Web Semantics: Science, Services and Agents on the World Wide Web 8 (2010) 342–347

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Invited paper

TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records

Chintan Patel*, Karthik Gomadam[1], Sharib Khan[1], Vivek Garg[2]

Applied Informatics Inc., New York, NY 10006, United States

ARTICLE INFO

Article history:
Received 10 February 2010
Received in revised form 25 June 2010
Accepted 10 August 2010
Available online 15 September 2010

Keywords:
Clinical trials
Healthcare informatics
Web 2.0
Semantic web applications

ABSTRACT

Clinical trials form a critical link in the translation of basic biological research into routine clinical practice. However, finding eligible participants for clinical trials is a critical hurdle and a frequent cause for delays in the completion of trials. The lack of comprehensive, efficient, and consumer-friendly online tools has limited consumers' ability to pro-actively find the right clinical trials and participate in them. The rising use of Internet as a medium for seeking health information and the advent of online Personal Health Records (PHRs) platforms such as Microsoft HealthVault (MHV) and Google Health (GH) have created the opportunity to change the status quo. These platforms are connected to databases of several healthcare organizations and enable patients to import their medical record and utilize it for self-management by using third-party applications built on these platforms.

In this paper we discuss TrialX, a consumer-centric tool that matches patients to clinical trials by extracting their PHR information and linking it to the most relevant clinical trials using semantic web technologies. We use an ontology, HealthOnt, that provides the underlying representation for integration and semantic retrieval of health data and clinical trials. TrialX is currently live and integrated with both MHV and GH and has matched thousands of patients to clinical trials in the last 12 months.

1. Introduction

Clinical trials are medical research studies performed on humans to evaluate the safety and efficacy of new drugs.[3] They form a critical and essential phase in the lifecycle of drug development. Every year more than 50,000 clinical trials are conducted in the United States alone. However, finding the right participants (patients or healthy volunteers) has been and continues to be a major bottleneck in the timely completion of these trials. Four out of five trials are delayed and among these, 50% of the trial delays are due to participant recruitment challenges [1,2]. The situation is particularly troublesome in the domain of oncology, where fewer than 3% of potentially eligible cancer patients enroll in trials [3]. In fact more than 75% of participants are not even aware of them, even though surveys have repeatedly shown that a majority of patients would be open to participating in trials if they had the right information.[4]

From a social and public health perspective, delays in clinical trials have several adverse consequences. One, it entails the loss of human life that could otherwise have benefited from the new intervention (particularly with conditions such as advanced cancers where experimental treatments are the last resort for some patients). The delay in trial completion postpones the entry of a drug in the general market thus postponing its benefits to the large population. Another consequence of the low recruitment rates is that finding patients for trials of uncommon conditions can be challenging and at times impossible, causing the failure of drugs to treat such conditions to make it to the market. The delay has economic consequences too; since a drugs patent period is limited, every days delay in clinical trial results leads to lost revenues in the range of $1–$8 million per day for pharmaceutical companies [2]. In total, delays in clinical trials result in over $10 billion/yr in estimated missed revenues for pharmaceutical companies.

Two fundamental hurdles prevent increased participation of patients in clinical trials.

1. The current process of finding eligible participants is largely driven by clinical trial investigators (physicians charged with conducting the trial) or specialized recruitment professionals with limited options for patients to be pro-active in the process of finding and enrolling in trials.

* Corresponding author.
E-mail addresses: chintan@trialx.com (C. Patel), karthik@trialx.com (K. Gomadam).

[1] Work done as consultant to Applied Informatics Inc.
[2] Work done when previously employed at Applied Informatics Inc.
[3] http://en.wikipedia.org/wiki/Clinical_trial Accessed October 03, 2009.
[4] http://www.ciscrp.org/information/facts.asp.

1570-8268/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2010.08.004

2. The existing participant-oriented tools such as websites that list clinical trials or websites designed to signup patients for clinical trials (such sites are called eRecruitment sites), have several limitations; they lack user-friendliness [4] and many patients dont know enough details of their health conditions (such as what medications they are taking) to find the most relevant trials from the thousands of trials active at any time.

