# Evaluation of code-based algorithms to identify pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension patients in large administrative databases

**Viviane P. Sprecher[1]** [iD]**, Eva-Maria Didden[1], Joel N. Swerdel[2] and Audrey Muller[1]**

[1]Actelion Pharmaceuticals Ltd, Allschwil, Switzerland; [2]Janssen Pharmaceuticals, Raritan, NJ, USA

## Abstract

Large administrative healthcare (including insurance claims) databases are used for various retrospective real-world evidence studies. However, in pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension, identifying patients retrospectively based on administrative codes remains challenging, as it relies on code combinations (algorithms) and the accuracy for patient identification of most of them is unknown. This study aimed to assess the performance of various algorithms in correctly identifying patients with pulmonary arterial hypertension or chronic thromboembolic pulmonary hypertension in administrative databases. A systematic literature review was performed to find publications detailing code-based algorithms used to identify pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension patients. PheValuator, a diagnostic predictive modelling tool, was applied to three US claims databases, yielding models that estimated the probability of a patient having the disease. These models were used to evaluate the performance characteristics of selected pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension algorithms. With increasing algorithm complexity, average positive predictive value increased (pulmonary arterial hypertension: 13.4–66.0%; chronic thromboembolic pulmonary hypertension: 10.3–75.1%) and average sensitivity decreased (pulmonary arterial hypertension: 61.5–2.7%; chronic thromboembolic pulmonary hypertension: 20.7–0.2%). Specificities and negative predictive values were high ($\geq$97.5%) for all algorithms. Several of the algorithms performed well overall when considering all of these four performance parameters, and all algorithms performed with similar accuracy across the three claims databases studied, even though most were designed for patient identification in a specific database. Therefore, it is the objective of a study that will determine which algorithm may be most suitable; one- or two-component algorithms are most inclusive and three- or four-component algorithms identify most precise pulmonary arterial hypertension or chronic thromboembolic pulmonary hypertension populations, respectively.

## Keywords

pulmonary arterial hypertension, chronic thromboembolic pulmonary hypertension, PheValuator, claims databases, validation

## Introduction

Pulmonary hypertension (PH) is a rare and progressive disease characterised by increased pulmonary vascular resistance that ultimately leads to right heart failure and death.[1,2] Pulmonary arterial hypertension (PAH) and chronic thromboembolic pulmonary hypertension (CTEPH) are

Corresponding author:
Audrey Muller, Actelion Pharmaceuticals Ltd, Gewerbestrasse 16, 4123 Allschwil, Switzerland.
Email: amuller9@ITS.JNJ.com

two groups of PH (groups 1 and 4, respectively),[2] and derived estimates from UK and Swedish centres show that PAH and CTEPH are estimated to affect 44.8–46.0 patients per million and 24.2–34.9 patients per million, respectively.[3–5] Patients with PAH or CTEPH are diagnosed on average between the age of 50 and 65 years and may present with similar symptoms, but there are differences in risk factors, pathologic mechanisms, diagnostic strategies and treatment approaches for the two diseases.[2]

Administrative databases include structured, coded information for health reimbursement (claims), hospital management and national vital statistics. These databases are widely used in epidemiology, health economics and public health surveillance to provide real-world data on disease burden, treatment patterns, healthcare resource utilisation and costs.[6] Claims databases cover a large and representative sample of the population, making them of particular interest in the rare disease context. These databases typically collect information on diagnoses (e.g. using International Classification of Diseases (ICD) codes), procedures and drug use. While research is not the primary purpose of such databases, their advantages include easy access at a relatively low cost, not being subject to recall bias and data availability in a well-structured format. Limitations include their lack of clinical data and a total reliance on coding accuracy.[7]

In PH, dedicated diagnosis codes for each group, including PAH and CTEPH, became available in the October 2017 ICD-10 update.[8] However, given the retrospective design of most claims database studies and a potential under-utilisation post-implementation, these new PH group-specific ICD-10 codes are still of limited usefulness for patient identification. Currently, identifying patients retrospectively relies on ICD-9 and pre-October 2017 ICD-10 codes, which differentiate only between primary (idiopathic/heritable PAH) and secondary PH (associated forms of PAH and PH groups 2–5), and do not appropriately reflect the current PH classifications.

To identify PAH or CTEPH patients from claims databases, different combinations of diagnosis, procedure and drug codes are used, herein collectively referred to as code-based algorithms. A large variety of code-based algorithms for PH have been published but the accuracy of very few has been assessed using clinical data.[6,9] Therefore, the reliability of these algorithms remains unclear, and there are calls for future studies to address this.[6] As such, the aim of this study was to assess which code-based algorithms most accurately identified patients with PAH or CTEPH in claims databases, using the diagnostic predictive modelling tool, PheValuator.[10]

## Methods

A systematic literature review was performed to identify published code-based algorithms developed for PAH and CTEPH. The performance characteristics (sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)) of the algorithms in identifying patients with PAH or CTEPH were then assessed using PheValuator.[10]

### Identification of published code-based algorithms

A systematic literature search was completed on 1 October 2019 using PubMed and Embase databases, in accordance with Preferred Reporting Items for Systematic reviews and Meta-Analyses guidelines (Supplementary Figure 1),[11] by the lead author (V.P.S.). The search was limited to English, German, French or Spanish language manuscripts published between 1 January 2000 and 30 September 2019, inclusive, without restrictions on type and origin of database, or age of the population studied. The definition and management of PAH and CTEPH have changed substantially in the past two decades. Thus, this time frame was chosen to capture the diversity of administrative algorithms used for patient identification. The full search strategy, including search terms, is shown in Supplementary Table 1. Further details on the screening of search results and extraction of information are provided in the Supplementary methods.

For PAH algorithms, priority for further evaluation was given to (1) algorithms that identified patients with PAH or any PAH subgroup in any administrative database, published between 2000 and 2019, for which the accuracy of patient identification was evaluated in the original publication, and (2) algorithms identifying patients with PAH in administrative databases for health reimbursement (claims databases) published between 2015 and 2019. The time frame was reduced in the second group to include only the most recently published algorithms, which were assumed to be based on a combination of ICD-9 and ICD-10 codes as well as potentially on all currently available PAH-specific drugs, and therefore represent an improved version of past algorithm application strategies. When several algorithms were similar, only the most inclusive algorithm was selected for further evaluation. Due to the limited number of CTEPH algorithms identified in the literature, all were selected for further evaluation.

