aided detection in chest radiography based on artificial intelligence: a survey.* Biomed Eng Online, 2018. **17**(1): p. 113.
43. Qin, Z.Z., 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**(1): p. 15000.
44. Himes, B.E., et al., *Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records.* J Am Med Inform Assoc, 2009. **16**(3): p. 371-9.
45. Min, X., B. Yu, and F. Wang, *Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD.* Sci Rep, 2019. **9**(1): p. 2362.

Figures Legends

Figure 1: NTMLD cases meeting inclusion/exclusion criteria. Dx: Diagnosis; HES: Hospital Episode Statistics; IMRD: IQVIA Medical Research Data; Rx: Treatment-identified.

Figure 2: Pre-diagnosis primary care patient journey in 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; LFT: Lung function tests; COPD: Chronic obstructive pulmonary diseases; MRC: Medical Research Council breathlessness score

Figure 3 Annual prevalence of diagnosed NTMLD. IMRD: IQVIA Medical Research Data; HES: Hospital Episode Statistics

Figure 4: Projected precision-recall curve and variable importance for bagged modelling cohort

Figure 4a: Projected precision-recall curve for the predictive algorithm

Figure 4b: Variable importance for the predictive algorithm

Figure 5 : Relative risk ratio of NTMLD for selected diagnoses and 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. COPD: Chronic obstructive pulmonary diseases; ICS: Inhaled corticosteroids; Immunosuppressive drugs (including, but not limited to systemic and inhaled corticosteroids, TNF-alfa inhibitors, calcineurin inhibitors, interleukin inhibitors).

Figure S1: Annual estimates for total NTMLD count (2014-2016)

Figure S2: Relative risk of NTMLD by time between first occurrence and index date. COPD: Chronic obstructive pulmonary diseases; ICS: Inhaled corticosteroids; Immunosuppressive drugs (including, but not limited to systemic and inhaled corticosteroids, TNF-alfa inhibitors, calcineurin inhibitors, interleukin inhibitors).

Tables

Table 1: Demographics

Proportion of NTMLD cases and controls by demographic segments of interest		
	NTMLD cases (N=741)	**Controls (N=112, 874)**
Age at Dx		
Mean age at diagnosis [SD]	59.8 years [19.2]	48.5 years [24.7]
Under 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%

Key: NTMLD: nontuberculous mycobacterial lung disease; BMI: body mass index; GP: general practitioner; Dx: diagnosis

Table 2: Top 10 most frequent predictors

Proportion and median event frequency for top 10 most frequent predictors in case cohort (N=741)					**Proportion and median event frequency for top 10 most frequent predictors in control cohort (N=112,874)**				
Predictor	**Type**	**% of cases**	**Median number of events pre-index**	**Median time of first exposure prior to index (years)**	**Predictor**	**Type**	**% of cases**	**Median number of events pre-index**	**Median time of first exposure prior to index (years)**
Penicillin	Rx	77.3%	4	2.3	Penicillin	Rx	54.9%	2	1.9
Macrolides	Rx	55.6%	3	1.8	Cough	Dx	40.7%	1	0.8
ICS	Rx	55.5%	20	2.9	LFTs	Test	19.0%	3	1.7
LFTs	Test	52.4%	6	2.3	Macrolides	Rx	18.0%	1	1.6
Systemic CTS	Rx	48.0%	5	2.2	Systemic CTS	Rx	17.0%	2	1.2
Cough	Dx	47.6%	2	1.7	Imaging	Test	15.6%	1	2.6
Imaging	Test	47.1%	2	1.2	ICS	Rx	15.6%	6	0.7
COPD	Dx	33.9%	5	2.3	Asthma	Dx	13.5%	4	2.1
Dyspnoea	Dx	31.8%	2	1.8	Fatigue	Dx	13.5%	1	0.5
Tuberculosis	Dx	28.1%	1	0	Depression	Dx	12.9%	2	0.7

Key: CTS: corticosteroids; ICS: inhaled CTS; LFT: lung function test; COPD: chronic obstructive pulmonary disease; Rx: drug-identified; Dx: diagnosis

