
Article:
Kiely, D.G., Doyle, O., Drage, E. et al. (2019) Utilising artificial intelligence to determine patients at risk of a rare disease : idiopathic pulmonary arterial hypertension. Pulmonary Circulation, 9 (4). ISSN: 2045-8932

https://doi.org/10.1177/2045894019890549

Research Article

Pulmonary Circulation

Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension

David G. Kiely[1,2,3], Orla Doyle[4], Edmund Drage[4], Harvey Jenner[4], Valentina Salvatelli[4], Flora A. Daniels[4], John Rigg[4], Claude Schmitt[5], Yevgeniy Samyshkin[5], Allan Lawrie[2,3,*] and Rito Bergemann[6,*]

[1]Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital, Sheffield, UK; [2]Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK; [3]INSIGNEO, University of Sheffield, Sheffield, UK; [4]Real-World & Analytical Solutions, IQVIA, London, UK; [5]GSK, Middlesex, UK; [6]Evalueserve UK Ltd, London, UK

Abstract

Idiopathic pulmonary arterial hypertension is a rare and life-shortening condition often diagnosed at an advanced stage. Despite increased awareness, the delay to diagnosis remains unchanged. This study explores whether a predictive model based on healthcare resource utilisation can be used to screen large populations to identify patients at high risk of idiopathic pulmonary arterial hypertension. Hospital Episode Statistics from the National Health Service in England, providing close to full national coverage, were used as a measure of healthcare resource utilisation. Data for patients with idiopathic pulmonary arterial hypertension from the National Pulmonary Hypertension Service in Sheffield were linked to pre-diagnosis Hospital Episode Statistics records. A non-idiopathic pulmonary arterial hypertension control cohort was selected from the Hospital Episode Statistics population. Patient history was limited to $\leq$5 years pre-diagnosis. Information on demographics, timing/frequency of diagnoses, medical specialities visited and procedures undertaken was captured. For modelling, a bagged gradient boosting trees algorithm was used to discriminate between cohorts. Between 2008 and 2016, 709 patients with idiopathic pulmonary arterial hypertension were identified and compared with a stratified cohort of 2,812,458 patients classified as non-idiopathic pulmonary arterial hypertension with $\geq$1 ICD-10 coded diagnosis of relevance to idiopathic pulmonary arterial hypertension. A predictive model was developed and validated using cross-validation. The timing and frequency of the clinical speciality seen, secondary diagnoses and age were key variables driving the algorithm's performance. To identify the 100 patients at highest risk of idiopathic pulmonary arterial hypertension, 969 patients would need to be screened with a specificity of 99.99% and sensitivity of 14.10% based on a prevalence of 5.5/million. The positive predictive and negative predictive values were 10.32% and 99.99%, respectively. This study highlights the potential application of artificial intelligence to readily available real-world data to screen for rare diseases such as idiopathic pulmonary arterial hypertension. This algorithm could provide low-cost screening at a population level, facilitating earlier diagnosis, improved diagnostic rates and patient outcomes. Studies to further validate this approach are warranted.

Keywords

predictive algorithm, machine learning, idiopathic pulmonary arterial hypertension (PAH), diagnosis

Date received: 24 May 2019; accepted: 20 October 2019

Pulmonary Circulation 2019; 9(4) 1–9
DOI: 10.1177/2045894019890549

Introduction

Idiopathic pulmonary arterial hypertension (iPAH) is a rare, progressive and life-shortening disease. It is characterised by a small vessel vasculopathy and elevated pulmonary artery pressure; and if it is untreated, it leads to right heart failure and death, with a median survival of less than three years.[1]

*Joint Senior Authors.

Corresponding authors:
David G. Kiely, Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital, Sheffield Teaching Hospitals NHS Foundation Trust, Glossop Road, Sheffield S10 2JF, UK.
Email: david.kiely1@nhs.net
Allan Lawrie, Infection, Immunity & Cardiovascular Disease, University of Sheffield Medical School, Beech Hill Road, Sheffield, S10 2RX, UK
Email: a.lawrie@sheffield.ac.uk

The annual incidence of iPAH has been estimated at 1–3.3 cases per million per year.[2–4] Estimates of UK prevalence of idiopathic, heritable or anorexigen-induced PAH range from 12.4 to 24.8 per million,[5] with recent published data from the national audit, identifying a prevalence of 15 per million of population in England.[6] The symptoms of iPAH are non-specific and clinical signs are subtle until the disease is advanced. Progressive shortness of breath and fatigue are common; and as the disease progresses, exertional chest tightness, pre-syncope and syncope may occur. Leg swelling is a late sign in young patients and reflects severely impaired right ventricular function.[7] Given the rarity of iPAH and the non-specificity of symptoms, patients are frequently misdiagnosed with other common cardiorespiratory diseases. A lengthy delay between the onset of symptoms and a definitive diagnosis is normal, typically around two years; and this delay is unchanged over the last two decades.[8] Consequently, iPAH is often diagnosed at an advanced stage in terms of symptom burden and haemodynamic severity.[9] In contrast, systemic sclerosis-associated pulmonary arterial hypertension (SSc PAH) is typically diagnosed earlier, as the high prevalence (9%) of PAH in SSc has led to the implementation of specific screening programmes in this high-risk group of patients.[10–12] Furthermore, an evidence-based algorithm for early SSc PAH detection has recently been developed.[13]

The application of artificial intelligence (AI) capabilities, and specifically machine learning algorithms, has created the opportunity to identify actionable healthcare insights from large and complex healthcare datasets.[14,15] One proposed application of such technologies is to use routinely collected patient data to screen for or predict those at high risk for disease to potentially improve patient outcomes.[16] Examples of such data include the National Health Service (NHS) Hospital Episode Statistics (HES) database in England, which provides close to full national coverage for a population of approximately 55 million, and medical insurance records in the United States, where coverage varies depending on provider and/or location.

