Machine Learning for Early Detection of Hidradenitis Suppurativa: A Feasibility Study Using Medical Insurance Claims Data

Waqar Ali[1], Jonathan Williams[2], Betty Xiong[3], James Zou[3] and Roxana Daneshjou[3,4]

JID Innovations (2025);**5:**100362 doi:10.1016/j.xjidi.2025.100362

Patients with hidradenitis suppurativa (HS) are often misdiagnosed and may wait up to 10 years to receive a diagnosis of HS. This study aimed to predict HS diagnosis prior to actual diagnosis on the basis of previous medical history using models developed with insurance claims data. Three machine learning models were compared with a model using features selected by a dermatologist (clinical baseline model). The study analyzed 5,900,000 United States individuals' insurance records over 13.5 years. The population included 13,886 patients with HS with at least 1 claim in each of the 2 years prior to their first HS diagnosis and 69,428 control patients with no HS diagnosis. The models aimed to classify HS diagnosis status on the basis of clinical features observed over 2 years. Model performance was assessed by area under the receiver operating characterisitic curve, F1-score, and precision and recall rates. The machine learning models (logistic regression, random forest, and XGBoost) showed a higher area under the receiver operating characterisitic curve than the clinical baseline model (logistic regression = 0.75, random forest = 0.79, XGBoost = 0.80, clinical = 0.71). In the clinical model and the best-performing XGBoost model, the top features associated with diagnosis were patient age at prediction and sex. The XGBoost model top features also included the use of sulfamethoxazole/trimethoprim and clindamycin phosphate and obesity.

Keywords: Hidradenitis suppurativa, Machine learning, Medical claims

INTRODUCTION

Hidradenitis suppurativa (HS) is a chronic, inflammatory disease that presents with painful and draining cysts, abscesses, and nodules in intertriginous areas such as the axilla and groin (Nguyen et al, 2021). HS is estimated to have a prevalence of approximately 1% in most countries where studies have been conducted (Sabat et al, 2020).

Because the individual symptoms of cysts and abscesses, particularly in the early stages of the disease, mimic those seen in more common conditions, lesions are often misdiagnosed as boils or furunculosis, and patients may wait up to 10 years to receive an actual diagnosis of HS (Kokolakis et al, 2020; Lee et al, 2017; Nguyen et al, 2021). Other causes of diagnostic delay include the stigmatizing nature of the disease, lack of awareness among patients and healthcare providers, and lack of access to a dermatologist (Snyder et al, 2023).

As HS progresses, patients can develop permanent scarring and sinus tracts; patients who have a delayed diagnosis often present with more severe disease (Kokolakis et al, 2020). Because treatments for HS can slow the progression of disease, early diagnosis and intervention are essential for improving patient QOL and outcomes (Snyder et al, 2023).

Patients with HS often have other comorbidities such as type 2 diabetes mellitus, hypertension, and dyslipidemia (Scala et al, 2021). Patients with HS utilize healthcare resources more often than those without (Strunk et al, 2017), and these interactions may serve as an opportunity for earlier identification and intervention. Algorithms to identify HS from medical codes in observational databases, including medical claims and electronic health records, have previously been developed, validated, and compared in terms of sensitivity and specificity (Hardin et al, 2022; Kim et al, 2014), in addition to detailed descriptive characterizations of the HS population and disease burden using real-world data sources (Marvel et al, 2019). There has also been promising recent research on the use of artificial intelligence approaches for the quantification of severity of HS focusing on the imaging modality (Hernández Montilla et al, 2023). In contrast, the question of earlier detection of the disease using historical patient-level data has been relatively less explored than in other disease areas.

Decision support tools aim to help healthcare workers identify subsets of patients that may benefit from a particular

[1]UCB Pharma, Slough, United Kingdom; [2]UCB Pharma, Brussels, Belgium; [3]Department of Biomedical Data Science, Stanford University, Stanford, California, USA; and [4]Department of Dermatology, Stanford University, Stanford, California, USA

Correspondence: Waqar Ali, UCB Pharma, 216 Bath Road, Slough SL1 3WE, United Kingdom. E-mail: waqar.ali@ucb.com

Abbreviations: AUC, area under the curve; HS, hidradenitis suppurativa; ICD, International Classification of Diseases; LR, logistic regression; RF, random forest; XGB, XGBoost

Received 3 July 2024; revised 5 February 2025; accepted 25 February 2025; accepted manuscript published online XXX; corrected proof published online XXX

Cite this article as: JID Innovations 2025;5:100362

intervention. Such an intervention may be an earlier referral to a dermatologist for patients who have not yet been diagnosed with HS but have a risk profile that suggests a possible diagnosis (Garg et al, 2021).

The aim of this research was to explore whether there are signals in a patient's past medical history that may aid an earlier HS diagnosis prior to actual diagnosis using a set of models developed with insurance claims data from the United States.

RESULTS

Study population

From the PatientSource database random sample, equating to 5,900,000 individuals, we identified 18,946 patients with a HS diagnosis code in their longitudinal record, giving an overall prevalence of 0.3%, which lies within statistics from published literature (Jfri et al, 2021). Of these, 13,886 had at least 1 claim in the 2 years prior to the index date (Figure 1), with a lookback period ≥24 months, and were therefore included in the case population. The control population included 69,428 patients with no HS diagnosis code who had a lookback period ≥24 months and at least 1 claim (diagnosis, procedure, surgery, or prescription) in this period. Random sampling was used to split the total patient pool, consisting of case and control patients, into the training (n = 53,362), validation (n = 16,611), and test (n = 13,341) cohorts (Figure 1).

The training cohort used for the modeling exercise (Figure 2) included 8929 patients with at least 1 HS diagnosis, of whom 77% were female. In this population, the average age was 35.6 years for females and 42.4 years for males (Table 1). The age and sex distributions of patients in the validation and test cohorts were comparable with those in the training cohort (Table 1).

Clinical baseline model

The clinical baseline model (Figure 3a) was developed using a set of clinician-selected features. Clinical Classifications Software Refined (Agency for Healthcare Research and Quality, 2017) was used to help refine the list of features and remove features that were not deemed relevant, leading to a total of 113 clinical features. The top 15 features identified by Shapley value analysis in the trained clinical