Supplementary Information Legends

Table S1: List of included drug regimens to identify NTMLD patients

Table S2: List of predictors included in the model

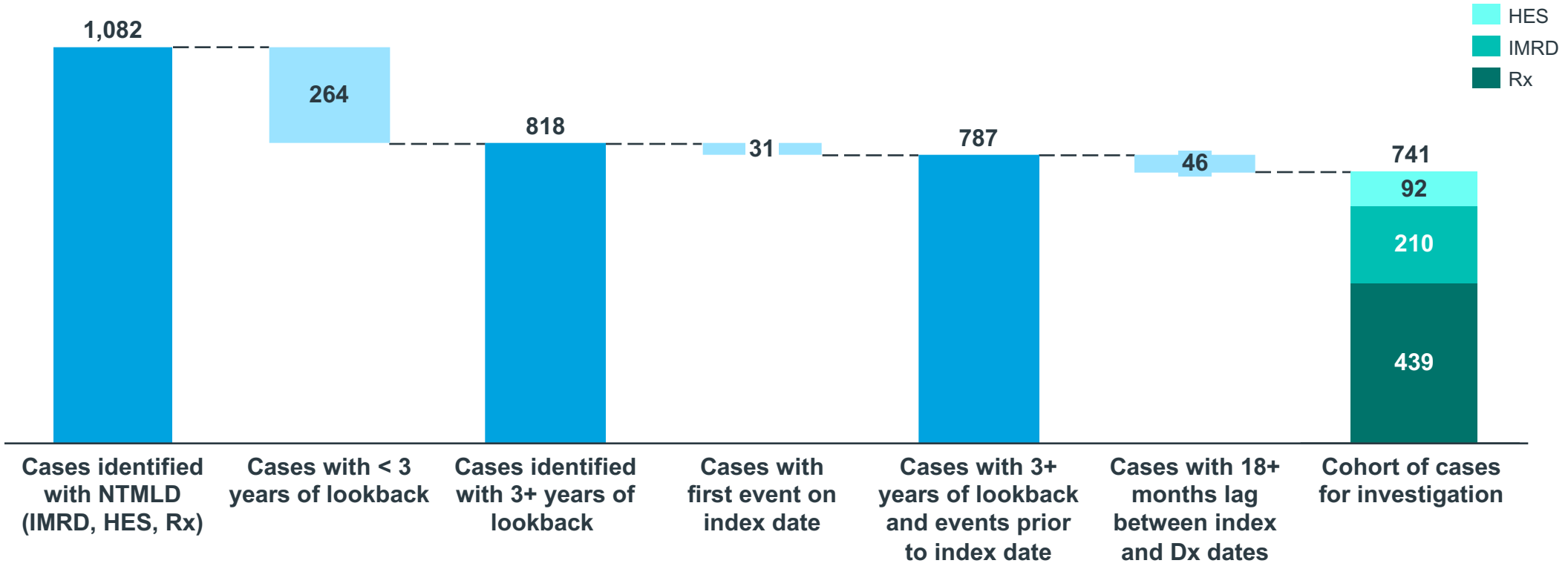

Figure 1: NTMLD cases meeting inclusion/exclusion criteria. Dx: Diagnosis; HES: Hospital Episode Statistics; IMRD: IQVIA Medical Research Data; Rx: Treatment-identified; NTMLD: Nontuberculous mycobacteria lung disease;