### Evaluation of code-based algorithm performance using the diagnostic predictive modelling tool, PheValuator

In order to cover the age groups mainly affected by PAH and CTEPH, the following three US claims datasets were used in the evaluation of algorithm performance: Optum® De-Identified Clinformatics® Data Mart Database (OptumInsight, Eden Prairie, MN), ages ≥18 years ('Optum'); IBM® MarketScan® Commercial Claims and Encounters Database, ages 18–62 years ('CCAE', limited to ≤62 years to observe a patient for at least three years post-inclusion); and IBM® MarketScan® Medicare Supplemental and Coordination of Benefits Database, ages

≥66 years ('Medicare', limited to ≥66 years to observe a patient for at least one year pre-inclusion). Data used included patient records from 1 January 2010 until 30 March 2019, inclusive. The extent of overlap between patients in Optum® and IBM® MarketScan® datasets is unknown and impossible to estimate due to data anonymisation. However, this potential duplication is likely to be sufficiently small to allow for the presentation of averages across all datasets. The Optum® and IBM® MarketScan® databases used in this study were reviewed by the New England Institutional Review Board. Studies conducted in these databases were determined to be exempt from ethics approval as they do not qualify as human subjects research.

Using the ATLAS tool within the Observational Health Data Sciences and Informatics (OHDSI) toolset,[12] PAH and CTEPH cohorts were created in the three selected datasets, based on the code-based algorithms identified in the literature. When available, the exact source codes were used for creating the cohorts, otherwise standardised coding concepts from the Common Data Model Vocabularies of the Observational Medical Outcomes Partnership[13] were applied. Diagnosis codes based on ICD-9 were translated into pre-October 2017 ICD-10 and vice versa per published conversion suggestions, i.e. 416.0 to/from I27.0, and 416.8 to/from I27.2.[8] This conversion was performed to allow PAH and CTEPH patient identification according to both of these past coding conventions.

The cohorts reflecting the code-based algorithms were then exported into R and evaluated using the OHDSI R package PheValuator.[10] Disease-specific diagnostic predictive models (i.e. for PAH and CTEPH separately) were developed for each dataset, using logistic regression with Least Absolute Shrinkage and Selection Operator L1-regularisation.[14] The models yielded sets of weighted predictors for the diseases of interest. Among the most important predictors from the PAH models were history of pulmonary heart disease or PH, echocardiography and phosphodiesterase type-5 inhibitors. For CTEPH, similar important predictors as for PAH were obtained, with the addition of embolism and pulmonary ventilation/perfusion imaging. These predictors were used to estimate the probability of a patient having PAH or CTEPH in a randomly selected cohort from each dataset. In the absence of a diagnostic standard defining patients' true disease status (i.e. based on clinical assessment results and laboratory values), these predictions defined a probabilistic gold standard for evaluation, with the assumption that the probability of an individual having PAH or CTEPH based on the predictive model corresponded to the true disease status of that individual.

The PAH and CTEPH algorithms selected for evaluation were compared with the respective probabilistic gold standard. The resulting estimates for the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) were used to calculate the following: (i) sensitivity (defined as TP/(TP + FN)), (ii) specificity (defined as TN/(TN + FP)), (iii) PPV (defined as TP/(TP + FP)) and (iv) NPV (defined as TN/(TN + FN)).

## Results

The systematic literature review returned a total of 59 publications describing code-based algorithms developed to identify patients with PAH ($n = 50$), with CTEPH ($n = 5$) or with either PAH or CTEPH ($n = 4$). The majority of the publications ($n = 43$ (72.9%)) identified patients in US databases, and others were from the UK ($n = 5$), Canada ($n = 4$), France ($n = 2$), Taiwan ($n = 2$), Norway ($n = 1$), Spain ($n = 1$) and South Korea ($n = 1$).

### PAH

*Identification of published code-based algorithms.* In total, 51 unique algorithms describing PAH in an administrative database were identified in the literature (Fig. 1; Supplementary Table 2). Of the 48 algorithms using diagnosis codes, 38 were based on ICD-9 (79.2%), five on ICD-10 (10.4%) and four on a combination of ICD-9 and ICD-10 (8.3%), as well as one algorithm on UK-specific Read codes (2.1%), which was subsequently excluded from this evaluation on the basis of being too database-specific and requiring substantial modification to convert into ICD-9/-10.

*Evaluation of code-based algorithm performance using the diagnostic predictive modelling tool, PheValuator.* Individually, each of the 15 PAH algorithms analysed using PheValuator produced similar results across the three claims databases studied (Supplementary Tables 3 and 5). In contrast, averages across the databases showed variability in accuracy between algorithms. Overall, average sensitivity ranged from 2.7 to 61.5% and average PPV from 13.4 to 66.0% (Table 1). Sensitivity tended to decrease with increasing PPV. Overall, average specificity and average NPV were ≥97.5% and ≥99.5%, respectively (Supplementary Table 5). Average PAH prevalence across databases was 0.47% for all algorithms, ranging from 0.16 to 0.87%, calculated per database.