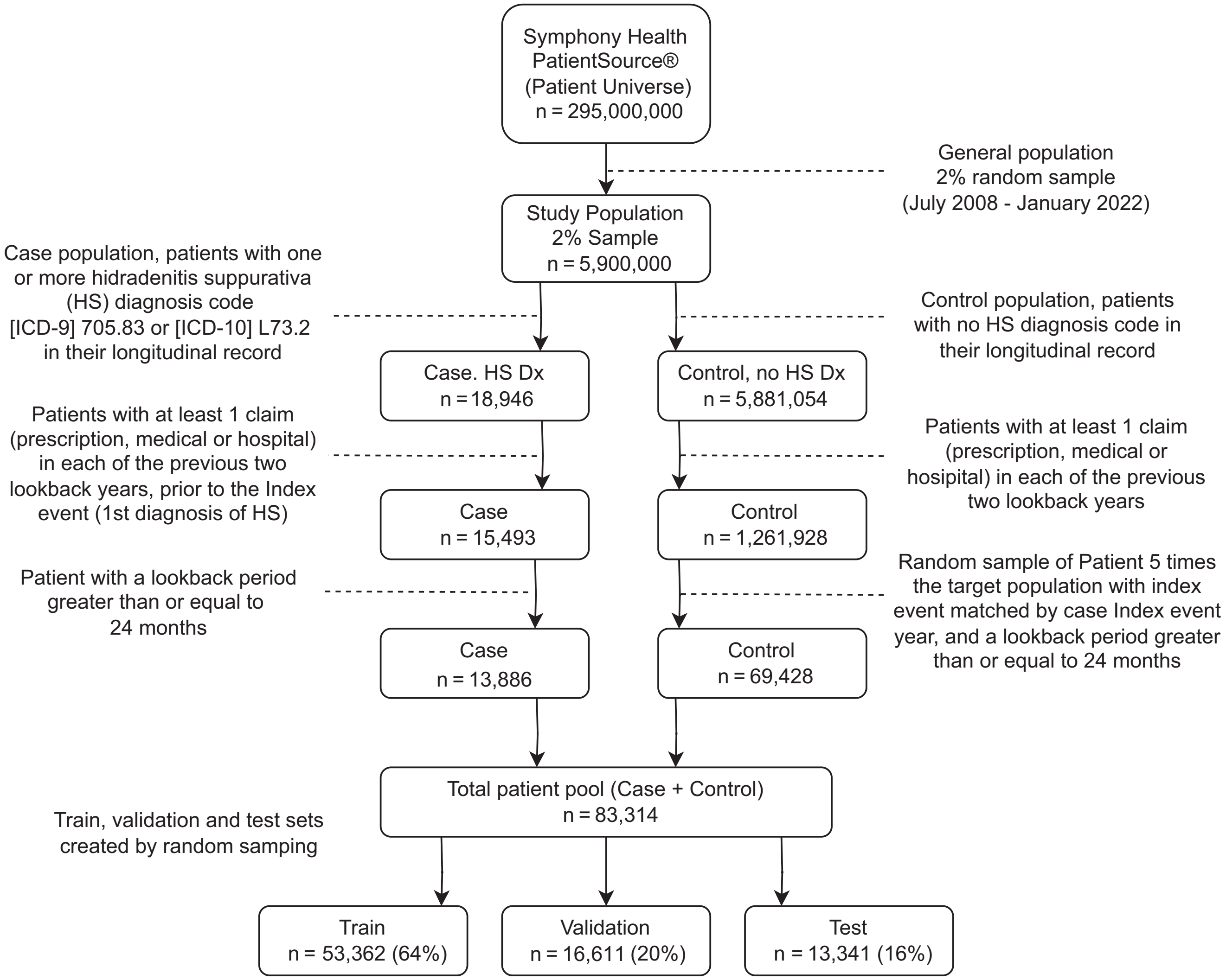

Symphony Health
PatientSource®
(Patient Universe)
n = 295,000,000
General population
2% random sample
(July 2008 - January 2022)
Study Population
2% Sample
n = 5,900,000
Case population, patients with one or more hidradenitis suppurativa (HS) diagnosis code [ICD-9] 705.83 or [ICD-10] L73.2 in their longitudinal record
Control population, patients with no HS diagnosis code in their longitudinal record
Case. HS Dx
n = 18,946
Control, no HS Dx
n = 5,881,054
Patients with at least 1 claim (prescription, medical or hospital) in each of the previous two lookback years, prior to the Index event (1st diagnosis of HS)
Patients with at least 1 claim (prescription, medical or hospital) in each of the previous two lookback years
Case
n = 15,493
Control
n = 1,261,928
Patient with a lookback period greater than or equal to 24 months
Random sample of Patient 5 times the target population with index event matched by case Index event year, and a lookback period greater than or equal to 24 months
Case
n = 13,886
Control
n = 69,428
Total patient pool (Case + Control)
n = 83,314
Train, validation and test sets created by random samping
Train
n = 53,362 (64%)
Validation
n = 16,611 (20%)
Test
n = 13,341 (16%)

Figure 1. Study population. Presented is a summary of the process of sampling anonymized US-based insurance records from Symphony Health PatientSource to generate training, validation, and test cohorts. Dx, diagnosis; HS, hidradenitis suppurativa; ICD, International Classification of Disease; US, United States.

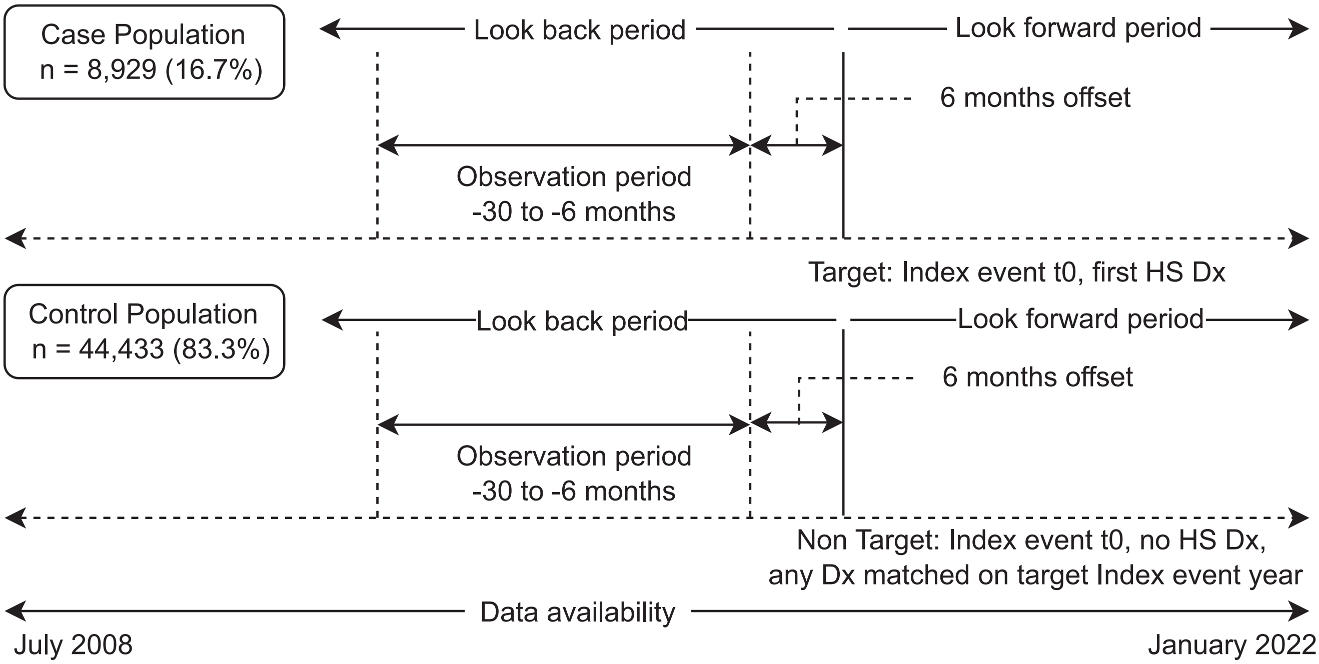


Figure 2. Modeling approach. The modeling approach involved generating a 24-month observation window, with a 6-month offset to the first diagnosis of HS/index event date in both the case and control populations. Dx, diagnosis; HS, hidradenitis suppurativa; t0, time of index event.