Recently, we published data from the Sheffield Pulmonary Hypertension IndeX (SPHInX) project, demonstrating that patients with iPAH have high levels of healthcare resource utilisation (HCRU) in the three years prior to diagnosis, with approximately 25 hospital visits.[17] We also demonstrated that national HES data can be linked to patient-level hospital diagnostic data in patients with iPAH in 99% of cases. Our analyses showed that HES data has the potential to support the development of a predictive model to screen for iPAH.[17] In this study, we now describe the development and internal validation of a predictive AI model to identify patients at risk of iPAH.

Methods

Construction of the SPHInX dataset

To identify HCRU patterns in the years prior to a diagnosis of iPAH, we obtained NHS HES patient records from April 2000 to March 2017 for all patients diagnosed with iPAH at the Sheffield Pulmonary Vascular Disease Unit (SPVDU) during 2008–2016. These HES data consisted of information relating to inpatient, outpatient and accident and emergency attendances. For a non-iPAH group, we identified a cohort of patients using codes from the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) that were associated with cardiorespiratory disease and frequently used in patients with iPAH.

NHS numbers were used to link the HES datasets with positive iPAH cases diagnosed at SPVDU. A diagnosis of iPAH was confirmed by medical expert, and the study included only those that had undergone detailed clinical assessment including blood testing, lung function testing, exercise testing, echocardiography, multi-modality imaging (nuclear medicine imaging, computed tomography, magnetic resonance imaging), right heart catheterisation and classification according to international guidelines and multidisciplinary assessment.[18] Patient linkage was quality controlled by comparing the consistency of gender, year of birth, general practitioner postcode and key dates (first diagnosis, first right heart catheterization and first visit at SPVDU). The initial non-iPAH cohort included all HES patients who had at least one primary or secondary diagnosis in ICD-10 codes relevant to cardiorespiratory disease, that would result in high levels of HCRU similar to iPAH but whose pattern of behaviour would ideally be distinguishable from iPAH. The list of pre-specified ICD-10 codes for the definition of the non-iPAH cohort can be found in Supplementary Table 1.

Selection of clinical variables for inclusion in the predictive model

We considered diagnoses (ICD-10 coding scheme), procedures codes (OPCS coding scheme) and the clinical specialty of the treating physician ('clinical specialty' codes) as potential variables for the predictive model. Diagnosis and procedure codes were labelled as either primary or secondary in the HES dataset; primary diagnosis referred to the main condition investigated, and primary procedure referred to the most resource-intensive procedure carried out. All other diagnoses and procedures contained within the episode were captured as secondary.

To select a set of diagnosis and procedure codes relevant to the iPAH HCRU footprint, a hybrid data- and clinically driven approach was used. First, all codes that appeared in $\geq$1% of the iPAH cohort (condition 1) or <1% of the iPAH cohort and >2% of the non-iPAH cohort (condition 2) were selected. The non-iPAH cohort in this selection step comprised 5630 patients confirmed to not have iPAH who attended SPVDU within the study window. This method ensured that variables found rarely in the iPAH cohort but more commonly in non-iPAH were retained for modelling (i.e. the anti-correlated events). To reduce the number

of variables further, variables were included only if they were: (i) definitely or possibly related to the iPAH journey (for those identified by condition 1) and (ii) definitely or possibly relevant to the exclusion of iPAH (for those identified by condition 2), following independent review by two clinical experts. For inclusion, a variable had to be selected by at least one of the experts. The experts were blinded to the prevalence of the codes. All clinical specialty codes appearing in at least 1% of the iPAH cohort were included in the model.

The selected variables were described using three metrics; frequency variables (e.g. the frequency of certain procedures), date difference variables (e.g. the number of days between a procedure and the index date) and aggregated time variables (e.g. the number of new diagnoses within 12 months of the index date). Clinical codes or events that were missing were assumed to represent an absence of the event and were encoded as a zero for count metrics. Data difference metrics for absent events were coded as missing and passed to the model directly.

Definition of index date and lookback period for development of predictive model

In this study, the pre-diagnosis history window was limited to a maximum of five years from the index date (Supplementary Fig. 1). For the non-iPAH cohort, the index date corresponded to the most recent relevant event in the patient's history. An event was considered relevant if it – (i) contained a diagnosis code belonging to the list of pre-specified ICD-10 codes relevant to iPAH and (ii) was a cardiology, respiratory or neurology clinical specialty. For the iPAH cohort, the index date was the most recent relevant event prior to the first visit at the SPVDU, ensuring that the pre-diagnosis history occurred prior to their referral to SPVDU and hence substantially prior to the date of confirmed diagnosis. Patients in both cohorts without a valid index date were excluded. The lookback period was defined as either five years or the entire length of a patient's history in the HES records, which ever was shortest.

Selection of population for development of predictive model

To build a robust predictive model for iPAH, it is crucial to ensure that the non-iPAH cohort is comprised of patients who have similar patterns of HCRU in the years leading up to diagnosis. That is, we want to ensure that the predictive model is being trained to learn an iPAH HCRU footprint rather than merely distinguishing patients who have low versus high HCRU. Stratification was applied to narrow the non-iPAH cohort to patients who more closely resemble patients with iPAH. Each patient was required to have at least one of the selected ICD-10 codes (see 'Selection of clinical variables for inclusion in the predictive model') in the primary diagnosis field.