1.1. Impact of technology and web

Two significant trends are converging to create an opportunity for disruption in the current patient recruitment landscape and to make it possible to create the kind of solution that addresses the problems described above. One is the rise of Internet as a health information-seeking medium for consumers. The data from the Pew Internet and American Life Project shows that the Internet has become a common source of health information for a large number of people; these studies estimate that between 75% and 80% of Internet users search for health information online, with a steady upward trend in this number [5]. On average, 8 million people search for health related information online every day. Websites such as WebMD are visited by millions of health consumers every month. In addition to general health searches, clinical trial information (or information about new treatments) ranks as one of the important health topics that users search for on the Web. According to the National Library of Medicine, its clinical trials listing website, ClinicalTrials.gov, receives close to 65,000 unique visitors a day and serves more than 50 million page views per month.[5]

The second converging factor is that the healthcare industry is on the verge of an unprecedented adoption of Healthcare Information Technology (HIT) solutions such as Electronic Health Records (EHRs) and Personal Health Records (PHRs). The recent American Recovery and Reinvestment Act of 2009, has specifically allocated more than $20 billion for incentivizing EHR adoption [6]. Riding this thrust in HIT and the consumer health trend, technology corporations including Google and Microsoft are building sophisticated PHR platforms (Google Health[6] and Microsoft HealthVault,[7] respectively) to enable consumers to have access to their health information from organizations such as Cleveland Clinic and be able to use that information for self-management of their health conditions.

1.1.1. Personal Health Records and their impact

The Markle foundations connecting for health collaborative, has defined a PHR as an "electronic application through which individuals can access, manage and share their health information and of others for whom they are authorized, in a private, secure and confidential environment" [7]. The US Health and Human Services and the Office of the National Coordinator for HIT have identified PHRs as a potentially disruptive technology and a national priority [8]. These PHR platforms could be the enablers of a radical paradigm shift in the health information economy; the shift of locus of control of health information to the patient (consumer).

The central goal of the new PHR platforms, GH and MHV is to let patients use their health information for better self-management by using third-party application services built on these PHRs. Third-party PHR applications can re-use the data within a patients PHR (with due patient consent and with required measures to protect patient privacy) and provide personalized and useful services to the patient. The provision in these PHRs to enable patients to control what they can do with their health information and who they can share the information with, has been described as a "tectonic shift in the health information economy" by Mandl and Kohane [9]. Such a platform provides an excellent opportunity to use semantic technologies to utilize the information in the PHR to match patients to relevant clinical trials. In fact, one of the important PHR use cases described by Mandl and Kohane, is the possibility of patients using their PHR information to find clinical trials that match their health profile. If realized, this new PHR model can have a potentially dramatic effect on addressing the current roadblocks that exist in utilizing medical records for research purposes, as patients can now give their consent to third-party application providers such as a clinical trial matching service, as described in this paper.

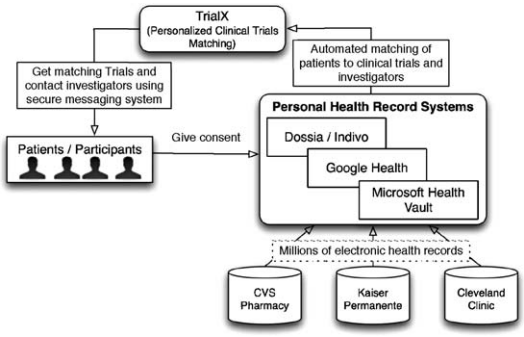

Fig. 1. Information flow between PHRs and TrialX.

We present the details of our system TrialX, a consumer-centric application that is built on top of PHRs and enables patients to discover trials that match their health conditions and connects them with the trial investigators. By leveraging PHRs, TrialX is poised to significantly alter the dynamics of the clinical trials recruitment landscape. PHRs could be adopted by millions of consumers in the coming years, thus significantly increasing awareness and outreach for clinical trials recruitment. If successful, our PHR-based approach could have major public health impact by increasing access to new treatments, reducing the delays in getting drugs to market and improving the likelihood of finding patients for trials of rare conditions. Moreover, the ability to pull patient-consented PHR information creates the possibility of mining de-identified data for obtaining surveillance statistics on clinical trials recruitment, or beyond, for surveillance of adverse events related to clinical trials or post-market drug surveillance.