The highest average sensitivities across the databases were 48.1% (range: 47.0–48.7%) using an algorithm published by Papani et al. (*at least one primary/secondary PH diagnosis code (outpatient setting)*)[15] and 61.5% (range: 60.1–64.2%) using the algorithm reported by Choi et al. (*at least one primary/secondary PH diagnosis code (any setting)*; Table 1).[16] The highest average PPVs across the databases were 63.3% (range: 47.6–83.8%) in the algorithm by Burke et al. (*at least one primary/secondary PH diagnosis code (outpatient) + at least one primary/secondary PH diagnosis code (inpatient) + at least one PAH drug code*),[17] and 66.0% (range: 49.8–83.9%) for another algorithm reported by Papani et al. (*at least one primary/secondary PH diagnosis code (outpatient) + at least two classes of PAH drug codes*).[15]

Fig. 2 presents the grouping of algorithms according to the number of algorithm components. Algorithms based on
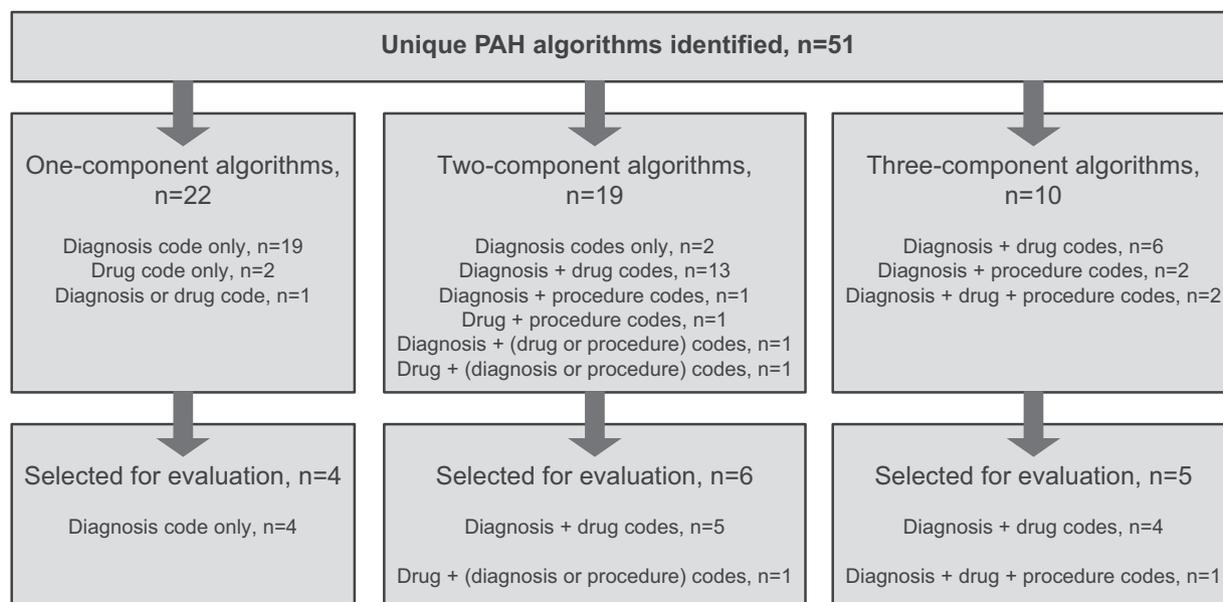
**Fig. 1.** Summary of PAH algorithms identified and selected for evaluation using PheValuator.
PAH: pulmonary arterial hypertension.

**Table 1.** Performance characteristics of code-based PAH algorithms: averages and ranges of sensitivity and PPV across Optum, CCAE and Medicare databases.

|  | Sensitivity (min, max) (%) | PPV (min, max) (%) |
|---|---|---|
| One-component algorithms | | |
| ≥1 primary PH Dx[23,24,28–35] | 24.6 (22.4, 27.4) | 18.3 (11.9, 29.7) |
| ≥1 primary PH Dx with exclusion criteria[25] | 21.4 (19.4, 23.5) | 19.8 (12.9, 32.4) |
| ≥1 primary/secondary PH Dx (outpatient)[15] | 48.1 (47.0, 48.7) | 14.0 (9.0, 22.1) |
| ≥1 primary/secondary PH Dx[16] | 61.5 (60.1, 64.2) | 13.4 (8.4, 21.2) |
| Two-component algorithms | | |
| (≥1 primary/secondary/other PH Dx or RHC) + ≥1 PAH-specific Rx[36] | 8.2 (5.6, 11.0) | 36.9 (29.8, 48.4) |
| (≥1 primary PH Dx or ≥1 secondary/other PH Dx with Dx for PH group 1 associated disease) + ≥1 Rx for daily PDE5i[26] | 3.4 (2.4, 5.0) | 41.8 (35.2, 54.5) |
| ≥1 primary/secondary PH Dx (outpatient) + ≥1 class of PAH-specific Rx[15] | 8.5 (5.8, 11.2) | 34.9 (28.0, 47.8) |
| ≥1 primary/secondary PH Dx + ≥1 PAH-specific Rx[37] | 9.2 (6.1, 12.5) | 32.5 (26.2, 44.8) |
| ≥1 primary/secondary PH Dx + ≥1 CCB or PAH-specific Rx[38] | 4.0 (2.7, 6.0) | 14.5 (8.2, 25.7) |
| ≥1 primary/secondary/other PH Dx + ≥1 PAH-specific Rx[39] | 7.5 (5.3, 9.7) | 43.2 (33.0, 58.2) |
| Three-component algorithms | | |
| ≥1 primary/secondary PH Dx (outpatient) + ≥2 classes of PAH-specific Rx[15] | 2.7 (2.2, 3.6) | 66.0 (49.8, 83.9) |
| ≥1 primary/secondary/other PH Dx (outpatient) + ≥2 PAH-specific Rx[22] | 7.4 (5.2, 9.7) | 39.2 (31.6, 53.2) |
| ≥2 primary/secondary/other PH Dx (outpatient) + ≥1 PAH-specific Rx[22] | 7.4 (5.2, 9.3) | 49.0 (36.0, 68.7) |
| ≥1 primary/secondary PH Dx (outpatient) + ≥1 primary/secondary PH Dx (inpatient) + ≥1 PAH-specific Rx[17] | 3.7 (3.0, 4.4) | 63.3 (47.6, 83.8) |
| ≥1 primary/secondary/other PH Dx + RHC + ≥1 PAH-specific Rx[22] | 3.7 (2.6, 5.0) | 61.8 (46.1, 81.0) |

