STUDY PROTOCOL

Protocol for the derivation and validation of a clinical prediction model to support the diagnosis of asthma in children and young people in primary care [version 1; peer review: 2 approved]

Luke Daines [1], Laura J. Bonnett [2], Andy Boyd [3], Steve Turner [4], Steff Lewis [1], Aziz Sheikh [1], Hilary Pinnock[1]

[1]Asthma UK Centre for Applied Research, Usher Institute, University of Edinburgh, Edinburgh, EH8 9AG, UK
[2]Department of Biostatistics, University of Liverpool, Liverpool, UK
[3]Institute of Population Health Science, University of Bristol, Bristol, UK
[4]Department of Child Health, University of Aberdeen, Aberdeen, UK

v1 **First published:** 24 Mar 2020, **5**:50
https://doi.org/10.12688/wellcomeopenres.15751.1

Latest published: 24 Mar 2020, **5**:50
https://doi.org/10.12688/wellcomeopenres.15751.1

Abstract

Background: Accurately diagnosing asthma can be challenging. Uncertainty about the best combination of clinical features and investigations for asthma diagnosis is reflected in conflicting recommendations from international guidelines. One solution could be a clinical prediction model to support health professionals estimate the probability of an asthma diagnosis. However, systematic review evidence identifies that existing models for asthma diagnosis are at high risk of bias and unsuitable for clinical use. Being mindful of previous limitations, this protocol describes plans to derive and validate a prediction model for use by healthcare professionals to aid diagnostic decision making during assessment of a child or young person with symptoms suggestive of asthma in primary care.

Methods: A prediction model will be derived using data from the Avon Longitudinal Study of Parents and Children (ALSPAC) and linked primary care electronic health records (EHR). Data will be included from study participants up to 25 years of age where permissions exist to use their linked EHR. Participants will be identified as having asthma if they received at least three prescriptions for an inhaled corticosteroid within a one-year period and have an asthma code in their EHR. To deal with missing data we will consider conducting a complete case analysis. However, if the exclusion of cases with missing data substantially reduces the total sample size, multiple imputation will be used. A multivariable logistic regression model will be fitted with backward stepwise selection of candidate predictors. Apparent model performance will be assessed before internal validation using bootstrapping techniques. The model will be adjusted for optimism before external validation in a dataset created from the Optimum Patient Care Research Database.

Discussion: This protocol describes a robust strategy for the derivation and validation of a prediction model to support the diagnosis of asthma in children and young people in primary care.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 1 24 Mar 2020	✓ report	✓ report

1 **Glen Martin**, University of Manchester, Manchester, UK

2 **Patrick J E Bindels**, Erasmus MC, Rotterdam, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

Keywords

Asthma, Diagnosis, Primary Care, Clinical Prediction Models, ALSPAC

This article is included in the Avon Longitudinal Study of Parents and Children (ALSPAC) gateway.

Corresponding author: Luke Daines (luke.daines@ed.ac.uk)

Author roles: Daines L: Conceptualization, Data Curation, Funding Acquisition, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Bonnett LJ**: Methodology, Software, Supervision, Writing – Review & Editing; **Boyd A**: Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Turner S**: Supervision, Writing – Review & Editing; **Lewis S**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Sheikh A**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Pinnock H**: Conceptualization, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project was funded by LD's CSO Clinical Academic Fellowship (CAF/17/01) and supported by the Asthma UK Centre for Applied Research. The UK Medical Research Council and Wellcome (Grant ref: 102215, 217065) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and Luke Daines and Hilary Pinnock will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf); ALSPAC's record linkage program was established using funding from Wellcome (Grant ref: 086118) and subsequently through the core support.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Daines L, Bonnett LJ, Boyd A *et al.* **Protocol for the derivation and validation of a clinical prediction model to support the diagnosis of asthma in children and young people in primary care [version 1; peer review: 2 approved]** Wellcome Open Research 2020, **5**:50 https://doi.org/10.12688/wellcomeopenres.15751.1

First published: 24 Mar 2020, **5**:50 https://doi.org/10.12688/wellcomeopenres.15751.1

Introduction

Asthma affects an estimated 339 million people worldwide[1] but is commonly mis-diagnosed in children and adults in primary care[2,3]. Incorrectly labelling an individual with asthma can result in the prescription of inappropriate treatment and the underlying cause of symptoms being missed. On the other hand, not recognising that an individual has asthma can lead to ongoing symptoms, reduced quality of life and risk of asthma attack.