Table 1. Descriptive Statistics in the Overall Study Cohort and for HS Cases and Controls in the Training, Validation, and Test Partition Cohorts

Overall HS study cohort (n = 18,946)		
	Count, n (%)	Proportion in data, %
Female	14,361 (76)	0.2
Male	4585 (24)	0.1
Total population	18,946 (100)	0.3
Training partition cohort		
	HS cases (n = 8929)	Controls (n = 44,433)
Female, n (%)	6876 (77)	25,763 (58)
Male, n (%)	2053 (23)	18,670 (42)
Age at index date, mean (SD)		
Female patients	35.6 (14.7)	45.7 (21.9)
Male patients	42.4 (16.7)	44.4 (23.9)
Validation partition cohort		
	HS cases (n = 2725)	Controls (n = 13,886)
Female, n (%)	2097 (77)	8,126 (59)
Male, n (%)	628 (23)	5,760 (41)
Age at index date, mean (SD)		
Female patients	36.4 (15.1)	45.2 (21.3)
Male patients	42.2 (17.0)	44.2 (23.9)
Test partition cohort		
	HS cases (n = 2232)	Controls (n = 11,109)
Female, n (%)	1710 (77)	6562 (59)
Male, n (%)	522 (23)	4547 (41)
Age at index date, mean (SD)		
Female patients	36.5 (14.9)	45.5 (21.7)
Male patients	43.3 (16.9)	44.3 (23.8)

Abbreviations: HS, hidradenitis suppurativa; US, United States.
Using data referenced from the US Census Bureau data from 2021, we confirmed that all 50 states in the US were represented in the data source used in this research. Details can be found in Supplementary Table S1.

baseline model are presented in Figure 4a. The top features associated with HS diagnosis were patient age at the time of prediction, patient sex, and use of doxycycline hyclate.

Machine learning models

Results for the 3 data-driven machine learning models are presented in Figures 3 and 5. In general, the 3 machine learning models were comparable in performance, although the tree-based models (random forest [RF] and XGBoost [XGB]) showed marginally better area under the curve (AUC) (logistic regression [LR] = 0.75, RF = 0.79, XGB = 0.80) (Figure 3b). Importantly, all 3 machine learning models showed an increase in AUC over the clinical baseline model (AUC = 0.71) (Figure 3a), the improvement in performance being highest for the tree-based models. The same trend was observed in other evaluation metrics, with the machine learning models exhibiting much higher values for the F1-score (LR = 0.45, RF = 0.46, XGB = 0.46), precision (LR = 0.30, RF = 0.33, XGB = 0.34), and recall (LR = 0.69, RF = 0.74, XGB = 0.75) than the baseline model (F1-score = 0.38, precision = 0.27, recall = 0.60).

In application areas where there is significant imbalance between the positive and negative classes, such as rare disease diagnosis, it has been suggested (Ozenne et al, 2015) that the precision—recall curve may be a better indicator of model performance. Although our aim in this study was not the development of a diagnostic tool, we also present the precision—recall results for the machine learning models in Figure 5. As a reference, the figure also shows the performance of a dummy ‘random’ classifier, which assigns the most frequent class label to each instance in the validation set. The results are in line with other metrics, with the XGB model outperforming others (Figure 5a), although we noted

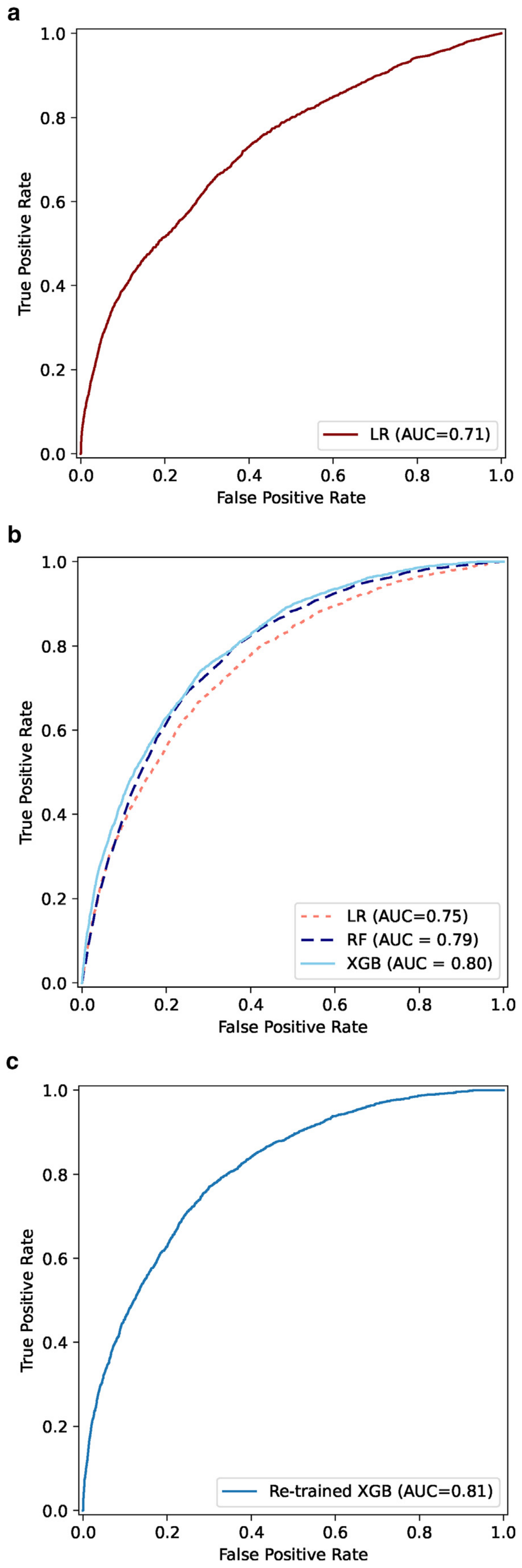


Figure 3. ROC curves. ROC curves for the (**a**) clinical baseline model, (**b**) machine learning models, and (**c**) final optimized model. The clinical baseline model was generated using a small set of clinical features selected by a dermatologist and refined using the Clinical Classifications Software for ICD-9 and ICD-10. Three machine learning models were developed using a large, unbiased set of features; the feature set contains all billing codes for each individual, including diagnosis, inpatient procedures, outpatient services and procedures, and generic drug name. XGB was the best performing model on the validation set. In **c**, the XGB model performance is evaluated on the independent test set. AUC, area under curve; ICD, International Classification of Diseases; LR, logistic regression; RF, random forest; ROC, receiver operating characteristic; XGB, XGBoost.

that the average precision is relatively low for all models owing to the class imbalance. Depending on the target population and clinical setting, the class imbalance may be even more severe in practice, leading to even lower precision and thus the need for model calibration and an operating point other than the default threshold of 0.5 used in this study. Under an assumed disease prevalence of 2% and therefore case/control ratio 1:50, we estimate the precision to fall to 0.05 (recall = 0.75) if a default operating point of 0.5 is used. This may lead to an unacceptably high false-positive rate; therefore, a more reasonable operating point choice in a general population setting could be 0.75, leading to a precision of 0.15 (recall = 0.35). A detailed analysis of appropriate operating point would of course be context dependent and may be the subject of future studies.

On the basis of the results mentioned earlier, the best performing model on the validation set (XGB) was subsequently retrained using data from the combined training and validation sets, and its performance was evaluated on the independent test set. As shown in Figures 3c and 5b, performance on the independent test set was in line with earlier results on the validation set (AUC = 0.81, F1-score = 0.47, precision = 0.34, recall = 0.75), indicating that the model generalizes beyond the training data, with the caveat that the test set belongs to the same underlying population and data source.

The top 15 features identified in the final machine learning model are presented in Figure 4b. The top features associated with eventual HS diagnosis were patient age at prediction, patient sex, use of sulfamethoxazole/trimethoprim, use of clindamycin phosphate, and obesity.