2. TrialX information flow

Fig. 1 describes the overall information flow within the TrialX system. TrialX is integrated with MHV, GH and Indivo PHRs. These systems in turn are integrated with databases of several healthcare organizations such as hospitals (Cleveland Clinic, NewYork-Presbyterian) and pharmacies (CVS). A patient can import his or her health record from any of the partner organizations of these platforms. After importing, they can choose to allow third-party applications built on these platforms (such as TrialX) to use their record to provide useful services. For example, an application in HealthVault can read the height, weight, age and gender of a patient and calculate their Body Mass Index (BMI) and explain the significance of the same to the patient.

We illustrate this integration with an example from Google Health (as shown in Fig. 2. The Google Health platform allows patients to add healthcare applications that utilize the PHR information. Patients can add the TrialX application from within the Google Health platform and after adding the application can generate matching trials by clicking the *show matching trials* link. This is illustrated in Fig. 2.

[5] http://clinicaltrials.gov/ct2/info/about.
[6] http://google.com/health. Accessed October 01, 2009.
[7] http://www.healthvault.com. Accessed April 04, 2009.

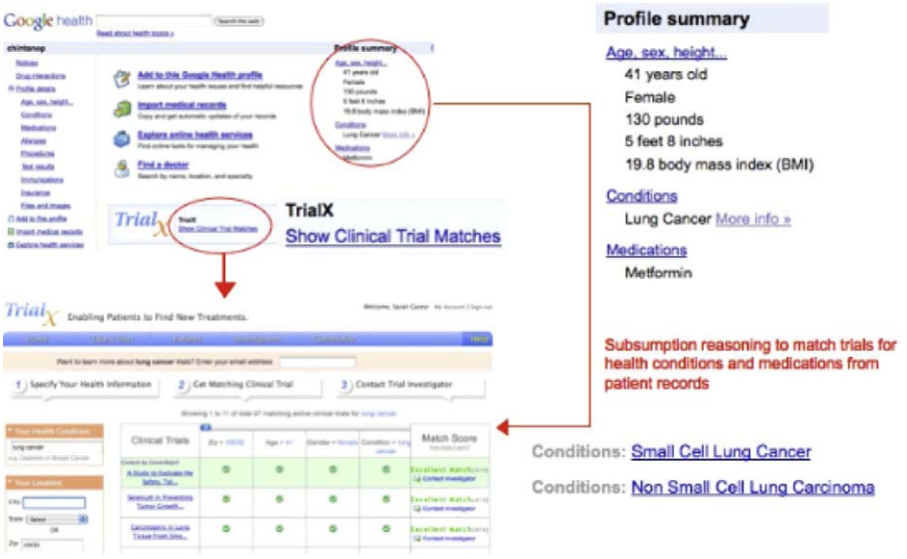

Fig. 2. Integration between Google Health and TrialX.

The TrialX system works in the backend to pull the patients condition, demographics and other information securely and uses this information to generate a list of matching trials on TrialX.com. The next section describes the components used in TrialX and the matching algorithm.

3. Semantics driven TrialX architecture

To enable integration, representation and subsequent querying of diverse data sources of patient records and clinical trials, TrialX uses three key components that are based on semantic web technologies.

3.1. Scalable semantic health data model

The representation of health care data is a challenging problem. The sheer number of data-types such as conditions, medications, allergies, laboratory results and corresponding values such as diabetes, metformin and blood glucose is enormous. To handle this problem, we developed *HealthOnt*, a generic top-level ontology to represent the health care data-types and values (see Fig. 3).