In primary care, a diagnosis of asthma is commonly based on clinical features, previous history together with evidence of variable airflow limitation or, more recently, airway inflammation[4,5]. However, as a heterogeneous and variable condition, the clinical features of asthma can differ according to phenotype, and due to the varying intensity of symptoms and signs over time[5,6]. In addition, no single investigation can confirm or refute asthma in every situation[4]. Consequently, the best strategy for confirming an asthma diagnosis in primary care remains unclear[7–9], and national/international guidelines recommend conflicting diagnostic strategies[4,5,10], which has led to confusion and uncertainty amongst health professionals[11].

A clinical prediction model for asthma diagnosis in primary care could support diagnostic decision making by providing the probability that asthma is present using information from clinical symptoms, signs, medical and family history or tests. A recent systematic review identified seven clinical prediction models for asthma diagnosis in primary care, six models were derived in adults, and one in children[12]. Unfortunately each of the prediction models had methodological limitations and were considered unsuitable for clinical practice[12]. Therefore, this protocol describes plans to derive and internally and externally validate a prediction model intended for use by a primary healthcare professional to aid their diagnostic decision making during the assessment of a child or young person with symptoms suggestive of asthma.

Methods

Source of data: derivation and internal validation

We will derive a clinical prediction model using participant-reported data from the Avon Longitudinal Study of Parents and Children (ALSPAC) and participant's linked primary care electronic health records (EHR). ALSPAC is a prospective observational study that recruited pregnant women resident in and around the City of Bristol, UK with expected dates of delivery 1st April 1991 to 31st December 1992. The initial (recruitment Phase I) number of pregnancies enrolled was 14,541 (14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age). When the children from Phase I were approximately seven years of age, attempts were made to bolster the recruited sample with eligible cases who had not joined the study originally. By the time children from Phase I were 24 years of age, a further 913 index children had been recruited (456, 262 and 195 recruited during Phases II, III and IV, respectively)[13–15]. The total recruited sample available for analyses is therefore 15,454 pregnancies, resulting in 15,589 foetuses. Of these 14,901 were alive at one year of age. Frequent assessments of participants have been conducted with 68 data collection points between birth and 18 years of age[13]. At Phase I enrolment, 49.7% of children were female and 96.1% were of White ethnicity[13]. The ALSPAC website (www.bristol.ac.uk/alspac/) contains details of all available data through a searchable dictionary and variable search tool. REDCap electronic data capture tools have been used to collect and manage ALSPAC study data since 2014[16,17]. To enhance the ALSPAC resource, the Project to Enhance ALSPAC through Record Linkage (PEARL) has established the requisite ethico-legal permissions to link to and extract participant's primary care EHR, allowing data from clinical consultations and prescribing to be utilised. When the index children reached legal adulthood (i.e. age 18), ALSPAC conducted a postal fair processing campaign which aimed to re-enrol them into the study in their own right and to inform participants about the planned incorporation of linked health and administrative records in the ALSPAC databank. Participants had the right to opt-out of this use of their information (applied both at a study-level and in accordance with the UK's National Opt-Out mechanisms). ALSPAC subsequently sought permission from general practitioner (GP) practices with registered ALSPAC participants to extract linked records. Participants were identified in the GP record using National Health Service (NHS) ID number and extracted by practice system software companies. Linkage to GP records was carried out following this campaign. Only coded data from the primary care records were extracted.

Participants

Study participants will be chosen from the ALSPAC dataset using pre-specified eligibility criteria. For the derivation dataset, the inclusion criteria will be participants:

- recruited into the original cohort (i.e. at Phase I, because children recruited at a later stage have missing data from birth to seven years of age)
- alive at one year
- with consent for the use of their linked EHR

Outcome

The outcome measure will be derived from linked primary care EHR. We will use prescribing data in combination with the presence of an asthma Read code (version 2) to identify the presence of asthma and the date of diagnosis. We will identify participants who received at least three prescriptions for an inhaled corticosteroid, as a single inhaler or in combination with a long-acting beta agonist, on separate days within a one-year period. From this group, we will select participants who have an asthma 'specific' Read code (according to the validated code list from Nissen *et al.*)[18] occurring at any time in their patient record. Individuals with at least three inhaled corticosteroid prescriptions in one year and a 'specific' asthma Read code will be designated as having asthma (see *Extended data*[19] for the Read code lists to be used in this study).