Subgroup analysis of model performance

As noted previously, sex and age at prediction are ranked as the most important features driving the model performance in our analysis. This is not surprising because HS is known to exhibit a higher prevalence in females, along with a lower tendency for disease diagnosis at the very early and later stages of life, consistent with the cohort statistics presented in Table 1. Therefore, we carried out additional analysis of the final model performance, looking at specific subgroups within the overall test set, namely, females only, males only, patients aged <40 years at prediction, and patients aged ≥40 years. For each subgroup, we provide performance metrics in the form of confusion matrices (Figure 6), precision−recall curves (Figure 7), and receiver operating characteristic (Figure 8). The results indicate that the model performs better in terms of precision and recall for the female subgroup and the subgroup aged <40 years than for the other subgroups. This variability in performance

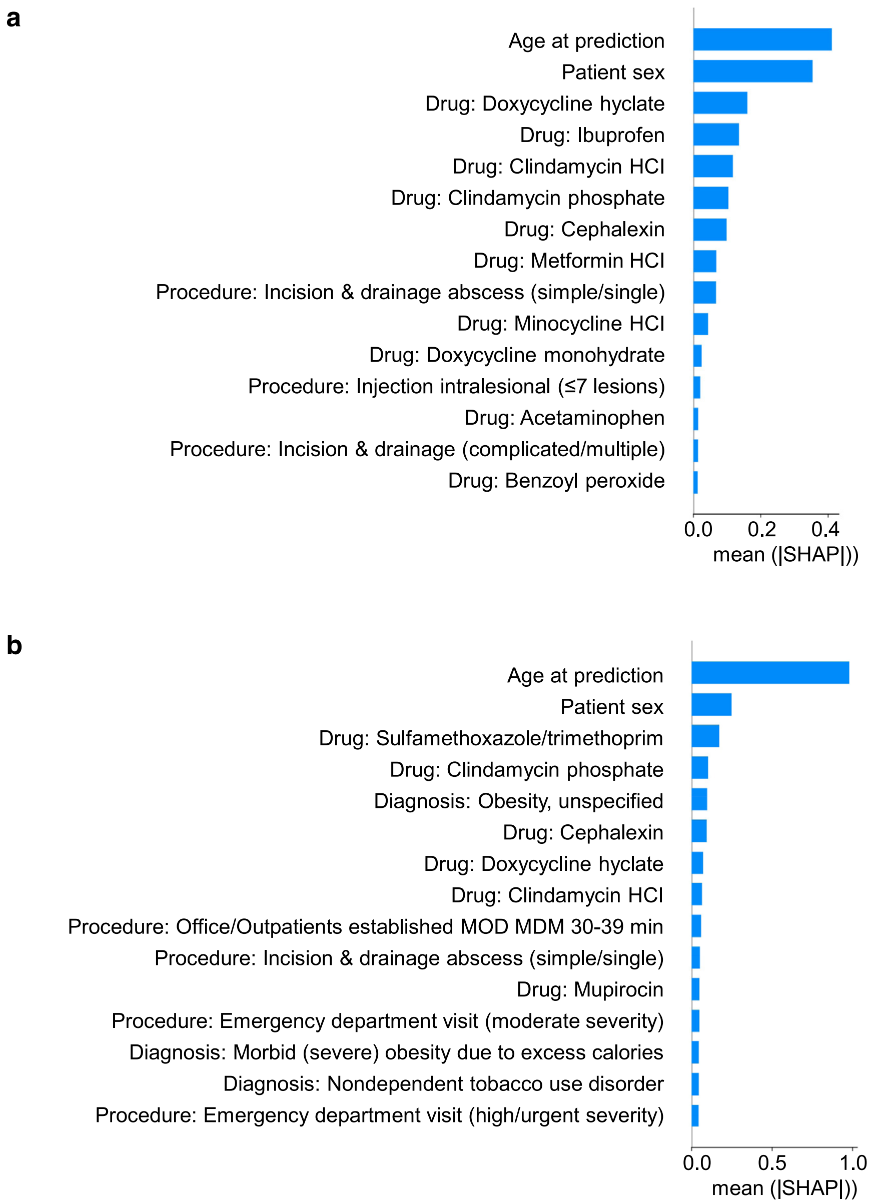


Figure 4. Shapley analysis. Shapley analysis for (**a**) clinical baseline model and (**b**) final machine learning model. The top 15 clinical features were identified using Shapley analysis, ranked by SHAP value. The prefixes Drug, Procedure, and Diagnosis refer to the coding systems from which the clinical features originated. The clinical baseline model (a logistic regression–based model) was generated using a small set of clinical features selected by a dermatologist and refined using the Clinical Classifications Software for ICD-9 and ICD-10. XGBoost was the best performing model on the validation set; this model was selected as the final machine learning model, retrained using the combined training and validation sets, and applied to the independent test set. AP, average precision; ICD, International Classification of Diseases; LR, logistic regression; RF, random forest; SHAP, Shapley Additive Explanations; XGB, XGBoost.

highlights the need for a considerate choice of an operating point in practice, depending upon the underlying population characteristics and the associated acceptable false-positive/-negative rates.

We note that the data source utilized for this study does not include other demographic information such as race and/or ethnicity, which may also significantly influence the risk of HS and therefore impact model performance.

Role of acne in HS misdiagnosis

A significant concern while evaluating any decision tool for clinical practice support is the effect of a high rate of false positives potentially driven by conditions similar in symptoms and presentation to the disease of interest. In the case of HS, it is possible that features associated with the more common condition of acne could lead to the inaccurate identification of individuals as patients with HS. Although this study was not aimed at developing a diagnostic support tool for clinical practice, we investigated the output from the final optimized model on the test set with a focus on acne. We first identified patients with an acne diagnosis in the overall study population using the International Classification of Diseases (ICD)-9 codes 706.0 and 706.1 and ICD-10 codes L70.1, L70.2, L70.3, L70.4, L70.5, L70.8, and L70.9. A total of 303,763 of 5,900,000 individuals (5.1%) had at least 1 of these diagnosis codes at any point in their medical claim history. Somewhat expectedly, within the HS cases, the frequency of acne diagnosis was much higher, at 19.5%, than in the controls (6.9%). We then analyzed the confusion matrix for the test set comprising true positives, false positives, true negatives, and false negatives from our machine learning model in terms of acne frequency. As shown in Figure 9, the fraction of false positives that had an acne diagnosis is only marginally higher (8.5%) than that of the controls, suggesting that this is likely not a driving factor for the misclassification.