AI methodology underpinning the predictive model

For rare disease detection based on historical HCRU, a predictive model design should be sensitive to interactions between HCRU events and avoid overfitting while leveraging the richness of the data available (Supplementary Fig. 2). To accommodate this, we utilised gradient boosting trees, a supervised machine learning algorithm, to develop our predictive model.[19] This algorithm is an ensemble of decision trees implemented using boosting, whereby the successive tree aims to reduce the error of the previous tree. The algorithm was embedded within a bootstrap aggregation framework whereby 100 base learners were trained on a bootstrapped sample of the training dataset where sampling was carried out with replacement.[20] The scores of all the learners were averaged to produce the prediction on the test set.[20] The base learner of the model was implemented using the XGBoost package.[21] Each gradient boosting tree model was a combination of 50 trees. All other XGBoost parameters were set to default values. XGBoost handles missing data by learning which branch of the node (pertaining to the missing variable) is optimal for a given observation. The analysis was performed on a local Dell PowerEdge R730xd Server with 2 × Intel Xeon E5-2695 v3 2.3 GHz processors and 64 Gb LRDIMM 2400MT/s RAM. Fig. 1(a) provides an overview of the key steps in the algorithm's development.

Validation of the predictive model

To assess model performance, data were partitioned into training and test sets. Training data were used to learn the parameters of the model while test data were used to estimate how well the model would generalise to new patients. Given the relatively small number of patients with iPAH available, a cross-validation strategy was used to assess model performance while providing predictions for all patients included in modelling.[22] Specifically, a five-fold cross-validation was used (Fig. 1(b)), in which patients were partitioned into five non-overlapping groups. Four groups were used for training of the model and the final group was used for testing. This process was iterated using each group served as a test set.

The contribution of each variable to the performance of the individual gradient boosting tree model was averaged across all learners in the bagged ensemble to provide a single view of variable importance. The output of the model, a risk score assigned to each patient that ranges from 0 to 1, was compared with a determined threshold to categorise patients predicted as iPAH-positive or iPAH-negative.

The performance metrics for the predictive model were based on conservative estimates of prevalence of iPAH from published data.[23,24] Rates of 1/1,000,000 (lower bound), 5.5/1,000,000 (middle) and 10/1,000,000 (upper bound) were used. These prevalence estimates provide guidance for how to scale the expected count of false positives in a real-world

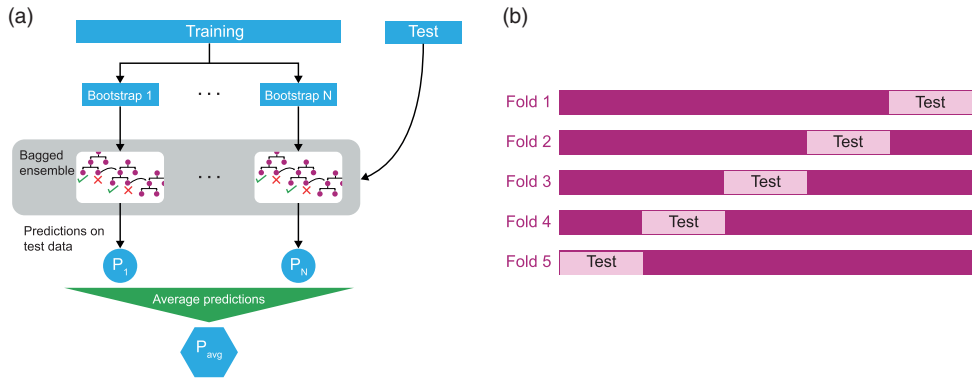

Fig. 1. (a) The bagging approach adopted in the predictive model. Each learner of the bagged ensemble is a gradient boosting tree model composed of 50 trees ($N = 100$). (b) Cross-validation strategy used for the training and the test of the model performance. This strategy has been preferred to a single hold-out set due to the modest size of the iPAH cohort.
iPAH: idiopathic pulmonary arterial hypertension.

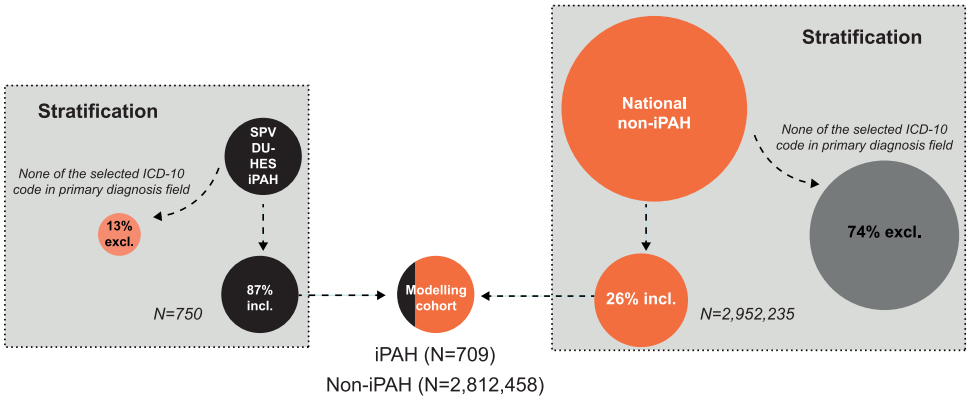

Fig. 2. Sampling strategy to select patients for the iPAH and non-iPAH cohorts used in the modelling procedure. Stratification of the non-iPAH cohort was used to narrow the cohort to patients who more closely resemble those with iPAH.
HES: Hospital Episode Statistics; ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision; iPAH: idiopathic pulmonary arterial hypertension; SPVDU: Sheffield Pulmonary Vascular Disease Unit.