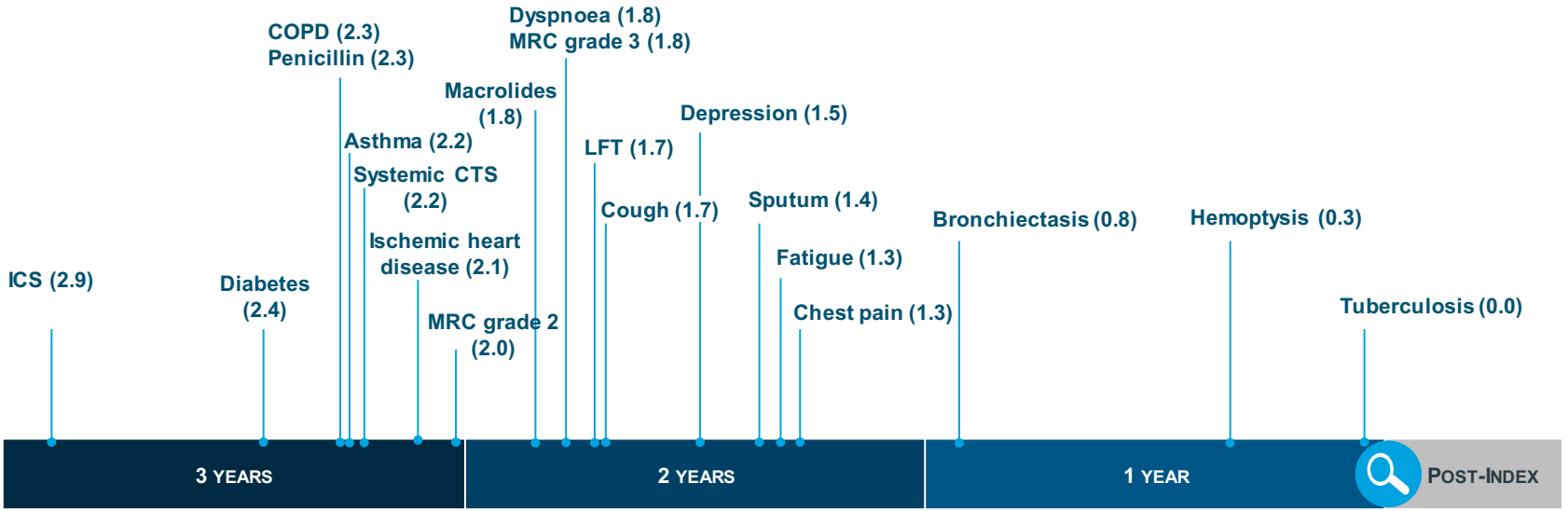

Figure 2: Pre-diagnosis primary care patient journey in NTMLD. Numbers in brackets indicate the median time since index date in years for the top 20 most frequent predictors in the NTMLD case cohort. ICS: Inhaled corticosteroids; LFT: Lung function tests; COPD: Chronic obstructive pulmonary diseases; MRC: Medical Research Council breathlessness score