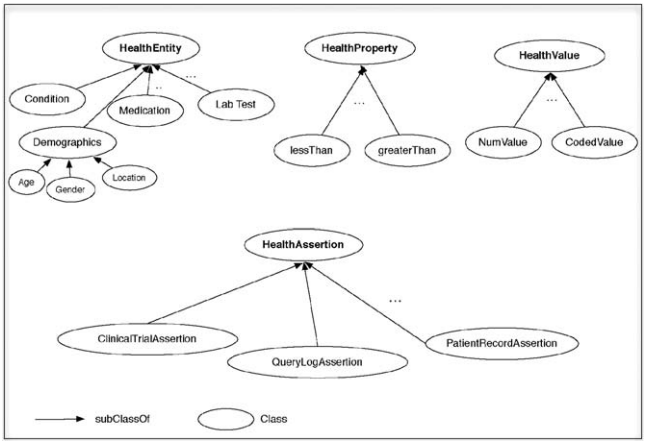

Fig. 3. HealthOnt ontology. The HealthEntity, HealthProperty and HealthValue are top-level concepts that are combined into an EAV-triplet under HealthAssertion. An example of PatientRecordAssertion: *Medication*, *taking*, *Vancomycin* and ClinicalTrialAssertion: *Medication*, *notTaking*, *Antibiotics*.

HealthOnt uses an Entity-Attribute-Value (EAV) based top-level representation to encode *HealthEntities* such as Condition, Medication, Laboratory Tests. The attributes (*HealthProperty*) include concepts such as lessThan, greaterThan, taking, notTaking and so on and these are used to represent granular information about an entity. For example an entity such as 'Age' can have an attribute 'lessThan' with value '65'. The *HealthValue* represents the actual numeric value or coded value from an external ontology such as the Unified Medical Language System (UMLS), RxNorm or International Classification of Diseases (ICD). These EAV triplets are used to create a *HealthAssertion*, which can be *PatientRecordAssertion*, for representing patient data from PHRs, *ClinicalTrialAssertion* for modeling the structured eligibility criteria, *QueryLogAssertion* for storing intermediate website logs and so on. The key idea is that this representation allows scalable representation of healthcare information obtained from patient records and clinical trial data sources for different tasks as follows:

1. *HealthOnt for semantic matching*. HealthOnt facilitates direct semantic matching of the patient data and clinical trial information since both share the same top-level concepts. Before the matching can occur, an intermediate transformation step is required and this is described in the next section on data integration.
2. *HealthOnt and RESTful API*. Each HealthEntity is associated with a REST-based web query parameter that allows direct translation of web requests to internal data-structures and algorithms for processing the requests through TrialX APIs.
3. *HealthOnt for dynamic web forms*. The HealthEntity are also associated with a *DisplayEntity* to enable generation and validation of dynamic web forms to collection patient information. TrialX allows domain experts to create eligibility questionnaires on the fly using this extensible representation.

3.2. Semantic data integration

The integration of diverse data sources is one of the core challenges addressed in the TrialX application. The patient data arrives from different sources such as PHRs (GH, MHV, Indivo), EHRs, and various health and wellness websites. On other hand, the clinical trial information is integrated from sources such as ClinicalTrials.gov, CenterWatch.com, Clinical Trial Management Systems in hospitals and so on. To handle multiple, growing data sources for

Table 1
The matching of patient record (*PatientRecordAssertion*) against structured clinical trial criteria (*ClinicalTrialAssertion*). The concepts from *ClinicalTrialAssertion* are semantically expanded to include all the children in the ontology to determine the match, the concept *Multiple Myeloma* is expanded to match *Stage II multiple myeloma* and *Revlimid* is identified as a type of *Immunomodulator*.

HealthEntity	PatientRecordAssertion	ClinicalTrialAssertion	Match
Age	51 yrs	18–100 yrs	Yes
Condition	Stage II Multiple Myeloma	Multiple Myeloma	Yes
Activity Level	Fully Active	NA	Yes
Received Previous Drug	Revlimid	Immunomodulators (exclusion criteria)	No
Zip	07310	10032 (25 miles)	Yes

obtaining patient records and clinical trials, we perform integration in 2 steps:

1. *HealthOnt Transform Layer*. Firstly, a wrapper is created for each source data system (PHRs, EHRs etc.) and the data items are mapped to the corresponding one or more *HealthEntity* in TrialX or new entities are added if required. After the mapping, the common TrialX REST API is used to POST the data processed by the wrapper to store the new data in TrialX. The patient data and clinical trial data are transformed into *PatientRecordAssertions* and *ClinicalTrialAssertions* respectively.
2. *Semantic Mapping Layer*. After the data is stored in TrialX, the value fields in the HealthAssertions are mapped to the UMLS concepts using lexical matching tools[10]. The UMLS integrates about 9 million terms from 140 biomedical terminologies into 2 million unique concepts (a unit of meaning). The UMLS also enables mapping codes across different terminologies. We use these cross-terminology mappings to directly transform coded data values from the input source.