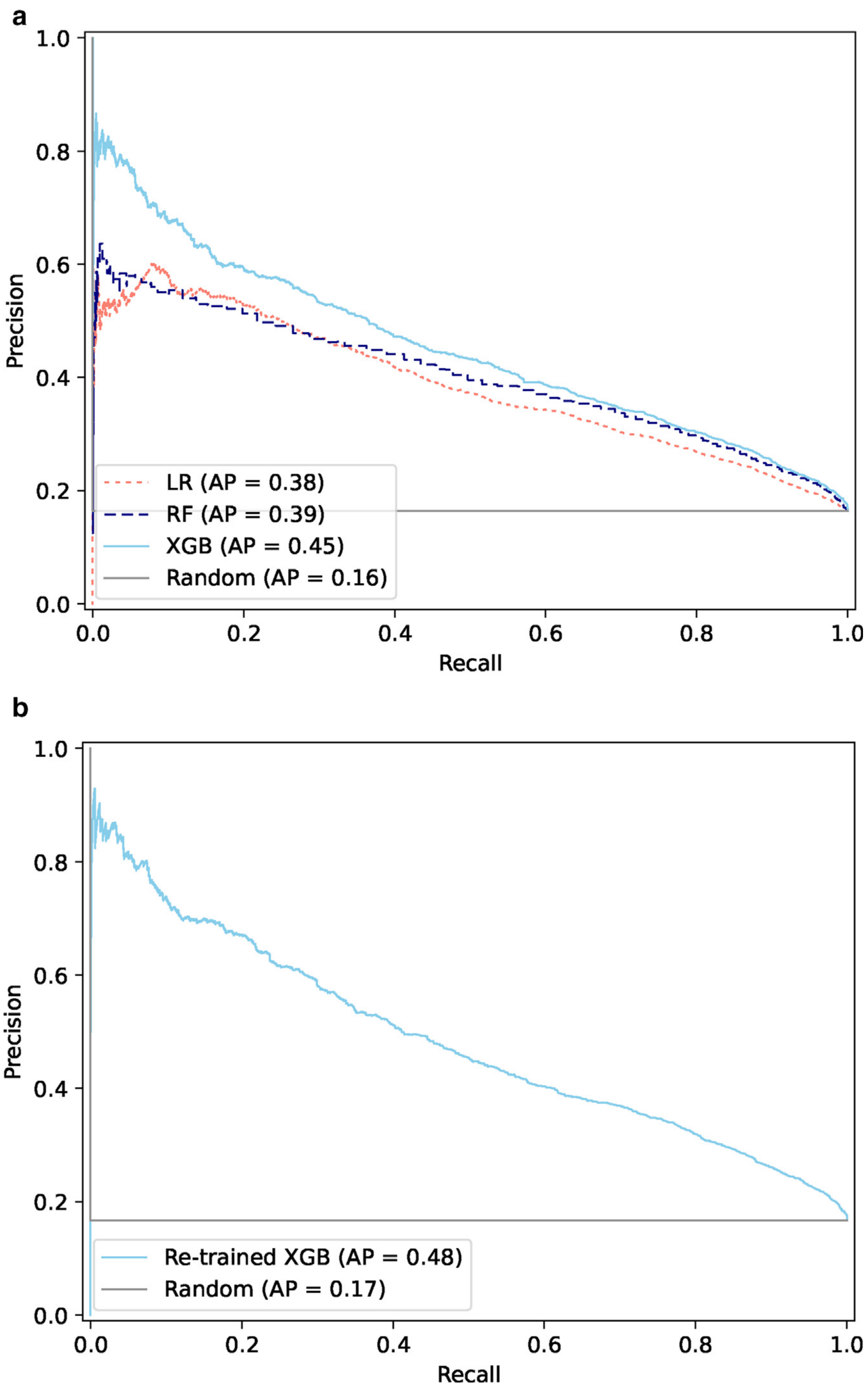


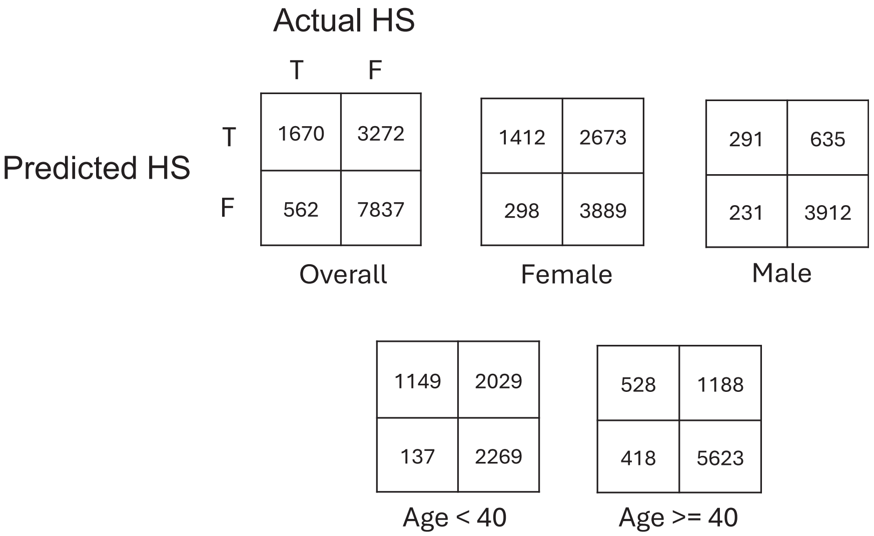
Figure 5. Precision−recall curves. Precision−recall curves for (**a**) machine learning models (validation set) and (**b**) final optimized model (test set). The horizontal gray line represents the performance of the random classifier as reference, which assigns the most frequent class label in the training set to each patient in the validation and test sets. SHAP, Shapley Additive Explanations;

DISCUSSION

HS can take years to be diagnosed, and earlier diagnoses can help patients get on the appropriate treatment, saving them from the added distress associated with disease progression. In this study, we investigated the potential of developing decision-support algorithms from insurance claims data by detecting signals of HS prior to actual diagnosis. Because there is currently no clinical guidance to identify patients at high risk for HS prior to actual diagnosis, there is no real “baseline” to compare with.

In this study, we compared a model using hand-selected features selected by a dermatologist with a data-driven machine learning approach. Notably, neither approach is the current standard of care. We found that the machine learning approach was superior, with XGB having the highest receiver operating characteristic AUC during validation (although all machine learning models had better performance than the clinical baseline model). The performance remained robust on an independent test set that was not accessed until the best model was chosen. Although insurance claims data are sparse, our study shows that a signal for future HS diagnosis can be detected at least 6 months prior to actual diagnosis.

Shapley analysis revealed the top features associated with HS diagnosis, and these corresponded to clinically



Actual HS / Predicted HS

Overall

Predicted HS	Actual HS T	Actual HS F
T	1670	3272
F	562	7837

Female

Predicted HS	Actual HS T	Actual HS F
T	1412	2673
F	298	3889

Male

Predicted HS	Actual HS T	Actual HS F
T	291	635
F	231	3912

Age < 40

Predicted HS	Actual HS T	Actual HS F
T	1149	2029
F	137	2269

Age >= 40

Predicted HS	Actual HS T	Actual HS F
T	528	1188
F	418	5623

Figure 6. Confusion matrices. Confusion matrices for the final model performance on the test set are presented. In addition to the overall test set, model performance is also presented for the following subgroups: females only, males only, patients aged <40 years, and patients aged ≥40 years.