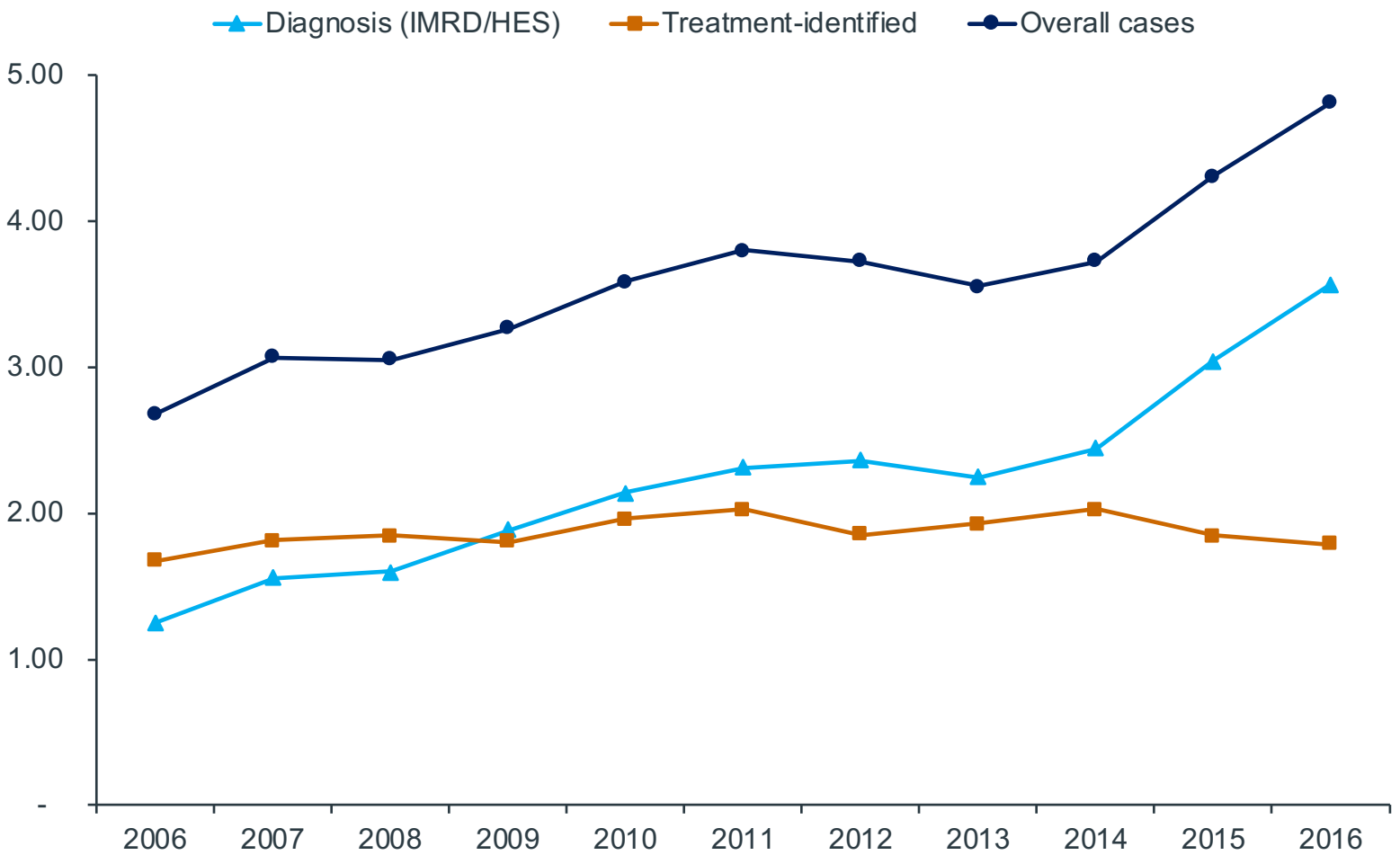

Figure 3 Annual prevalence of diagnosed NTMLD. IMRD: IQVIA Medical Research Data; HES: Hospital Episode Statistics