3.3. Real-time semantic matching and query

At the heart of the TrialX platform is the matching module. This module performs full clinical record based computational matching of participants with clinical trials using semantic and natural language processing (NLP) techniques. We analyze key sections of the patient's health record including demographics, health conditions, laboratory results, medications, and procedures. This information is used in matching the patient with the eligibility criteria information obtained from a trial description. One of the key features of the matching is that it is performed at a semantic concept level (biomedical meaning) rather than checking for absence or presence of a criterion at the lexical level. Consider a patient taking the drug *Vancomycin*. This patient will be matched to a clinical trial that requires patients taking *antibiotics*. This is because, the matching algorithm can infer using semantic conceptual knowledge that *Vancomycin* is an *antibiotic*. By using a semantic approach to trial matching, we provide more meaningful results as compared to lexical (term level) search results. The matching is done by comparing the HealthAssertions of each patient medical record with the assertions for the database of assertions for the trial criteria (see example in Table 1)

4. Semantic search versus keyword based search

There are several ways in which the results of a semantic-based search as performed by the TrialX system differs from a traditional keyword based search. To illustrate the differences we provide the results of a search performed for a hypothetical patient record using TrialX and ClinicalTrials.gov, consider the following patient scenario:

55 y.o. male with lung cancer, living in Baltimore, Maryland (USA) looking to enroll in a clinical trial at the Johns Hopkins Medical Center.

ClinicalTrials.gov search parameters: We entered the keywords lung cancer, location term as Johns hopkins, country as United States and restricted the search to open studies only (currently active studies). ClinicalTrials.gov offered no fields to enter patient demographics. *TrialX search parameters*: We entered medical condition = lung cancer, demographics such as age = 55, gender = male, city= Baltimore and State = MD and study site = Johns hopkins.

We manually evaluated the accuracy of the retrieved clinical trials from ClinicalTrials.gov and TrialX. The ClinicalTrials.gov returned 30 clinical trials and TrialX generated 5 matching clinical trials. All 5 studies produced by TrialX were also present in the (top 6) CT.gov results. The additional results (25 studies) generated on ClinicalTrials.gov were excluded from the TrialX search results for several reasons. Nine trials in the ClinicalTrials.gov set contained trials with conditions that were either broader or narrower than 'lung cancer'. Additionally, the pure keyword based searching in ClinicalTrials.gov produced several false positives (43%). For example, a false positive trial was shown because it matched the keyword 'lung' in sponsor institution field, *National Heart, Lung, and Blood Institute* and 'cancer' on *National Cancer Institute*. Three trials in the ClinicalTrials.gov data set were not included in the TrialX set because they did not contain any email address (this is a technical exclusion because TrialX only stores trials that one at-least one email contact associated with them). The email information is used to enable patients to contact the investigator of the trial through the TrialX messaging system.

5. User interface and visualization

Complexity of medical information can often be overwhelming to users. Rather than just listing a set of matching trials as blue links on a white page, TrialX results are shown in a matrix that highlights the match across different trial criteria and patient conditions. This view allows users to get detailed description of the match results and allows them to explore/tune their search by adjusting the values for the different facets. These facets might vary across different conditions and are captured in the TrialX semantic model. An illustration of the matrix view is provided in Fig. 4. The order of columns in this view is determined by the importance of a particular attribute to the condition for which the user is searching for trials.

6. Sharing the goodness

TrialX provides multiple ways for third-party websites to access clinical trial information. We have created a RESTful API through which clinical trials information can be obtained in an RDF output. The RDF information allows general health and wellness websites or bloggers to incorporate clinical trial information enriched with semantic metadata. For example, Web resources that have content on diabetes would automatically be able to pull related clinical trial content from TrialX. Fig. 5 illustrates the RDF export of an Asthma clinical trial.

