Implementation of prognostic machine learning algorithms in paediatric chronic respiratory conditions: a scoping review

Nicole Filipow,1,9 Eleanor Main,1 Neil J Sebire,2,3 John Booth,2,3 Andrew M Taylor,3,4 Gwyneth Davies,2,3 Sanja Stanojevic5

ABSTRACT

Machine learning (ML) holds great potential for predicting clinical outcomes in heterogeneous chronic respiratory diseases (CRD) affecting children, where timely individualised treatments offer opportunities for health optimisation. This paper identifies rate-limiting steps in ML prediction model development that impair clinical translation and discusses regulatory, clinical and ethical considerations for ML implementation. A scoping review of ML prediction models in paediatric CRDs was undertaken using the PRISMA extension scoping review guidelines. From 1209 results, 25 articles published between 2013 and 2021 were evaluated for features of a good clinical prediction model using the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines.

Most of the studies were in asthma (80%), with few in cystic fibrosis (12%), bronchiolitis (4%) and childhood wheeze (4%). There were inconsistencies in model reporting and studies were limited by a lack of validation, and absence of equations or code for replication. Clinician involvement during ML model development is essential and diversity, equity and inclusion should be assessed at each step of the ML pipeline to ensure algorithms do not promote or amplify health disparities among marginalised groups. As ML prediction studies become more frequent, it is important that models are rigorously developed and assessed at each step of the ML pipeline to ensure algorithms do not promote or amplify health disparities among marginalised groups. As ML prediction studies become more frequent, it is important that models are rigorously developed using published guidelines and take account of regulatory frameworks which depend on model complexity, patient safety, accountability and liability.

INTRODUCTION

The rapidly expanding field of machine learning (ML) has created widespread promise in healthcare for the diagnosis, prognosis and management of disease to ultimately enrich personalised medicine. ML is a broad field that uses statistics and algorithms to acquire knowledge from existing data, with the aim of predicting a future outcome for a set of similar circumstances, and the opportunity for an ongoing process of updating and fine tuning when new data are available. A rapid expansion in the application of ML in medicine has been fuelled by vast amounts of data captured through clinical records, imaging, diagnostic investigations, patient registries and more recently electronic health records (EHRs) and wearable devices. As automated data capture becomes more widespread in routine care, so too does the potential for ML models to diagnose disease or predict disease trajectories.

Machine learning

A branch of artificial intelligence, ML uses algorithms to identify patterns in often large and complex datasets that traditional statistical methods can have difficulty uncovering.1 Broadly, ML is separated into supervised, unsupervised or deep learning; each is employed depending on the objective of the analysis and the information presented in the data (figure 1).2 Supervised methods form prediction models based on data with labelled outcomes, for example, when the disease severity of patients is known. Unsupervised methods are used to identify the shared characteristics between similar groups of data where the outcomes are not labelled or defined, for example, to identify clinically meaningful subgroups of disease when the relative disease severities of patients are unknown. Deep learning may be supervised or unsupervised and uses artificial neural networks (ANNs) to learn from data. ANNs are complex models that use many interconnected layers of processing units, termed neurons, which extract levels of information from raw data to generate a set of rules for predictions.3

Predicting clinical outcomes in paediatric chronic respiratory diseases

ML predictive algorithms are particularly attractive within the field of chronic respiratory diseases (CRD), which present with heterogeneous clinical outcomes from diagnosis across the life course. In CRDs that affect children such as asthma, cystic fibrosis (CF), primary ciliary dyskinesia (PCD),
bronchopulmonary dysplasia (BPD) and children’s interstitial lung disease (chILD), the prediction of clinical outcomes is especially important, where timely individualised treatment regimes offer opportunities for health maintenance before symptoms of the disease become severe and irreversible.4 CRDs in children often involve longitudinal follow-up over multiple years with complex outcomes from clinical encounters which may be captured repeatedly through patient registries, cohort studies, or EHRs. These large datasets have driven the development of ML algorithms to predict likelihood of unfavourable clinical outcomes common in paediatrics such as respiratory exacerbation, hospitalisation, or accelerated lung function decline, with the aim of supporting early treatment decisions to prevent severe outcomes such as lung transplant or death.5 6 The adoption of ML predictive models in clinical care is rare however, which is discouraging given the increase of ML publications in respiratory medicine in the last decade.7 A series of recent reviews in other disease areas has highlighted inaccuracies and failures in reporting standards of prognostic models generally as the major constraint to clinical translation.8-10

Objectives
While opportunities exist for ML prediction models to impact clinical care, challenges to implementation remain a barrier to clinical use. To explore the gap between model development and clinical application specific to CRDs affecting children, we carried out a scoping review of the available literature to evaluate the reporting of ML prediction models and identify the rate-limiting steps in model development that impair clinical translation. We further discuss regulatory, clinical and ethical considerations for implementation and the future opportunities for EHRs to influence ML prediction models in clinical care.

