Prediction of persistent chronic cough in patients with chronic cough using machine learning

Wansu Chen[1], Michael Schatz[2,3], Yichen Zhou[1], Fagen Xie[1], Vishal Bali[4], Amar Das[4], Jonathan Schelfhout[4], Julie A. Stern[1] and Robert S. Zeiger[2,3]

[1]Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA, USA. [2]Department of Allergy, Kaiser Permanente Southern California, San Diego, CA, USA. [3]Department of Clinical Science, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, USA. [4]Center for Observational and Real-World Evidence (CORE), Merck & Co., Inc., Kenilworth, NJ, USA.

Corresponding author: Wansu Chen (wansu.chen@kp.org)

Shareable abstract (@ERSpublications)
Prediction of persistent chronic cough https://bit.ly/3V3vVzf

Cite this article as: Chen W, Schatz M, Zhou Y, *et al*. Prediction of persistent chronic cough in patients with chronic cough using machine learning. *ERJ Open Res* 2023; 9: 00471-2022 [DOI: 10.1183/23120541.00471-2022].

Received: 12 Sept 2022
Accepted: 11 Dec 2022

Abstract

Introduction The aim of this study was to develop and validate prediction models for risk of persistent chronic cough (PCC) in patients with chronic cough (CC). This was a retrospective cohort study.

Methods Two retrospective cohorts of patients 18–85 years of age were identified for years 2011–2016: a specialist cohort which included CC patients diagnosed by specialists, and an event cohort which comprised CC patients identified by at least three cough events. A cough event could be a cough diagnosis, dispensing of cough medication or any indication of cough in clinical notes. Model training and validation were conducted using two machine-learning approaches and 400+ features. Sensitivity analyses were also conducted. PCC was defined as a CC diagnosis or any two (specialist cohort) or three (event cohort) cough events in year 2 and again in year 3 after the index date.

Results 8581 and 52 010 patients met the eligibility criteria for the specialist and event cohorts (mean age 60.0 and 55.5 years), respectively. 38.2% and 12.4% of patients in the specialist and event cohorts, respectively, developed PCC. The utilisation-based models were mainly based on baseline healthcare utilisations associated with CC or respiratory diseases, while the diagnosis-based models incorporated traditional parameters including age, asthma, pulmonary fibrosis, obstructive pulmonary disease, gastro-oesophageal reflux, hypertension and bronchiectasis. All final models were parsimonious (five to seven predictors) and moderately accurate (area under the curve: 0.74–0.76 for utilisation-based models and 0.71 for diagnosis-based models).

Conclusions The application of our risk prediction models may be used to identify high-risk PCC patients at any stage of the clinical testing/evaluation to facilitate decision making.

Introduction

Chronic cough (CC) is defined as cough lasting >8 weeks [1–3]. With a prevalence of 1–13% [1, 3–9], CC is a common reason for both primary and specialist visits. Patients may suffer lower quality of life [10, 11] as cough may occur a hundred or thousand times daily and persist for years [12, 13]. This can cause significant physical, social and psychological consequences.

CC has been found to be associated with frequent comorbidities, narcotic use and healthcare resource utilisation [7, 14], and is most frequent in patients with both respiratory disease and gastro-oesophageal reflux disease (GERD) [14]. Emergency department visits (33.8%), hospitalisations (14.5%), ⩾2 different specialty department visits (19.4%), chest radiographs (41.9%), advanced chest imaging (15.6%), antitussives including codeine (43.4%), systemic respiratory antibiotics (62.8%), proton pump inhibitors (31.70%), antidepressants (27.5%) and neuromodulators (15.3%) have been found to be common among CC patients in the follow-up period [14].

Although approaches to evaluate and manage CC have been well described [15–19], a large group of patients have persistent chronic cough (PCC) due to the challenges of CC management. A 7-year follow-up of 42 patients with unexplained CC after extensive evaluation noted that the mean±SD duration of cough was 11.5±4.5 years at the time of final assessment, and just over half of the patients had either no change or worsening of cough after more than a decade [20]. Up to 40% of patients seen in a cough clinic were found to have unexplained CC [9]. In two previous electronic health record (EHR)-based studies conducted using Kaiser Permanente Southern California (KPSC) EHR data, 11.3% of CC patients had repeated CC within 1 year after the index visit [7]. Repeated CC occurred in 40.6% of CC patients cared for by specialists [14]. Understanding the most influential predictors of PCC and then stratifying CC patients based on risks of PCC can facilitate clinical decision making and adequate management of these patients. Here we use the term PCC (instead of chronic refractory cough), defined as having repeated evidence of CC (see Materials and methods section for details) in the 2nd year and again in the 3rd year after the initial evidence of CC, to indicate that the cough is refractory to conventional treatment of cough-associated conditions or traits.

The emergence of comprehensive EHR and machine learning offers an opportunity to facilitate management of CC patients. Deep learning and more traditional machine-learning models have been developed using both structured and unstructured (clinical notes) data to classify CC and non-CC patients with accuracy [21]. To date, we are not aware of any risk prediction models to predict the risk of PCC. There is a critical need for novel and accurate risk stratification tools for prediction of patients at increased risk of PCC. The risk prediction models developed in this study can facilitate identification of high-risk patients for PCC at any stage of the clinical testing/evaluation and can help in proper monitoring of these patients.

The aim of the present study was to develop and validate prediction models for risk of PCC within a large health system. More specifically, we sought to apply machine-learning techniques to high-dimensional clinical and healthcare utilisation data in EHR to predict the risk of PCC in patients whose CC was diagnosed by specialists and in a more generalised population in which CC was not necessarily diagnosed by specialists.

