Using Information Entropy to Monitor Chief Complaint Characteristics and Quality

Shaun Grannis*[1, 2], Brian Dixon[4, 2], Yuni Xia[3] and Jianmin Wu[4, 2]

[1]Indiana University School of Medicine, Indianapolis, IN, USA; [2]Regenstrief Institute, Indianapolis, IN, USA; [3]Indiana University Purdue University Indianapolis, Indianapolis, IN, USA; [4]Indiana University School of Informatics, Indianapolis, IN, USA

Objective

We describe how entropy, a key information measure, can be used to monitor the characteristics of chief complaints in an operational surveillance system.

Introduction

Health care processes consume increasing volumes of digital data. However, creating and leveraging high quality integrated health data is challenging because large-scale health data derives from systems where data is captured from varying workflows, yielding varying data quality, potentially limiting its utility for various uses, including population health. To ensure accurate results, it's important to assess the data quality for the particular use. Examples of sub-optimal health data quality abound: accuracy varies for medication and diagnostic data in hospital discharge and claims data; electronic laboratory data used to identify notifiable public-health cases shows varying levels of completeness across data sources; data timeliness has been found to vary across different data sources. Given that there is clear increasing focus on large health data sources; there are known data quality issues that hinder the utility of such data; and there is a paucity of medical literature describing approaches for evaluating these issues across integrated health data sources, we hypothesize that novel methods for ongoing monitoring of data quality in rapidly growing large health data sets, including surveillance data, will improve the accuracy and overall utility of these data.

Methods

Our analysis used chief complaint data derived from the original real-time HL7 registration transactions for ED encounters over a 3-year study period between January 1, 2008 and December 30, 2010 from over 100 institutions participating in the Indiana Public Health Emergency Surveillance System (PHESS) [1]. We used the following syndrome categories based on various definitions: respiratory, influenza like illness, gastrointestinal, neurological, undifferentiated infection, skin, and lymphatic. We calculated entropy for chief complaint data [2]. Entropy measures uncertainty and characterizes the density of the information contained in a message, commonly measured in bits. We analyzed entropy stratified a) by syndrome category, b) by syndrome category and time, and c) by syndrome category, time, and source institution.

Results

Analysis of more than 7.4 million records revealed the following: First, overall information content varied by syndrome, with "neurological" showing greatest entropy (Figure 1). Second, entropy measures followed consistent intraorganizational trends: information content varied less within an organization than across organizations (Figure 2). Third, information entropy enables detection of otherwise unannounced changes in system behavior. Figure 3 illustrates the monthly entropy measures for the respiratory syndrome from 5 sources over 36 months. One source changed registration software. Their visit volume didn't change, but the information content of the chief complaint changed, as indicated by a substantial shift in entropy.

Conclusions

As we face greater data volumes, methods assessing the data quality for particular uses, including syndrome surveillance, are needed. This analysis shows the value of entropy as a metric to support monitoring of surveillance systems. Future work will refine these measures and further assess the inter-organizational variations of entropy.

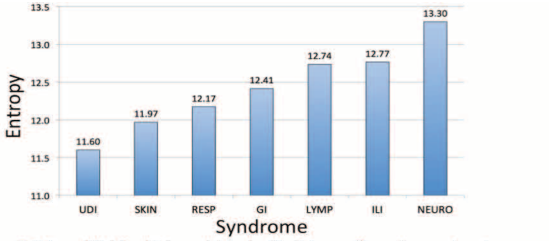

Figure 1: Entropy (bits) for chief complaints classified into specific syndrome categories.

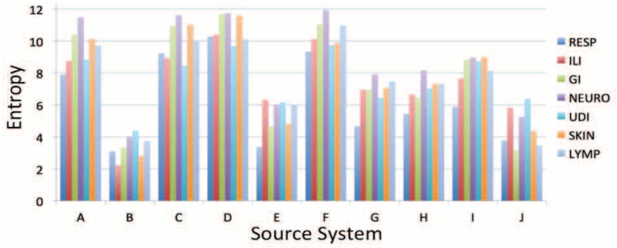

Figure 2: Entropy (bits) for chief complaints classified into specific syndrome categories stratified by source system for 10 high-volume emergency departments.

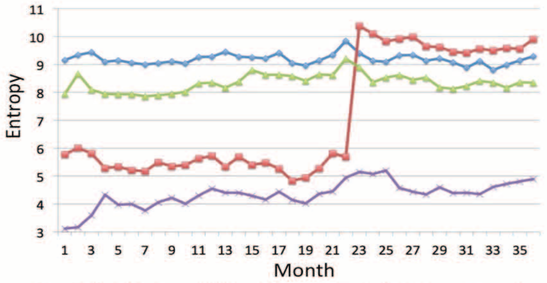

Figure 3: Monthly entropy (bits) for chief complaints classified into specific syndrome categories, stratified by source system for 5 high-volume emergency departments. Note the shift in values for one source that changed registration software.

Keywords

analytics; data quality; surveillance; system monitoring; information theory

Acknowledgments

This work was performed at the Regenstrief Institute in Indianapolis, Indiana. This study is supported in part by the CDC through the Indiana Center of Excellence in Public Health Informatics (1P01HK000077-01).

References

1. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of Emergency Department care delivered in Indiana. AMIA Annu Symp Proc. 2011; 409-16.

2. Shannon, Claude E. (July/October 1948). “A Mathematical Theory of Communication”. Bell System Technical Journal 27 (3): 379–423.

*Shaun Grannis
E-mail: sgrannis@regenstrief.org