METHODS
Overview
We carried out a scoping review using the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for scoping reviews guidelines11 to identify prognostic ML algorithms in CRDs that affect children, including but not limited to CF, bronchiectasis, asthma, PCD, BPD and chILD. The purpose of the review was not to provide a summary of models in specific diseases, but rather to investigate the rate limiting steps to clinical implementation of ML predictive models generally across paediatric CRDs, which have in common similar predictors and outcomes.

To identify barriers to clinical implementation within model development, relevant ML models were evaluated with reference to the key recommendations for model reporting specific for respiratory, sleep and critical care studies, summarised below.12 These metrics were summarised from the published guidelines for the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD).13 14

Search criteria
A search for published articles was performed in the Medline database using a previously curated list of phrases to identify prediction studies,15 and included the updated phrase suggestions.16 To filter for studies that used ML, the following MeSH terms and keywords were included: (Unsupervised Machine Learning/ or unsupervised learning.mp.) or (machine learning.mp. or Machine Learning/) or (artificial intelligence.mp. or Artificial Intelligence/) or (Supervised Machine Learning/ or supervised learning.mp.) or (deep learning.mp. or Deep Learning/) or (Neural Networks, Computer/ or neural network*.mp.) or ((cluster analysis or clustering).mp. or Cluster Analysis/) or ((support vector machine or SVM). mp. or Support Vector Machine/) or random forest*.mp. or (decision tree*.mp. or Decision Trees/) or Bayesian. mp.

Respiratory MeSH terms and keywords included (cystic fibrosis.mp. or Cystic Fibrosis/) or (Asthma/ or asthma. mp.) or (Bronchiectasis/ or bronchiectasis.mp.) or (Bronchopulmonary Dysplasia.mp. or Bronchopulmonary Dysplasia/) or (primary ciliary dyskinesia.mp. or Ciliary Motility Disorders/) or (interstitial lung disease.mp. or Lung Diseases, Interstitial/) or (chronic respiratory disease or chronic respiratory illness or chronic respiratory condition).mp.

Paediatric studies were identified from the patient ages in the study data rather than included as a search term to not exclude articles that did not specifically mention paediatrics. The search was limited to publications in the past decade (2011-15 October 2021), since it was anticipated that most ML prediction models would have been recently published given the rise in ML studies in respiratory medicine in the past decade.7 Furthermore, EHR systems were not implemented widely into healthcare systems prior to 2010.17 Any subsequent related studies of relevant articles were searched for to ensure all aspects of model development and validation were captured.
**Table 1** Key reporting elements evaluated in this scoping review to infer features of a good clinical prediction model

<table>
<thead>
<tr>
<th>Features</th>
<th>Definition</th>
<th>Key reporting elements*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalisability</td>
<td>How well the model works in populations external to the study population, and as such can be used to infer performance in a clinical setting</td>
<td>Data source, Participants, Validation (internal/external)</td>
</tr>
<tr>
<td>Biasedness</td>
<td>Occurs when certain elements are more heavily weighted than others, or with inconsistency or subjectivity in defining the outcome.</td>
<td>Missing Data, Outcomes</td>
</tr>
<tr>
<td>Interpretability</td>
<td>How well the model is understood by clinicians</td>
<td>Predictors</td>
</tr>
<tr>
<td>Replicability</td>
<td>The ability to replicate the model in the same or independent population</td>
<td>Model specification, Model structure</td>
</tr>
<tr>
<td>Performance</td>
<td>Whether the model provides benefit to patients</td>
<td>Prospective study, Randomised controlled trial</td>
</tr>
</tbody>
</table>

*Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis guidelines summarised in Leisman et al. 12*

Articles were excluded from review based on the following criteria: (1) not a primary journal article, (2) irrelevant (ie, in vitro studies, pharmacokinetic models, ML model not developed, not a CRD), (3) diagnostic or disease differentiation models, (4) descriptive models, (5) not predictive of clinical outcomes (ie, predictive of cost of care), (6) did not use primarily paediatric data, (7) did not report the age of study participants. The initial search results were filtered through a title search, and potential articles were further screened through a review of abstracts and full text.

