VALUE IN HEALTH ▮ (2018) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Methodological Article

Development and Validation of an Algorithm for Identifying Patients with Hemophilia A in an Administrative Claims Database

Jennifer Lyons, PhD, MPH[1,*], *Vibha Desai, PhD*[1], *Yaping Xu, MD, MPH*[2], *Greg Ridgeway, MS, PhD*[3,4], *William Finkle, PhD*[3], *Paul Solari, MD, FACP*[2], *Sean Sullivan, MS, PhD*[5], *Stephan Lanes, PhD, MPH*[1]

[1]*HealthCore, Wilmington, DE, USA;* [2]*Genentech, South San Francisco, CA, USA;* [3]*Consolidated Research, Inc., Los Angeles, CA, USA;* [4]*University of Pennsylvania, Philadelphia, PA, USA;* [5]*University of Washington, Seattle, WA, USA*

ABSTRACT

Background: The accuracy with which hemophilia A can be identified in claims databases is unknown. **Objective:** Develop and validate an algorithm using predictive modeling supported by machine learning to identify patients with hemophilia A in an administrative claims database. **Methods:** We first created a screening algorithm using medical and pharmacy claims to identify potential hemophilia A patients in the US HealthCore Integrated Research Database between January 1, 2006 and April 30, 2015. Medical records for a random sample of patients were reviewed to confirm case status. In this validation sample, we used lasso logistic regression with cross-validation to select covariates in claims data and develop a predictive model to estimate the probability of being a confirmed hemophilia A case. **Results:** The screening algorithm identified 2,252 patients and we reviewed medical records for 400 of these patients. The screening algorithm had a positive predictive value (PPV) of 65%. The predictive model identified 18 predictors of being a hemophilia A case or noncase. The strongest predictors of case status included male sex, factor VIII therapy, office visits for hemophilia A, and hospitalizations for hemophilia A. The strongest predictors of noncase status included hospitalizations for reasons other than hemophilia A and factor VIIa therapy. A probability threshold of ≥0.6 resulted in a PPV of 94.7% (95% CI: 92.0–97.5) and sensitivity of 94.4% (95% CI: 91.5–97.2). **Conclusions:** We developed and validated an algorithm to identify hemophilia A cases in an administrative claims database with high sensitivity and high PPV.

Keywords: algorithm, hemophilia A, predictive model, validation.

Introduction

Hemophilia A is a rare, X-linked chromosomal disorder that causes a deficiency in clotting factor VIII, leading to bleeding complications and subsequent morbidity [1]. In the United States, there are fewer than 20,000 diagnosed cases of hemophilia A, with male patients representing the majority of these (>90%) [2,3]. Although hemophilia A affects a small number of patients, the disease contributes to a large economic and clinical burden because of the lifelong infusions and high healthcare utilization required [2,4], which vary by disease severity and treatment type [5,6]. Owing to the rarity of the condition, large populations are needed to identify sufficient numbers of patients with hemophilia A to understand natural history, comorbidities, treatment patterns, and treatment costs.

Administrative databases offer large populations for study, but no validated algorithm exists to identify patients with hemophilia A [2,4]. In studies using unvalidated algorithms, the proportion of hemophilia cases identified by the algorithm (sensitivity) and the proportion of patients identified by the algorithm who have hemophilia (positive predictive value [PPV]) are unknown. Unlike studies using previous algorithms, we sought to include all patients with hemophilia regardless of age or sex (high sensitivity), while avoiding the inclusion of many noncases (high PPV). In the current study, we used administrative claims data supplemented by data from medical records to develop and validate an algorithm to predict hemophilia A case status using predictive modeling and machine learning methods.

J. Lyons, V. Desai, and S. Lanes were employees of HealthCore during the development of this manuscript. Y. Xu is an employee of Genentech. P. Solari was an employee of Genentech during the development of this manuscript. G. Ridgeway is an employee of the University of Pennsylvania and Consolidated Research. W. Finkle is an employee of Consolidated Research. S. Sullivan reports grants and personal fees from Genentech.

* *Address correspondence to:* Jennifer Lyons, 300 Brickstone Square, Suite 801A, Andover, MA 01810, USA.
E-mail: jglyons@bu.edu

1098-3015$36.00 – see front matter

https://doi.org/10.1016/j.jval.2018.03.008

Methods

This retrospective cohort study collected information from the HealthCore Integrated Research Database (HIRD) between January 1, 2006 and April 30, 2015. The HIRD is an insurance claims database comprising enrollment data as well as medical and pharmacy claims data from health plan members across the United States dating from 2006. We created the algorithm in two steps. First, we developed a screening algorithm designed to define an enriched sample containing the hemophilia A cases (which necessarily included some patients who did not have hemophilia A [false positives]). Then we obtained medical records for a sample of the screening population to review medical records to confirm true cases of hemophilia A, and we constructed a predictive model using variables in claims data to discriminate true-positive from false-positive cases.