clinical setting. Sensitivity (true positives/(true positives + false negatives)) and specificity (true negatives/(true negatives + false positives)) were calculated. Positive predictive values (PPV) and negative predictive values (NPV) were calculated for the three levels of stratified prevalence whereby the count of false positives was projected to the level expected to be observed at a stratified population level. That is, performance metrics were scaled so that they are representative of what would be expected in a real world clinical setting. The likelihood of a positive test (sensitivity/(1 – specificity)) was calculated as a measure of how frequently a positive diagnosis prediction is made for those with compared with iPAH versus those without iPAH.

Results

Sample population

Fig. 2 summarises the sampling strategy used to identify patients within the iPAH and non-iPAH cohorts. A total of 864 patients with a confirmed iPAH diagnosis at the SPVDU were initially identified. A comparison of the SPVDU and HES datasets revealed that 13 patients had duplicate database IDs, resulting in an initial group of 852 patients in the iPAH cohort. After application of the stratification criteria, designed to ensure that the variable distributions of the two cohorts closely resembled one another, this was reduced to 750 patients. The initial non-iPAH cohort consisted of 11,354,750 patients, and was reduced to a cohort of 2,952,235 patients after application of the stratification criteria. Patients without a valid index date or at least one month of history prior to the index date were removed, resulting in 709 and 2,812,458 patients within the iPAH and non-iPAH cohorts, respectively. The demographics for the iPAH and non-iPAH cohort are shown in Supplementary Table 2, and the baseline phenotypic characteristics of patients with iPAH in Supplementary Table 3. Patients with iPAH had a lower median age (60 years versus 71 years) and had a lower rate of systemic hypertension (48% versus 60%) than

those without iPAH. For the iPAH cohort, the average time between the first visit at SPVDU and the index date was 76 ± 272 d.

Clinical variables for inclusion in the predictive model

Following variable selection, a total of 141 clinical variables were initially identified for inclusion in the model: 23 primary diagnoses, 74 secondary diagnoses, 24 primary procedures and 19 secondary procedures, plus the age at the index date (see Supplementary Table 4). After an initial analysis, the ICD-10 codes I270 (primary pulmonary hypertension) and I272 (other secondary pulmonary hypertension) were excluded from the model variables. These predictors are strongly related to receiving a subsequent diagnosis of iPAH and could therefore artificially inflate the performance of the model, and are typically coded for patients shortly before their referral to tertiary care for the iPAH cohort. A total of 142 clinical specialty codes were contained in the two cohorts; of these, the 52 clinical specialty codes (see Supplementary Table 5) appearing in at least 1% of the iPAH cohort were included as variables in the model.

Validation of the predictive model

In Fig. 3(a), the PPV is plotted as a function of sensitivity and specificity at three different levels of iPAH prevalence. Fig. 3(b) shows the PPV and NPV as a function of iPAH prevalence. Assuming an iPAH prevalence of 5.5:1,000,000, the predictive model would need to screen 969 patients to identify 100 patients with iPAH. At a prevalence of 10:1,000,000, the number of patients required to identify the 100 patients with iPAH would drop to 587. Based on the conservative prevalence estimate of 5.5 per million, the