**Evaluating ML prediction studies**

Using hallmarks of ML prediction studies identified in Leisman et al, 12 models were evaluated for their reporting of metrics that infer features of a good clinical prediction model: generalisability, biasedness, interpretability, replicability and clinical performance 13-18 (table 1).

Generalisability was assessed through investigation of study characteristics including the location, data source, number of centres, dates of investigation, patient characteristics, as well as evidence of internal and/or external validations. To investigate potential sources of bias, the handling of missing data, sample size and the definition of the outcome was noted. Interpretability was inferred through methods of predictor selection, numbers of predictors, as well as any methods to meaningfully describe the resulting model. Replicability was inferred if the model structure/specificity was provided through equations or shared code.

Traditional measures of model performance involve a range of metrics that assess how well the model classifies data compared with the labelled classifications during internal and external validations, such as area under the receiver operator curve, specificity/sensitivity, accuracy, or precision–recall curves. 12 However, high performance by these measures does not inevitably represent clinical efficacy or patient benefit. 18 As such, this review focused on whether a prognostic study in a randomised controlled trial (RCT) has been carried out to evaluate clinical performance of the ML model.

**RESULTS**

**Study selection**

A flow chart of the scoping review process is displayed in figure 2. There were 1209 results, 243 abstracts were screened and 25 articles were included in the review, which are summarised in table 2.

The studies selected for review were published between 2013 and 2021; 72% were published since 2018 (figure 2). Most of the studies were related to asthma (80%), with few in CF (12%), bronchiolitis (4%) and childhood wheeze (4%). While a small number of studies using ML were identified in BPD, PCD, chronic cough and bronchiectasis, they were either diagnostic or disease differentiation ML models, or were carried out in adults and were excluded. The majority of studies used disease exacerbation or hospitalisation as an outcome, with other predictions including risk of lung damage (through quantification of imaging with CT), development of comorbidities (eg, developing asthma in childhood wheeze), disease-specific measures such as asthma control, or positive clinical outcomes such as asthma remission or response to treatment. One study examined risk of critical care, which was defined as admission to ICU or death; otherwise, risk of death or lung transplant was not assessed as a primary outcome which was expected, given their rare occurrence in paediatrics.

A range of ML algorithms were used, with more studies using supervised (72%) over unsupervised (20%) and deep learning (8%) methods. Some studies employed multiple ML methods to identify the optimal model while others focused on the development of a single model. Random forest was the most widely used supervised method (n=10), followed by decision trees (n=5), Bayesian models (n=4), support vector machines (n=4), Lasso (n=3), various boosting methods (n=3), and combined models (ie, autoML, ensemble learning,
predictor pursuit), where predictions are made from multiple sequential methods (n=3). ANNs were the only deep learning methods used and cluster analysis was the only unsupervised method. Many of the descriptive studies excluded from review used cluster analysis to define the characteristics of subgroups of disease, without predicting future outcomes.

Generalisability
There were 19 studies (76%) that reported each of the data and patient metrics used to infer generalisability (figure 3). Often the data were described from previous studies but it was not always clear if the original data exclusions also applied to the present study. Clarity on these details should be included in the main text. Most studies originated from a single centre (52%) rather than multicentre or a national database (44%). There were more studies with data from North America (68%) than Europe (24%), Australia (4%), or the Middle East (4%). Data collected during studies (i.e., cross-sectional, longitudinal cohort) were the most common sources of data (56%), followed by registry data from either routine EHR (24%), regional or national databases (12%), or clinical records (4%). One study did not report the source of data for model development (4%).

Participant ages within a study ranged from infants less than a year old, to 2–22 years. Dates of study data ranged from 1993 to 2019, 6 studies (24%) did not report any study dates. There was no evidence of any model being updated over time. Year of publication did not always correlate with years of study data.