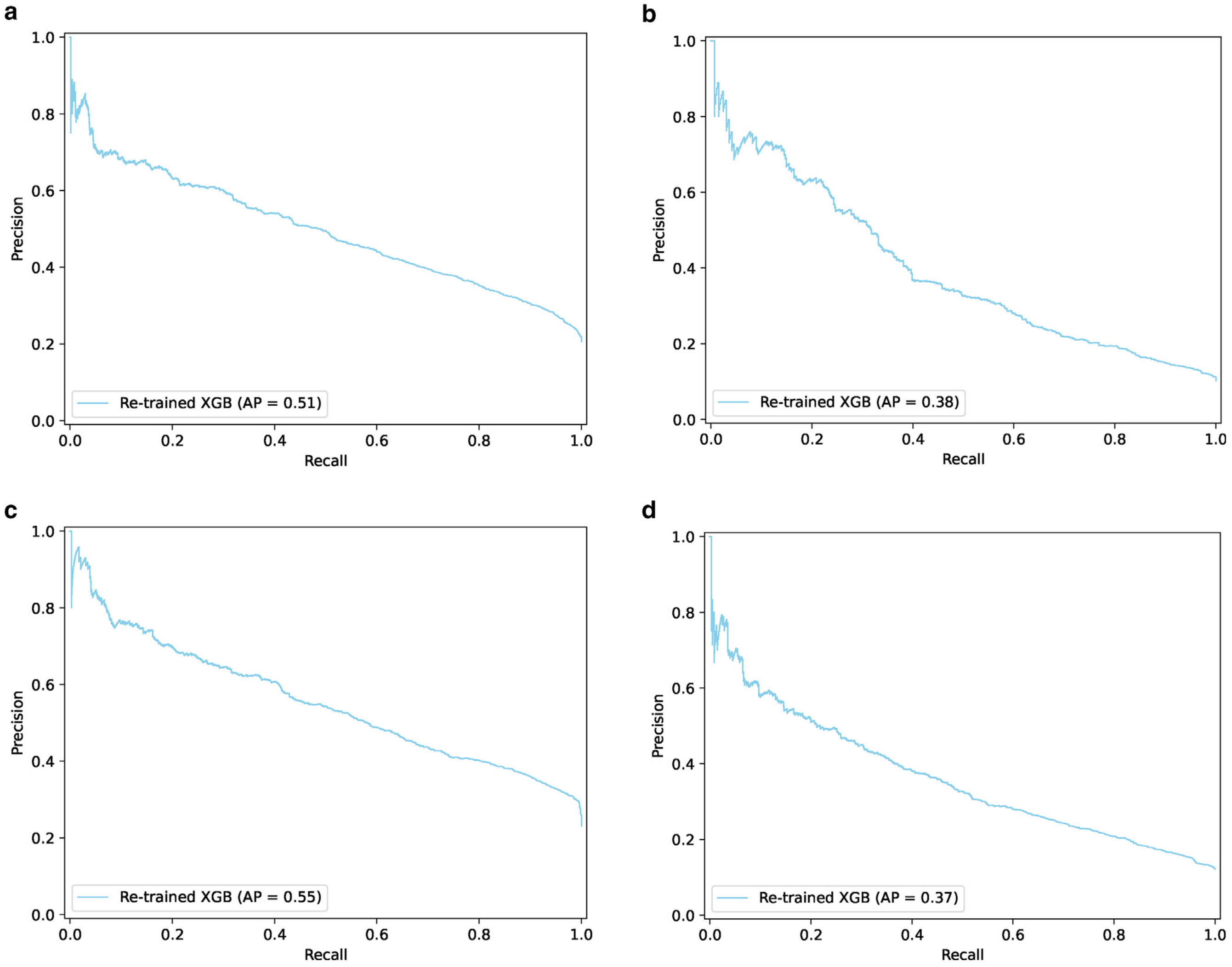


Figure 7. Precision–recall curves for subgroups. Precision–recall curves for test set subgroups: (**a**) females only, (**b**) males only, (**c**) patients aged <40 years, and (**d**) patients aged ≥40 years.

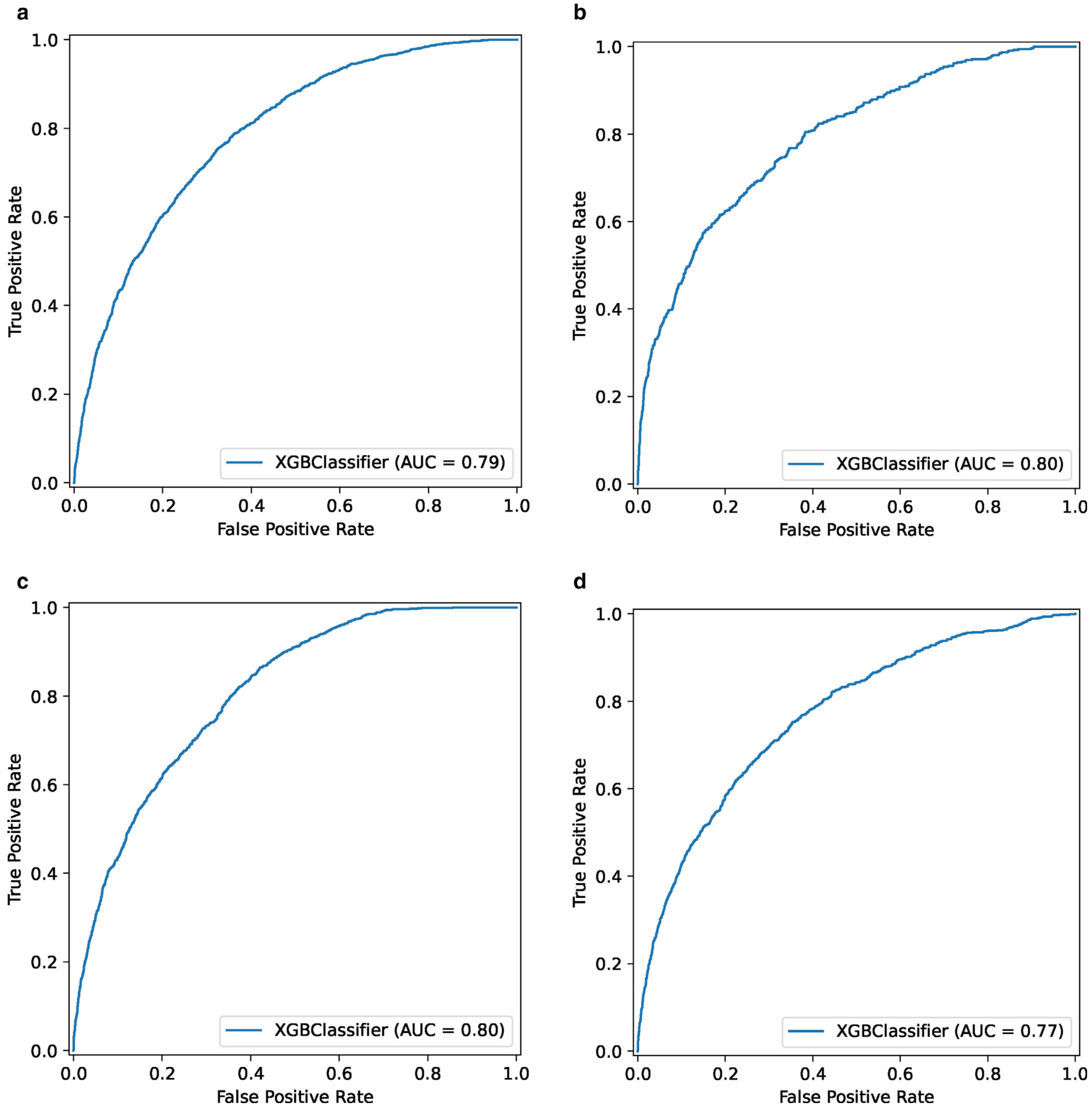
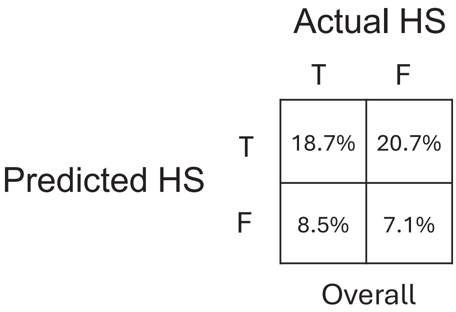


Figure 8. ROC curves for subgroups. ROC curves for test set subgroups (**a**) females only, (**b**) males only, (**c**) patients aged <40 years, and (**d**) patients aged ≥40 years. AUC, area under the curve; ROC, receiver operating characteristic.

expected features. Age and sex were the 2 top demographic factors in this study, and clinically, female sex and age between 30 and 39 years have been associated with HS diagnosis (Garg et al, 2017). The other factors in the top 15 features were also clinically related to HS symptoms. Many patients experience painful cysts and abscesses prior to actual HS diagnosis, and treatments for these appeared in the feature list: antibiotics (doxycycline, clindamycin, and minocycline), pain medications (ibuprofen and acetaminophen), incision and drainage, and intralesional injections. For treatment, these patients might be presenting to different physicians each time or to different care settings (eg, to urgent care or emergency rooms vs outpatient clinics); thus, a clinician may not have the complete picture. However, an algorithm assessing claims data can detect patterns in clinical features, as noted in this study. Other features in the top 15 were linked to comorbidities of HS or treatment of those comorbidities, such as diabetes (which is treated with metformin) and acne (which is treated with benzoyl peroxide).