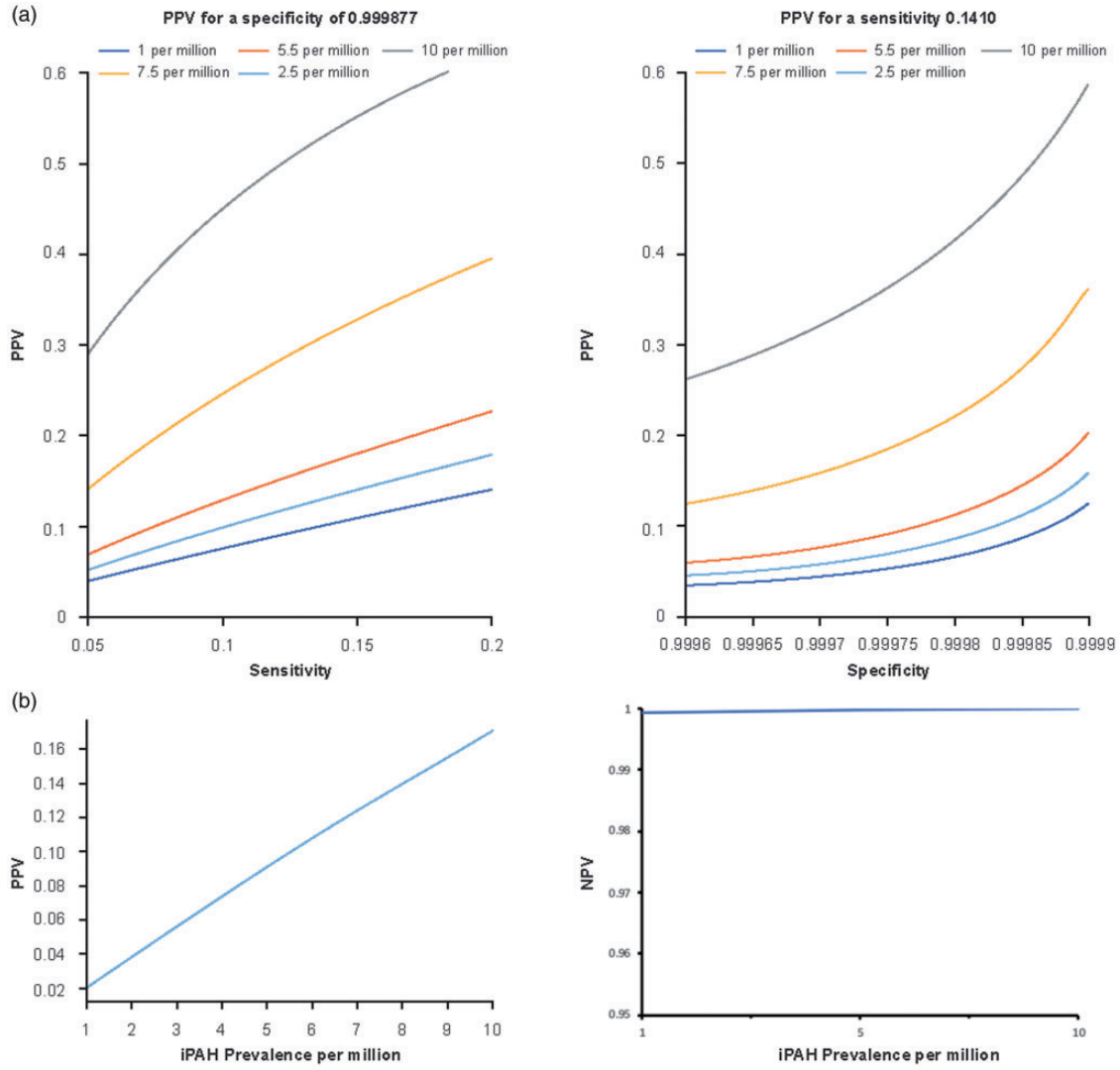

Fig. 3. (a) PPV as a function of sensitivity (left) and specificity (right) for different levels of prevalence. (b) PPV and NPV as a function of iPAH prevalence per million in the full population when the model is optimized to find 100 patients with iPAH.
iPAH: idiopathic pulmonary arterial hypertension; NPV: negative predictive value; PPV: positive predictive value.

Table 1. A 2 × 2 contingency table of the model output when optimised to find 100 patients with iPAH. It is assumed that the prevalence of the disease in the stratified population is 1:10,000.

	Predicted: non-iPAH	Predicted: iPAH	Total
Actual: non-iPAH	TN = 7,089,131	FP = 869	*7,090,000*
Actual: iPAH	FN = 609	TP = 100	*709*
	7,089,740	*969*	

FN: false negative; iPAH: idiopathic pulmonary arterial hypertension; TN: true negative; FP: false positive; TP: true positive.

model has 99.99% specificity, 14.10% sensitivity with 10.32% PPV and 99.99% NPV. This corresponds to a likelihood ratio of a positive test of 1151. A 2 × 2 contingency table of the model when optimized to identify 100 true positive patients with iPAH is shown in Table 1. To contextualise these results, the stratified population with this conservative estimate of prevalence in the absence of the predictive model would be expected to contain one patient with iPAH in every ~10,000 screened. This corresponds to a PPV of 0.01%, as compared with a PPV of 10.32% for the predictive model.

The specificity and sensitivity of the model vary according to the risk score threshold used to determine whether a patient is classified as iPAH or non-iPAH. This threshold can be lowered in order to identify higher numbers of patients with iPAH. Similarly, the less prevalent the disease, the more patients that must be screened to identify the same number of true positive patients with iPAH. A detailed breakdown of the number of positively-identified patients who would need to be screened to identify a certain number of patients with iPAH is shown in Table 2.

Fig. 4 shows the top 15 most important variables for model performance. The timing and frequency of the clinical speciality seen, the burden of co-morbidities and patient age were found to be the key variables driving the performance of the algorithm. To evaluate whether the model is selecting patients for clinically meaningful patients for iPAH screening, we examined the profiles of the top 100 patients with iPAH predicted as such by the model (i.e. the true positives: patients with iPAH and the highest scores) and the top 500 non-iPAH patients predicted as iPAH (i.e. the false positives; patients receiving high score that are not affected by the disease) (Supplementary Table 6). We observed similar frequencies of physician clinical specialties across these two groups, particularly for Respiratory Medicine, Cardiology and General Medicine. The top false positives were observed to have similar or higher proportions of patients with secondary ICD codes. This suggests that the false positives identified by the model are indeed patients that experience a pre-diagnosis HCRU footprint similar to that of patients with confirmed iPAH.

Table 2. Performance of the model. The performance of the model is expressed in terms of patients who would need to be screened (patients identified as positive by the model) in order to find a certain number of patients with iPAH (true positive patients). The number of patients to be screened also depends on the population/stratified prevalence of the disease, as indicated in the heading rows.

	Number of model-identified patients to screen for iPAH		
Stratified prevalence:	*1:5600*	*1:10,000*	*1:56,000*
Population prevalence:	*10:1 m*	*5.5:1 m*	*1:1 m*
iPAH Patients Identified			
10	44	70	346
25	109	175	864
50	212	340	1672
75	383	624	3151
100	**587**	**969**	**4965**
200	1911	3256	17,312
250	3163	5453	29,385
300	5970	10,426	57,004
350	9616	16,897	93,011
400	19,189	33,953	188,295

iPAH, idiopathic pulmonary arterial hypertension; *m*, million.