Materials and methods

Study design and setting

This retrospective cohort study was conducted utilising multi-ethnic health plan enrollees of KPSC. KPSC is an integrated healthcare system that provides comprehensive healthcare services for >4.8 million patients across 15 medical centres and ∼250 medical offices throughout the Southern California region. The race/ethnicity distribution, demographics and socioeconomic status of KPSC health plan enrollees are comparable to those of the residents in the Southern California region [22]. The study protocol was approved by the KPSC's institutional review board.

Study subjects

We identified two cohorts of patients 18–85 years of age using two distinct definitions of CC published previously [7, 23]. Specialist-defined CC patients (referred to as specialist cohort) had an KPSC internal encounter code of CC (529563) based on an outpatient visit to a specialist (pulmonologist, allergist, head and neck surgeon or gastroenterologist) between 1 January 2011 and 31 December 2016. For patients who had multiple encounters with CC diagnosis, the first one was selected as the index date. Event-defined CC patients (referred to as event cohort) had at least three cough events [7]. A cough event was defined as a cough diagnosis (Ninth Revision of International Classification of Diseases (ICD-9): 786.2 or Tenth Revision of International Classification of Diseases (ICD-10): R05), dispensing of cough medication or any indication of active cough in clinical notes [7]. The methods used to extract cough information from clinical notes were previously described [7] and can be accessed using the link https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7849260/table/T1/. The 3rd event of the first qualifying trio was defined as the index date. Patients without continuous health plan enrolment and pharmacy benefit or use of an angiotensin-converting enzyme inhibitor (ACE-I) on the index date or in the 12 months prior to or 3 years after the index date were excluded. Patients who disenrolled from the health plan or died in the 3 years after the index date were also excluded. The consort diagram for specialist and event cohorts can be found in figure 1.

Outcome identification

CC during follow-up was defined as having diagnosis of CC (internal code 529563) or the following:

1) specialist cohort: any two cough events that were 56–120 days apart
2) event cohort: any three cough events [7]

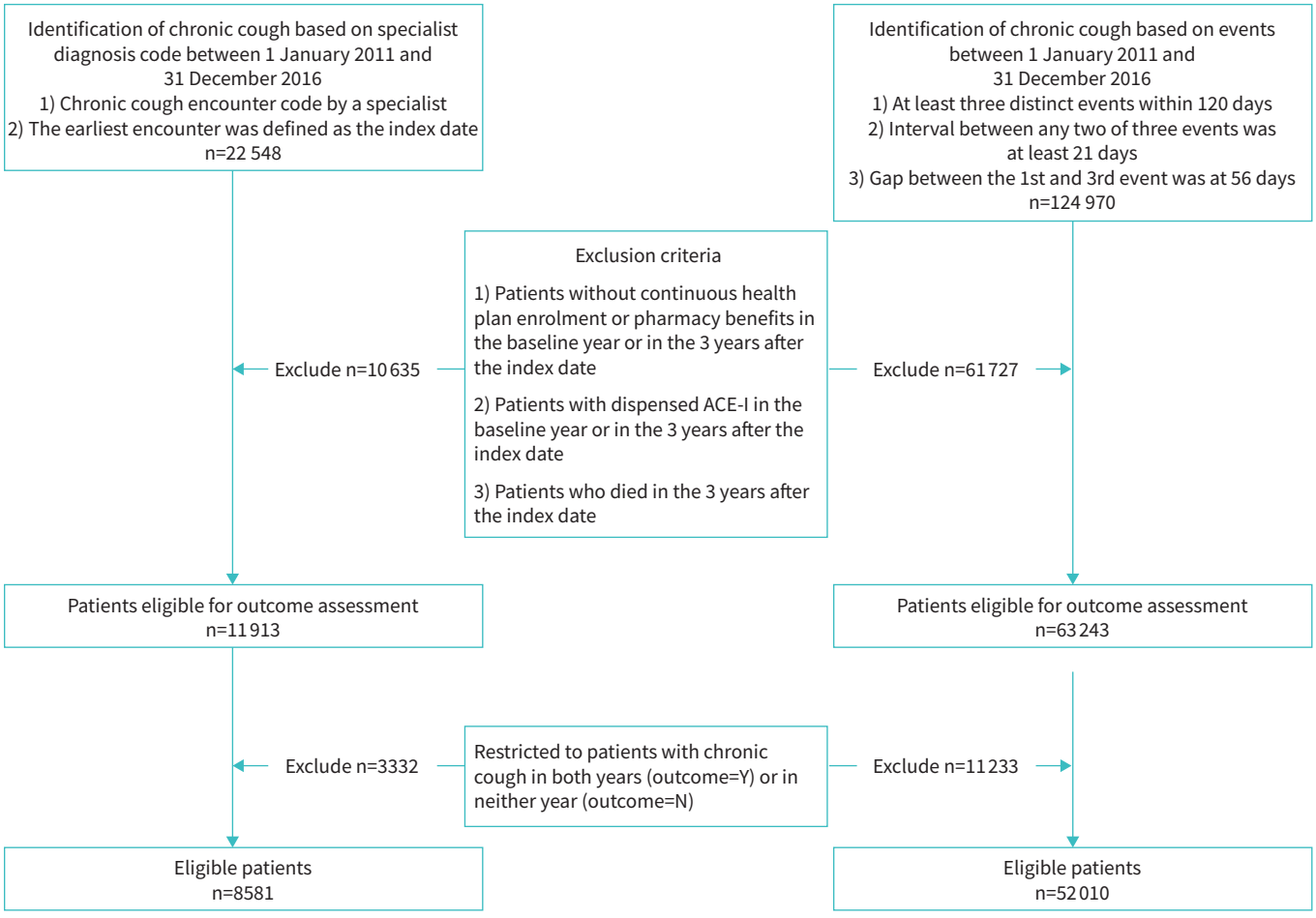

FIGURE 1 Consort diagram for specialist and event cohorts. Specialist cohort: any chronic cough diagnosis or any two cough events 56 days apart. Event cohort: any chronic cough diagnosis or any three cough events within 120 days, with the 1st and the last at least 56 days apart and any two of the three at least 21 days apart. ACE-I: angiotensin-converting enzyme inhibitor.