There are some limitations to this work. For most results presented in this analysis, the performance metrics were



		Actual HS	
		T	F
Predicted HS	T	18.7%	20.7%
	F	8.5%	7.1%

Figure 9. Percentage of patients with acne diagnosis. Presented is a confusion matrix for the final optimized model on the test set, representing the percentage of patients having an acne diagnosis in each group (true positive, false positive, true negative, and false negative). The patients in the test set with an actual HS diagnosis also have a much higher rate of acne diagnosis. Patients in the false-positive group (bottom left) have only a marginally higher rate of acne diagnosis than those in the control group overall. HS, hidradenitis suppurativa.

calculated using the default probability threshold of 0.5 for the prediction of output class (cases or control patients). As discussed in the results section, this does result in a relatively high false-positive rate for all models developed in this study. This choice was solely for model comparisons and may not be the optimal threshold in a real-world setting, especially in the context of predicting uncommon and rare conditions. The goal of decision support in HS is to provide a tool that can help identify patients who should be referred for specialized care earlier, not to make the diagnosis. Thus, an operating point that favors finding cases may be preferred because false positives will be further assessed by a specialist before any interventions such as onerous confirmatory diagnostic tests or treatment decisions.

In addition, we note that the test set used the same source of data as the training and validation sets, with a similar underlying patient population and the same inherent strengths and weaknesses of the healthcare data capture methods. For example, one limitation of the claims database is that it does not contain race and/or ethnicity data as well as some relevant clinical information that may be captured in unstructured clinical notes, such as location of abscesses, a factor that may be a strong predictor of an HS diagnosis. Further assessing model generalizability and potential refinement with more informative features in other richer data sources can be the subject of future research.

We emphasize that the aim of this exploratory study is not the development of a clinical diagnostic tool, a goal that would need further validation of the approach in diverse data sources as well as model calibration to ensure acceptable false-positive and false-negative rates in the intended target population. Still, as an initial proof-of-concept exercise, our findings demonstrate the potential to detect HS earlier using decision-support algorithms developed with longitudinal healthcare data such as medical insurance claims. Earlier diagnosis and intervention have the potential to improve patient QOL and treatment outcomes.

MATERIALS AND METHODS

Data source

The study utilized anonymized United States–based insurance records from Symphony Health's (an ICON plc company) PatientSource database between July 1, 2008 and January 31, 2022. This interconnected source of healthcare data included prescription, medical, and hospital claims across payers, pharmacies, hospitals, and clinics. The proprietary process utilized by Symphony Health ensured complete removal of personal health information, in compliance with the Health Insurance Portability and Accountability Act.

The PatientSource database includes data from 295 million individuals. We had access to a 2% random sample from the database, equating to 5.9 million individuals, which was used to identify patients with HS. Using data referenced from the US Census Bureau (United States Census Bureau, 2021) data from 2021, we confirmed that all 50 United States were represented in the data source used in this research (Supplementary Table S1).

The dataset did not contain details of race or ethnicity, and demographic information was limited to the patient's year of birth, sex, 2-digit Zone Improvement Plan code, and patient state (2-digit abbreviation). For Health Insurance Portability and Accountability Act compliance, the birth year of patients aged >80 years was changed to the current year minus 80.

Case and control populations

The study population was split into case and control populations (Figure 1). The case population was comprised of individuals with 1 or more HS diagnosis code in their longitudinal patient record to provide the broadest population of patients with HS for analysis. A case fulfilled the cohort filter criteria if the following conditions were met: 1 or more instances of an HS diagnosis code using the ICD-9 705.83 or 1 or more instance of ICD-10 L73.2 (Strunk et al, 2017). The choice of requiring only a single HS diagnosis for defining cases is motivated by the high sensitivity of this approach (Hardin et al, 2022) leading to a larger case cohort. To increase the likelihood of capturing initial HS diagnoses and individuals with a more complete claims history, the case population contained individuals with at least 1 claim of any type (prescription, medical, or hospital) in each of the 2 years prior to the first identified HS diagnosis code (index event).

The control population included individuals with no diagnosis of HS in their patient record. A control fulfilled the filter criteria if the following conditions were met: no occurrences of HS diagnoses using ICD-9 705.83 and ICD-10 L73.2 (Strunk et al, 2017). These individuals were matched with the case population on index event year by identifying a random diagnosis code in that year and then checking that they had at least 1 non-HS–related claim record of any type (prescription, medical, or hospital) in each of the 2 years prior to the diagnosis code (index event). No further matching between cases and controls (eg, propensity score) was carried out on any other variables except the index year because the aim of this study was to identify all possible HS discriminating features, including the available demographics data. This would not be the case if the aim were to identify potential causal factors for HS, which would necessitate controlling for potential confounders. To reduce class imbalance during the modeling steps, the control population was created with a ratio of 5 controls to 1 case record by randomly undersampling the controls. This ratio of case and controls ensures that the data are not overly unbalanced for the model training, while restricting the loss of too many control data points, keeping in view the large feature space of the machine learning models.

The characteristics of the overall study cohort as well as the training, validation, and test partitions are summarized in Table 1. The descriptive analysis that was performed identified an even distribution of index events through the data period (July 2008–January 2022). The female-to-male case population distribution in the dataset was 3 to 1.

Modeling approach

In the modeling approach, a 24-month observation window was generated, with a 6-month offset to the first diagnosis of HS (case population) or index event date (control population) (Figure 2). Only data from the 24-month observation window for each patient were used for training all models in predicting HS status as a binary classification task. No data from either the 6-month preindex period or the postindex period were used for model training, to ensure that the models were not inadvertently gaining an unfair advantage by accessing medical events related to the actual diagnosis

Modeling activities

To carry out the HS prediction task, 2 key sets of models were generated: a baseline model using a relatively small number of clinically relevant codes (prescription, medical, and hospital) identified by a dermatologist to create features and a set of machine learning models based on a more extensive, minimally filtered list of features. Details for both sets of models are given below.

Clinical baseline model

An initial baseline LR-based model was trained utilizing a set of features handpicked by a board-certified dermatologist (RD), with the aim of understanding how a model would perform using a small set of clinically understood features.

This included diagnosis codes for conditions such as diabetes mellitus and skin ulcers, specific prescriptions for antibiotics (including clindamycin phosphate, mupirocin calcium, doxycycline hyclate, minocycline hydrochloride, and clindamycin hydrochloride) and analgesics (including acetaminophen and ibuprofen), and procedures (including incision and drainage abscess).