There were 18 studies (72%) that carried out an internal validation. This was most often accomplished by splitting the dataset into a test and training set. Only a single study (4%) carried out an external validation in geographically different data (figure 3).

Bias
Sample sizes ranged considerably across studies, from small scale studies (n=49 people) to larger analyses (n=52 037 people). It was often not clear whether large-scale studies included data of repeated measures, or if they were independent records. In handling missing values, 32% of studies used complete case analysis, 28% imputed missing values, and one used a combination of both; however, 36% of studies did not define any explicit methods. In defining outcomes, proxy measures were often used, for example exacerbation was often recorded as requirement of a medication, which can be biased towards clinician or centre treatment preferences.

Interpretability
The number of predictors ranged from 9 to 648. Studies using large numbers of predictors (n>50) did not typically rely on any variable reduction techniques or they did not describe if or which variables were included in the final prediction model if variable reduction was considered. These models are uninterpretable as it is unknown which of the hundreds of variables influenced clinically relevant poor outcomes for a particular person. Studies using smaller numbers often used clinical knowledge to select variables, excluded those with high missingness, or used various statistical techniques to ensure included variables were clinically relevant, which may allow for more interpretability.

Repeatability
None of the studies shared any code or equations for their predictive models.

Clinical performance
Two studies carried out prospective studies to assess the performance of a ML model in a clinical setting. One
<table>
<thead>
<tr>
<th>Author</th>
<th>Disease</th>
<th>Data source</th>
<th>Centres</th>
<th>Study dates</th>
<th>Age range</th>
<th>Primary outcome(s)</th>
<th>Missing data</th>
<th>ML method (best model)</th>
<th>Validation</th>
<th>Prospective study/RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hogan et al 2021</td>
<td>Asthma</td>
<td>Registry (National Database)</td>
<td>NR</td>
<td>2013</td>
<td>5–18</td>
<td>Hospitalisation</td>
<td>Complete case analysis</td>
<td>ANN</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Filipow et al 2021</td>
<td>CF</td>
<td>Registry (Regional Database)</td>
<td>1</td>
<td>2000–2018</td>
<td>2–18</td>
<td>Hospitalisation, exacerbation</td>
<td>Complete case analysis</td>
<td>Cluster Analysis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Bose et al 2021</td>
<td>Asthma</td>
<td>Registry (EHR)</td>
<td>30+</td>
<td>2005–2016</td>
<td>2–5</td>
<td>Asthma persistence</td>
<td>Mixed</td>
<td>XGBoost</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>van Horck et al 2021</td>
<td>CF</td>
<td>Study data</td>
<td>3</td>
<td>NR</td>
<td>5–18</td>
<td>Exacerbation</td>
<td>NR</td>
<td>Random Forest</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Raita et al 2021</td>
<td>Bronchiolitis</td>
<td>Study data</td>
<td>17</td>
<td>2011–2014</td>
<td>&lt;1</td>
<td>Wheeze, asthma</td>
<td>Imputed</td>
<td>Cluster Analysis</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sills et al 2021</td>
<td>Asthma</td>
<td>Registry (EHR)</td>
<td>5</td>
<td>2009–2013</td>
<td>2–21</td>
<td>Hospitalisation</td>
<td>Complete case analysis</td>
<td>AutoML</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Seol et al 2021</td>
<td>Asthma</td>
<td>NR</td>
<td>1</td>
<td>NR</td>
<td>&lt;18</td>
<td>Exacerbation</td>
<td>NR</td>
<td>Naïve Bayes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Lovric et al 2021</td>
<td>Asthma</td>
<td>Study data</td>
<td>1</td>
<td>NR</td>
<td>2–22</td>
<td>Response to treatment</td>
<td>Imputed</td>
<td>AdaBoost</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Caparros-Martin et al 2020</td>
<td>CF</td>
<td>Study data</td>
<td>NR</td>
<td>NR</td>
<td>2.5–4.4</td>
<td>Lung damage (quantified CT)</td>
<td>NR</td>
<td>Cluster Analysis</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Wang et al 2019</td>
<td>Asthma</td>
<td>Study data</td>
<td>8</td>
<td>1993–1995</td>
<td>5–12</td>
<td>Asthma remission</td>
<td>Complete case analysis</td>
<td>Decision Tree</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Messinger et al 2019</td>
<td>Asthma</td>
<td>Registry (EHR)</td>
<td>1</td>
<td>2016–2017</td>
<td>2–18</td>
<td>Paediatric Asthma Score</td>
<td>Complete case analysis</td>
<td>ANN</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Khasha et al 2019</td>
<td>Asthma</td>
<td>Study data</td>
<td>1</td>
<td>NR</td>
<td>&gt;5</td>
<td>Asthma control</td>
<td>Imputed</td>
<td>Ensemble Learning</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Goto et al 2019</td>
<td>Asthma</td>
<td>Registry (National Database)</td>
<td>NR</td>
<td>2007–2015</td>
<td>2–14 (IQR)</td>
<td>Critical care (admission to intensive care unit or death), hospitalisation</td>
<td>Imputed</td>
<td>Decision Tree</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Patel et al 2018</td>
<td>Asthma</td>
<td>Registry (EHR)</td>
<td>2</td>
<td>2012–2015</td>
<td>2–18</td>
<td>Hospitalisation</td>
<td>Complete case analysis</td>
<td>Gradient Boosting</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ross et al 2018</td>
<td>Asthma</td>
<td>Study data</td>
<td>8</td>
<td>1993–1995</td>
<td>5–12</td>
<td>Asthma control</td>
<td>Imputed</td>
<td>Predictor Pursuit</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Spyroglou et al 2018</td>
<td>Asthma</td>
<td>Study data</td>
<td>1</td>
<td>2008–2016</td>
<td>1–14.5</td>
<td>Exacerbation</td>
<td>NR</td>
<td>Bayesian</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Huffaker et al 2018</td>
<td>Asthma</td>
<td>Study data</td>
<td>1</td>
<td>2016–2017</td>
<td>5–17</td>
<td>Exacerbation</td>
<td>Complete case analysis</td>
<td>Random Forest</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Table 2  Continued