PCC was defined as having CC in both years 2 and 3 after the index date. In contrast, non-PCC was defined as having CC in neither year. Figure 2 illustrates the patient accrual and outcome identification windows for the specialist cohort (top) and the event cohort (bottom).

Patient demographic and clinical features at baseline

Patient demographics including behavioural characteristics (*e.g.*, smoking status), diagnosis-based comorbidities, laboratory tests, medication dispensing, medical procedures and healthcare utilisation on or in the 12 months prior to the index date were extracted (supplementary table S1). The ICD-9/ICD-10 codes used to define the comorbidities can be found in supplementary table S2. Medical conditions used to define respiratory-related diseases are listed in supplementary table S3. We also included the diagnosis groups defined by Rochester Epidemiology Program (https://www.rochesterproject.org/portal/). Missing values were imputed [24] if the frequency of missing was <60%. We used predictive mean matching method [25, 26] with k=5 for imputation. 10 imputed datasets were generated.

Model training, validation and testing

Data from all but one KPSC medical service area formed the training/validation dataset and the omitted KPSC medical service area served as a testing dataset. Using the 10 imputed training/validation datasets, we first applied gradient boosting model (GBM) implemented in "LightGBM" [27] to determine the relative importance (measured by mean information gain) of all the potential features. Random Forest (RF) [28] with five-fold cross validation was then applied to the top 30 important features. Age was forced into the model. The 30 features were added one at a time. Each time, the feature that yielded the maximum improvement of area under the curve (AUC) was selected. This iterative process continued until AUC

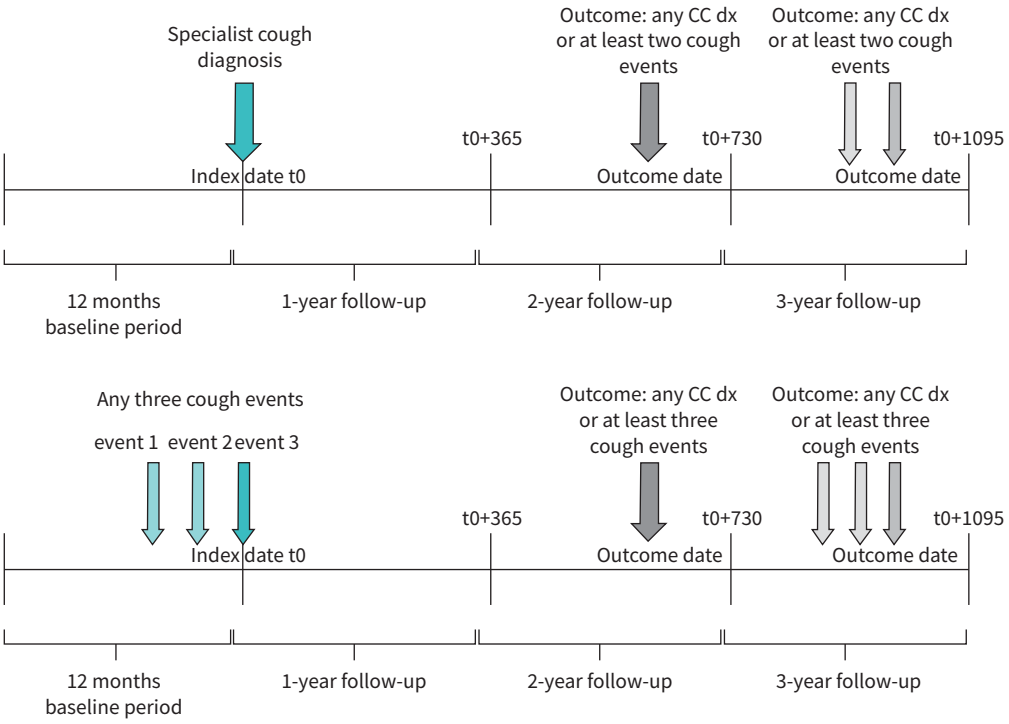

FIGURE 2 Cohort identification and outcome definition. Top: specialist cohort; bottom: event cohort. t0: date of 1st qualifying visit in the accrual period 2011–2016; dx: diagnosis; CC: chronic cough; cough events: a cough event was defined as a cough diagnosis (ICD-9: 786.2 or ICD-10: R05), dispensing of cough medication or any indication of cough in clinical notes; ICD: International Classification of Diseases.

increased <0.004. The hyperparameters were tuned for each model and the technical details can be found in the supplementary eMethods.

The final models were applied to the testing datasets. The discriminative power was evaluated by AUC. For each cohort, calibration was assessed by calibration plots. Patients were grouped by cut-offs at 20th, 40th, 60th and 80th percentiles of the predicted probability (five-group definition), as well as by specific risk thresholds (six-group definition; specialist cohort: 0.2, 0.3, 0.4, 0.5, 0.6; event cohort: 0.1, 0.2, 0.3, 0.4, 0.5). Each point estimate on a calibration plot reflects both predicted risk and the observed risk for a specific risk group. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F-1 score, the harmonic mean of sensitivity and PPV, were also estimated.