This prediction model is intended for use by health professionals at the point of asthma diagnosis. It is therefore important to ensure that any information that occurs after the diagnosis is excluded from the development dataset. To do this, an event date is required. The event date for those with the outcome will be taken as the date at which the first of the inhaled corticosteroid prescriptions was made. Those without the outcome have no equivalent event date. Therefore, participants without an outcome will be assigned an event date at random. To do so participants with an outcome will be grouped by their age (years) at event date and the proportion of individuals within each year age group will be taken. Then participants without an outcome will be randomly sorted into age-at-event groups so that the same proportion of individuals will be placed in the age-at-event groups. Other than age-at-event, (which will no longer be available for modelling), the outcome measure will be developed blind to information about the predictors.

Predictors

Potential candidate predictors were identified based on the results from our systematic review of prediction models for the diagnosis of asthma in primary care[12] and based on clinical usefulness decided after discussion within the research team (including GPs, respiratory paediatricians and statisticians). We will choose the final list of candidate predictors from the following: gender, social class, wheeze, cough, night cough, breathlessness, eczema, hay fever, allergy to food or drink, allergy to substance other than food or drink, maternal asthma, maternal atopy, maternal cigarette smoking during pregnancy, childhood exposure to cigarette smoke, mould in the participants house, lung function indices from spirometry, fractional exhaled nitric oxide (FeNO), skin prick testing results, immunoglobulin E (IgE) serum samples, evidence of lung function or reversibility testing in the patient EHR and prescription of a short-acting beta agonist (SABA).

Sample size

The number of candidate predictor variables will be restricted to a minimum of 10 events per variable. Following preliminary analysis of the ALSPAC dataset, 11972 participants met the eligibility criteria, with 994 participants having the outcome of interest. Taking into account the 22 candidate predictors (27 parameter levels) and the number of outcome events, the events per variable (36.8) far exceeds recommendations for sample sizes[20].

Missing data

Different approaches to dealing with missing data will be considered. Variables with over 40% missingness will be excluded. Given the large number of participants, conducting a complete case analysis will be considered by running estimations using the full list of candidate predictors. If the exclusion of cases with missing data substantially reduces the total sample size, then multiple imputation by creating up to 20 imputed datasets via chained equations[21] will be considered[22].

Statistical analysis methods

Statistical analysis will be conducted using R (version 3.5.3), and SPSS (version 26).

Handling of predictor variables. Where possible, variables will be used in the form in which they were collected – for example, a continuous variable will not be split into categories unless necessary[23]. Continuous variables will be checked for linearity and if necessary, fractional polynomials used to improve the fit of non-linear relationships[24].

Variables relating to participant characteristics and symptoms were captured in the ALSPAC study by child and parent completed questionnaires. As questionnaires were completed on a number of occasions, many symptoms/exposures (such as wheeze) were collected on more than one occasion. Fortunately, ALSPAC used the same (or very similar) questions making it possible to combine responses collected at different time points. When computing variables into candidate predictors these data will initially be considered as counts; describing the number of times a symptom/exposure occurred before the event date. However, if there are a large proportion of zeros, we will recode the variable into a binary category, thus capturing the presence or absence of a symptom/exposure prior to the event date.

Two candidate predictors will be derived from linked EHR. To capture information relating to the evidence of lung function or reversibility testing, a list of relevant codes was compiled. The occurrence of any lung function or reversibility Read code (see *Extended data*)[19] will be extracted from the linked EHR together with an anonymised identifier and event date. From these data a binary variable will be created describing the presence or absence of a ‘lung function/reversibility’ code occurring before the event date. A binary variable identifying the presence or absence of a SABA prescription prior to the event date will be similarly constructed, though the code list will be compiled from the Systematized Nomenclature of Medicine (SNOMED) prescribing terms[25].

Type of model. We intend to use multivariable logistic regression as this is an appropriate method where outcome measures are binary and candidate predictor variables are categorical, continuous, or combined[26,27].

