Journal of Biomedical Informatics 82 (2018) 41–46

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Automatic address validation and health record review to identify homeless Social Security disability applicants

Jennifer Erickson*, Kenneth Abbott, Lucinda Susienka

Minnesota Disability Determination Services, 121 7th Place E, Saint Paul, MN 55101, United States

ARTICLE INFO

Keywords:
Health records
Natural language processing
Social security
Disability
Homeless

ABSTRACT

Objective: Homeless patients face a variety of obstacles in pursuit of basic social services. Acknowledging this, the Social Security Administration directs employees to prioritize homeless patients and handle their disability claims with special care. However, under existing manual processes for identification of homelessness, many homeless patients never receive the special service to which they are entitled. In this paper, we explore address validation and automatic annotation of electronic health records to improve identification of homeless patients.

Materials and Methods: We developed a sample of claims containing medical records at the moment of arrival in a single office. Using address validation software, we reconciled patient addresses with public directories of homeless shelters, veterans' hospitals and clinics, and correctional facilities. Other tools annotated electronic health records. We trained random forests to identify homeless patients and validated each model with 10-fold cross validation.

Results: For our finished model, the area under the receiver operating characteristic curve was 0.942. The random forest improved sensitivity from 0.067 to 0.879 but decreased positive predictive value to 0.382.

Discussion: Presumed false positive classifications bore many characteristics of homelessness. Organizations could use these methods to prompt early collection of information necessary to avoid labor-intensive attempts to reestablish contact with homeless individuals. Annually, such methods could benefit tens of thousands of patients who are homeless, destitute, and in urgent need of assistance.

Conclusion: We were able to identify many more homeless patients through a combination of automatic address validation and natural language processing of unstructured electronic health records.

1. Objective

Homeless patients face a variety of obstacles in pursuit of basic social services. Acknowledging this, the Social Security Administration (SSA) directs employees to prioritize homeless patients and handle their disability claims with special care [1]. However, under existing manual processes for identification of homelessness, many homeless patients never receive the special service to which they are entitled. In this paper, we explore address validation and automatic annotation of electronic health records in order to improve identification of homeless patients.

2. Background and significance

The Social Security Administration (SSA) is a significant source of support for disabled Americans, providing cash benefits for more than 14 million people [2]. Each year, millions of new Social Security disability claims make their way through federal and state agencies, with a final award rate of 35% [3]. Some of these disability applicants qualify as homeless under SSA policy ("A claimant is homeless if he or she does not have a fixed, regular, and adequate nighttime residence") [1]. The SSA supports the Supplemental Security Income Outreach Access and Recovery (SOAR) program [4] and other initiatives [5–8] to improve homeless patient access to Social Security disability benefits; these efforts suggest that psychiatric problems present significant obstacles to pursuit of public assistance. The SSA is also a member of the United States Interagency Council on Homelessness (USICH), which leads federal efforts to prevent and end homelessness [9]. Homeless research by another major federal agency, the Department of Veterans Affairs (VA), has shown that receipt of disability benefits is associated with

Abbreviations: DDS, Disability Determination Services; FO, Field Office; MeSH, Medical Subject Headings; NLP, Natural Language Processing; OCR, Optical Character Recognition; ROC, Receiver Operating Characteristic; SOAR, Supplemental Security Income Outreach Access and Recovery; SSA, Social Security Administration; UMLS, Unified Medical Language System; USICH, United States Interagency Council on Homelessness; VA, Department of Veterans Affairs

* Corresponding author.
E-mail address: jennifer.erickson@ssa.gov (J. Erickson).

https://doi.org/10.1016/j.jbi.2018.04.012
Received 18 August 2017; Received in revised form 30 January 2018; Accepted 24 April 2018
Available online 26 April 2018
1532-0464/

decreased risk of homelessness [10].

Homeless patients qualify for special service under SSA policy, including extra reminder phone calls and letters, if they do not respond to initial requests for evidence or action [1]. The SSA homeless case folder flag exists to identify homeless patients. Typically, patients receive the homeless flag only when they report homelessness to SSA field office (FO) staff or state Disability Determination Services (DDS) employees, who may recognize a given address as corresponding with a homeless shelter. However, applicants may fail to report homelessness, possibly due to embarrassment, social stigma, or unfamiliarity with one agency's particular definition of homelessness. For some disability claims, there is no mention of homelessness outside of electronic medical records. Research has established the feasibility of text mining and natural language processing (NLP) for analysis of these records, providing insight into risk factors [11–13], diagnosis [14], treatment [15–17], and other concepts [18–29], including homelessness [30,31]. Other research suggests that it is possible to identify some homeless patients via analysis of street addresses [32,33]. We found no studies combining these NLP and address validation strategies, though it is certainly possible that a combination of these methods might outperform either method in isolation, due to complex interactions between predictor variables. With this study, we attempt to improve identification of homeless patients by using both address validation and NLP (Fig. 1).