Utilisation-based model versus diagnosis-based model

For each of the cohorts, two models with different input features were developed and validated. The utilisation-based model was supplied with all the available patient characteristics listed in supplementary table S1, while the diagnosis-based model was developed without medication and healthcare utilisation-related variables.

Sensitivity analysis

Alternately to RF, "LightGBM" was also applied to develop and validate risk prediction models based on the top 30 important features described in the Model training, validation and testing section. The model training and validation process was the same as described above.

Statistical analysis

All descriptive analyses were performed using SAS (Version 9.4 for Unix; SAS Institute, Cary, NC, USA). Model development and validation was conducted using Python (Version 3.7.9; Python Software Foundation, Fredericksburg, VA, USA) for both GBM (LightGBM, Microsoft Research, Redmond, WA, USA) package version 3.2.1) and RF (RandomForestClassifier, Scikit-Learn library, Version 0.24.1 [29]). The calibration plots were also produced in Python.

Results

Characteristics of the study cohorts

8581 and 52 010 patients met the eligibility criteria for the specialist and event cohorts, respectively (figure 1). In the specialist cohort, 66.8% of patients were females, 50.6% were white people and 25.3% were Hispanic (table 1). On average, patients in the specialist cohort were 60.0 years of age, with mean membership length of 10.7 years. 38.2% of the patients were obese and an additional 34.5% were overweight. Gastro-oesophageal reflux, asthma, and allergic or chronic rhinitis were frequent (>30%).

Compared to patients in the specialist cohort, patients in the event cohorts seemed to be 5 years younger on average, and the percentage of Hispanic patients and current smokers was higher (table 1). GERD, asthma, allergic or chronic rhinitis, COPD, and post-nasal drip appeared to be less common. Both cohorts had extremely high healthcare utilisation in the baseline year (table 1).

Frequency of PCC

3279 (38.2%) and 6460 (12.4%) patients in the specialist and event cohorts, respectively, developed PCC. 4927 (57.4% of patients in the specialist cohort) also appeared (overlapped) in the event cohort, of which 1948 (39.5% of 4972 patients) developed PCC. This indicates that the risk of PCC in the specialist cohort is high, and the risk is not impacted by patient's qualification for the event cohort.

Model training, validation and testing

The sizes of the training/testing datasets were 7454/1127 and 43 642/8363, respectively, for the specialist cohort and the event cohort. The 10 imputation datasets used to pre-select the top 30 most important features yielded the same list of 30 features (see the 30 features in supplementary table S1), and none of the 30 features contained missing values. Therefore, the original dataset (unimputed) was used for algorithm training, validation and testing.

For the specialist cohort, the final utilisation-based model contained age, number of clinic encounters with a respiratory diagnosis, narcotics or codeine medication dispensing (y/n), number of clinic encounters with a CC diagnosis and number of clinic encounters with a pulmonologist in the 12 months prior to the index date (table 2). For the event cohort, the final utilisation-based model covered the same features except that 1) number of non-urgent clinic encounters instead of number of clinic encounters with a CC diagnosis was selected, and 2) number of antitussive codeine medication dispensing was added. The AUC in the testing dataset reached 0.739 and 0.758, respectively, for the utilisation-based models of the specialist and event cohorts.

Features being chosen for the final diagnosis-based models were the same for the two cohorts (table 2). They included age and indicators of the following comorbid conditions: asthma, pulmonary fibrosis, COPD diagnosis, GERD, hypertension, bronchiectasis and depression. The AUCs were 0.711 and 0.706, respectively, when the algorithms were validated based on the testing dataset.

The calibration plots based on the five groups of equal group size are displayed in figure 3. It appears that the utilisation-based model for the event cohort fits the data well, while the other three models slightly under- or over-estimated the risk of PCC in some risk groups. The calibration plots based on groups defined by risk thresholds demonstrated similar patterns (supplementary figure S1).

Sensitivity, specificity, PPV, NPV and F-1 score at five risk thresholds (0.2 to 0.6 for the specialist cohort, and 0.1 to 0.5 for the event cohort) are displayed in table 3. As expected, sensitivity/NPV measures decreased while PPV/specificity measures increased with the increase of risk threshold (table 3). Taking the utilisation-based model for the specialist cohort as an example, patients with at least 30% predicted risk of PCC constituted 82.0% of the total PCC cases (sensitivity). Meanwhile, in patients with predicted risk of PCC of at least 30%, 50.6% truly developed PCC (PPV). When the risk threshold increased to 60%, the sensitivity dropped to 23.1%, while PPV climbed to 70.9%. Supplementary figure S2 shows sensitivity, PPV and F-1 score curves. It appears that F-1 scores are maximised in the window of 20–40% and 10–30% for the models developed based on the specialist and the event cohorts, respectively.

Sensitivity analysis

The GBM-based models selected the same or similar features (supplementary table S4). AUC measures between the GBM and the RF models were almost identical.