CCB: calcium-channel blocker; Dx: diagnosis code; PAH: pulmonary arterial hypertension; PDE5i: phosphodiesterase type-5 inhibitor; PH: pulmonary hypertension; PPV: positive predictive value; RHC: right-heart catheterisation; Rx: drug code.
Note: PAH-specific Rx: sildenafil, tadalafil, bosentan, ambrisentan, macitentan, epoprostenol, iloprost, treprostinil, selexipag and riociguat.
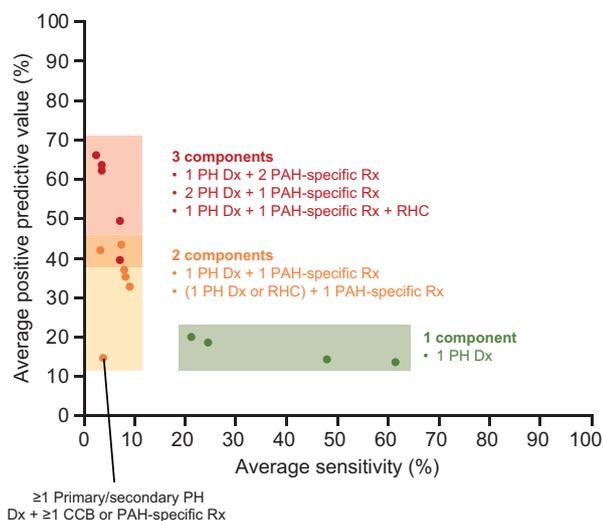
**Fig. 2.** PAH algorithms: graphical presentation of average positive predictive value and average sensitivity across the three databases examined (Optum, CCAE and Medicare). Code-based algorithms including a single diagnosis code are shown as green dots, those based on two components as orange dots and those based on three components as red dots.

CCB: calcium-channel blocker; Dx: diagnosis code; PAH: pulmonary arterial hypertension; PH: pulmonary hypertension; RHC: right-heart catheterisation; Rx: drug code.

a single diagnosis code had comparatively moderate PPVs and differing sensitivities. The algorithms that were based on two components yielded sensitivities below 10%, and higher PPVs, with the exception of one algorithm that included a therapy not specific to PAH (calcium-channel blockers) that generated an outlier in terms of PPV. Algorithms based on three components also returned low sensitivities and comparatively high PPVs.

Code-based PAH algorithm performance results obtained via predictive modelling were compared with previously published performance evaluations in Table 2. Specificity and NPV were generally similar between PheValuator results and previously published performance results. PPV results from PheValuator and previous evaluations were more similar for two- and three-component algorithms than one-component algorithms, while results for sensitivity varied greatly.

## CTEPH

*Identification of published code-based algorithms.* A total of 11 unique CTEPH algorithms were identified (Supplementary Table 4), of which four applied diagnosis codes only, six applied a combination of diagnosis and procedure codes, and one applied a diagnosis code plus either a procedure or a drug code. Diagnosis codes were based on ICD-9 in seven algorithms (63.6%), on ICD-10 in two (18.2%) and a combination of both in two (18.2%).

*Evaluation of code-based algorithm performance using the diagnostic predictive modelling tool, PheValuator.* The performance of

individual CTEPH algorithms was similar across the three claims databases used (Supplementary Table 6). Using PheValuator, average sensitivity ranged from 0.2 to 20.7%, and was inversely related to average PPV, which ranged from 10.3 to 75.1% for the 11 CTEPH algorithms (Table 3). Average specificity and average NPV were $\geq 99.5\%$ for all algorithms (Supplementary Table 6). Average CTEPH prevalence across databases was 0.25% for all algorithms, ranging from 0.11 to 0.36% as calculated per database.

The highest average sensitivities across databases were 20.4% (range: 18.9–22.1%) and 20.7% (range: 17.5–22.6%) observed for two algorithms reported by Tapson et al. (*at least one pulmonary embolism (PE) diagnosis code + at least one PH-related symptom + at least one code for PH-specific procedure*; and *at least one PE diagnosis code + at least one PH-related symptom*, respectively).[18] The highest PPV averages across databases were 55.1% (range: 33.0–68.1%) for the algorithm described by Martinez et al. (*at least one venous thromboembolism diagnosis code + at least one primary PH diagnosis code + at least one therapy (lung transplant | pulmonary endarterectomy | PAH drug code)*)[19] and 75.1% (range: 67.4–87.9%) for the algorithm by Teal et al. (*at least one PE diagnosis code + PH-related symptoms + at least one diagnosis code for chronic PE or PH without primary PH + code for PH-unspecific procedure + code for PH-specific procedure*).[20]

The algorithm providing the highest accuracy when considering all four performance characteristics was reported by Said et al. (*at least one PE diagnosis code + first diagnosis code for PH + procedure code for echocardiography or right-heart catheterisation (RHC) + second diagnosis code for PH*), which had an average sensitivity of 6.0% (range: 4.4–6.9%), PPV of 48.7% (range: 28.4–66.7%), with specificity and NPV of $\geq 99.8\%$.[21]

Fig. 3 presents the grouping of algorithms according to their complexity. Algorithms based on two components had differing PPVs and low to moderate sensitivities. Algorithms based on three components yielded differing sensitivities together with higher PPVs. The most complex algorithms ($\geq 4$ components) had very low sensitivities and comparatively high PPVs.

For comparison, code-based CTEPH algorithm performance results obtained via predictive modelling were tabulated with previously published performance evaluations (Table 4). Specificity and NPV results from previous publications were similar to the results of PheValuator, yet sensitivity results from the PheValuator-based evaluation were generally much lower than in previously published evaluations.

## Discussion

The aim of this study was to evaluate the performance characteristics of published code-based algorithms identifying patients with PAH or CTEPH in administrative databases.