TrialX also provides a widget generation service that uses mined trial information along with external semantic knowledge to identify the minimum matching information for a clinical trial. The user interface elements (a dropdown for trial phase) are cap-

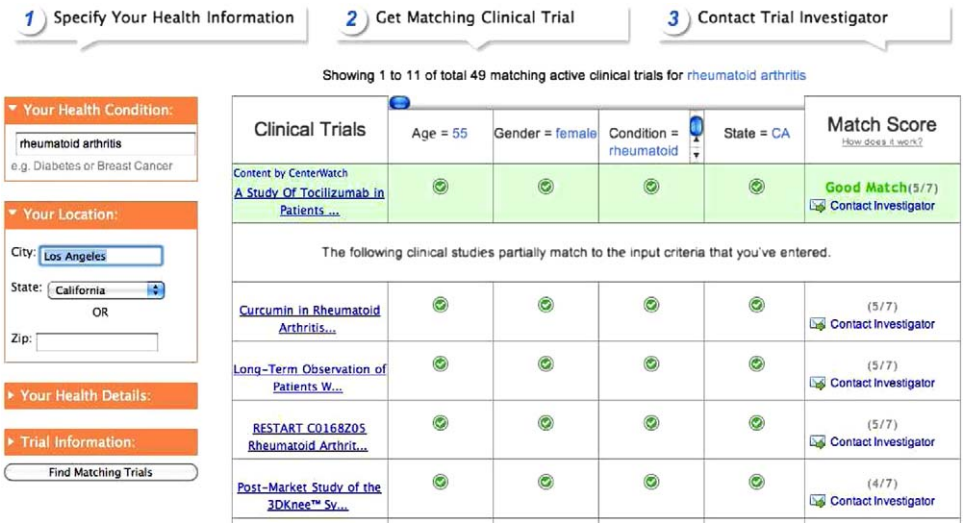

Fig. 4. Search results matrix view.

tured into the triple store and are incorporated into the widgets based on the requirements. This service can be accessed from http://trialx.com/widget. This widget can also be embedded across other Web applications. Fig. 6 illustrates a widget generated for diabetes.

TrialX also has an iPhone application and we are currently developing an enterprise grade health care platform that can be deployed in clinical sites.

7. Discussion and conclusion

In this paper we have described the need for building online consumer-centric technologies to connect patients to relevant clinical trials. We are leveraging paradigm changing technologies such as PHRs in combination with semantic web technologies to match patients based on the information in their record and the information provided for a clinical trial.

The approach utilized to build TrialX provides a mechanism for connecting patients to relevant and personalized information. We have created a fundamentally different type of solution that is designed to empower patients and give them the tools to learn about clinical trials, obtain recommendations for the most suitable trials that match their health conditions without having to fill complex questionnaires and lets them have direct access to the clinical trials investigator, without having to go through multiple layers or navigate complex and poorly designed sites.

7.1. Impact of TrialX

We envision that TrialX will have a positive impact at two different levels. One is the immediate impact that a PHR-based clinical trials recruitment solution can have on increasing access to new treatments and reducing the time needed to find eligible patients, thus speeding up clinical trials and the availability of new treatments. Another major impact of such a solution would be its potential to make clinical trials of rare conditions more viable; several trials of rare condition fail to complete because of the lack of reach to such populations or because the cost of doing so with current methods is prohibitive.

Beyond the immediate benefits, TrialX and other PHR based applications in general open up the possibility of new data sources for population-based surveillance. One of the key functions and core components of a public health agency is surveillance and epidemiology. Several public health agencies routinely conduct

```
<owl:Thing rdf:about="http://trialx.com/#913"/>
- <rdf:Description rdf:about="http://trialx.com/#Asthma">
  - <rdf:type>
      <owl:Class rdf:about="http://trialx.com/#Title"> </owl:Class>
    </rdf:type>
    <rdfs:label>Persistent Allergic Asthma Trial CIGE025AUS33</rdfs:label>
  </rdf:Description>
- <rdf:Description rdf:about="http://trialx.com/#CUI391267">
  - <rdf:type>
      <owl:Class rdf:about="http://trialx.com/#Conditions"> </owl:Class>
    </rdf:type>
    <rdfs:label>Persistent Asthma</rdfs:label>
  </rdf:Description>
- <rdf:Description rdf:about="http://trialx.com/#NCTID193">
  - <rdf:type>
      <owl:Class rdf:about="http://trialx.com/#ClinicalTrial"> </owl:Class>
    </rdf:type>
  </rdf:Description>
- <rdf:Description rdf:about="http://trialx.com/#site1524">
  - <rdf:type>
      <owl:Class rdf:about="http://trialx.com/#SiteName"> </owl:Class>
    </rdf:type>
    <rdfs:label>Bernstein Clinical Research Center, LLC</rdfs:label>
  </rdf:Description>
```

Fig. 5. RDF export of Asthma trial information.