TABLE 1 Characteristics of study subjects at baseline by study cohort

Patient characteristics	Specialist cohort	Event cohort
Subjects n	8581	52 010
Demographics and lifestyle variables		
Age years, mean±SD	60.0±14.3	55.5±16.2
Female sex	5730 (66.8)	35 855 (68.9)
Ethnicity		
White	4339 (50.6)	22 460 (43.2)
Black	790 (9.2)	5826 (11.2)
Hispanic	2167 (25.3)	16 656 (32.0)
Asian/Pacific Islanders	1093 (12.7)	5810 (11.2)
Others/multiple/unknown	192 (2.2)	1258 (2.4)
Family education ⩽grade 12 (%) geocoded	15.9 (13.1)	18.3 (14.0)
Years of health plan enrolment, mean±SD	21.0±10.7	19.7±10.4
Medical insurance (mutually inclusive)		
Commercial	4730 (55.1)	32 095 (61.7)
Medi-CAL[#]	260 (3.0)	2932 (5.6)
Medicare	3391 (39.5)	16 479 (31.7)
Private pay	2319 (27.0)	11 566 (22.2)
Smoking status (based on the last measure prior to the index date)		
Current	203 (2.4)	3041 (5.8)
Passive	62 (0.7)	398 (0.8)
Quit	2659 (31.0)	15 412 (29.6)
Never	5652 (65.9)	33 057 (63.6)
Unknown	5 (0.1)	102 (0.2)
BMI[¶], mean±SD	29.1±6.3	29.7±7.0
BMI category[¶]		
Underweight (<18.5 $kg \cdot m^{-2}$)	124 (1.4)	798 (1.5)
Normal weight (18.5–24.9 $kg \cdot m^{-2}$)	2211 (25.8)	12 698 (24.4)
Overweight (25–29.9 $kg \cdot m^{-2}$)	2957 (34.5)	16 751 (32.2)
Obese (⩾30 $kg \cdot m^{-2}$)	3279 (38.2)	21 622 (41.6)
Unknown (BMI missing)	10 (0.1)	141 (0.3)
Comorbidities		
Charlson Comorbidity Index (weighted), mean±SD	1.5±1.7	1.4±1.7
Gastro-oesophageal reflux	3631 (42.3)	14 670 (28.2)
Asthma	2688 (31.3)	15 391 (29.6)
COPD clinical diagnosis	1028 (12.0)	5623 (10.8)
Allergic rhinitis	2904 (33.8)	13 814 (26.6)
Chronic rhinitis	2696 (31.4)	8129 (15.6)
Post-nasal drip (upper airway cough syndrome)	1715 (20.0)	5526 (10.6)
Chronic sinusitis	2137 (24.9)	15 056 (28.9)
Obstructive sleep apnoea	646 (7.5)	3172 (6.1)
Pneumonia and influenza and other acute lower respiratory functions	1599 (18.6)	13 190 (25.4)
Anxiety	1418 (16.5)	10 211 (19.6)
Depression	1642 (19.1)	10 876 (20.9)
Potential cough complications		
Costochondritis	152 (1.8)	1255 (2.4)
Subconjunctival haemorrhage	94 (1.1)	554 (1.1)
Stress incontinence	492 (5.7)	3515 (6.8)
Sleep disturbance	805 (9.4)	5009 (9.6)
Vomiting	71 (0.8)	543 (1.0)
Rib fracture	41 (0.5)	228 (0.4)
Any complication	1513 (17.6)	10 112 (19.4)
Healthcare utilisation		
Any event ⩾1 ED visit	2367 (27.6)	18 108 (34.8)
Any event ⩾1 hospitalisation	757 (8.8)	6541 (12.6)
ED visit with a respiratory dx[+]	992 (11.6)	7786 (15.0)
Hospital admission with a respiratory dx[+]	424 (4.9)	3743 (7.2)
Clinic encounters with a respiratory dx[+], mean±SD	2.95±3.05	2.61±2.89
Clinic encounters with a chronic cough dx, mean±SD	0.485±0.98	0.169±0.64
Non-urgent care clinic encounters with a chronic cough dx, mean±SD	0.468±0.96	0.164±0.63
Urgent care clinic encounters with a chronic cough dx, mean±SD	0.017±0.14	0.005±0.08

Continued

TABLE 1 Continued

Patient characteristics	Specialist cohort	Event cohort
Specialist visits		
Pulmonologist	5566 (64.9)	9317 (17.9)
Allergist	3017 (35.2)	6218 (12.0)
Head and neck surgery	2101 (24.5)	7943 (15.3)
Gastroenterology	1468 (17.1)	7506 (14.4)
Urology	666 (7.8)	3808 (7.3)
Outpatient pharmacy dispensing (all oral)		
Narcotics, no codeine	2827 (32.9)	18 770 (36.1)
Narcotics, including codeine	5251 (61.2)	37 334 (71.8)
Codeine	3947 (46.0)	29 989 (57.7)
Antitussives, including codeine	4993 (58.2)	36 100 (69.4)
Antitussives, no codeine	2776 (32.4)	16 679 (32.1)

Data are presented as n (%) unless otherwise stated. BMI: body mass index; ED: emergency department; dx: diagnosis. [#]: Medi-CAL or other state programmes; [¶]: last measure prior to the index date; [+]: refer to supplementary table S3 for the definition of respiratory dx. Chronic cough dx was not included in the list of respiratory dx.

Model application/implementation

To facilitate external application of the two diagnosis-based RF models, we plan to develop a publicly available web-based tool. As an example, entering data into the models for the specialist cohort and the event cohort for a hypothetical 70-year-old patient with asthma, COPD and hypertension yielded an estimated risk of PCC at 64.0% and 23.0%, respectively. As a demonstration, decision rules based on one of the trees built for the diagnosis-based RF model are displayed in supplementary figure S3 (a: left side; b: right side).