<table>
<thead>
<tr>
<th>Author</th>
<th>Disease</th>
<th>Data source</th>
<th>Centres</th>
<th>Study dates</th>
<th>Age range</th>
<th>Primary outcome(s)</th>
<th>Missing data</th>
<th>ML method (best model)</th>
<th>Validation</th>
<th>Prospective study/RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shin et al 2018</td>
<td>Asthma</td>
<td>Registry (EHR)</td>
<td>1</td>
<td>2016</td>
<td>6–18</td>
<td>Hospitalisation</td>
<td>NR</td>
<td>Random Forest</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Das et al 2017</td>
<td>Asthma</td>
<td>Registry (EHR)</td>
<td>1</td>
<td>2013–2014</td>
<td>≤18</td>
<td>Frequent</td>
<td>NR</td>
<td>Logistic Regression</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Pite et al 2016</td>
<td>Wheeze</td>
<td>Study data</td>
<td>1</td>
<td>1993–2006</td>
<td>≤7</td>
<td>Asthma</td>
<td>Complete case analysis</td>
<td>Cluster Analysis</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Van Vliet et al</td>
<td>Asthma</td>
<td>Study data</td>
<td>1</td>
<td>NR</td>
<td>6–18</td>
<td>Asthma control</td>
<td>NR</td>
<td>Random Forest</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Luo et al 2015</td>
<td>Asthma</td>
<td>Study data</td>
<td>4</td>
<td>2011–2012</td>
<td>2–18</td>
<td>Asthma control 1 week prior</td>
<td>NR</td>
<td>Multiboost with Decision Stumps</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Howrylak et al 2014</td>
<td>Asthma</td>
<td>Study data</td>
<td>8</td>
<td>1993–1995</td>
<td>5–12</td>
<td>Exacerbation</td>
<td>Imputed</td>
<td>Cluster Analysis</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Farion et al 2013</td>
<td>Asthma</td>
<td>Registry (clinical records)</td>
<td>1</td>
<td>2000–2004</td>
<td>1–17</td>
<td>Exacerbation severity</td>
<td>Imputed</td>
<td>Naive Bayes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Robroeks et al 2013</td>
<td>Asthma</td>
<td>Study data</td>
<td>1</td>
<td>NR</td>
<td>6–16</td>
<td>Exacerbation</td>
<td>NR</td>
<td>SVM</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

