SHAREABLE PDF

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

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.