TABLE 2 Baseline predictors and performance based on training and testing datasets for both utilisation-based and diagnosis-based models

Baseline predictors	Specialist cohort	Event cohort
Utilisation-based model		
Age	X	X
Clinic encounters with a respiratory dx[#]	X	X
Narcotics or codeine medication dispensing	X	X
Clinic encounters with a chronic cough dx	X	
Non-urgent clinic encounters with a chronic cough dx		X
Clinic encounters with a pulmonologist	X	X
Antitussive codeine medication dispensing		X
AUC training dataset	0.740	0.770
AUC testing dataset	0.739	0.758
Diagnosis-based model		
Age	X	X
Asthma	X	X
Pulmonary fibrosis	X	X
COPD clinical diagnosis	X	X
GERD	X	X
Hypertension	X	X
Bronchiectasis	X	X
AUC training dataset	0.701	0.713
AUC testing dataset	0.711	0.706

AUC: area under the curve; GERD: gastro-oesophageal reflux disease. [#]: refer to supplementary table S3 for the definition of respiratory dx. Chronic cough dx was not included in the list of respiratory dx.

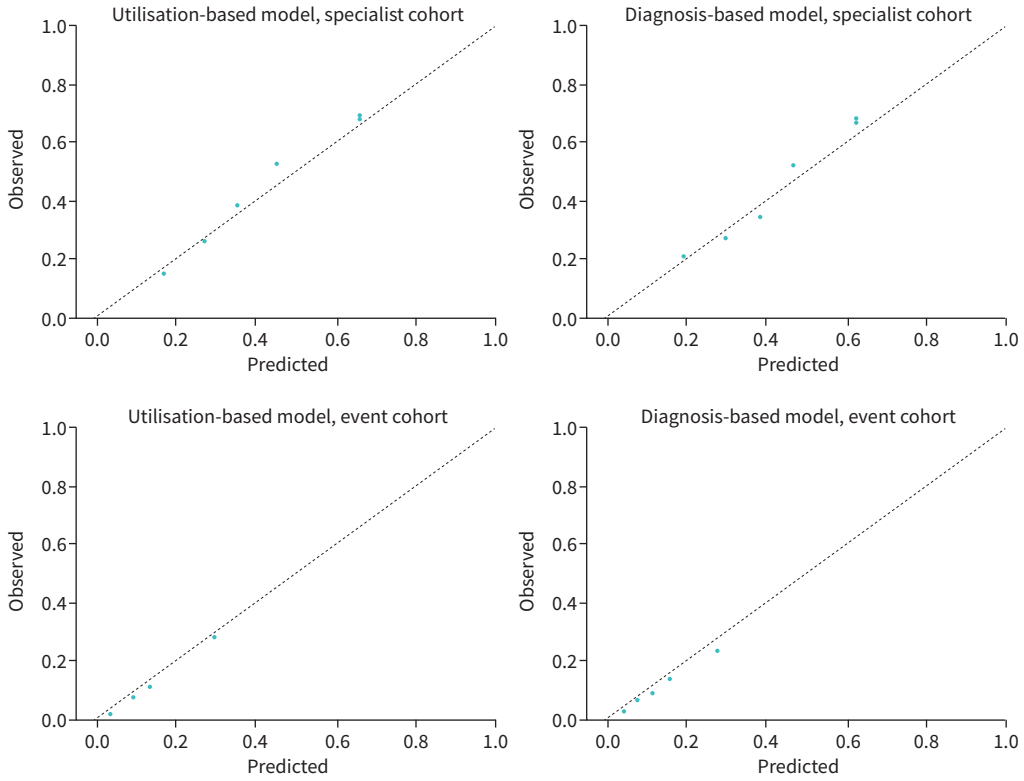

FIGURE 3 Calibration plot based on groups defined by percentiles. For both specialist and event cohorts, patients were separated into five groups based on 20th, 40th, 60th and 80th quantities.

Discussion

We applied machine-learning methods to derive and validate clinical prediction models for PCC within a large integrated healthcare system. Despite the inclusion of over 400 potential features in the candidate pool, the utilisation-based models were mainly based on baseline healthcare utilisations associated with CC or respiratory diseases, while the diagnosis-based models incorporated traditional parameters including age, asthma, pulmonary fibrosis, COPD, GERD, hypertension and bronchiectasis. Final models were all parsimonious and moderately accurate in the internal validation.

The two types of models (utilisation-based *versus* diagnosis-based) could be used in different scenarios. Large healthcare systems can implement the utilisation-based risk models within their EHR to automatically calculate the risk of PCC for care providers. However, individual physicians who work in small clinics can benefit from a web-based tool generated from the diagnosis-based models to estimate risk of PCC based on physician's and patient's input at the time of clinical care. When we developed the utilisation-based models, all the diagnosis codes were in the feature candidate pool. However, none of the diagnosis codes was selected in the final models. Clinicians may consider the specialist model for patients who see specialists or those who should have been referred to specialists. Although there is no restriction on patient referral, KPSC has practice guidelines on patient referral and encourages primary care physicians to undertake the initial workups.

The selection of risk threshold should be based on the type of intervention being implemented. For example, if a healthcare organisation or a provider considers a very expensive treatment, the risk threshold may be set high (*e.g.*,50%). However, if the intervention being considered is less expensive, such as a follow-up visit in 6 months, a lower risk threshold may be considered. The F-1 score suggests a balance between sensitivity and PPV; however, it may not provide the most sensible threshold for a given clinical situation.