Predictor selection before modelling. From the list of candidate predictors, predictors missing in more than 40% of participants will be excluded. Where inter-relatedness between predictors exists, the predictors that best capture the information sought will be retained by choosing firstly based on clinical relevance, and secondly (if variables are equally relevant), the predictor with least missing data. We will not use univariate analysis to screen for significant associations between potential predictors and the outcome, as predictors behave differently in a multivariable model[28].

Predictor selection during modelling. Backward step-wise selection based on the Akaike’s Information Criterion (AIC) will be used to select predictor variables during modelling[29].

Model performance. The apparent performance of the final model will be calculated in the original sample. Discrimination will be reported using the concordance statistic (c-statistic). Calibration will be assessed visually using a calibration plot and by calculating the calibration slope, calibration-in-the-large (CITL) and ratio of expected and observed number of events (E/O)[30].

Internal validation

Rather than reduce the sample size by using split-sample or cross-validation, bootstrapping techniques for model validation will be used[30]. The modelling process including variable selection will be repeated in up to 500 samples drawn with replacements from the original sample. The bootstrap performance of the model will be assessed in each bootstrap sample using the c-statistic. We will determine the performance of the bootstrap model in the original sample (test performance) and calculate the optimism as the difference between the bootstrap and test performance. We will average the estimates of optimism from each bootstrap sample and subtract the value from the apparent performance to generate an optimism-corrected estimate of performance. The optimism adjusted calibration slope will be used as the shrinkage factor to adjust the regression coefficients of the developed model for optimism.

External validation

External validation of the model will be conducted in a dataset created from routinely collected anonymised primary care records from the Optimum Patient Care Research Database. Using the model derived in the ALSPAC dataset (and adjusted for shrinkage), we will calculate the linear predictor and predicted probability of the outcome for each individual in the external validation dataset. Model performance will be assessed using measures of discrimination (c-index) and calibration (calibration plot, calibration slope, CITL and E/O)[30]. Recalibration of the model will be completed if there is a systematic under- or over-prediction of risk in the external validation dataset.

Risk groups

The potential for generating risk groups using the probabilities from the multivariable model will be considered. Currently, the British Thoracic Society/Scottish Intercollegiate Guideline Network recommend that clinicians weigh up the probability of an asthma diagnosis into high, intermediate and low probability[4]. Therefore, defining the probability generated from the prediction model in such terms may appeal to clinicians. However, the range of predicted probabilities incorporated within a group, and the benefit of constructing risk groups will require further consideration and the involvement of patients and clinicians[31].

Discussion

Current clinical prediction models for the diagnosis of asthma are at high risk of bias and not recommended for use in clinical practice[12]. This protocol builds on the findings from our systematic review to derive and validate a clinical prediction model for primary healthcare professionals to support their decision making during the assessment of a child or young person with symptoms to suggest asthma. The protocol has been guided by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)[26]. We intend to include a broad range of predictor variables, including demographics, symptoms, medical and family history, and ideally results from clinical tests. We have a robust plan to deal with missing data, assess model performance, internally validate the model and take account of model optimism.

ALSPAC has only been given permissions to extract coded information from the medical records due to concerns that free-text notes may pose a confidentiality risk (relating to third-parties rather than the participant themselves). This means the prediction model development will only make use of coded information, and thus cannot take advantage of the increase in algorithm sensitivity that using free-text brings[32]. The use of coded diagnosis in combination with prescription records mitigates the risk of 'false positive' identification of cases resulting from GPs recording practice where a code is entered for a tentative diagnosis or a coded diagnosis entry is added to the record along with free-text information indicating the patient does not have the condition[33].

Following the derivation and internal validation of a clinical prediction model in the ALSPAC dataset, we will externally validate the model in routinely collected data and explore the implementation of the model into clinical practice.

Declarations

Ethics approval and consent to participate

Ethical approval for the derivation and internal validation was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. ALSPAC's use of linked health records is based on approvals from the ALSPAC Ethics and Law Committee, Health Research Authority Research Ethics Committee and Confidentiality Advisory Group. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Ethical approval to complete external validation of a clinical prediction model for the diagnosis of asthma in primary care was obtained from the Anonymised Data Ethics and Protocol Transparency Committee (ADEPT) (Approval Reference: ADEPT0320).