ANN, artificial neural network; EHR, electronic health record; NR, not reported; RCT, randomised controlled trial; SVM, support vector machine.
study assessed the accuracy of a naïve Bayes model compared with both a standard score and physician decisions in a prospective study to predict severity of asthma exacerbation in the emergency department. The naïve Bayes performed with less accuracy than both.24

In an RCT, the second study provided an asthma exacerbation prediction model to physicians in the intervention group, while standard care was maintained in the control group. There was no difference in prevalence of exacerbation within 1 year for patients in either group, although the physicians in the intervention group had a reduced time in reviewing EHRs for asthma management.29

**DISCUSSION**

The 25 prognostic ML studies assessed in this scoping review were overwhelmingly focused on asthma and the majority were supervised models. The studies were mainly limited by a lack of validation or prospective study, and the absence of equations or code for replication, which are major steps required for clinical implementation. Some recent studies used data from 1 to 2 decades ago, which may have limited relevance to current populations for which treatments and care have changed. Some of the models were opaque, uninterpretable models that used high numbers of predictors and did not explain the resulting predictions. This is especially important in healthcare since a clinician needs to know not only who is at risk, but also what they can do to change the outcome.

A large proportion of studies did not report on the handling of missing data, which does not provide transparency to evaluate whether sample populations are under-represented, for example, towards those who are sicker and have more data. Smaller datasets were typically derived from research studies, where there is greater control over the variables collected or the inclusion criteria for the study. However, ML methods were typically developed for large datasets, and studies using national/regional databases, EHRs, or data from daily home monitoring benefit from large samples likely more representative of wider populations.

External validations are necessary to understand the generalisability of the predictions; however, only one was conducted. In the study, similar clusters of children with CF developed from data in Canada were identified in data from the UK, providing evidence for the generalisability of the model.25 Internal validations were frequent, but their performance relies heavily on the definition of the outcome. If the outcome is somewhat subjectively captured, for example, prescription of medication, the resulting predictions are biased towards the subjective. This is highlighted in the two prospective studies that identified no patient benefit despite good model performance during development.24 29 If the models are trained on data where the outcome is influenced by clinician decision, it is unsurprising that the models would not outperform a clinician. While these models may benefit areas of healthcare such as easing/increasing clinician workflow, objectively captured outcomes such as chest imaging, lung function, or physiological data may result in models with greater patient benefit.

This scoping review was limited in that the studies were not assessed with the full TRIPOD guidelines, and bias and clinical applicability were not assessed with the full Prediction model Risk Of Bias ASeessment Tool46 guidelines. A summarised reporting checklist was instead used, which investigated the articles at an overarching level rather than a granular level to identify key themes. Even without detailed assessment using the full reporting checklists, the summarised checklist revealed that studies still largely failed to report on or carry out key metrics, and thus more granular investigation at this point was not required to identify shortcomings in model reporting. Development of ML prediction models is still an unexplored area of research in paediatric CRDs other than asthma, highlighted here by a lack of studies identified in other respiratory conditions. As research into these areas continues, and as ML prediction studies in paediatric CRDs are becoming more frequent (72% published since 2018), it is important that the models are rigorously developed. A quality assessment tool for artificial intelligence-centered diagnostic studies is currently being developed, and combined with the TRIPOD guidelines for prediction studies will be useful for designing future ML prediction models with clinical implications.43

**Further considerations**

The lack of model implementation is a point of discussion in healthcare generally, and in addition to model development and reporting require regulatory, clinical and ethical frameworks.18 48–51 A hypothetical pathway...
for ML model development using these frameworks is summarised in figure 4.

Regulatory
The level of regulation and approval required for a prediction model can depend on its complexity, where more complicated, uninterpretable models are classed as a medical device and must be approved by relevant governing bodies, such as the Food & Drugs Administration (FDA) in the USA, or the Medicines and Healthcare products Regulatory Agency; PROBAST, Prediction model Risk Of Bias A SSESSment Tool; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Clinical
Implementation also requires the confidence of clinicians, and clinician involvement during model development is essential. Especially in respiratory disease, prior research has generated ample knowledge on contributors of poor outcomes, which should not be ignored in model development or assessment. Combining clinical knowledge with ML may improve both performance and clinical trust in models, better facilitating their adoption in clinical care.