Discussion

To our knowledge, this is the first study to describe an AI approach using routinely collected data on HCRU to develop a screening algorithm to identify patients at high risk of iPAH. Key variables for model performance were the timing and frequency of clinical specialities, secondary diagnoses and procedures. The promising results reported in this study indicate the potential role for the application of AI to routinely collected healthcare data for population health screening in iPAH and other rare diseases.

The screening algorithm has been developed through numerous iterative steps by a multi-disciplinary team, including clinical and AI experts, and specifically to account for the epidemiology and confounding conditions related to diagnosing iPAH. We have also accounted for the diagnostic service model for iPAH in the NHS in England, which is delivered by a network of specialist pulmonary hypertension centres; this is reflected in the iPAH population index date definition (i.e. only using patient data prior to referral to a specialist centre for modelling).

In the present study, to identify 100 patients with iPAH (true positives), 969 patients identified by the model as being at a high risk of iPAH would be needed to be screened (based on a prevalence of 5.5 per million). This corresponds to a specificity of 99.99%, a sensitivity of 14.10%, PPV of 10.32% and NPV of 99.99%. The likelihood ratio of a positive test is 1151, meaning that the model would identify a patient as being at a high risk of iPAH 1151 times more often in patients who do have iPAH than in patients who do not. These performance metrics represent a conservative

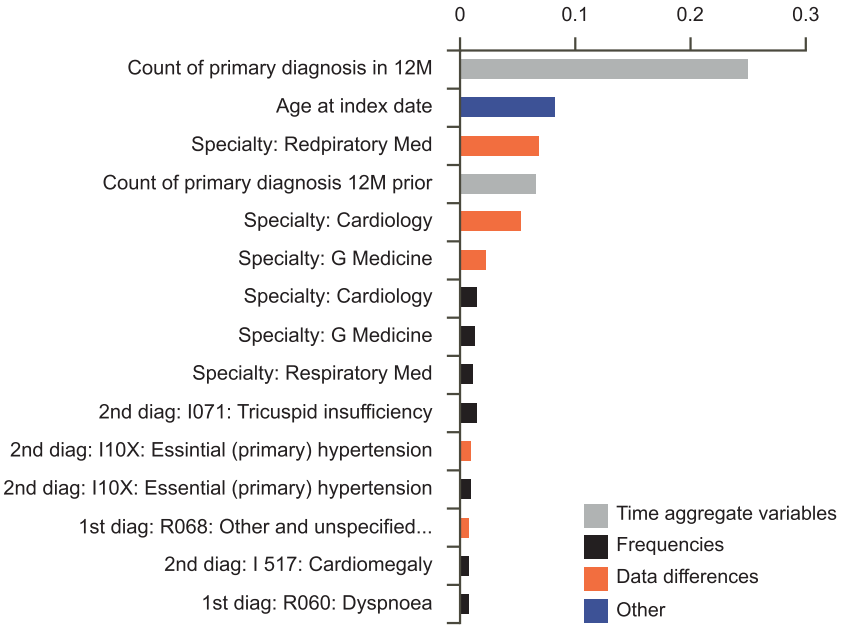

Fig. 4. The 15 most important variables of the model, ranked by average rank across the 100 bags and five groups. The importance of the variables is expressed in terms of a normalized value between 0 and 1 that corresponds to how much each variable contributes to the performance of the gradient boosting tree. Each colour corresponds to the variable class (see key).

view of the likely prevalence of iPAH in England. For the purpose of this study, we focussed on the lower bound of prevalence found in the published literature (5–6 cases per million[18]), whereas in England national audit data have indicated a prevalence of 15 per million.[6] As we demonstrate, the performance of the model would improve with increased prevalence levels; and based on a prevalence of 10 per million, 587 patients flagged as being at high risk by the model would need to be screened to identify 100 patients with iPAH.

We acknowledge that the PPV of 100 per 969 (10.32%) may appear to be low; however, this represents a significant step change when compared with the estimated prevalence of iPAH. This algorithm therefore identifies patients for screening for iPAH at much higher rate than the background prevalence by a factor of ~10,000. Indeed, the performance of this algorithm is similar to the prevalence of PAH in SSc. The benefits of using screening algorithms to identify patients with iPAH has been demonstrated in context of SSc using the DETECT screening algorithm.[13] In contrast with iPAH, where the prevalence of disease is low in the general population, PAH occurs in approximately 9% of patients with SSc.[12] The DETECT model has been demonstrated to effectively diagnose patients with SSc PAH in a clinical setting, showing that a targeted approach that identifies patients at high risk of a rare diseases is feasible.[13] Evidence suggests that earlier treatment is associated with improved outcomes in patients with iPAH[25–27]; and in SSc PAH, a comparison of contemporaneous cohorts of patients diagnosed from screening versus symptomatic presentation demonstrated that those patients identified from screening had less severe haemodynamic disease and better survival.[28] A criticism of these studies is the potential for lead time bias to influence outcomes, and no studies in PAH have unequivocally demonstrated that earlier intervention alters the natural history of disease.[29] Given the success of PAH screening in SSc, even at current performance, the model would identify patients for screening at a manageable level, where investigative approaches to diagnose pulmonary hypertension could be deployed.