Consent for publication

ALSPAC participants have been provided with fair processing materials describing the studies use of the data they have provided or those collected through record linkage and about the legal basis under which the study operates: this includes the sharing of de-identified data with researchers and the publishing of research findings. Study members have the right to withdraw from elements of the study or from the study entirely at any time. Full details of the ALSPAC consent procedures are available from the study website.

Data availability

Underlying data

No underlying data are associated with this article.

ALSPAC data access is through a system of managed open access. The steps below highlight how to apply for access to the data referred to in this article and all other ALSPAC data. The datasets presented in this article are linked to ALSPAC project number B2830, please quote this project number during your application. The ALSPAC variable codes highlighted in the dataset descriptions can be used to specify required variables.

- 1. Please read the ALSPAC access policy (https://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf) which describes the process of accessing the data and samples in detail, and outlines the costs associated with doing so.
- 2. You may also find it useful to browse our fully searchable research proposals database (https://proposals.epi.bristol.ac.uk/), which lists all research projects that have been approved since April 2011.
- 3. Please submit your research proposal for consideration by the ALSPAC Executive Committee. You will receive a response within 10 working days to advise you whether your proposal has been approved.

If you have any questions about accessing data, please email alspac-data@bristol.ac.uk.

The study website also contains details of all the data that is available through a fully searchable data dictionary: http://www.bristol.ac.uk/alspac/researchers/data-access/data-dictionary/.

Extended data

Open Science Framework: Clinical prediction model for the diagnosis of asthma in children and young people in primary care. https://doi.org/10.17605/OSF.IO/FU4GN[19].

This project contains the following extended data:

- AsthmaSpecific_ReadcodeList (TXT). (Asthma-specific read codes.)
- LungFunctionAndReversibility_ReadCodeList (TXT). (Lung function/reversibility testing read codes.)

Author contributions

LD, HP, SL and AS conceived the study with contributions to the design from LJB, AB and ST. LD wrote the first draft of the manuscript. All authors contributed to manuscript review and final approval.

Acknowledgements

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. We are also indebted to John Henderson who was instrumental in establishing the collaboration between the Asthma UK Centre for Applied Research and ALSPAC, and to Richard Thomas for his ongoing support relating to the use of linked medical records.

References

1. Global Asthma Network: **The Global Asthma Report.** 2018; [accessed March 2020].
Reference Source
2. Looijmans-van den Akker I, van Luijn K, Verheij T: **Overdiagnosis of asthma in children in primary care: a retrospective analysis.** *Br J Gen Pract.* 2016; **66**(644): e152–7.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Aaron SD, Vandemheen KL, FitzGerald JM, *et al.*: **Reevaluation of Diagnosis in Adults With Physician-Diagnosed Asthma.** *JAMA.* 2017; **317**(3): 269–79.
PubMed Abstract | Publisher Full Text
4. Health Improvement Scotland: **BTS/SIGN British Guideline for the management of asthma.** SIGN 158. 2019.
Reference Source
5. Global Initiative for Asthma: **Global Strategy for Asthma Management and Prevention.** 2019; [accessed March 2020].
Reference Source
6. Pavord ID, Beasley R, Agusti A, *et al.*: **After asthma: redefining airways diseases.** *Lancet.* 2018; **391**(10118): 350–400.
PubMed Abstract | Publisher Full Text
7. White J, Paton JY, Niven R, *et al.*: **Guidelines for the diagnosis and management of asthma: a look at the key differences between BTS/SIGN and NICE.** *Thorax.* 2018; **73**(3): 1–5.
Publisher Full Text
8. Murray C, Foden P, Lowe L, *et al.*: **Diagnosis of asthma in symptomatic children based on measures of lung function: an analysis of data from a population-based birth cohort study.** *Lancet Child Adolesc Health.* 2017; **1**(2): 114–23.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Drake SM, Simpson A, Fowler SJ: **Asthma Diagnosis: The Changing Face of Guidelines.** *Pulm Ther.* 2019; **5**(2): 103–115.
PubMed Abstract | Publisher Full Text | Free Full Text
10. National Institute for Health and Care Excellence: **Asthma: diagnosis, monitoring and chronic asthma management**. NG80, 2017; [accessed March 2020].
Reference Source
11. Keeley D, Baxter N: **Conflicting asthma guidelines cause confusion in primary care.** *BMJ.* 2018; **360**: k29.
PubMed Abstract | Publisher Full Text
12. Daines L, McLean S, Buelo A, *et al.*: **Systematic review of clinical prediction models to support the diagnosis of asthma in primary care.** *NPJ Prim Care Respir Med.* 2019; **29**(1): 19.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Boyd A, Golding J, Macleod J, *et al.*: **Cohort Profile: The 'Children of the 90s'--the index offspring of The Avon Longitudinal Study of Parents and Children.** *Int J Epidemiol.* 2013; **42**(1): 111–127.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Fraser A, Macdonald-Wallis C, Tilling K, *et al.*: **Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort.** *Int J Epidemiol.* 2013; **42**(1): 97–110.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Northstone K, Lewcock M, Groom A, *et al.*: **The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019 [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2019; **4**: 51.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Harris PA, Taylor R, Thielke R, *et al.*: **Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support.** *J Biomed Inform.* 2009; **42**(2): 377–81.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Harris PA, Taylor R, Minor BL, *et al.*: **The REDCap consortium: Building an international community of software platform partners.** *J Biomed Inform.* 2019; **95**: 103208.
PubMed Abstract | Publisher Full Text
18. Nissen F, Morales DR, Mullerova H, *et al.*: **Validation of asthma recording in the Clinical Practice Research Datalink (CPRD).** *BMJ Open.* 2017; **7**(8): e017474.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Daines L: **Clinical prediction model for the diagnosis of asthma in children and**

