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

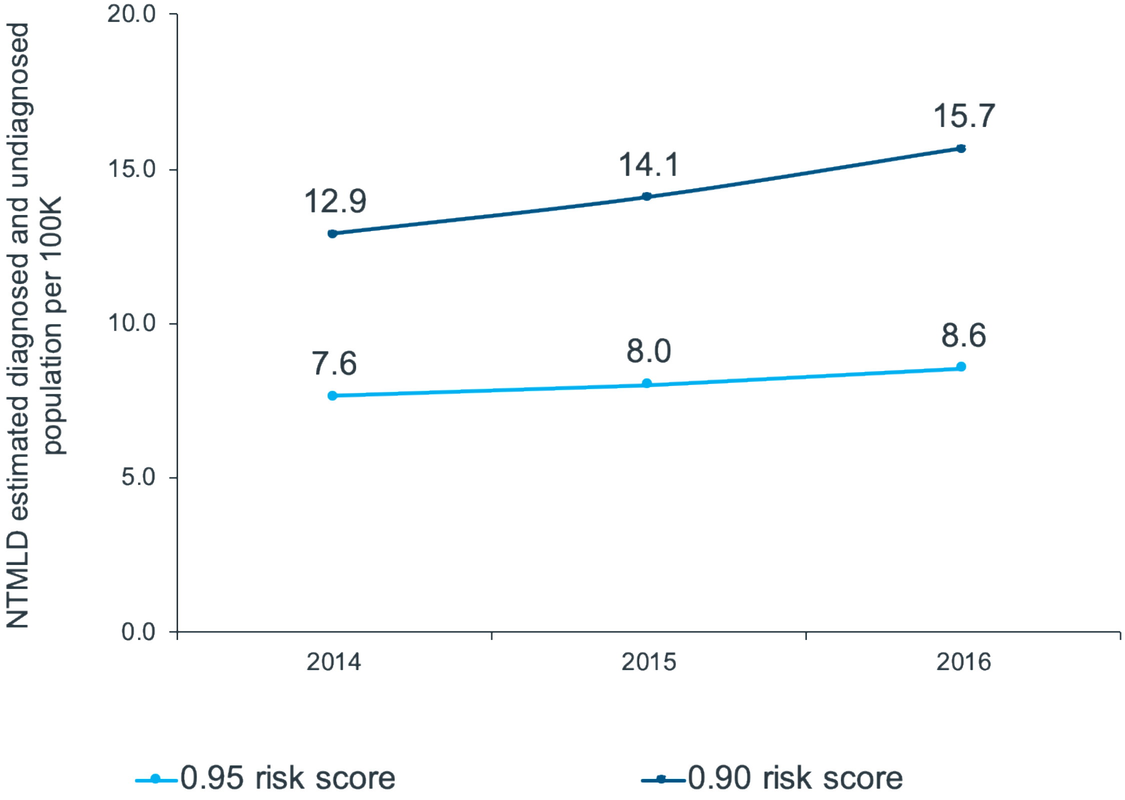

Figure S 1 Annual estimates for total prevalence of NTMLD cases including both diagnosed and undiagnosed cases

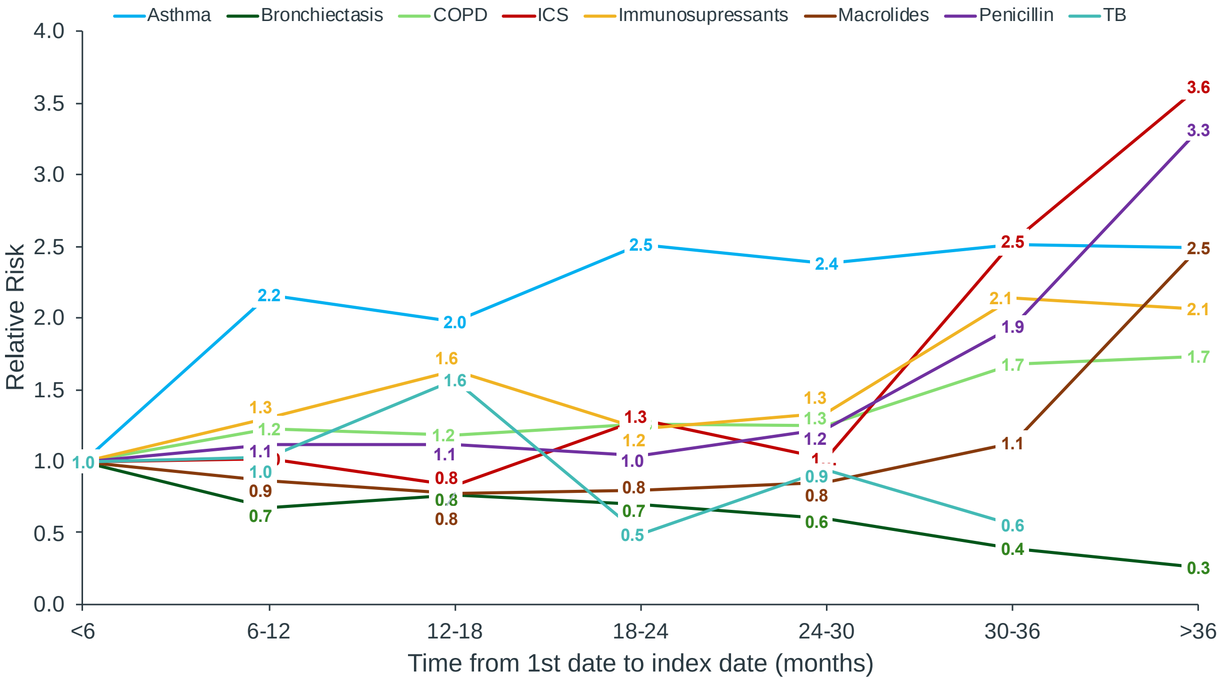

Figure S 2 Relative risk ratios of NTMLD by time between first occurrence and index date. COPD: Chronic obstructive pulmonary diseases; ICS: Inhaled corticosteroids; Immunosuppressive drugs (including, but not limited to systemic and inhaled corticosteroids, TNF-alfa inhibitors, calcineurin inhibitors,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