**Table 2.** Comparison of published PAH algorithm performance results with presented PheValuator averages across Optum, CCAE and Medicare databases.

| | PheValuator evaluation results (min, max) (%) | Published evaluation results (%) |
|---|---|---|
| One-component algorithms | | |
| ≥1 primary PH Dx[23,24] | | |
|     Sensitivity | 24.6 (22.4, 27.4) | 63.1[24] |
|     Specificity | 99.3 (98.6, 99.9) | 96.7[24] |
|     PPV[a] | 18.3 (11.9, 29.7) | 6.7, 33.3[a,23]; 66.7[24] |
| ≥1 primary PH Dx with exclusion criteria[25] | | |
|     PPV | 19.8 (12.9, 32.4) | 3.3 |
| ≥1 primary/secondary PH Dx (outpatient)[15] | | |
|     PPV | 14.0 (9.0, 22.1) | 9.3, 15.8[a] |
| Two-component algorithms | | |
| (≥1 primary PH Dx or ≥1 secondary/other PH Dx with Dx for PH group 1 associated disease) + ≥1 Rx for daily PDE5i[26] | | |
|     PPV | 41.8 (35.2, 54.5) | 41.5 |
| ≥1 primary/secondary PH Dx (outpatient) + ≥1 class of PAH Rx[15] | | |
|     Sensitivity | 8.5 (5.8, 11.2) | 64.3, 67.4[a] |
|     Specificity | 99.9 (99.8, 99.9) | 81.9, 86.9[a] |
|     PPV | 34.9 (28.0, 47.8) | 34.7, 40.0[a] |
|     NPV | 99.6 (99.2, 99.9) | 92.4, 96.3[a] |
| Three-component algorithms | | |
| ≥1 primary/secondary PH Dx (outpatient) + ≥2 classes of PAH Rx[15] | | |
|     Sensitivity | 2.7 (2.2, 3.6) | 28.2, 42.9[a] |
|     Specificity | 99.9 (99.9, 99.9) | 94.0, 98.6[a] |
|     PPV | 66.0 (49.8, 83.9) | 57.1, 66.9[a] |
|     NPV | 99.5 (99.1, 99.8) | 89.7, 93.0[a] |

[a]Performance evaluated twice in the same publication.
Dx: diagnosis code; NPV: negative predictive value; PAH: pulmonary arterial hypertension; PDE5i: phosphodiesterase type-5 inhibitor; PH: pulmonary hypertension; PPV: positive predictive value; Rx: drug code.
Note: PAH-specific Rx: sildenafil, tadalafil, bosentan, ambrisentan, macitentan, epoprostenol, iloprost, treprostinil, selexipag and riociguat.

In total, 15 PAH and 11 CTEPH algorithms were included in the evaluation, and a complete set of performance characteristics for each were provided, making this the first study of its kind. PPV increased with increasing algorithm complexity and was inversely related to sensitivity. Since specificity and NPV were similarly high (≥97.5%) for all algorithms, these accuracy measures were not the main focus of the study.

In the current study, a PPV of ≥60% was reported for three of the PAH algorithms evaluated but these also reported very low sensitivity (2.7–3.7%).[15,17,22] Conversely, the highest sensitivity obtained using PheValuator was 61.5% in an algorithm reporting a low PPV (13.4%).[16] For the six PAH algorithms with previously published performance results,[15,23–26] increasing algorithm complexity and recentness of published evaluation were factors that contributed to achieving similar PPV results between the PheValuator predictive modelling and the published performance results. Overall, algorithm performance depended on the number of components used. For PAH, none of the evaluated algorithms achieved an overall best performance in accurately identifying patients. Thus, despite the large number of published algorithms, there may be an unmet need for an algorithm that accurately identifies PAH patients with a good balance of all parameters.

Two of the CTEPH algorithms evaluated reported a PPV of ≥50%, the highest sensitivity according to PheValuator was 21%.[18–20] The algorithm with the best performance results across all parameters was 'at least one PE diagnosis code + first diagnosis code for PH + procedure code for echocardiography or RHC + second diagnosis code for PH',[21] with a PPV of 48.7% and a sensitivity of 6.0%. Comparison of the three CTEPH algorithms with previously published performance results[19,20,27] to PheValuator predictive modelling was difficult due to the incomplete set of performance characteristics published. Where results were available, specificity, PPV and NPV were reasonably similar but sensitivity estimates obtained using PheValuator were much lower than the values reported in the original publication.

When comparing PheValuator results with published algorithm evaluations, differences in coding and reporting practices as well as in the amount of information available – between countries, healthcare centres and systems, database

**Table 3.** Performance characteristics of code-based CTEPH algorithms: averages and ranges of sensitivity and PPV across Optum, CCAE and Medicare databases.

| | Sensitivity (min, max) (%) | PPV (min, max) (%) |
|---|---|---|
| **Two-component algorithms** | | |
| ≥1 primary PH Dx + ≥1 PE Dx[40] | 6.3 (4.7, 7.3) | 41.4 (23.1, 57.3) |
| ≥1 PE Dx + ≥1 PH-related symptom[18] | 20.7 (17.5, 22.6) | 10.3 (5.5, 12.9) |
| ≥1 PE Dx + ≥1 primary/secondary PH Dx[18] | 12.7 (11.3, 14.7) | 31.0 (15.0, 41.1) |
| ≥1 VTE Dx + ≥1 primary/secondary/other PH Dx[41] | 6.8 (5.4, 8.3) | 22.3 (10.5, 28.9) |
| ≥1 primary PH Dx (inpatient) + ≥1 CTEPH-related procedure or hospital stay for/history of PE[27] | 14.8 (14.0, 16.2) | 27.0 (14.4, 37.5) |
| **Three-component algorithms** | | |
| ≥1 PE Dx + ≥1 PH-related symptom + ≥1 V/Q scan/Echo/CTA/RHC/PA[18] | 20.4 (18.9, 22.1) | 13.9 (7.1, 19.0) |
| ≥1 PE Dx + ≥1 primary/secondary PH Dx + ≥1 V/Q scan/Echo/CTA/RHC/PA[18] | 17.4 (15.0, 19.1) | 31.4 (15.6, 42.1) |
| ≥1 VTE Dx + ≥1 primary PH Dx + ≥1 therapy (LT/PEA/PAH-specific Rx)[19] | 0.6 (0.4, 0.8) | 55.1 (33.0, 68.1) |
| **At least four-component algorithms** | | |
| ≥1 PE Dx + first primary/secondary PH Dx + Echo/RHC + second primary/secondary PH Dx[21] | 6.0 (4.4, 6.9) | 48.7 (28.4, 66.7) |
| ≥1 PE Dx + first primary/secondary PH Dx + Echo (specialist)/RHC + second primary/secondary PH Dx[42,43] | 1.8 (1.6, 2.3) | 44.1 (22.3, 63.1) |
| ≥1 PE Dx + PH-related symptoms + ≥1 Dx for chronic PE or PH without primary PH + ECG/Echo/MRI/HRCT + V/Q scan/RHC/CTA/PA[20] | 0.2 (0.1, 0.2) | 75.1 (67.4, 87.9) |