young people in primary care. 2020.
http://www.doi.org/10.17605/OSF.IO/FU4GN

20. Peduzzi P, Concato J, Kemper E, *et al.*: **A simulation study of the number of events per variable in logistic regression analysis.** *J Clin Epidemiol.* 1996; **49**(12): 1373–9.
PubMed Abstract | Publisher Full Text

21. van Buuren S, Groothuis-Oudshoorn K: **mice: Multivariate imputation by chained equations in R.** *J Stat Softw.* 2011; **45**(3): 1–68.
Publisher Full Text

22. Graham JW, Olchowski AE, Gilreath TD: **How many imputations are really needed? Some practical clarifications of multiple imputation theory.** *Prev Sci.* 2007; **8**(3): 206–13.
PubMed Abstract | Publisher Full Text

23. Altman DG, Royston P: **The cost of dichotomising continuous variables.** *BMJ.* 2006; **332**(7549): 1080.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Royston P, Sauerbrei W: **Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables.** Chichester, England: John Wiley & Sons'. 2008.
Publisher Full Text

25. NHS Digital: **The NHS Digital SNOMED CT Browser.** [Accessed March 2020].
Reference Source

26. Collins GS, Reitsma JB, Altman DG, *et al.*: **Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD).** *Ann Intern Med.* 2015; **162**(10): 735–6.
PubMed Abstract | Publisher Full Text

27. Midi H, Sarkar SK, Rana S: **Collinearity diagnostics of binary logistic regression model.** *J Interdisc Math.* 2010; **13**(3): 253–67.
Publisher Full Text

28. Harrell FE Jr: **Regression modelling strategies with applications to linear models, logistic and ordinal regression and survival analysis 2nd Edition**. Switzerland: Springer; 2015.
Publisher Full Text

29. Akaike H: **A new look at the statistical model identification.** *IEEE Trans Automat Contr.* 1974; **19**(6): 716–23.
Publisher Full Text

30. Steyerberg E: **Clinical prediction models: a practical approach to development, validation, and updating.** New York, USA: Springer Science & Business Media; 2009.
Publisher Full Text

31. Bonnett LJ, Snell KIE, Collins GS, *et al.*: **Guide to presenting clinical prediction models for use in clinical settings.** *BMJ.* 2019; **365**: l737.
PubMed Abstract | Publisher Full Text

32. Ford E, Carroll JA, Smith HE, *et al.*: **Extracting information from the text of electronic medical records to improve case detection: a systematic review.** *J Am Med Inform Assoc.* 2016; **23**(5): 1007–15.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Nicholson A, Tate AR, Koeling R, *et al.*: **What does validation of cases in electronic record databases mean? The potential contribution of free text.** *Pharmacoepidemiol Drug Saf.* 2011; **20**(3): 321–4.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 15 July 2020

https://doi.org/10.21956/wellcomeopenres.17271.r39328

✓

Patrick J E Bindels
Department of General Practice, Erasmus MC, Rotterdam, The Netherlands