The screening algorithm was developed in consultation with clinical experts and included the following criteria:

1. At least one medical or pharmacy claim for factor VIII therapy, factor eight inhibitor bypassing agent (FEIBA), or recombinant factor VIIa therapy regardless of hemophilia A diagnosis history; or
2. At least one medical or pharmacy claim for complex factor VIII/von Willebrand factor therapy and no medically attended visits with a diagnosis of von Willebrand's disease (International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] codes beginning with 286.4); or
3. At least one medical or pharmacy claim for desmopressin and at least one medically attended visit with a diagnosis of hemophilia A (ICD-9-CM codes beginning with 286.0); and
4. At least 6 months of continuous enrollment in the database or, for patients 6 months of age or younger, at least one diagnosis of hemophilia A (ICD-9-CM 286.0)

A medically attended visit was defined as an encounter with any type of provider that included a Current Procedural Terminology (CPT) code for physician evaluation and management. The index date for each cohort member was the earliest date the patient fulfilled any of the above conditions.

We randomly selected a list of fully insured patients who met the screening criteria to obtain medical records for 400 patients. For each patient, we obtained and abstracted the medical records from their providers for 2 years before to 2 years after the date the patient met the screening criteria. Each medical record was abstracted by nurses who received training on the study design and overall study objectives. The medical record abstraction form included information about confirmed hemophilia A diagnosis, alternative medical diagnoses (for cases in which hemophilia A was not confirmed), and hemophilia A treatment patterns.

PPV for the screening algorithm was calculated as the number of cases confirmed by medical record review divided by the total number of patients for whom medical records were sufficient for clinical review and determination of case status. A 95% confidence interval was calculated for the PPV estimate.

Predictors of having and not having hemophilia A were identified for possible inclusion in the model using literature review, consultation with clinical experts, and various healthcare utilization variables available in the claims database. These variables included exposure to treatments for hemophilia A and related conditions, hemophilia A characteristics, diagnosis of other bleeding disorders, healthcare utilization, and patient demographics.

We conducted a least absolute subset selection and shrinkage operator (lasso) logistic regression analysis to develop a model for identifying cases of hemophilia A [7]. Lasso regression was selected to account for the large number of potentially correlated predictors [8]. To avoid overfitting the model, we used 20-fold cross-validation in both the model selection process and model evaluation [9]. In the 20-fold cross-validation, the sample of 400 patients with medical records was randomly divided into 20 subsets. For each of 20 training iterations, each subset was excluded once for use in validation, and the model was fit with the remaining 19 subsets. The results from the 20 training iterations were used to select the shrinkage operator and the best-fitting model to minimize prediction error rate. Using the values of the predictors in the model, we estimated for each patient in the screening cohort the probability of being a true case of hemophilia A. The PPV and sensitivity of the algorithm depend on the estimated probability cutoff or threshold. We evaluated the performance of the algorithm by calculating the area under the receiver operator characteristic (ROC) curve [10]. As lasso is a main effects model that may not account for the more complicated relationships within these data, the performance of the lasso model was compared with a more complex machine learning method, generalized boosted modeling, which allowed for higher order interaction terms and nonlinearities [11,12].

Results

The screening cohort included 2,252 patients who met the screening algorithm as potentially having hemophilia A, including

Table 1 – Formation of screening hemophilia A cohort, HIRD, January 1, 2006–April 30, 2015.

Criterion no.	Criteria	N
1	Patients with at least one medical or pharmacy claim for factor VIII therapy, factor eight inhibitor bypassing agent (FEIBA), or factor VIIa therapy*	2,059
2	Patients with at least one medical or pharmacy claim for factor VIII/von Willebrand complex therapy AND no medically attended visits with a diagnosis of von Willebrand's disease†	49
3	Patients with at least one medical or pharmacy claim for desmopressin AND at least one medically attended visit with a diagnosis of hemophilia‡	452
4	Total patients with criteria 1, 2, or 3	2,560
5	Patients retained after including only patients with 6 months of eligibility surrounding index§	2,127
6	Patients of age ≤6 months with a diagnosis of hemophilia A	125
	Total patients	2,252

* Factor VIII therapy was identified using GPI code 85100010x and HCPCS codes J7182, J7185, J7186, J7187, J7190, J7191, J7192, C9136, Q9975, and Q2023; FEIBA was identified using GPI code 85100020x and HCPCS code J7198; factor VIIa therapy was identified using GPI code 85100026x and HCPCS code J7189; and desmopressin was identified using GPI code 30201010x and HCPCS code J2597.

† von Willebrand's disease identified using ICD-9-CM code 286.4x.

‡ Desmopressin was identified using GPI code 30201010x and HCPCS code J2597; medically attended visits for hemophilia A were identified using ICD-9-CM code 286.0x and CPT codes, 99201-99499.

§ Index date is the *earliest* date the patient meets the cohort entry criteria.