CTA: computed tomography angiography; CTEPH: chronic thromboembolic pulmonary hypertension; Dx: diagnosis code; ECG: electrocardiography; Echo: echocardiography; HRCT: high-resolution computed tomography; LT: lung transplant; MRI: magnetic resonance imaging; PA: pulmonary angiography; PAH: pulmonary arterial hypertension; PE: pulmonary embolism; PEA: pulmonary endarterectomy; PH: pulmonary hypertension; PPV: positive predictive value; RHC: right-heart catheterisation; Rx: drug code; V/Q scan: ventilation/perfusion scan; VTE: venous thromboembolism.
Note: PAH-specific Rx: sildenafil, tadalafil, bosentan, ambrisentan, macitentan, epoprostenol, iloprost, treprostinil, selexipag and riociguat.
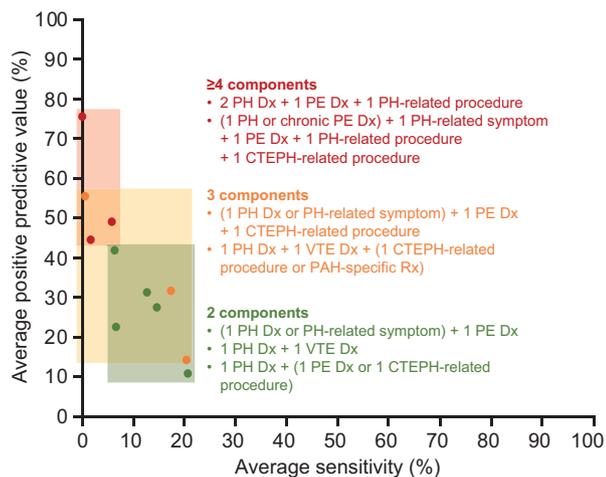


**Fig. 3.** CTEPH algorithms: graphical presentation of average positive predictive value and average sensitivity across the three databases examined (Optum, CCAE and Medicare). Code-based algorithms including two components are shown as green dots, those based on three components as orange dots and those based on at least four components as red dots.
CTEPH: chronic thromboembolic pulmonary hypertension; Dx: diagnosis code; PAH: pulmonary arterial hypertension; PE: pulmonary embolism; PH: pulmonary hypertension; Rx: drug code; VTE: venous thromboembolism.

types and over time – may play a role in algorithm performance variability between different databases. This has been demonstrated in studies where identical algorithms were evaluated in two different settings and returned different results.[15,23,24] In addition, the lack of exact codes in the original publications limited the recreation of identical cohorts.

Disease prevalence in the study population influences the PPV, such that the lower the prevalence in the study population, the lower the PPV. Future work should therefore consider evaluating PAH and CTEPH algorithms in a PH population rather than the general population captured in US claims databases. Furthermore, algorithms may identify only patients with severe disease, who have all the codes in the algorithm present in their claims records. Whether this is the case for more complex algorithms with low sensitivity needs to be assessed with patient-level clinical data, such as comorbidities, symptoms and PAH-/CTEPH-related procedure results.

## Strengths and limitations

An important strength of the methodology applied is its extendibility and adaptability to various types of administrative databases and databases from different countries,

**Table 4.** Comparison of published CTEPH algorithm performance results with presented PheValuator averages across Optum, CCAE and Medicare databases.

| | PheValuator evaluation results (min, max) (%) | Published evaluation results (%) |
|---|---|---|
| **Two-component algorithms** | | |
| ≥1 primary PH Dx (inpatient) + ≥1 CTEPH-related procedure or hospital stay for/history of PE[27] | | |
|     Sensitivity | 14.8 (14.0, 16.2) | 70.4 |
|     Specificity | 99.9 (99.8, 99.9) | 95.0 |
|     PPV | 27.0 (14.4, 37.5) | 40.9 |
|     NPV | 99.8 (99.7, 99.9) | 98.5 |
| **Three-component algorithms** | | |
| ≥1 VTE Dx + ≥1 primary PH Dx + ≥1 therapy (LT/PEA or PAH-specific Rx)[19] | | |
|     Sensitivity | 0.6 (0.4, 0.8) | 85.3, 85.8[a] |
|     Specificity | 99.9 (99.9, 99.9) | 99.2, 100.0[a] |
| **At least four-component algorithms** | | |
| ≥1 PE Dx + PH-related symptoms + ≥1 Dx for chronic PE or PH without primary PH + ECG/Echo/MRI/HRCT + V/Q scan/RHC/CTA/PA[20] | | |
|     Sensitivity | 0.2 (0.1, 0.2) | 0.4 |

[a]Performance evaluated twice in the same publication. In the original publication, CTEPH patients were identified applying a code-based algorithm in addition to screening complementary general practitioner information and clinical notes, which may increase the probability of identifying CTEPH patients. This additional information was not available for evaluation in the present study.
CTA: computed tomography angiography; CTEPH: chronic thromboembolic pulmonary hypertension; Dx: diagnosis code; ECG: electrocardiography; Echo: echocardiography; HRCT: high-resolution computed tomography; LT: lung transplant; MRI: magnetic resonance imaging; NPV: negative predictive value; PA: pulmonary angiography; PAH: pulmonary arterial hypertension; PE: pulmonary embolism; PEA: pulmonary endarterectomy; PH: pulmonary hypertension; PPV: positive predictive value; RHC: right-heart catheterisation; Rx: drug code; V/Q scan: ventilation/perfusion scan; VTE: venous thromboembolism.
Note: PAH-specific Rx: sildenafil, tadalafil, bosentan, ambrisentan, macitentan, epoprostenol, iloprost, treprostinil, selexipag and riociguat.