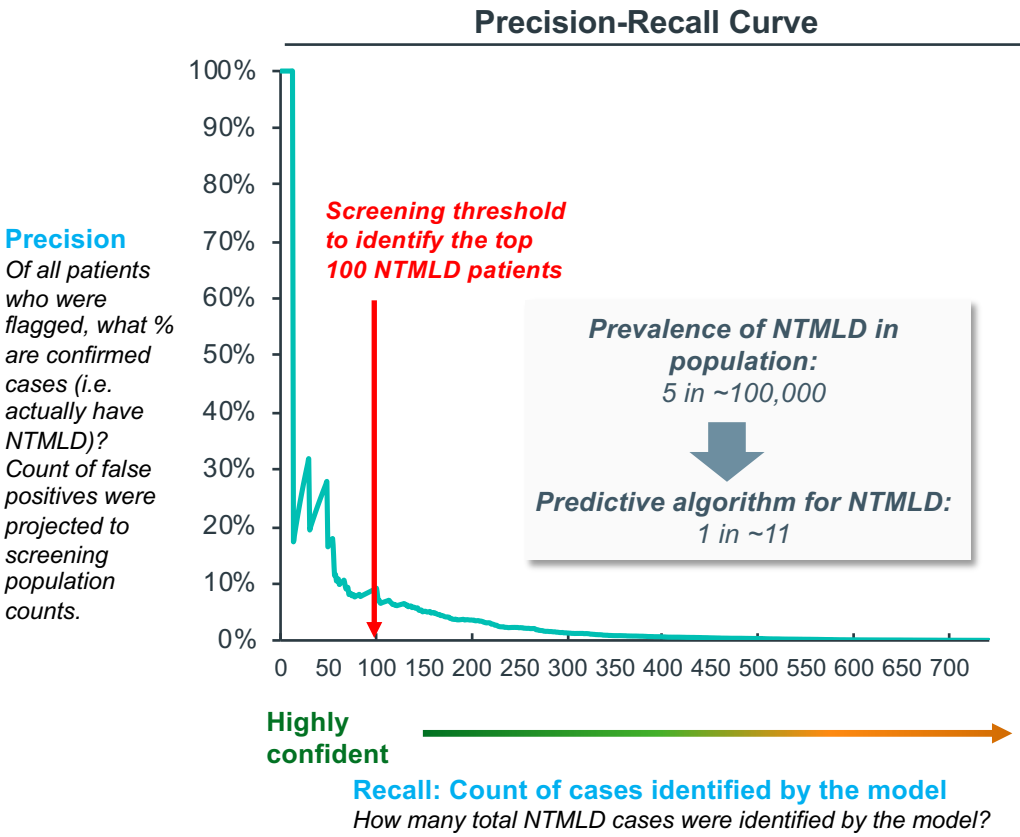

Figure 4a: Projected precision-recall curve for the predictive algorithm

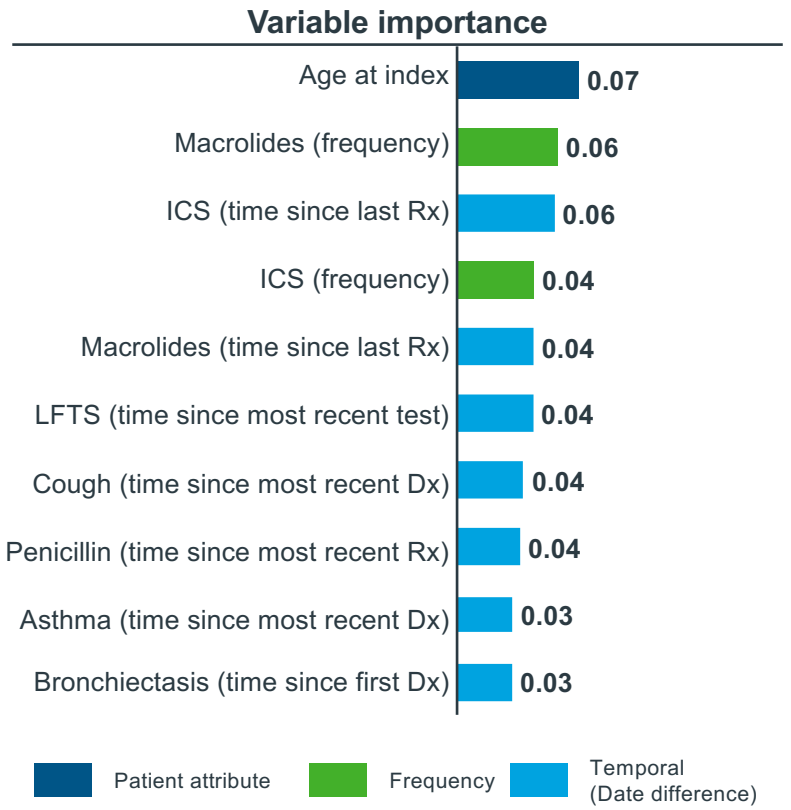

Figure 4b: Variable importance for the predictive algorithm

Figure 4: Projected precision-recall curve and variable importance for the predictive algorithm