In contrast with SSc, iPAH has no known associated risk factors that would facilitate such an accurate predictive model. However, patients with iPAH do have high levels of HCRU prior to diagnosis, with recent work by our group identifying an average of 25 hospital interactions in the three years prior to diagnosis.[17] The present study demonstrates that we can identify patients with a high risk of iPAH at a similar rate to that of PAH in patients with SSc. The current economic burden of iPAH is high, with patients presenting with more severe disease requiring more inpatient admissions, longer lengths of stay and more emergency department visits.[17,30] As the SPHInX predictive model is based upon existing, accessible and routinely-collected healthcare data, the cost of identifying patients at high risk of iPAH would be relatively small, and could therefore be of value despite the low sensitivity for iPAH. However, the health economic impact of investigating patients identified at high risk of iPAH and approaches to contacting these patients would require further exploration. Developing predictive models that identify patients at high risk of specific diseases using routinely collected HCRU data provides an opportunity to design studies that can explore the health economic impact of diagnostic and treatment interventions in these high-risk patients. This would allow the

development of novel study designs randomising high-risk patients to integrated diagnostic and treatment strategies that would allow a comprehensive health technology assessment. In addition, this would facilitate a comparison of long-term outcomes eliminating the potential lead time bias of historic studies comparing earlier treatment interventions in unmatched cohorts.

This study has a number of limitations. First, iPAH is a rare condition and the methodological approach used due to the number of patients meant that we used a cross-validation approach rather than having separate training and test cohorts. Second, the HES dataset is an example of a system that records secondary care HCRU from a national cohort; but the data fields and type of activity that are recorded are specific to this system. However, the general principles underpinning these datasets are similar to those used in other countries and the concept is therefore potentially translatable, but requires further validation. Finally, the performance of any predictive model depends on the population in which the model is deployed. However, one of the benefits of using an AI approach is the ability of the model to learn and be adapted based on the characteristics of the population studied. Although the algorithm was developed on an English population, confirmed iPAH cases were obtained from a single UK centre. However, the Sheffield centre provides population coverage for over 15 million people,[9,17] representing approximately one-third of the English population, and the 864 patients identified over a 16-year period equates to an estimated annual incidence of 3.6 per million per year and an estimated prevalence of 19 per million, in keeping with the published national data. The confirmed iPAH cases were also demographically similar to that reported in other registries.

In conclusion, this study highlights the potential application of AI using existing and routinely collected data to identify patients at high risk of rare conditions such as iPAH. Studies to further validate this approach to screen for iPAH in the general population are now warranted.

Author contributions

All authors contributed to the conception or design of the study and were involved in analysing or interpreting the data. DGK, AL, OD, VS, FAD, ED, and HJ also contributed to the acquisition of the data; and all authors contributed to the writing of the manuscript.

Acknowledgements

The authors acknowledge the support of the wider Sheffield Pulmonary Hypertension IndeX (SPHInX) project team who have contributed toward the collection of data. Medical writing assistance, including development of the initial draft based on author direction, assembling tables and figures, collating authors' comments, grammatical editing, and referencing, was provided by Liam Campbell, PhD, of Fishawack Indicia Ltd, UK, funded by GlaxoSmithKline (GSK).

Conflict of interest

DGK declares grants and personal fees from Actelion, Bayer, GSK and MSD. YS is an employee and shareholder of GSK. OD, HJ, FAD, VS, JR and ED are employees of IQVIA. CS was an employee of GSK at the time of the study, and is now an employee of Viiv Healthcare, a company partly owned by GSK. AL declares grants and personal fees from GSK and Actelion, including travel support from Actelion, and has received research grants fellowships from the British Heart Foundation, and the Medical Research Council. AL also reports collaboration with Kymab Ltd. RB was an employee and shareholder of GSK at the time of the study.

Data sharing statement

Information on data sharing commitments for GSK-sponsored studies and requesting access to anonymized individual participant data and associated documents can be found at www.clinicalstudydatarequest.com. Specifically, the datasets reported in this publication are not publicly available due to restrictions of the licence granted for use of National Health Service Hospital Episode Statistics. However, de-identified data used for the purpose of this study are available from the corresponding authors upon reasonable request and subject to permission from National Health Service Digital for access to the Hospital Episode Statistics data, Sheffield Teaching Hospitals National Health Service Foundation Trust information governance authorities for access to Sheffield Teaching Hospitals National Health Service Foundation Trust data as well as the Sheffield Pulmonary Hypertension IndeX (SPHInX) project team.

Ethics approval

Relevant permissions and approvals were sought and obtained from the East Midlands – Derby Research Ethics Committee (ref: 16/EM/0286), and Confidentiality Advisory Group (CAG), for the linkage of datasets under Section 251 of the Health and Social Care act 2014 (ref: 16CAG0091). The Independent Group Advising on the Release of Data (IGARD) at NHS Digital approved the use of Hospital Episode Statistics data for this study. The process to receive these permissions required research approvals from the Sheffield Teaching Hospitals National Health Service Foundation Trust Caldicott Guardian. We also sought and received a letter of support for the research from the Pulmonary Hypertension Association UK (PHA UK) patient advocacy group. Any patient who had opted out of research was removed from our analyses.

Funding

These studies were funded by GlaxoSmithKline (GSK; HO-17-18229), who were involved in the study design, analysis and interpretation of data. AL is supported by a British Heart Foundation Senior Basic Science Research Fellowship (FS/13/48/30453 and FS/18/52/33808). Employees of GSK are authors of the article and were therefore involved in the writing and final decision to submit for publication.

Guarantor

Prof David Kiely.