There is currently a lack of knowledge translation and implementation science between data scientists and clinicians, which are needed to be integrated into model development. Qualitative research may be necessary to gauge acceptance and potential utility of predictive models before they are developed.
Ethical
ML algorithms have been known to amplify or create health disparities among marginalised groups. Ethical concerns can arise at every step of ML model development, including the selection/funding of the problem, collection of data, definition of the outcome, algorithm development and algorithm monitoring post deployment. These issues can arise from inconsistencies in access to healthcare or under representation of certain groups in particular centres, which is reflected in the data used to train models. Including variables directly in the model to account for marginalised groups, such as gender or ethnicity, is not always the best practice and may perpetuate the biases. A review detailing a roadmap for responsible and ethical ML in healthcare is useful for addressing some of these concerns. Diversity, equity and inclusion should be considered at every step of ML model development.

Opportunities with EHRs
The opportunity for ML to support clinical decisions has been pronounced through the adoption of EHR in healthcare systems. EHRs are often unstructured and inconsistently captured; however, they are a rich, real-world source of vast amounts of clinical data useful for uncovering meaningful patterns. Data infrastructure plays a key role in harnessing EHRs to enable the extraction, processing and analysis of large volumes of data. Feasibility and interoperability between data systems are important for this process, and standards such as fast healthcare interoperability resources (FHIR) should be considered (https://www.hl7.org/fhir/).

With appropriate infrastructure, a streamlined process between data capture, analytics and implementation can exist to predict outcomes for patient data at a new clinical encounter or visualise patient trajectories over time to support or inform clinical practice (figure 5). As EHR data grow large over time, the algorithms can and should be updated to reflect newer cohorts or include new information. The process is easily severed if steps for implementation are not considered or followed through, which risks an abundance of models that fail to be implemented into clinical practice. It is therefore necessary that models are developed to be generalisable, unbiased and interpretable with good clinical performance, and consider regulatory, clinical and ethical frameworks for implementation.

CONCLUSIONS
The 25 prognostic ML algorithms in CRDs affecting children assessed in this scoping review were most notably limited by a lack of validations and replicability. For ML to enhance personalised medicine and influence clinical care, it is important that the models are rigorously developed and that the regulatory, clinical and ethical frameworks for implementation are considered at every step of the ML pipeline—from predevelopment to post implementation. This is especially important as EHRs become more widespread and facilitate the integration of ML algorithms directly into clinical care.

Figure 5 Opportunity for machine learning (ML) in clinical care with the availability of electronic health records (EHR) systems. Patient data at each clinical encounter is stored in secured databases or analytical platforms, which have the capacity or potential to be accessed by researchers. Open-source software in healthcare means collaboration is more feasible in developing ML models. Following the appropriate steps for regulation and implementation, the resulting algorithms can be fed back into EHR systems to calculate the risk of poor outcomes for incoming data from a new clinical encounter. This can be accessed by clinicians to support or inform clinical care and decision-making. The analytical approach also has the potential to merge data from external sources, such as research or wearable devices to improve model performance. Furthermore, patients can often now access their own EHR data through apps and online patient portals. This has the potential to display the results of individual calculated risks should it be considered appropriate and the appropriate regulatory or governance processes applied.

Contributors GD, SS, EM and NF designed the study. NF reviewed the articles and wrote the initial draft. EM, GD and SS edited the manuscript. NJS, AMT and JB reviewed the final manuscript. All authors reviewed and approved the final version of the manuscript.

Funding NF received funding from a UCL, GOSH and Toronto SickKids studentship. GD is funded by a UK Research and Innovation Future Leaders Fellowship (Grant reference MR/T041285/1). SS received funding from a SickKids CFIT Synergy grant, Vertex Pharmaceutical Research Innovation grant, and the European Respiratory Society. All research at Great Ormond Street Hospital (GOSH) NHS Foundation Trust and UCL Great Ormond Street Institute of Child Health is made possible by the NIHR GOSH Biomedical Research Centre.

Disclaimer The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests GD reports personal fees from Chiesi Ltd and Vertex Pharmaceuticals, outside the submitted work. SS reports personal fees from Chiesi Pharmaceuticals and the European Respiratory Society, and leadership roles for the American Thoracic Society PFT committee, and the European Respiratory Society: Global Lung Function Initiative Network. NS reports Elsevier paediatric pathology textbook royalties (author), outside the submitted work.

Patient consent for publication Not applicable.

Ethics approval Not applicable.
REFERENCES