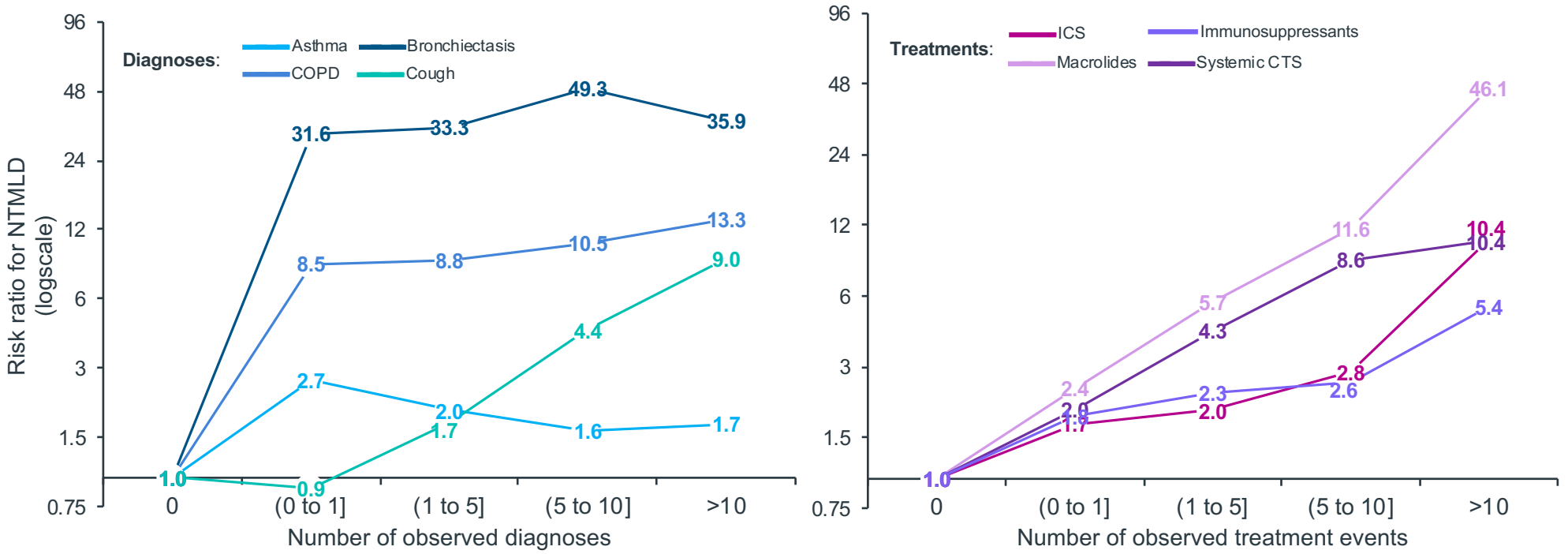

Figure 5: Risk ratios for selected diagnoses and treatments. The change in risk ratio is illustrated for groups of patients with a varying number of diagnostic or treatment events in proportion to the risk observed in patients with no observed record for the diagnostic or treatment event. COPD: Chronic obstructive pulmonary diseases; ICS: Inhaled corticosteroids; Immunosuppressive drugs (including, but not limited to systemic and inhaled corticosteroids, TNF-alfa inhibitors, calcineurin inhibitors, interleukin inhibitors).

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

Database Description

IQVIA Medical Research Data (IMRD), formally known as The Health Improvement Network (THIN) Research Database

IMRD is a large UK primary care database containing EMR information. As of September 2019, IMRD contained non-identified primary care medical records from over 18 million patients, of which approximately 2.9 million are currently active, representing over 4% of the UK population. Data are available from 1990 onwards for many patients, with summarised medical information detailed prior to that. The database holds all prescribed medication, signs, diagnoses, lab tests and additional information such as lifestyle factors, BMI and vaccinations. It is possible to obtain additional information from the healthcare team, patients and their carers.

IMRD have been shown to be generally representative of the UK in terms of age and gender comparisons, and QOF chronic disease prevalence [1, 2]. In addition, a study has been performed which compares IMRD data with practices using a different general practice software system (EMIS), and it was shown to match closely with these data, with the main exception that IMRD data patients are slightly more representative of the most affluent social classes. As this socioeconomic information is available in IMRD data, researchers are able to adjust for it in analyses. Studies using IMRD require review by the Scientific Review Committee (SRC) with no requirement for publication.

Data files in IMRD are arranged in standardised tables. Diagnoses are coded in hierarchical Read codes which are grouped in themed "chapters" and include terms relating to symptoms, diagnoses, procedures, and laboratory tests. Prescription items are coded using Gemscript codes, based on NHS dictionary of medicines and devices and linked to BNF chapters

The list of risk factors was derived from literature sources including British Thoracic Society and American Thoracic Society guidelines alongside input from a clinical expert. The Data Science team in IQVIA responsible for generating the code lists has a process in place for the derivation of relevant and accurate codes for databases utilising Read codes, including IMRD: Broad search terms based on the predictor (comprised of diagnoses and /or tests) were developed by an epidemiologist familiar with the coding structure using medical terms and associated synonyms. These were then confirmed before use by review of a qualified medical practitioner familiar with GP systems used in the UK.

