ELSEVIER

Journal of Clinical Epidemiology ■ (2015) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Migraineurs were reliably identified using administrative data

Carl van Walraven[a,b,c,*], Ian Colman[a]

[a]*Department of Medicine/School of Epidemiology, Public Health, and Preventive Medicine, University of Ottawa, 451 Smyth Road, Ottawa ON K1N 6N5, Canada*

[b]*Ottawa Hospital Research Institute*

[c]*ICES uOttawa*

Accepted 8 September 2015; Published online xxxx

Abstract

Background: Migraine is a common and important source of pain and disability in society. Accurately identifying such people using routinely collected health data would be beneficial for health services research.

Objective: Externally validate a previously published method to identify migraineurs using health administrative data; and determine if a better model can be derived using data-mining techniques.

Methods: Migraine status was determined for Ontarians participating in a population-based, cross-sectional survey. Consenting participants were linked to population-based health administrative data to identify age, sex, and coded diagnoses. Discrimination and calibration measures were used to appraise the models. A de novo technique we term ''double threshold analysis'' was used to determine optimal lower and upper expected probabilities to identify migraine status in the newly derived model.

Results: A total of 1,01,114 people (mean age 46 years, 46% male) were included in the study, of which 11,314 (11.2%) had migraines. Using data-driven parameter estimates, the previous model to identify migraineurs had adequate discrimination (c-statistic 0.707 [95% CI 0.701−0.712]) and calibration (Hosmer−Lemeshow [H−L] statistic 20.8). A new model that included diagnostic code scores for physician visits, emergency visits, and hospitalizations with nonlinear terms for age and interactions significantly improved the model (c-statistic 0.724 [0.716−0.733], 16.4). Categorizing all people with a predicted migraine probability less than 10% or greater than 90% as without and having the disease, respectively, resulted in a sensitivity of 3.1%, a specificity of 99.96%, and a positive predictive value of 81.0% while capturing 57.0% of the cohort and 29.3% of migraineurs.

Conclusion: A previously derived model to identify migraineurs was improved using data-mining techniques permitting accurate cohort identification using routinely collected health administrative data.

Keywords: Administrative data; Multivariable logistic regression; Claims signature model; Validation; Migraine; Data mining

1. Introduction

Routinely collected health care administrative data (''claims data'') are commonly used to study particular diseases. Diseases are identified in such data using diagnostic codes whose accuracies can vary extensively for the conditions they are supposed to represent. Despite the knowledge that these codes may be inaccurate, the accuracy of codes used to identify diseases in health administrative databases are not measured or reported in the vast majority of published studies using administrative data [1].

To increase the accuracy of identifying conditions using administrative data, some researchers have created multivariate models to predict the probability that any particular patient has the disease of interest [2−5]. These models allow researchers to use all data available to them—rather than the presence or absence of a single diagnostic code—to identify people in a population with a particular disease or condition.

Kolodner et al. [6] published a series of such multivariate models to identify patients in the community with migraine. Migraine is a common condition that is associated with notable disability and increased health utilization. Being able to accurately identify such people in a population could permit the study of treatment, outcomes, and epidemiology of migraineurs; for example, migraine has been associated with suicidal ideation and suicide attempt [7]. Kolodner et al. presented six different models which included different combinations of terms for patient demographics (age and sex); migraine diagnostic codes from

Conflict of interest: None.

* Corresponding author. ASB1-003 1053 Carling Avenue, Ottawa, Ontario K1Y 4E9, Canada. Tel.: 613-761-4903; fax: 613-761-5492.

E-mail address: carlv@ohri.ca (C. van Walraven).

http://dx.doi.org/10.1016/j.jclinepi.2015.09.007

0895-4356/

What is new?

Key findings

- A previously derived model to identify people with migraine using administrative data sets had adequate discrimination and calibration in an external population.

What this adds to what was known?

- Using data-mining techniques, our ability to identify migraineurs was improved significantly. Unique analytic methods explicitly determined the influence of various probability thresholds on the model's accuracy for identifying migraineurs using administrative data.

What is the implication and what should change now?

- Researchers can use multivariate models and double-threshold analyses to precisely determine how accurately cases are identified using administrative data.

medical claims; comorbidity codes from medical claims; and medication exposure (including drugs used as migraine abortants, analgesics, and prophylaxis; other analgesics; and anti-emetics). However, to our knowledge, these models have never been validated in an external population. As such, its accuracy is unknown.

In this study, we externally validated the Kolodner model in a random sample of Ontarians who participated in a large community health survey. In addition, we determined whether we could improve the model by using data-mining techniques to see if other patient-centered information within administrative data sets could be used to identify migraineurs.

2. Methods

2.1. Data sources

This study used the Canadian Community Health Survey (CCHS) cycle 2.1, cycle 3.1, and the 2007−2008 iteration administered between 2001 and 2005. The CCHS is a cross-sectional survey conducted biennially by Statistics Canada that collects data on health status. The survey uses a complex multistage sampling strategy to randomly select households in each health region. Individuals in each household were then randomly selected to participate in the survey. The target population included individuals aged 12 years and older with the following individuals being excluded: those living on Indian Reserves; those living within institutions; those who are full-time members of the Canadian Forces; or those in extremely remote areas. For this study, Ontarians were included if they had a valid health card number and had consented to have their survey responses made linkable to other administrative data.

To create study covariates, we linked these data to anonymized population-based health administrative databases in Ontario, Canada in which the costs for all hospital and physician services are covered by a universal health care system. Databases used in this study included the Registered Persons Database, which captures each person's date of birth and sex; Discharge Abstract Database (DAD), which captures all hospitalizations; the National Ambulatory Care Registry System (NACRS), which captures all visits to any emergency department; and the Physician Services Database (PSD) which captures claims for all health services submitted for remuneration. Medication costs for people under 65 years are not publically funded; therefore, medication utilization in all patients is not captured by the data sets and was not included in this analysis. All databases were linked deterministically via encrypted health care numbers.

2.2. Migraine status determination

Within each CCHS cycle, we determined each person's migraine status using responses from the question dealing with migraine: "Remember, we're interested in conditions diagnosed by a health professional. Do you have migraine headaches?" People who responded positively to this question were categorized as migraineurs.

2.3. Model covariates

As we stated previously, medication information is unavailable for Ontarians younger than 65 years. For that reason, we applied the model from Kolodner et al. which included terms for patient age (expressed as both a linear and a squared term), patient gender, and diagnoses of migraine or depression in medical claims. To replicate the Kolodner model, we determined each person's age (at the time of the survey) and sex from registered persons database. We linked to DAD, NACRS, and PSD to retrieve all diagnostic codes for migraine or depression (See Appendix at www.jclinepi.com) from all encounters for each person during the 2 years before the interview date.

To improve the prediction of migraine status, we developed "diagnostic code scores" for each data set (ie, DAD, NACRS, and PSD). Within each data set, these diagnostic code scores measured the strength of each code's independent association with migraine status. To create the scores, we first clustered codes together for ICD-10-CA codes (used in DAD, NACRS), we used the first three alphanumerics; for ICD-9 codes (used in PSD), we used the first three numbers. We then created binary variables indicating the presence or absence of each code for each person. Patients with no records in the data set were assigned a value of "0" for each binary covariate. We then used univariate binary logistic