This is a protocol for the derivation and validation of a clinical prediction model to support the diagnosis of asthma in children and young people. The prediction model will be developed for use in primary care. As the authors stated the diagnosis of asthma in primary care is based predominantly on clinical features in combination with demographics, medical history and ideally results from clinical tests. It would be of great help for general practitioners if a prediction model could assist them in classifying children in low, intermediate and high risk of developing asthma. Especially in the very young children (< 6 years) a probable diagnosis of asthma is a challenge in general practice.
In this protocol, the researchers are going to derive a model from the ALSPAC study in combination with data form the primary care EHR of the participants in the ALSPAC study. External validation is planned in the Optimum Patient Care Research Database.
It is an interesting protocol and any attempt to develop a prediction model for general practice is highly relevant. I do have some comments and I hope these comments will help to improve the protocol further.

- Especially in pre-school children, a presumptive asthma diagnosis is a true challenge for general practitioners. Clinical features are not always very specific and the more ‘classical’ symptoms like shortness of breath (or the presence of inhalation allergy) not present. From the age of 6 onwards, the clinical features of asthma-like shortness of breath and wheeze are more explicit. It makes a diagnosis of asthma in older children less a challenge for most GP’s (a diagnosis is often not the problem anymore, undertreatment however is). Moreover, from the age of 5-6 years, lung function testing is possible in children. Besides clinical symptoms, a more objective measurement (FEV1 and reversibility testing) is available to support an asthma diagnosis. Will it be possible to construct an overall prediction model for young children and young people? Or will it be necessary to construct prediction models for specific age groups (e.g. < 6 years, 6-12 years, and adolescents) where each age specific prediction model will have its own specific variables?

- In the majority of patients, allergic asthma starts at a young age and persists during adolescence and adulthood. A first clinical presentation of asthma at adolescence (without any symptoms as a child) is to my knowledge rare. I am sure the authors do have a specific reason for also developing this prediction score for young people?

- Clinical symptoms will probably be important variables in the prediction model (wheeze, cough, night cough, breathlessness). The variables will be captured form self-reported questionnaires by the parent or child. It is well known that there can be a considerable difference in self-reported symptoms (in questionnaires) and symptoms registered in an EHR by a GP. Especially for wheeze. Is the information on the validity of the self-reported symptoms in ALSPAC available?

- When computing variables into candidate predictors these data will be considered as counts (responses at different time points will be combined). The number of times a symptom occurred before the event date will be used as a candidate variable. However, a child will not have visited the GP each time a period of cough or wheeze occurred. Let's presume having had three periods of wheeze in the last year is a variable in the final model. The information is based on questionnaires data and not on clinical presentation. Can a GP rely on self-counted episodes of symptoms by the parent, especially in very young children?

- Did the authors consider starting the follow up of patients at the first possible clinical presentation of asthma (preferably the first presentation in a clinical setting because that is the moment a GP might start to consider a possible diagnosis of asthma)?

- Two candidate predictors will be derived from the linked EHR (lung function and SABA prescriptions). In the list of potential candidate predictors also skin prick testing, IgE en FeNO are considered (but not through linkage with the EHR). Are these variables available in ALSPAC?

- The outcome measure will be derived from prescribing data in combination with the presence of an asthma Read code. The authors will only identify participants who received at least three prescriptions for an inhaled corticosteroid (within one year). This might mean a selection of more severe asthma (or children who are therapy compliant!) and children with mild asthma (one or two prescriptions in combination with SABA prescriptions or children who are not compliant but do have asthma) will be excluded. Compliance will be a problem that has to be taken into account especially in adolescents?

- External validation of the model will be conducted in the Optimum Patient Care Research Database. The information on this database is limited in the protocol. What kind of database is it? An EHR based database? How are symptoms (and other potential variables) registered in this database? Is it done in the same way as in ALSPAC? For instance, in ALSPAC symptoms are based on self-reported questionnaires, how are symptoms registered in Optimum?