Drug Regimens for Case Selection

Table S1: List of included drug regimens to identify NTMLD patients

1.	Rifampicin	Isoniazid/Ethambutol
2.	Rifabutin	Isoniazid/Ethambutol
3.	Isoniazid	Amikacin
4.	Isoniazid	Streptomycin
5.	Isoniazid	Azithromycin
6.	Isoniazid	Clarithromycin
7.	Isoniazid	Ethambutol
8.	Isoniazid	Linezolid
9.	Isoniazid	Moxifloxacin
10	Isoniazid	Rifabutin
11	Isoniazid	Rifampicin
12	Isoniazid	Cotrimoxazole
13	Ethambutol	Amikacin
14	Ethambutol	Streptomycin
15	Ethambutol	Azithromycin
16	Ethambutol	Clarithromycin
17	Ethambutol	Linezolid
18	Ethambutol	Moxifloxacin
19	Ethambutol	Rifabutin
20	Ethambutol	Rifampicin
21	Ethambutol	Rifampicin/Isoniazid
22	Ethambutol	Cotrimoxazole
23	Amikacin	Azithromycin
24	Amikacin	Clarithromycin
25	Amikacin	Clofazimine
26	Streptomycin	Azithromycin
27	Streptomycin	Clarithromycin
28	Streptomycin	Clofazimine
29	Tigecycline	Clarithromycin
30	Rifabutin	Clarithromycin
31	Clofazimine	Azithromycin
32	Clofazimine	Clarithromycin
33	Azithromycin	Moxifloxacin
34	Azithromycin	Ciprofloxacin
35	Clarithromycin	Moxifloxacin
36	Ethambutol	Ciprofloxacin
37	Clarithromycin	Prothionamide
38	Rifampicin	Clarithromycin
39	Azithromycin	Rifampicin
40	Clarithromycin	Ciprofloxacin

Note: Patients on any of the above combination regimens (including those also on additional antibiotics) were included.

Selection of Predictors

The list of risk factors was derived from literature sources including British Thoracic Society and American Thoracic Society alongside input from clinical key opinion leader. The Data Science team in IQVIA responsible for generating the code lists has a process in place for the derivation of relevant and accurate codes for databases utilising Read codes, including IMRD: Broad search terms based on the predictor (comprised of diagnoses and /or tests) were developed by an epidemiologist familiar with the coding structure using medical terms and associated synonyms. These were then confirmed before use by review of a qualified medic familiar with GP systems used in the UK.

Table S2: List of predictors included in the model

Predictors
Age at index
Alcohol use: hazardous (Moderate alcohol use)
Alcohol use: harmful (Excessive alcohol use)
Arrhythmia
Arteries
Aspergillosis
Asthma
Autoimmune disorders
Biopsy (lung-related only)
Body mass index
Bronchiectasis
Bronchoscopy/Endoscopy/Tracheostomy
Cerebrovascular Disease
Chemical fumes exposure
Chest adenopathy
Chest pain

Congenital respiratory malformations
COPD
Cough
Crackles/rales
Crohn's / Ulcerative colitis / Irritable bowel disease
Cystic fibrosis
Dementia
Depression
Diabetes
Dyspnea
Emphysema
Family number (members of the same postcode or address are given the same family number)
Fatigue
Fever
Gastroesophageal reflux disease
Heart failure
Heart valve disorder
Haemoptysis
HIV
Hyperlipidemia
Idiopathic pulmonary fibrosis
Imaging (X-ray / CAT scan / Fluoroscopy / MRI)
Immune deficiency
Immunosuppressants prescription
Inhaled corticosteroids prescription
Ischemic heart disease
Liver cirrhosis
Lung cancer
Lung function test
Macrolides prescription

Malignancy
Mediastinum test
Metastatic carcinoma
MRC Dyspnoea scale 1
MRC Dyspnoea scale 2
MRC Dyspnoea scale 3
MRC Dyspnoea scale 4
MRC Dyspnoea scale 5
Multiple Sclerosis
Obesity
Organ Transplant
Pulmonary alveolar proteinosis
Primary ciliary dyskinesia
Pectus Excavatum
Penicillin prescription
Pneumoconiosis
Pneumocystis pneumonia
Pneumonia
Pneumonitis
Psoriasis
Psychosis
Pulmonary alveolar proteinosis
Respiratory failure
Respiratory syncytial virus
Rheumatic disease
Scoliosis
Sex
Sjogren’s syndrome
Smoking status at index: Current smoker
Smoking status at index: Ex-smoker

Smoking status at index: Never smoker
Smoking status at index: unknown
Stem cell transplant
Systemic corticosteroids prescription
Systemic lupus erythematosus
TNF inhibitors prescription
Weight loss

Comorbidities and medication use were included in the model using metrics describing their frequency (count divided by length of history) and their timing (days since first and last exposure).