For completeness, the initial list of baseline features selected by the dermatologist was further extended and refined using the Clinical Classifications Software for ICD-9 Clinical Modification and Clinical Classifications Software ICD-10 Clinical Modification. The Clinical Classifications Software, used up to October 2015, aggregates 14,000 illnesses and conditions into 285 mutually exclusive categories, whereas the Clinical Classifications Software Refined system, valid from October 2015, aggregates >70,000 diagnosis codes into over 530 clinically meaningful categories. These categories allowed the researchers to extend the initial feature list by identifying other relevant features belonging to the same categories. The final list of 114 individual features used in the baseline model can be found in Supplementary Materials and Methods. This model was used as a baseline comparator for the machine learning models.

Machine learning models

In addition to the baseline model, machine learning models were also developed in a data-driven approach using a large set of features without human selection.

The feature set contains all individual unaggregated billing codes for each patient: diagnosis (ICD-9/10 Clinical Modification), inpatient procedures (ICD-9/10 Procedure Code System), outpatient services and procedures (Health Common Procedure Coding System), and generic drug name. These were used to create binary features x, where x = 1 if that particular code was observed 1 or more times in the 24-month observation lookback period for a particular patient. Features with a frequency <1% in the overall cohort (cases and control patients) were then filtered out, resulting in a total of 963 features. No further feature selection (backward elimination, least absolute shrinkage and selection operator, etc) was carried out in subsequent modeling steps.

The list of features described earlier was used to train 3 different machine learning models for the classification task: LR, RF, and XGB. Each of these models were trained using the training dataset, with hyperparameter tuning through grid search and 5-fold cross-validation.

The following standard performance metrics were computed for each of the trained models, using the validation dataset: AUC, F1 score, precision (also known as positive-predictive value), and recall. Although we present in this study the results for all performance metrics for each model for completeness, it should be noted that in a practical setting (eg, patient screening), it may be preferred to achieve a higher recall rate at the expense of lower precision. The best performing model on the validation set in terms of AUC was then selected and evaluated for performance on the independent, held-out test set.

ETHICS STATEMENT

The analysis was based on an anonymized dataset, so it was not necessary to obtain consent. All patient data were fully anonymized.

DATA AVAILABILITY STATEMENT

The data source data for this study was the Symphony Health PatientSource database and was licensed from a third party. The authors are not permitted to share the licensed data publicly. The original data used in the study are available for others to license by contracting with the database owners (Symphony Health) for replicating or building upon the results from this study.

ORCIDs

Waqar Ali: http://orcid.org/0009-0008-8733-5044
Jonathan Williams: http://orcid.org/0009-0005-6855-4098
Betty Xiong: http://orcid.org/0000-0002-2047-7696
James Zou: http://orcid.org/0000-0001-8880-4764
Roxana Daneshjou: http://orcid.org/0000-0001-7988-9356

CONFLICT OF INTEREST

RD and JZ received research funding from UCB Pharma. JW and WA are employees of UCB Pharma. The remaining author states no conflict of interest.

ACKNOWLEDGMENTS

This study was sponsored by UCB Pharma. Support for third-party writing assistance for this article was funded by UCB Pharma in accordance with Good Publication Practice (GPP2022) guidelines (https://www.ismpp.org/gpp-2022). The authors acknowledge Susanne Wiegratz (UCB Pharma, Monheim, Germany) for publication coordination as well as Jack Wardle (Costello Medical) for medical writing and editorial assistance.

AUTHOR CONTRIBUTIONS

Conceptualization: RD, BX, JZ, JW, WA; Formal Analysis: RD, BX, JZ, JW, WA; Writing - Original Draft Preparation: RD, BX, JZ, JW, WA; Writing - Review and Editing: RD, BX, JZ, JW, WA

DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE (AI) OR LARGE LANGUAGE MODELS (LLMs)

The authors did not use AI/LLM in any part of the research process and/or manuscript preparation.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at https://doi.org/10.1016/j.xjidi.2025.100362.

REFERENCES

Agency for Healthcare Research and Quality. HCUP CCS. Healthcare Cost and Utilization Project (HCUP). https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp; 2017 (accessed 10 May 2024).

Garg A, Kirby JS, Lavian J, Lin G, Strunk A. Sex- and age-adjusted population analysis of prevalence estimates for hidradenitis suppurativa in the United States. JAMA Dermatol 2017;153:760–4.

Garg A, Reddy S, Kirby J, Strunk A. Development and validation of HSCAPS-1: A clinical decision support tool for diagnosis of hidradenitis suppurativa over cutaneous abscess. Dermatology 2021;237:719–26.

Hardin J, Murray G, Swerdel J. Phenotype algorithms to identify hidradenitis suppurativa using real-world data: development and validation study. JMIR Dermatol 2022;5:e38783.

Hernández Montilla I, Medela A, Mac Carthy T, Aguilar A, Gómez Tejerina P, Vilas Sueiro A, et al. Automatic International Hidradenitis suppurativa Severity Score System (AIHS4): A novel tool to assess the severity of hidradenitis suppurativa using artificial intelligence. Skin Res Technol 2023;29:e13357.

Jfri A, Nassim D, O'Brien E, Gulliver W, Nikolakis G, Zouboulis CC. Prevalence of hidradenitis suppurativa: a systematic review and meta-regression analysis. JAMA Dermatol 2021;157:924–31.

Kim GE, Shlyankevich J, Kimball AB. The validity of the diagnostic code for hidradenitis suppurativa in an electronic database. Br J Dermatol 2014;171:338–42.

Kokolakis G, Wolk K, Schneider-Burrus S, Kalus S, Barbus S, Gomis-Kleindienst S, et al. Delayed diagnosis of hidradenitis suppurativa and its effect on patients and healthcare system. Dermatology 2020;236:421–30.

Lee EY, Alhusayen R, Lansang P, Shear N, Yeung J. What is hidradenitis suppurativa? Can Fam Physician 2017;63:114–20.

Marvel J, Vlahiotis A, Sainski-Nguyen A, Willson T, Kimball A. Disease burden and cost of hidradenitis suppurativa: a retrospective examination of US administrative claims data. BMJ Open 2019;9:e030579.

Nguyen TV, Damiani G, Orenstein LAV, Hamzavi I, Jemec GB. Hidradenitis suppurativa: an update on epidemiology, phenotypes, diagnosis, pathogenesis, comorbidities and quality of life. J Eur Acad Dermatol Venereol 2021;35:50–61.

Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015;68:855–9.

Sabat R, Jemec GBE, Matusiak Ł, Kimball AB, Prens E, Wolk K. Hidradenitis suppurativa. Nat Rev Dis Primers 2020;6:18.

Scala E, Cacciapuoti S, Garzorz-Stark N, Megna M, Marasca C, Seiringer P, et al. Hidradenitis suppurativa: where we are and where we are going. Cells 2021;10:2094.

Snyder CL, Chen SX, Porter ML. Obstacles to early diagnosis and treatment of hidradenitis suppurativa: current perspectives on improving clinical management. Clin Cosmet Investig Dermatol 2023;16:1833–41.

Strunk A, Midura M, Papagermanos V, Alloo A, Garg A. Validation of a case-finding algorithm for hidradenitis suppurativa using administrative coding from a clinical database. Dermatology 2017;233:53–7.

United States Census Bureau. Annual estimates of the resident population for the United States, regions, states, District of Columbia, and Puerto Rico. April 1, 2020 to July 1, 2021, https://data.census.gov/table?tid=PEPPOP2021.NST_EST2021_POP; 2021 (accessed 10 May 2024).