- It is an interesting protocol and I am looking forward to the model that will be derived.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Research in general practice

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 02 July 2020

https://doi.org/10.21956/wellcomeopenres.17271.r39329

✓

Glen Martin
Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

In this study protocol, the authors describe the plans for developing and validating (both internally and externally) a prediction model that will aim to estimate the risk of an asthma diagnosis. To develop and internally validate the model, the authors plan to use data from the Avon Longitudinal Study of Parents and Children (ALSPAC), which will be linked with primary care data. External validation will be based on a dataset created from the Optimum Patient Care Research Database.

The motivation of the work is clear, based on a previous systematic review, which identified that previous models in this area are at high risk of bias, and have not been deemed suitable for clinical practice.

From a statistical viewpoint, I think that the protocol is very clear and includes a detailed description of the methods that will be used. It was good to see that the protocol (and subsequent study) follows the TRIPOD guidelines. The methodology that will be used to develop the prediction model is sound and considers important aspects such as correctly handling continuous variables (including consideration of non-linear associations), predictor/ outcome definitions, and handling of missing data.

I have several comments, which the authors might like to consider and are intended to be constructive in helping to improve the protocol further. I list them here in the order that they occurred to me when reading the protocol:

1. I found the description of the outcomes slightly unclear, particularly the first paragraph of the subsection. The authors say that they will use “three prescriptions for an inhaled corticosteroid” and “the presence of an asthma Read code” to define the outcome. What motivated the choice of using three prescriptions of an inhaled corticosteroid in a single year? Is this the standard way of defining an asthma diagnosis in routine data? As I understand from the text, the authors will only look for the asthma Read codes in the subgroup identified as having the three prescriptions (i.e. anyone with an asthma Read code but who doesn’t have at least three prescriptions for an inhaled corticosteroid in a single year will be defined as not having asthma). While I think that this would reduce any false positives of the outcome definition, it might help the protocol if the authors justified the strict use of three prescriptions in a single year (since this could increase false negatives).

2. In the Sample size subsection, the authors justify the sample size based on events per variable (using the commonly applied EPV>10 rule-of-thumb). The justification for the sample size could be further improved if the authors also applied recent formal sample size criteria by Riley *et al.* (e.g. https://doi.org/10.1002/sim.7992 for logistic models)[1]. These criteria would allow the authors to directly show that the sample size is sufficient to minimise overfitting, for a given number of candidate predictors. These criteria require pre-specification of the anticipated Cox-Snell R2, so the authors could use their previous systematic review of existing asthma diagnosis risk models to do this. I suspect, given the relatively large sample size (EPV of 36.8), that the data will exceed these calculations, but determining this a priori would be a valuable addition to the protocol, in my view.

3. I valued the plans to use both clinical insights and data-driven methods to guide predictor selection. I just have one minor suggestion here: rather than using backward stepwise selection based on AIC for predictor selection during modelling, I wonder if using LASSO would be beneficial since this can select variables, while also penalising/shrinking parameter estimates. The latter characteristic would further minimise the risk of overfitting to the development data.

4. Model performance subsection: How will the calibration plots be constructed: will this be based on the decile method or based on a flexible calibration plot (https://pubmed.ncbi.nlm.nih.gov/26772608/)[2]?

5. Internal validation: “The bootstrap performance of the model will be assessed in each bootstrap sample using the c-statistic” – presumably, the end of this sentence should read “…using the c-statistic, calibration intercept and calibration slope”, since one needs to estimate the optimism in all of the performance measures? I think this is simply an omission of text since the authors refer to the “optimism adjusted calibration slope” later in the paragraph.

6. External validation: are all the variables that are available in the ALSPAC dataset (linked with EHR) and will be considered for inclusion in the model also available in the Optimum Patient Care Research Database that will be used for external validation? If not, how will the authors handle the situation where a variable in the developed model is not in the external validation dataset? Either way, it might be useful to add some additional detail of the Optimum Patient Care Research Database, including how the authors will define the cohort used for external validation (which will presumably be the same as that described for model development/ internal validation).

Overall, a great protocol for a study that has clear motivation. I hope my comments are useful in refining the protocol further.

References

1. Riley R, Snell K, Ensor J, Burke D, et al.: Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*. 2019; **38** (7): 1276-1296 Publisher Full Text
2. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, et al.: A calibration hierarchy for risk models was defined: from utopia to empirical data.*J Clin Epidemiol*. **74**: 167-76 PubMed Abstract | Publisher Full Text

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics/ Risk prediction modelling/ Prognostic research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.