MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-1056

Curating and Integrating Data from Multiple Sources to Support Healthcare Analytics

Kenney Ng[a], Chris Kakkanatt[b], Michael Benigno[b], Clay Thompson[b], Margaret Jackson[b], Amos Cahan[a], Xinxin Zhu[a], Ping Zhang[a], Paul Huang[c]

[a] *IBM T.J. Watson Research Center, Yorktown Heights, NY, USA*
[b] *Pfizer Inc., New York, NY, USA*
[c] *Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA*

Abstract

As the volume and variety of healthcare related data continues to grow, the analysis and use of this data will increasingly depend on the ability to appropriately collect, curate and integrate disparate data from many different sources. We describe our approach to and highlight our experiences with the development of a robust data collection, curation and integration infrastructure that supports healthcare analytics. This system has been successfully applied to the processing of a variety of data types including clinical data from electronic health records and observational studies, genomic data, microbiomic data, self-reported data from surveys and self-tracked data from wearable devices from over 600 subjects. The curated data is currently being used to support healthcare analytic applications such as data visualization, patient stratification and predictive modeling.

Keywords:
Data Collection; Data Curation; Automatic Data Processing.

Introduction

The volume and variety of healthcare related data continues to grow, spurred on by the increasing adoption and use of electronic health records (EHRs), the explosion of omics data and the proliferation of actigraphy data from wearable self-tracking devices and mobile applications [1]. The use of this data for healthcare analytics such as data visualization, patient stratification, predictive modeling, personalized medicine and drug discovery will increasingly depend on the ability to appropriately collect, curate and integrate disparate data from many different sources [2]. The linking of many different types of data will allow the analysis and modeling of complex multi-dimensional interactions which can enable deeper insights.

Materials and Methods

Our data consist of the following diverse types and sources for over 600 subjects: (1) clinical data from EHRs and observational studies: demographics, family history, labs, vitals, diagnoses, medications; (2) genomic data: 500K single nucleotide polymorphisms (SNPs); (3) microbiomic data: abundances of gut microbial taxa; (4) self-reported data from surveys: behavioral, lifestyle, diet; and (5) self-tracked data from wearables: activity, sleep, diet.

To process these data, we developed a robust data collection, curation and integration infrastructure composed of the following stages: (1) **Data Collection**: gathering raw data from different sources. (2) **Data Understanding**: characterizing the data fields, types, and values. (3) **Data Validation**: checking the data against known quantitative relationships and expected values and for consistency across data types. Examples include checking lab values against standard ranges and using system physiology models to identify potential outliers in reported caloric intake data. (4) **Data Cleaning**: cleaning, normalizing and mapping data values and tagging the data with confidence indicators. Examples include using natural language processing to extract and normalize noisy medication values and map them to standard ontologies and flagging suspicious self-reported and self-tracked data based on compliance definitions. (5) **Data Integration**: merging data from different sources, resolving ambiguities, deduplication, normalizing units and dates and linking variables to the same subject. (6) **Data Enrichment**: creating new variables from the original data that are potentially more informative. Examples include the computation of genetic risk scores from the SNPs and the computation of scores from survey responses.

The data are then stored in a database using an n-tuple structure: subject identifier, feature identifier, feature value, confidence and event date that accomodates heterogenous data and supports the data retrieval needs of analytic applications.

Results

We found that a semi-automated approach, where most of the processing is automated but unhandled issues and errors can be flagged for human intervention, was important. Because some data were human reported and entered, they contained errors. As a result, it was important to have validation methods to check, clean, normalize and map the data when possible. In addition, we tagged the data instances with confidence indicators to allow analytic applications to select appropriate data subsets for their own use. When possible, we leveraged the different data types to perform cross type consistency checking. Incremental update capability was also needed to support the addition of new data to the database. Finally, we used an iterative development and refinement process in order to accommodate enhancements resulting from new data types, new instances of existing data types and feedback from the analytic applications that consume the curated data.

Conclusions

A robust infrastructure was successfully developed and used to collect, curate and integrate EHR, genomic, microbiomic, self-reported and self-tracked data to support healthcare analytics. Future work will enhance and extend the system to handle additional types of data including metabolomics, continuous biomarker data streams and medical imaging data.

References

[1] Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D.. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. Biomed Res Int. 2014.

[2] Stonebraker M, Bruckner D, Ilyas IF, Beskales G, Cherniack M, Zdonik SB, Pagan A, Xu S. Data Curation at Scale: The Data Tamer System. CIDR 2013.

Address for correspondence:

Kenney Ng (kenney.ng@us.ibm.com)
IBM Research, 1 Rogers Street, Cambridge MA 02142, USA.