Comparing the Effectiveness of a Clinical Registry and a Clinical Data Warehouse for Supporting Clinical Trial Recruitment: A Case Study

Chunhua Weng, PhD[1], J Thomas Bigger, MD[2], Linda Busacca, BA[3]
Adam Wilcox, PhD[1], Asqual Getaneh, MD, MPH[2]
[1]Department of Biomedical Informatics; [2]Department of Medicine; [3]The Clinical Trials Office
Columbia University, New York, NY, 10032

Abstract

This paper reports a case study comparing the relative efficiency of using a Diabetes Registry or a Clinical Data Warehouse to recruit participants for a diabetes clinical trial, TECOS. The Clinical Data Warehouse generated higher positive predictive accuracy (31% vs. 6.6%) and higher participant recruitment than the Registry (30 vs. 14 participants) in a shorter time period (59 vs. 74 working days). We identify important factors that increase clinical trial recruitment efficiency and lower cost.

Introduction

As recently pointed out by Dr. Barbara Alving, Director of The National Center for Research Resources (NCRR), "*participant recruitment continues to be a significant barrier to the completion of research studies nationwide — recent NIH data indicates that just 4% of the U.S. population has participated in clinical trials.*"[1] Effective use of electronic patient information to identify potentially eligible clinical trial participants has great potential for streamlining the national clinical research enterprise.

Personal health records are used increasingly to match patients to clinical trials, as exemplified by TrialX (http://trialx.com/). ResearchMatch (https://www.researchmatch.org/), a large national research registry, was also established to boost clinical trial participation. Both methods use minimal patient information to match patients to studies. They facilitate two-way communications between research teams and patients, but have not yet leveraged the huge amounts of electronic patient information in clinical registries, electronic health records (EHR), or clinical data warehouses to realize the great potential for electronic screening (E-screening) to identify trial participants more efficiently and at lower cost.

Creation of clinical registries has long been a standard procedure for quality improvement for chronic diseases, such as hypertension, diabetes, and cancer. Such registries allow clinicians to efficiently monitor and treat patients with specific diseases by keeping current information for a relatively narrow range of key disease phenotypes. They often do not collect and provide access to detailed data for individual patients. If their variables are aligned with research recruitment criteria, registries can be more efficient for identifying research participants than classic retrospective screening activities, such as reviewing paper records or EHRs. Therefore, registries aligned with research aims should expedite the identification of potential participants, utilizing less time from the research team.

Meanwhile, as more institutions adopt EHR, clinical data warehousing has become a more popular data integration technology for supporting intelligent data analysis and strategic business decisions. A clinical data warehouse can organize data from disparate EHRs that reflect many aspects of an organization's operations, into a standardized data source to facilitate aggregated queries of large patient populations. We have previously demonstrated the value of using our Clinical Data Warehouse to improve recruitment efficiency of a multi-site, randomized clinical trial, ACCORD.[2]

At present, there is no standard way of using these technologies for improving clinical trial recruitment. In this study, we report one of the earliest case studies comparing the efficiency of a Diabetes Registry with a Clinical Data Warehouse for recruiting patients for an ongoing clinical trial, with a goal to better understand the tradeoffs in existing clinical trial recruitment methods and identify best practices to support clinical research.

The Case Study

1. The Clinical Trial: TECOS

The Trial Evaluating Cardiovascular Outcomes with Sitagliptin (TECOS)[3] is a large (N ≈ 14,000) multinational, placebo-controlled, double-blind, randomized, parallel-group clinical trial. TECOS is conducted at the New York Presbyterian Hospital Ambulatory Care Network (ACN)[4], which, through six primary care clinics, provides adult, pediatric and Ob/Gyn services to Northern Manhattan. One clinic

serves primarily the geriatric population. TECOS eligibility criteria are listed below.

Table 1. Variables Available for Electronic Query (R: Registry, W: Warehouse; 1: Present; 0: Absent)

TECOS Inclusion/Exclusion Variables	R	W
Inclusion Criteria		
1. Age >= 50	1	1
2. Type 2 diabetes mellitus	1	1
3. A1C 6.5 - 8.0 %	1	1
4. pre-existing ischemic vascular disease	0	1
5. stable diabetes regimen	0	1
Exclusion Criteria		
1. Type 1 diabetes	1	1
2. Insulin or sitagliptin therapy	0	1
3. Cirrhosis of the liver	0	1
4. Known allergy or intolerance to sitagliptin	0	1
5. Enrolled in another experimental protocol	0	0
6. Planned revascularization procedure	0	0
7. Medical condition that limits life expectancy /pose risk to the patient/patient cannot comply	0	0
8. GFR of <30 mL/min/1.73 m (calculation= serum creatinine via the MDRD)	0	0

It should be noted that it is not just the number of the variables that needs to be considered but also the data density of the variable in the data set that determines the magnitude of its effect at identifying potentially eligible participants.