**Table 5.** Summary of recommended published algorithms according to research objective.

| Research objective | Disease | Recommended algorithm |
|---|---|---|
| Identify patients at risk of disease | PAH | ≥ 1 primary/secondary PH Dx[16] |
| | CTEPH | ≥ 1 PE Dx + ≥ 1 primary/secondary PH Dx[18] |
| Analyse treatment patterns and healthcare resource utilisation | PAH | ≥ 1 primary/secondary/other PH Dx + ≥ 1 PAH-specific Rx[39] |
| | | ≥ 2 primary/secondary/other PH Dx (outpatient) + ≥ 1 PAH-specific Rx[22] |
| | CTEPH | ≥ 1 PE Dx + ≥ 1 primary/secondary PH Dx + ≥ 1 V/Q scan/Echo/CTA/RHC/PA[18] |
| | | ≥ 1 PE Dx + first PH Dx + procedure code for Echo/RHC + second PH Dx[21] |
| Estimate prevalence | PAH | ≥ 1 primary/secondary PH Dx (outpatient) + ≥ 1 primary/secondary PH Dx (inpatient) + ≥ 1 PAH-specific Rx[17] |
| | CTEPH | ≥ 1 PE Dx + first PH Dx + procedure code for Echo/RHC + second PH Dx[21] |

CTA: computed tomography angiography; CTEPH: chronic thromboembolic pulmonary hypertension; Dx: diagnosis code; Echo: echocardiography; PA: pulmonary angiography; PAH: pulmonary arterial hypertension; PE: pulmonary embolism; PH: pulmonary hypertension; RHC: right-heart catheterisation; Rx: drug code; V/Q scan: ventilation/perfusion scan.

provided that these databases are available in coded structured format. Despite not being validated yet, PheValuator facilitated the evaluation of several code-based algorithms across different databases, which, given the large number of algorithms evaluated, would not have been as efficient via a manual medical chart review. Thus, it provides a useful

alternative for algorithm evaluation. PheValuator will likely be updated over time, allowing for more flexibility in the underlying factors and assumptions that drive the predictive modelling.[10] Should this evaluation be repeated in the future, the results may vary. In addition, the accuracy of the algorithms depends not only on correct coding by healthcare

professionals, but also on the accuracy of the codes within the established coding systems. Analysing algorithms that did not include the new disease-specific ICD-10 codes (October 2017 update) in more recent datasets may have limited the overall accuracy of the algorithms evaluated.

### Recommendations based on published algorithms

Future studies wishing to utilise code-based algorithms should consider their objective when selecting an algorithm (Table 5). For research aiming to analyse treatment patterns and healthcare resource utilisation, one- or two-component algorithms based on diagnosis codes may be applied to capture all potential cases of PAH or CTEPH, respectively. However, to limit the number of false-positive patients, adding PH-specific drugs for PAH and PH-specific procedures for CTEPH to the algorithm is recommended. Studies aiming to estimate PAH or CTEPH prevalence may use three- or four-component algorithms, respectively, to only select true-positive patients, at the risk of missing patients.

## Conclusion

This study evaluated code-based algorithms used to identify patients with PAH or CTEPH against administrative databases and provided an approximate measure of their accuracy. Published algorithms were able to identify PAH and CTEPH patients with high specificity and NPV, but with widely variable low-to-moderate sensitivity and PPV. The objective of a study will determine the best-suited algorithm for patient identification in large administrative databases, from one- or two-component algorithms that produce the most inclusive study population to three- or four-component algorithms for identification of the most precise study population, for PAH and CTEPH, respectively.

### ORCID iD

Viviane P. Sprecher https://orcid.org/0000-0002-1558-6687

### Supplemental material

Supplemental material for this article is available online.

### References

1. Rose-Jones LJ and McLaughlin VV. Pulmonary hypertension: types and treatments. *Curr Cardiol Rev* 2015; 11: 73–79.
2. 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.
3. Kjellström BWM, Dahlerup H, Hesselstrand R, et al. Swedish Pulmonary Arterial Hypertension Registry. Annual Report, https://ucr.uu.se/spahr/arsrapporter (2018, accessed January 2020).
4. National Pulmonary Hypertension Audit, https://digital.nhs.uk/data-and-information/publications/statistical/national-pulmonary-hypertension-audit/national-pulmonary-hypertension-audit-2015 (2015, accessed July 2020).
5. National Audit of Pulmonary Hypertension. 9th Annual Report, https://digital.nhs.uk/data-and-information/publications/statistical/national-pulmonary-hypertension-audit/2018/2018 (accessed July 2020).
6. Mathai SC, Hemnes AR, Manaker S, et al. Identifying patients with pulmonary arterial hypertension using administrative claims algorithms. *Ann Am Thorac Soc* 2019; 16: 797–806.
7. Johnson EK and Nelson CP. Values and pitfalls of the use of administrative databases for outcomes assessment. *J Urol* 2013; 190: 17–18.
8. Mathai SC and Mathew S. Breathing (and coding?) a bit easier: changes to International Classification of Disease coding for pulmonary hypertension. *Chest* 2018; 154: 207–218.
9. Gillmeyer KR, Lee MM, Link AP, et al. Accuracy of algorithms to identify pulmonary arterial hypertension in administrative data: a systematic review. *Chest* 2019; 155: 680–688.
10. Swerdel JN, Hripcsak G and Ryan PB. PheValuator: development and evaluation of a phenotype algorithm evaluator. *J Biomed Inform* 2019; 97: 103258.
11. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; 339: b2535.
12. Observational Health Data Sciences and Informatics. Observational Health Data Sciences and Informatics ATLAS

tool, https://ohdsi.org/analytic-tools/ (2019, accessed 12 December 2019).

13. Observational Health Data Sciences and Informatics. OMOP Common Data Model, https://ohdsi.org/data-standardization/the-common-data-model/ (2019, accessed 12 December 2019).

14. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996; 58: 267–288.

15. Papani R, Sharma G, Agarwal A, et al. Validation of claims-based algorithms for pulmonary arterial hypertension. *Pulm Circ* 2018; 8: 2045894018759246.