Supplemental Figures

Figure S 1 Annual estimates for total prevalence of NTMLD cases including both diagnosed and undiagnosed cases

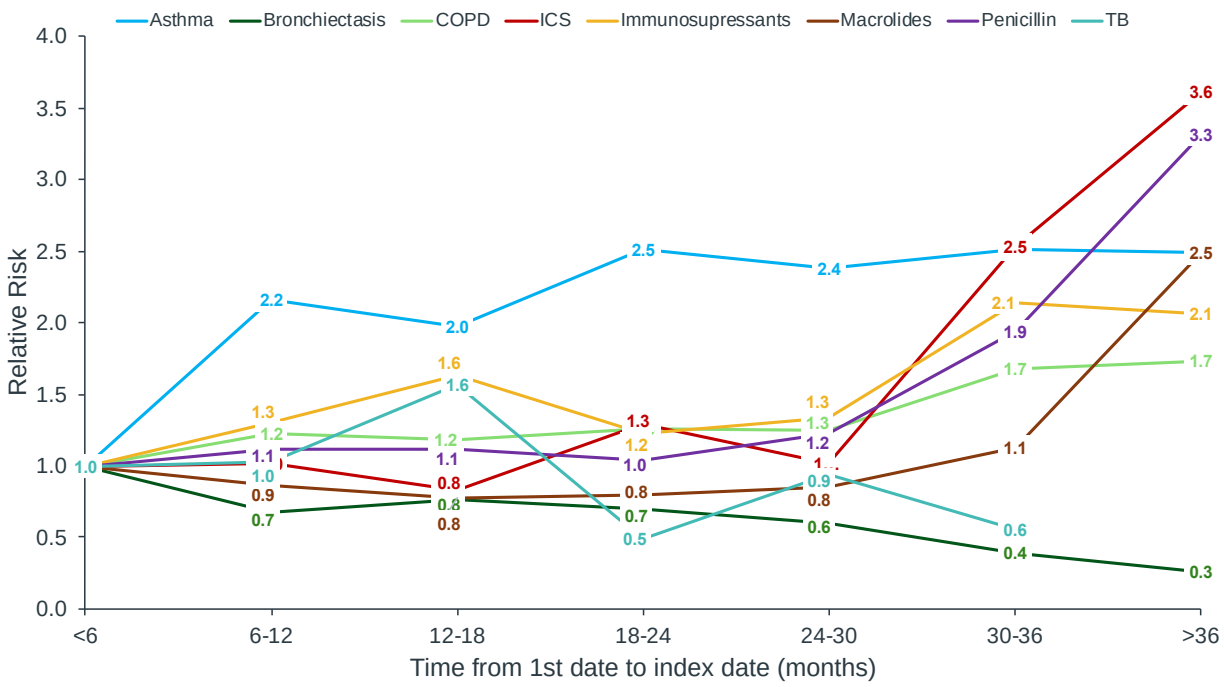

Figure S 2 Relative risk ratios of NTMLD by time between first occurrence and index date. COPD: Chronic obstructive pulmonary diseases; ICS: Inhaled corticosteroids; Immunosuppressive drugs (including, but not limited to systemic and inhaled corticosteroids, TNF-alfa inhibitors, calcineurin inhibitors, interleukin inhibitors).

1. Denburg, M.R., et al., *Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease.* Pharmacoepidemiol Drug Saf, 2011. **20**(11): p. 1138-49.
2. Lewis, J.D., et al., *Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research.* Pharmacoepidemiol Drug Saf, 2007. **16**(4): p. 393-401.