3. Materials and methods

3.1. Sample development

We focused on new disability claims arriving at Minnesota Disability Determination Services, which handles ~1.3% of the national SSA disability claim workload [34]. Our sample included claims arriving within a year of January 10, 2014; while our office received 33,420 new disability claims during this period, we studied only the 4628 claims arriving with at least one medical record in TIFF file format. We considered a variety of risk factors, including substance abuse, mental health diagnoses, veteran status, educational status, sex, and race [35]. We gathered information about each claim, including gender, age, education, years since cessation of paid employment, and whether the patient claimed disability solely under Title XVI of the Social Security Act, which features eligibility restrictions for income and resources [36]. We extracted both mailing addresses and residence addresses, if applicable, and we noted addresses for medical sources such as hospitals and clinics. We also extracted text representing each patient's stated reasons for claiming disability. We found only 26 patients who received an official homeless designation from an SSA employee. However, a cursory review of the remaining 4602 claims revealed an additional 357 homeless patients. The final total of 383 homeless patients did not include those who appeared to have stable housing when our office began work on their claim—even if their file contained references to previous or subsequent periods of homelessness. The homeless patients were more likely than non-homeless patients to have a psychiatric primary diagnosis (Supplemental Data), and this included one homeless patient who died after contracting influenza while living in a shelter.

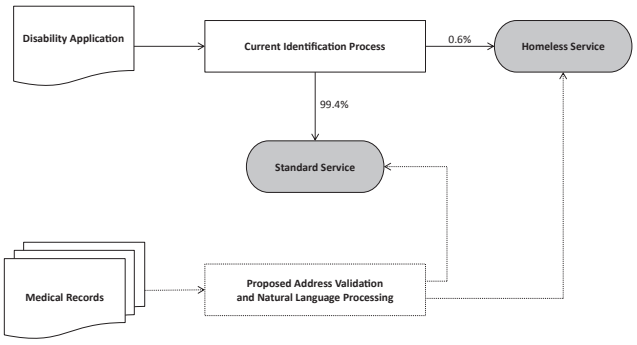

Fig. 1. Proposed combination of existing homeless identification process with address validation and natural language processing (NLP) of medical records.

3.2. Model development

We began by extracting addresses (Fig. 2), which we validated with the ZP4 [37] postal service database. We counted the number of times each patient's street address, geographic ID, block number, ZIP code, city, or county appeared in 10 years of historical data for claims receiving an official SSA homeless designation. We also compiled lists of shelters using a public directory [38] and recorded whether the claimant's mailing or residence address matched a known shelter, or whether the address failed to validate against any known address. We reviewed state and federal disability training materials to develop a list of words and phrases commonly appearing in the mailing addresses of homeless individuals, including *homeless, transient, unknown, general delivery*, or a variant of *C/O,* which signifies that mail intended for one person is entrusted to the care of another person, possibly at a different address; we then noted whether each patient's address contained any of these homelessness-related words or phrases. Similarly, we assembled lists of VA facilities [39], federal prisons [40], state correctional facilities [41], and county jails [42], counting the number of corresponding matches in the list of medical facilities accompanying each patient's application.

Next, we annotated medical records and disability applications. First, we used optical character recognition (OCR) software to extract text from 6984 faxed documents for all 4628 patients, including both medical and nonmedical records. Then, we used MetaMap [43], which includes an option for NegEx [44] negation detection, to annotate the text of these patient records and text from all 4628 disability applications. Annotations consisted of Medical Subject Heading (MeSH) [45] concepts from the Unified Medical Language System (UMLS) Metathesaurus [46]. To reduce the risk of spurious findings, we only included annotation codes from 18 of 134 UMLS Metathesaurus categories (Supplementary Data), which we selected based upon a review of category descriptions, searching for possible concordance with the demographic, medical, psychological, social, educational, occupational, and financial information typically present in our patients' records. The GeneralConText [47] package allowed us to identify and eliminate annotations involving hypotheticals or references to persons other than the patient. We counted the number of times each annotation code appeared in each patient's list of alleged medical conditions, as well as the number of times each annotation code appeared in each patient's electronic health record.

We used R [48] to complete model development. After stratifying our data by homeless designation [49], we eliminated annotation codes that did not appear in at least 5% of 383 homeless cases. We imputed page count, which was missing for 189 patients—this amounted to 4% of our sample. We trained five random forests [50] on our dataset: The first random forest used all predictor variables, and the remaining random forests used only subsets of predictor variables, to help characterize relative contributions of classes of variables to our primary model. To guide eventual interpretation of our model, we recorded both mean decrease in accuracy and mean decrease in Gini impurity, which show different responses to predictor correlation and scales of measurement [51]. We used the results from each model's 10-fold cross-validation to generate and analyze receiver operating characteristic (ROC) curves [52]. This approach allowed us to leverage all of our data for model training; however, it also produced meaningful validity statistics, since the cross-validation process generated separate training and testing subsets.