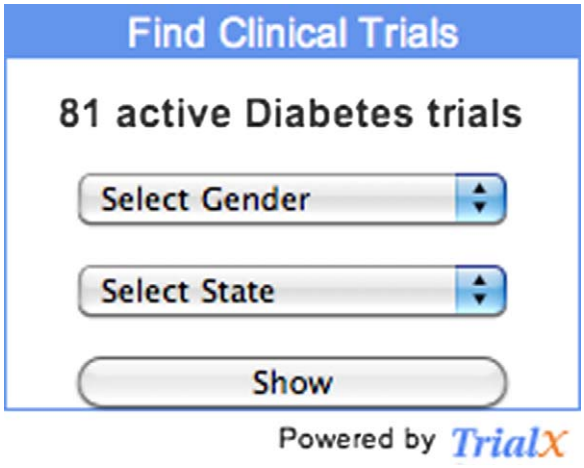

Fig. 6. TrialX widget for diabetes.

surveillance for several notifiable conditions. Some do more specialized surveillance in the areas of chronic diseases, syndromic surveillance and behavioral risk factors assessments (such as the Behavioral Risk Factor Surveillance System surveys conducted by the CDC and the state public health agencies). An emerging area of epidemiology and surveillance is Infodemiology and Infoveillance. Infoveillance, is defined as the epidemiology of health information needs and entails tracking trends in health information search behavior across different conditions, demographics and geography. The concept has being utilized to trend, in real-time, the spread of Influenza in a given population and nationally though Google Flu Trends.[8] It is possible to apply the concept of Infodemiology to do surveillance of clinical trial recruitment information and search behavior. For instance, we can measure metrics, based on user search behavior, on how many patients search for trials of a given condition, stratified by demographics and geography. Similar metrics can be created to track enrollment status of patients in different trials. To our knowledge, this would be the first of its kind real-time data related to consumers information needs about clinical trials and recruitment, thus providing a new type of infoveillance. Collecting such data would provide a new method to use PHR information for public health surveillance. Since PHRs could be potentially used by millions of consumers in the years to come, it provides an unprecedented amount of data that can be reused (after proper de-identification) for public health surveillance purposes.

The system has received enthusiastic response from the users and is actively growing. It provides the framework and the core technologies that can be used to extend the concept a generalized system HealthX, for connecting health consumers to personalized health information resources based on their PHR.

References

[1] Centerwatch, An Industry in Evolution, Centerwatch, 2003.
[2] M. Barrett, eRecruiting For Clinical Trials, Forrester.
[3] R. Winn, Obstacles to the accrual of patients to clinical trials in the community setting, Semin. Oncol. 21 (4) (1994) 112–117.
[4] V. Monaco, S.K. Krills, Online information about cancer clinical trials: evaluating the web sites of comprehensive cancer centers, AMIA Annu. Symp. Proc. (2003) 470–474.
[5] L.R. Susannah Fox, The online health care revolution: how the web helps Americans take better care of themselves, Pew Internet and American Life Project.
[6] D. Blumenthal, Stimulating the adoption of health information technology, N. Engl. J. Med. 360 (15) (2009) 1477–1479.
[7] M. Foundation, Connecting Americans to their healthcare, Markle Foundation, Connect. Health (2004) 48.
[8] P. Tang, J. Ash, D. Bates, J. Overhage, D. Sands, Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption, J. Am. Med. Inform. Assoc. 13 (2006) 121–126.
[9] K.D. Mandl, I.S. Kohane, Tectonic shifts in the health information economy, N Engl. J. Med. 358 (2008) 1732–1737.
[10] D.A. Lindberg, B.L. Humphreys, The unified medical language system, Methods Inf. Med. 32 (1993) 281–291.

[8] http://www.google.org/flutrends/.