Journal of Biomedical Informatics 55 (2015) 231–236

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Federated queries of clinical data repositories: Scaling to a national network

Griffin M. Weber *

Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, United States
Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, United States

ARTICLE INFO

Article history:
Received 3 January 2015
Revised 26 April 2015
Accepted 28 April 2015
Available online 6 May 2015

Keywords:
Algorithms
Hospital shared services
Medical record linkage
Medical records systems, computerized
Search engine

ABSTRACT

Federated networks of clinical research data repositories are rapidly growing in size from a handful of sites to true national networks with more than 100 hospitals. This study creates a conceptual framework for predicting how various properties of these systems will scale as they continue to expand. Starting with actual data from Harvard's four-site Shared Health Research Information Network (SHRINE), the framework is used to imagine a future 4000 site network, representing the majority of hospitals in the United States. From this it becomes clear that several common assumptions of small networks fail to scale to a national level, such as all sites being online at all times or containing data from the same date range. On the other hand, a large network enables researchers to select subsets of sites that are most appropriate for particular research questions. Developers of federated clinical data networks should be aware of how the properties of these networks change at different scales and design their software accordingly.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Federated query tools enable researchers to search the medical records of millions of patients across multiple hospitals, while allowing the hospitals to retain control over their data. In 2008, the Shared Health Research Information Network (SHRINE) gave investigators, for the first time, access to the full patient populations at four Harvard-affiliated hospitals. Since then, multiple hospital networks have emerged throughout the United States based on SHRINE and similar platforms like PopMedNet and FACE [1–3]. The Patient-Centered Outcomes Research Institute (PCORI) has accelerated the growth of these networks by recently awarding $100 million to 29 health data networks to create PCORnet: The National Patient-Centered Clinical Research Network, which will connect around 100 hospitals across the country [4–15]. By giving investigators unprecedented access to large populations, these networks are already having an impact on biomedical research [16,17].

There is no reason to think that the growth of federated data networks will end with PCORnet. As an increasing number of health centers adopt electronic health records, someday soon nearly all 5700 hospitals in the United States may be connected to a data network. However, is the software powering these networks ready for such growth? SHRINE was originally created for four hospitals. Today, even the largest networks have only a few dozen sites. Are future networks with 100 or 1000-fold as many sites simply bigger versions of what we currently have, or will we need to approach such networks in a fundamentally different way? This study seeks to answer this question by first defining a set of attributes for evaluating federated clinical data networks, and then using this as a conceptual framework for predicting what a future 4000 site network would look like. The starting point is actual data from a four site SHRINE network at Harvard. The current Harvard SHRINE sites are Partners Healthcare (Brigham and Women's Hospital and Massachusetts General Hospital), Beth Israel Deaconess Medical Center, Boston Children's Hospital, and Dana Farber Cancer Institute.

2. Materials and methods

2.1. Conceptual framework

The purpose of the conceptual framework is not to evaluate the performance of any particular software program in terms of speed or resource requirements, but rather to determine if certain fundamental properties of a network change as the number of sites

* Address: Center for Biomedical Informatics, 10 Shattuck St, Room 316A, Boston, MA 02115, United States. Tel.: +1 617 432 6134.
E-mail address: weber@hms.harvard.edu

http://dx.doi.org/10.1016/j.jbi.2015.04.012
1532-0464/© 2015 Elsevier Inc. All rights reserved.

increases, which could affect how the networks are built or used. Eight properties are considered in this study:

1. *Functional equivalence*: Sites in a network are functionally equivalent if they can process the same types of queries, such as temporal queries or queries that require natural language processing.
2. *Temporal equivalence*: Sites that are temporally equivalent have patient data covering the same date range. "Complete coverage" means that all data for those patients are available for that date range. In other words, the patients did not receive care at facilities outside the network during that time.
3. *Data release cycle synchronicity*: Typically, hospitals do not connect their live clinical systems directly to the federated research networks. The data are first copied into separate research data repositories, which are then exposed to the network. Unless all sites update their repositories at the same time, some sites will have more recent data than others.
4. *Ontological equivalence*: Sites that are ontologically equivalent can map their local coding systems to a shared ontology (e.g., standard vocabularies).
5. *Semantic discernibility*: Even when sites use the same ontology, they might use a given code in different ways. For example, there might be a preference to use one billing code over another at a particular site, or a diagnosis date might be when the code was recorded rather than when the patient was seen. The semantic discernibility of a network describes whether these differences can be detected, either directly from the ontology or indirectly from analysis of the results.
6. *System availability*: The availability of a network is the fraction of time when sites are running properly.
7. *Population overlap*: Different hospitals might have data about the same patient. This can either lead to over-counting the number of patients in a network (e.g., two hospitals count the same patient) or under-counting (e.g., a patient matches a complex query, but no single site has enough data to know it). The more the patient populations in a network overlap, the greater the uncertainty in the results [18].
8. *Data access restrictions*: A researcher can query all sites in a network only if he or she meets all the requirements needed to access those sites (e.g., human subjects training).