The two cohorts being studied are quite different. The differences in patient demographics and other clinical characteristics were similar to what was reported in a previous study [7]. As expected, patients in the specialist cohort were older and exhibited a more severe phenotype than patients in the event cohort.

https://doi.org/10.1183/23120541.00471-2022

TABLE 3 Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F-1 score based on selected risk thresholds for specialist and event cohorts, and for utilisation-based and diagnosis-based models

	Utilisation-based model, specialist cohort, risk threshold					Diagnosis-based model, specialist cohort, risk threshold					Utilisation-based model, event cohort, risk threshold					Diagnosis-based model, event cohort, risk threshold				
	0.2	0.3	0.4	0.5	0.6	0.2	0.3	0.4	0.5	0.6	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
n#	962	709	454	251	148	1016	768	517	300	108	3438	1397	570	284	101	4926	1427	555	122	29
Sensitivity %	0.960	0.820	0.607	0.371	0.231	0.969	0.820	0.646	0.437	0.163	0.729	0.460	0.267	0.158	0.065	0.812	0.371	0.198	0.056	0.014
Specificity %	0.219	0.500	0.735	0.878	0.936	0.144	0.412	0.668	0.850	0.949	0.628	0.869	0.956	0.981	0.994	0.439	0.854	0.950	0.990	0.998
PPV %	0.454	0.526	0.608	0.673	0.709	0.434	0.486	0.569	0.663	0.685	0.194	0.301	0.428	0.507	0.584	0.150	0.238	0.326	0.418	0.448
NPV %	0.891	0.804	0.734	0.674	0.642	0.874	0.772	0.736	0.690	0.626	0.950	0.929	0.914	0.905	0.897	0.950	0.917	0.906	0.895	0.892
F-1 score	0.617	0.641	0.607	0.479	0.348	0.600	0.610	0.605	0.527	0.263	0.306	0.364	0.329	0.241	0.116	0.254	0.290	0.247	0.099	0.028

Assessment was based on testing dataset only. #: number of eligible patients whose risk was above each risk threshold.

Healthcare systems might differ in terms of referrals for diagnosed or possible CC. Therefore, application of our cohort definition to any external organisations may identify a group of patients with different demographics and clinical characteristics.

The risk of PCC is high in CC patients, especially in patients diagnosed with CC by specialists. In a survey conducted in CC patients seen by specialists, patients reported an average of 9 years of CC history and a significant burden in terms of healthcare utilisation [23]. In two previous EHR-based studies in which CC patients were defined in similar approaches as they were in the current study, 40.6% and 11.3% had repeated CC within 1 year after the index visit, respectively, for patients diagnosed by specialists and patients defined by CC events [7, 14]. In the current study, the corresponding percentages were 38.2% for patients in the specialist cohort and 12.4% for patients in the event cohort in both years 2 and 3, demonstrating the persistent nature of CC, which requires careful consideration and management.

The disease burden in patients with PCC compared with those without PCC was previously studied [30]. Comorbidities, potential cough complications (particularly stress incontinence and sleep disturbances), antitussive medication use and healthcare utilisations were more frequent in patients with PPC [30]. Many risk factors were reported to be associated with PCC [30]; however, most of them were not selected into the final risk prediction models in the current study as the most influential predictors.

CC did not have a specific ICD code until recently. Research studies examining prevalence or burden of care of CC have previously relied on collection of repeated evidence of cough using natural language processing of clinical notes, repeated encounter cough diagnosis codes and medication prescriptions/dispensing records [4, 7] or an internal diagnosis code of CC specific to a healthcare organisation [7]. Starting from year 2022, the ICD-10 billing code for cough code (R05) is replaced by six new and more specific cough codes including R05.3 (CC) and R05.9 (unspecified cough) [31]. R05.3 is applicable to persistent cough, refractory cough and unexplained cough. The accuracy of the new ICD-10 CC codes needs to be validated against other validated CC identification methods before they are applied to future research studies. Our previous research suggested that the majority of patients meeting the definition of CC are not diagnosed with the internal encounter diagnosis code, although the specific CC code has been available to use [7]. Education should be provided to physicians for proper use of the new CC code (R05.3) and the unspecified cough code (R05.9).

There are several strengths to the current study including a comprehensive, data-driven approach to model development, use of high-dimensional data elements, development of diagnosis-based models in addition to utilisation-based models for ease of user implementation, and verification of results by adding another machine-learning approach for model development. This study has several limitations. First, information used in this study is entirely electronically collected from EHR, and therefore, the quality depends on the accuracy of physician coding, which may vary depending on physician's expertise in coding. KPSC offers coding courses at least annually for physicians to refresh their coding skills. Second, important features such as duration, severity and triggers of cough were not included. CC is a patient-reported condition, and collecting these self-reported characteristics *via* a survey could improve the model accuracy. Third, we included indicators (y/n) for lung function test, blood eosinophil test and methacholine challenge test during the study period; however, the results of these tests were not included due to the high percentage of patients without each test. For chest radiograph, the results are not easily obtainable unless a chart review or natural language process is performed, and thus, it was included as only an indicator of the test being performed without the actual test results. Fourth, no external validation of the developed models in another healthcare organisation was performed. Transportability of a prediction model is an important aspect when the utility of the model is assessed. We encourage others to test our models using various data sources. Fifth, some cough medication can be purchased over the counter outside of KPSC without prescription and thus are not included in our pharmacy database. Finally, given the chronic nature of PCC, a longer observation period (*e.g.*, 5+ years instead of 3 years) might reveal different insights about the risk of PCC.

Conclusions

We applied machine-learning methods to derive and validate prediction models for PCC within a large integrated healthcare system. The application of risk prediction models based on healthcare utilisation or clinical parameters can facilitate identification of high-risk patients for PCC at any stage of the clinical testing/evaluation and suggest more frequent monitoring of these patients due to high risk of persistent CC. Users are encouraged to further investigate, validate or recalibrate our models in different healthcare systems and databases. Should findings from further studies confirm or improve the accuracy of the proposed models, this could provide a framework for a systematic approach to target patients with high-risk of PCC.

Provenance: Submitted article, peer reviewed.

