Emerging Health Threats Journal 2011, **4**:s31. **doi:** 10.3134/ehtj.10.031

www.eht-journal.org

ABSTRACT

A bootstrapping method to improve cohort identification

S Kandula[1], Q Zeng-Treitler[1], L Chen[2], W Salomon[3], and BE Bray[1]

[1]*Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA;* [2]*Scientific Systems Company Inc., Woburn, MA, USA; and* [3]*Clinical Metrics LLC, Poland, ME, USA*
E-mail: sasi.kandula@utah.edu

Objective

ICD-9 codes are commonly used to identify disease cohorts and are often found to be less than adequate. Data available in structured databases—lab test results, medications etc.—can supplement the diagnosis codes. In this study, we describe an automated method that uses these related data items, and no additional manual annotations to more accurately identify patient cohorts.

Introduction

The research reported in this paper is part of a larger effort[1] to achieve better signal-to-noise ratio, hence accuracy, in pharmacovigilance applications. The relatively low frequency of occurrence of adverse drug reactions (ADRs) leads to weak causal relations between the reaction and any measured signal.[2,3] We hypothesize that by grouping related signals, we can enhance detection rate and suppress false alarm rate.

Methods

The proposed method has the following steps:

1. Identify findings related to the diagnosis of interest and calculate the corresponding values for each instance in patient sample, *W*. Let X_j be the vector of related findings for patient *j* and D_j the number of his/her encounters with related ICD-9 codes.
2. Identify a training set *T*, from *W*, containing positive and negative instances.
3. Set C_j (the class of patient *j*) to D_j
4. Begin iterative process (superscript denotes iteration number):
    a. Using *T*, build support vector machine based classification models to obtain non-linear relationship $\hat{C}_j^i = f^i(X_j)$, where $\hat{C}_j^i$ is the model's estimate for C_j.
    b. Apply $f^i(X_j)$ to *W* and generate histogram over $\hat{C}_j^i$ for all patients in *W*.
    c. Select a cut-off threshold η^i to separate the positive and negative populations in the histogram (e.g., η^i can be the global minima of the function describing the histogram).
    d. Set $L_j^i = 1$ if $\hat{C}_j^i > \eta^i$ and $L_j^i = 0$ otherwise; here, L_j^i is our estimate of the patient's label and a positive label indicates the patient is positive for the condition.
    e. Compare L_j^i to L_j^{i-1} and compute F^i, the percentage of patients for whom the label has changed in the current iteration.
    f. If $F^i < \delta$ (where δ is an acceptable threshold), return L_j^i and terminate; else, update $\hat{C}_j^{i+1} = [(i-1) \times \hat{C}_j^i \ + D]/i$ for all *j* in T.

It can be seen from the above definition that the influence of D_j, the number of ICD-9 codes of the given diagnosis, tapers down as training progresses, while the learned relationship dominates.

Results

We applied the method described to identify diabetes and hyperlipidemia patient cohorts in a Logician database containing structure data for 800,000 patients. Relevant features (Step 1) were identified by consultation with clinicians and domain experts.

For diabetes, the feature set included the number of abnormal hemoglobin A1C tests, number of anti-diabetic medications (insulin, insulin supplements, biguanides, sulphonylureas, alpha-glucosidase inhibitors and so on) and abnormal blood glucose tests. Although only 15,000 patients in the database had diabetes related ICD codes (250.*), the method described here identified 22000 patients as diabetic.

In the case of hyperlipidemia, the number of abnormal lipid panel tests and related medications (HMG-CoA reductase inhibitors, intestinal cholesterol absorption inhibitors and so on) were used as features. The model labeled 76,000 patients as positive for hyperlipidemia which is almost twice the number of patients who could have been identified using ICD codes alone (272.*).

For both conditions, clinician evaluation was conducted on 100 cases. The values for recall and f-measure observed with the bootstrapping algorithm were found to be higher than those observed with ICD-9 codes.

Acknowledgements

This paper was presented as an oral presentation at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA on 1–2 December 2010.

References

1 Zeng Q, Cimino JJ, Zou KH. Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. *J Am Med Inform Assoc* 2002;**9**:294–305.
2 Noren GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug-drug interaction surveillance. *Stat Med* 2008;**27**:3057–70.
3 Noren GN, Edwards IR. Modern methods of pharmacovigilance: detecting adverse effects of drugs. *Clin Med (London, England)* 2009;**9**:486–9.