2.2. Data from the Harvard SHRINE network

Data from the Harvard SHRINE network was used to predict what a future national network would look like. It is certainly a great leap to use data from only four sites to envision a network with four thousand hospitals. However, the fact that Harvard SHRINE, as one of the earliest federated networks, has had more than five years to mature means that it may be one of the best available sources from which to predict a future national network.

To study temporal equivalence, the Harvard SHRINE query tool was used to determine the number of patients with any of 40 common International Classification of Diseases (ICD-9) codes at each site by year from 2000 through 2013. The codes, which are listed in Table 1, correspond to the most frequent diagnosis categories as reported in the National Ambulatory Medical Care Survey. Because the codes cover a wide range of diseases, including both adult and pediatric diagnoses, the fraction of patients with these codes should be relatively stable over short periods of time. Therefore, if sites had complete data and were temporally equivalent, then number of patients at each site matching the 40 codes would roughly follow population growth, which was only about 10% in Boston from 2000 to 2013 [19]. Note that the purpose of this query is to estimate data completeness across all diseases over time—it does not reflect the typical use of SHRINE, which is to study a single disease.

Table 1
Top 40 ICD-9 diagnosis codes. Two frequently used ICD-9 codes in each of the top 20 primary diagnosis groups for physician office visits in the United States in 2012.

Diagnosis group	Top ICD-9 codes	
Acute upper respiratory infections, excluding pharyngitis	465.9	466.0
Allergic rhinitis	477.9	477.0
Arthropathies and related disorders	719.46	719.41
Asthma	493.90	493.92
Benign neoplasms	211.3	216.9
Cataract	366.9	366.16
Diabetes mellitus	250.00	250.01
Disorders of lipoid metabolism	272.0	272.4
Essential hypertension	401.9	401.1
Follow up examination	V67.09	V67.2
General medical examination	V70.0	V70.7
Gynecological examination	V72.31	V72.32
Heart disease, excluding ischemic	424.0	427.31
Malignant neoplasms	174.9	185
Normal pregnancy	V22.1	V22.0
Otitis media and eustachian tube disorders	382.9	381.81
Rheumatism, excluding back	729.5	729.1
Routine infant or child health check	V20.2	V20.0
Specific procedures and aftercare	V50.2	V58.66
Spinal disorders	724.2	724.5

As an example of semantic discernibility, a SHRINE query was run to determine the number of patients between 0 and 17 years old. A second query was then run to determine the number of patients between 0 and 17 years old from 2005 through 2009. This was an actual query that initially caused confusion as we were developing Harvard SHRINE. Despite each site mapping its local codes for age to the same common ontology (i.e., ontological equivalence), the query unexpectedly returned wildly different results across sites. This was later discovered to be due to subtle differences in how sites interpreted this query, rather than true differences in patient populations.

The Harvard SHRINE network has an automated monitoring tool that sends a test query to each site every two hours and generates an email alert if a site did not respond. All email alerts from 1/1/2013 through 12/31/2013 were collected to determine the availability of each site's system.

3. Results

3.1. Functional equivalence

The four Harvard SHRINE sites use an open source clinical data repository platform called Informatics for Integration Biology & the Bedside (i2b2). Since Harvard SHRINE's launch in 2008, i2b2 has had five major software updates (versions 1.3 through 1.7), or approximately one per year. Each site has its own timeframe for updating the software, and nationally there are many sites still using version 1.3. In just a four site network, if each version is equally likely, the probability that all sites are using the same version is just $0.2^3 = 0.008$. With 4000 sites, the probability of functional equivalence is negligible. Also, i2b2 is just one of many similar software programs used across the country, which makes it even less likely that all sites in a large network can support the exact same types of queries.

3.2. Temporal equivalence

PCORNet requires sites to identify patients with "complete data" over a longitudinal timespan; and, the Harvard SHRINE website states that it has a "complete set" of diagnosis data from each of its participating hospitals, starting from January 1, 2001.

Download English Version:

https://daneshyari.com/en/article/6928188

Download Persian Version:

https://daneshyari.com/article/6928188

Daneshyari.com