Acknowledgement: The authors thank Sole Cardoso (Kaiser Permanente Southern California (KPSC), Pasadena, CA, USA) for the assistance with formatting the manuscript and Botao Zhou (KPSC, Pasadena, CA, USA) for the additional analyses.

Support statement: Merck Sharpe & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, funded a research grant to the Southern California Permanente Medical Group (SCPMG) Research and Evaluation Department to perform the study. SCPMG investigators developed the protocol, performed the analyses, and wrote the manuscript. The sponsor participated in the study discussions and provided comments to the protocol, data analysis, and manuscript. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: All authors declare no conflict of interest.

References

1 Pratter MR. Overview of common causes of chronic cough: ACCP evidence-based clinical practice guidelines. *Chest* 2006; 129: 59S-62S.
2 McGarvey L, Gibson PG. What is chronic cough? Terminology. *J Allergy Clin Immunol Pract* 2019; 7: 1711–1714.
3 Smith JA, Woodcock A. Chronic cough. *N Engl J Med* 2016; 375: 1544–1551.
4 Weiner M, Dexter PR, Heithoff K, *et al.* Identifying and characterizing a chronic cough cohort through electronic health records. *Chest* 2021; 159: 2346–2355.
5 Ford AC, Forman D, Moayyedi P, *et al.* Cough in the community: a cross sectional survey and the relationship to gastrointestinal symptoms. *Thorax* 2006; 61: 975–979.
6 Koo HK, Jeong I, Lee SW, *et al.* Prevalence of chronic cough and possible causes in the general population based on the Korean National Health and Nutrition Examination Survey. *Medicine (Baltimore)* 2016; 95: e4595.
7 Zeiger RS, Xie F, Schatz M, *et al.* Prevalence and characteristics of chronic cough in adults identified by administrative data. *Perm J* 2020; 24: 1–3.
8 Meltzer EO, Zeiger RS, Dicpinigaitis P, *et al.* Prevalence and burden of chronic cough in the United States. *J Allergy Clin Immunol Pract* 2021; 9: 4037–4044.e2.
9 Chung KF, Pavord ID. Prevalence, pathogenesis, and causes of chronic cough. *Lancet* 2008; 371: 1364–1374.
10 French CT, Fletcher KE, Irwin RS. Gender differences in health-related quality of life in patients complaining of chronic cough. *Chest* 2004; 125: 482–488.
11 Chamberlain SA, Garrod R, Douiri A, *et al.* The impact of chronic cough: a cross-sectional European survey. *Lung* 2015; 193: 401–408.
12 Kelsall A, Decalmer S, McGuinness K, *et al.* Sex differences and predictors of objective cough frequency in chronic cough. *Thorax* 2009; 64: 393–398.
13 Sunger K, Powley W, Kelsall A, *et al.* Objective measurement of cough in otherwise healthy volunteers with acute cough. *Eur Respir J* 2013; 41: 277–284.
14 Zeiger RS, Schatz M, Butler RK, *et al.* Burden of specialist-diagnosed chronic cough in adults. *J Allergy Clin Immunol Pract* 2020; 8: 1645–1657.e7.
15 Irwin RS, Baumann MH, Bolser DC, *et al.* Diagnosis and management of cough executive summary: ACCP evidence-based clinical practice guidelines. *Chest* 2006; 129: 1S-23S.
16 Morice AH, McGarvey L, Pavord I, British Thoracic Society Cough Guideline Group. Recommendations for the management of cough in adults. *Thorax* 2006; 61: Suppl 1, i1–24.
17 Morice AH, Fontana GA, Sovijarvi AR, *et al.* The diagnosis and management of chronic cough. *Eur Respir J* 2004; 24: 481–492.
18 Gibson PG, Chang AB, Glasgow NJ, *et al.* CICADA: Cough in Children and Adults: diagnosis and assessment. Australian cough guidelines summary statement. *Med J Aust* 2010; 192: 265–271.
19 Morice AH, Millqvist E, Bieksiene K, *et al.* ERS guidelines on the diagnosis and treatment of chronic cough in adults and children. *Eur Respir J* 2020; 55: 1901136.
20 Yousaf N, Montinero W, Birring SS, *et al.* The long term outcome of patients with unexplained chronic cough. *Respir Med* 2013; 107: 408–412.
21 Xiao L, Gandhi P, Zhang P, *et al.* Applying interpretable deep learning models to identify chronic cough patients using EHR data. *Comput Methods Programs Biomed* 2021; 10: 106395.
22 Koebnick C, Langer-Gould AM, Gould MK, *et al.* Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data. *Perm J* 2012; 16: 37–41.
23 Zeiger RS, Schatz M, Hong B, *et al.* Patient-reported burden of chronic cough in a managed care organization. *J Allergy Clin Immunol Pract* 2021; 9: 1624–1637.e10.
24 Wright M, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* 2017; 77: 1–17.

25 Ishwaran H, Kogalur U, Blackston E, *et al.* Random survival forests. *Ann Appl Stat* 2008; 3: 841–860.
26 Little R. Missing-data adjustments in large surveys. *J Bus Econ Stat* 1988; 6: 287–296.
27 Ke G, Meng Q, Finley T, *et al.* Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017; 30: 3146–3154.
28 Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
29 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–2830.
30 Zeiger RS, Schatz M, Zhou Y, *et al.* Risk factors for persistent chronic cough during consecutive years: a retrospective database analysis. *J Allergy Clin Immunol Pract* 2022; 10: 1587–1597.
31 Moore K. New diagnosis codes effective Oct. 1. Here are some family physicians should know. https://www.aafp.org/pubs/fpm/blogs/gettingpaid/entry/new_diagnosis_codes.html Date last updated: 1 August 2021. Date last accessed: 30 January 2023.