ORCID iDs

Yevgeniy Samyshkin https://orcid.org/0000-0003-3561-5585
Allan Lawrie https://orcid.org/0000-0003-4192-9505

Supplemental Material

Supplemental material for this article is available online.

References

1. D'Alonzo GE, Barst RJ, Ayres SM, et al. Survival in patients with primary pulmonary hypertension. Results from a national prospective registry. *Ann Intern Med* 1991; 115: 343–349.
2. Peacock AJ, Murphy NF, McMurray JJ, et al. An epidemiological study of pulmonary arterial hypertension. *Eur Respir J* 2007; 30: 104–109.
3. Humbert M, Sitbon O, Chaouat A, et al. Pulmonary arterial hypertension in France: results from a national registry. *Am J Respir Crit Care Med* 2006; 173: 1023–1030.
4. Strange G, Playford D, Stewart S, et al. Pulmonary hypertension: prevalence and mortality in the Armadale echocardiography cohort. *Heart* 2012; 98: 1805–1811.
5. The NHS Information Centre. *National Audit of Pulmonary Hypertension*. Leeds, UK: The NHS Information Centre, 2011.
6. The NHS Information Centre. *National Audit of Pulmonary Hypertension*. Leeds, UK: The NHS Information Centre, 2015.
7. Hoeper MM, Bogaard HJ, Condliffe R, et al. Definitions and diagnosis of pulmonary hypertension. *J Am Coll Cardiol* 2013; 62: D42–D50.
8. Strange G, Gabbay E, Kermeen F, et al. Time from symptoms to definitive diagnosis of idiopathic pulmonary arterial hypertension: the delay study. *Pulm Circ* 2013; 3: 89–94.
9. Hurdman J, Condliffe R, Elliot CA, et al. ASPIRE registry: assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre. *Eur Respir J* 2012; 39: 945–955.
10. Sanchez-Roman J, Opitz CF, Kowal-Bielecka O, et al. Screening for PAH in patients with systemic sclerosis: focus on Doppler echocardiography. *Rheumatology* 2008; 47: v33–v35.
11. Kiely DG, Elliot CA, Sabroe I, et al. Pulmonary hypertension: diagnosis and management. *BMJ* 2013; 346(1): f2028.
12. Avouac J, Airo P, Meune C, et al. Prevalence of pulmonary hypertension in systemic sclerosis in European Caucasians and metaanalysis of 5 studies. *J Rheumatol* 2010; 37: 2290–2298.
13. Coghlan JG, Denton CP, Grunig E, et al. Evidence-based detection of pulmonary arterial hypertension in systemic sclerosis: the DETECT study. *Ann Rheum Dis* 2014; 73: 1340–1349.
14. Ashrafian H and Darzi A. Transforming health policy through machine learning. *PLOS Med* 2018; 15: e1002692.
15. Ngiam KY and Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019; 20: e262–e273.
16. Chen JH and Asch SM. Machine learning and prediction in medicine – beyond the peak of inflated expectations. *N Engl J Med* 2017; 376: 2507–2509.
17. Bergemann R, Allsopp A, Jenner H, et al. Using real-world data, can we diagnose iPAH earlier? An overview of the SPHInX project. *Pulm Circ* 2018; 8: 1–9.
18. Galie N, Humbert M, Vachiery JL, et al. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: the Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Heart J* 2016; 37: 67–119.
19. Friedman J. Greedy function approximation: a gradient boosting machine. *IMS 1999 Reitz Lecture*. Ann Stat 1999; 29: 1189–1232.
20. Breiman L. Bagging predictors. *Mach Learn* 1996; 24: 123–140.
21. Chen T and Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining -KDD '16*, 785–794. New York, USA, 2016: ACM Press.
22. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence – volume 2*, Montreal, 1995, pp.1137–1143. Quebec, Canada: Morgan Kaufmann Publishers Inc.
23. Lai Y-C, Potoka KC, Champion HC, et al. Pulmonary arterial hypertension: the clinical syndrome. *Circ Res* 2014; 115: 115–130.
24. McGoon MD, Benza RL, Escribano-Subias P, et al. Pulmonary arterial hypertension: epidemiology and registries. *J Am Coll Cardiol* 2013; 62: D51–D59.
25. Burger CD, Ghandour M, Padmanabhan Menon D, et al. Early intervention in the management of pulmonary arterial hypertension: clinical and economic outcomes. *Clinicoecon Outcomes Res* 2017; 9: 731–739.
26. Lau EM, Humbert M and Celermajer DS. Early detection of pulmonary arterial hypertension. *Nat Rev Cardiol* 2015; 12: 143–155.
27. Galie N, Rubin L, Hoeper M, et al. Treatment of patients with mildly symptomatic pulmonary arterial hypertension with bosentan (EARLY study): a double-blind, randomised controlled trial. *Lancet* 2008; 371: 2093–2100.
28. Humbert M, Yaici A, de Groote P, et al. Screening for pulmonary arterial hypertension in patients with systemic sclerosis: clinical characteristics at diagnosis and long-term survival. *Arthritis Rheum* 2011; 63: 3522–3530.
29. Hopkins WE, Ochoa LL, Richardson GW, et al. Comparison of the hemodynamics and survival of adults with severe primary pulmonary hypertension or Eisenmenger syndrome. *J Heart Lung Transplant* 1996; 15: 100–105.
30. Dufour R, Pruett J, Hu N, et al. Healthcare resource utilization and costs for patients with pulmonary arterial hypertension: real-world documentation of functional class. *J Med Econ* 2017; 20: 1178–1186.