16. Choi YM, Famenini S and Wu JJ. Incidence of pulmonary arterial hypertension in patients with psoriasis: a retrospective cohort study. *Perm J* 2017; 21: 16–073.

17. Burke JP, Hunsche E, Régulier E, et al. Characterizing pulmonary hypertension-related hospitalization costs among Medicare Advantage or commercially insured patients with pulmonary arterial hypertension: a retrospective database study. *Am J Manag Care* 2015; 21: S47–S58.

18. Tapson VF, Platt DM, Xia F, et al. Monitoring for pulmonary hypertension following pulmonary embolism: the INFORM study. *Am J Med* 2016; 129: 978–985.e2.

19. Martinez C, Wallenhorst C, Teal S, et al. Incidence and risk factors of chronic thromboembolic pulmonary hypertension following venous thromboembolism, a population-based cohort study in England. *Pulm Circ* 2018; 8: 2045894018791358.

20. Teal S, Auger WR, Hughes RJ, et al. Validation of a claims-based algorithm to identify patients with chronic thromboembolic pulmonary hypertension using electronic health record data. *Pulm Circ* 2019; 9: 2045894018814772.

21. Said Q, Martin BC, Joish VN, et al. The cost to managed care of managing pulmonary hypertension. *J Med Econ* 2012; 15: 500–508.

22. Burger CD, Pruett JA, Lichert CA, et al. Prostacyclin use among patients with pulmonary arterial hypertension in the United States: a retrospective analysis of a large health care claims database. *J Manag Care Spec Pharm* 2018; 24: 291–302.

23. Link J, Glazer C, Torres F, et al. International Classification of Diseases coding changes lead to profound declines in reported idiopathic pulmonary arterial hypertension mortality and hospitalizations: implications for database studies. *Chest* 2011; 139: 497–504.

24. Chang WT, Weng SF, Hsu CH, et al. Prognostic factors in patients with pulmonary hypertension – a nationwide cohort study. *J Am Heart Assoc* 2016; 5: e003579.

25. Fox BD, Azoulay L, Dell'Aniello S, et al. The use of antidepressants and the risk of idiopathic pulmonary arterial hypertension. *Can J Cardiol* 2014; 30: 1633–1639.

26. Kim D, Lee KM, Freiman RM, et al. Phosphodiesterase-5-inhibitor therapy for pulmonary hypertension in the US: actual vs recommended use. *AnnalsATS* 2018; 15: 693–701.

27. Cottin V, Avot D, Levy-Bachelot L, et al. Identifying chronic thromboembolic pulmonary hypertension through the French national hospital discharge database. *PLoS One* 2019; 14: e0214649.

28. Agarwal M, Agrawal S, Garg L, et al. Relation between obesity and survival in patients hospitalized for pulmonary arterial hypertension (from a Nationwide Inpatient Sample Database 2003 to 2011). *Am J Cardiol* 2017; 120: 489–493.

29. Anand V, Roy SS, Archer SL, et al. Trends and outcomes of pulmonary arterial hypertension-related hospitalizations in the United States: analysis of the Nationwide Inpatient Sample Database from 2001 through 2012. *JAMA Cardiol* 2016; 1: 1021–1029.

30. Davis KK, Lilienfeld DE and Doyle RL. Increased mortality in African Americans with idiopathic pulmonary arterial hypertension. *J Natl Med Assoc* 2008; 100: 69–72.

31. De-Miguel-Diez J, Lopez-de-Andres A, Hernandez-Barrera V, et al. Retrospective observational analysis of hospital discharge database to characterize primary pulmonary hypertension and its outcomes in Spain from 2004 to 2015. *Medicine (Baltimore)* 2019; 98: e15518.

32. Henriques-Forsythe M, Annangi S and Farber HW. Prevalence and hospital discharge status of human immunodeficiency virus-associated pulmonary arterial hypertension in the United States. *Pulm Circ* 2015; 5: 506–512.

33. Lilienfeld DE and Rubin LJ. Mortality from primary pulmonary hypertension in the United States, 1979–1996. *Chest* 2000; 117: 796–800.

34. Rush B, Biagioni BJ, Berger L, et al. Mechanical ventilation outcomes in patients with pulmonary hypertension in the United States: a national retrospective cohort analysis. *J Intensive Care Med* 2017; 32: 588–592.

35. Tripathi A, Black GB, Park YM, et al. Factors associated with the occurrence and treatment of supraventricular tachycardia in a pediatric congenital heart disease cohort. *Pediatr Cardiol* 2014; 35: 368–373.

36. 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.

37. Papani R, Duarte AG, Lin YL, et al. Pulmonary arterial hypertension associated with interferon therapy: a population-based study. *Multidiscip Respir Med* 2017; 12: 1.

38. Song S, Lee SE, Oh SK, et al. Demographics, treatment trends, and survival rate in incident pulmonary artery hypertension in Korea: a nationwide study based on the health insurance review and assessment service database. *PLoS One* 2018; 13: e0209148.

39. Studer S, Hull M, Pruett J, et al. Treatment patterns, healthcare resource utilization, and healthcare costs among patients with pulmonary arterial hypertension in a real-world US database. *Pulm Circ* 2019; 9: 2045894018816294.

40. Peacock AJ, Murphy NF, McMurray JJ, et al. An epidemiological study of pulmonary arterial hypertension. *Eur Respir J* 2007; 30: 104–109.

41. Wijeratne DT, Lajkosz K, Brogly SB, et al. Increasing incidence and prevalence of World Health Organization groups 1 to 4 pulmonary hypertension: a population-based cohort study in Ontario, Canada. *Circ Cardiovasc Qual Outcomes* 2018; 11: e003973.

42. Kirson NY, Birnbaum HG, Ivanova JI, et al. Prevalence of pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension in the United States. *Curr Med Res Opin* 2011; 27: 1763–1768.

43. Kirson NY, Birnbaum HG, Ivanova JI, et al. Excess costs associated with patients with chronic thromboembolic pulmonary hypertension in a US privately insured population. *Appl Health Econ Health Policy* 2011; 9: 377–387.