2. Data Collection

The Columbia University Medical Center Clinical Data Warehouse was established in 1994 and contains longitudinal medical records for about 2.7 million patients seen at the NewYork Presbyterian Hospital, including 15,172 diabetic patients in ACN clinics. It is equipped with an advanced data warehousing design[5] and informatics tools for semantic integration[6]. It contains rich information, including lab test results, imaging reports, and ancillary clinical notes, to facilitate patient care, administration, and clinical research. The ACN Diabetes Registry was created in 2005 to support the diabetes disease management effort at the ACN. It lists about 5,000 diabetes patients receiving care at the primary care clinics. For each patient, the Registry provides the following key variables used for diabetes care quality improvement: dates and values for A1C, urine micro- albumin, and LDL cholesterol. The diabetes registry is updated quarterly using data from our Clinical Data Warehouse.

Given their familiarity with the clinical uses of the Registry and prior experience with labor-intensive manual screening, the TECOS investigators at the ACN first used the Registry starting 7/16/2009. Although better than completely manual screening, the Registry generated more cases (n = 2,033) than the research team could efficiently review. Consequently, the investigator sought other methods to improve screening efficiency and found, through a consultation with a colleague, the Warehouse and began this assisted screening on 11/1/2009. Additionally it should be noted that although almost all of the inclusion criteria variables were present in both repositories (the Warehouse contained all five whereas the Registry three), the exclusion criteria were not similarly represented (the Warehouse had more than twice as many of these variables as the Registry).

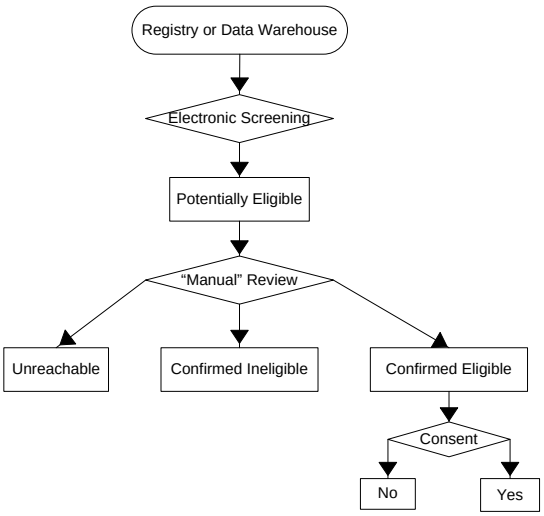

Figure 1. Determining eligibility and obtaining consent.

Potentially Eligible patients are those that met all the eligibility criteria during E-screening. The research team then checked exclusion criteria using data that were not queriable in the two digital sources. Patients who were still eligible after all records were reviewed were interviewed to determine eligibility and to obtain informed consent. Possible outcomes of this process are: ***Unreachable*** - Patients who were eligible by electronic query, but were not available for either "manual record review" or patient interview; ***Confirmed Ineligible*** – Patients excluded by "manual record review" or patient interview; and ***Confirmed Eligible*** – Patients who met all the protocol criteria and as determined by the research team. The consent process divided confirmed eligible patients into two categories: ***No*** – patients who were eligible, but declined to participate or ***Yes*** – Patients who were eligible and agreed to participate in TECOS.

The Registry query included the criteria "age ≥ 50", A1C (6.5-8.0%), and ICD-9 codes for ischemic vascular and other arterial diseases and symptoms (i.e., 412, 413, 414, 447.1, 414, 433, 437.9, 435.9,

411.81, V00.61-66, and V 12.54), and a list of clinics for which the investigator had recruitment permission. The Warehouse query applied not only all these variables but also others not available in the Registry (see Table 1).

Results

1. Comparative Screening Efficiency

Table 2 shows the eligibility status for patients identified using the Registry or the Warehouse. The Registry contained 2,033 potentially eligible patients of whom, the research team was able to manually review only 437 between 7/16/2009 and 10/31/2009 due to the time-consuming review process. In contrast, the Warehouse query retrieved only 100 "potentially eligible" patients from 15,172 diabetics in our Data Warehouse. All of the 100 patients were manually reviewed by the research team between 11/01/2009 and 01/31/2010.

Table 2. Comparative Screening & Enrollment Results.

Patient Eligibility Status	Registry	Warehouse
Initial Population	N ≈ 5,000	N ≈ 15,172
Potentially Eligible	N=2,033	N=100
Manually Reviewed	N=437	N=100
Confirmed Ineligible	355 (81.2%)	48 (48%)
Confirmed Eligible	29 (6.6%)	31 (31%)
Unreachable	43 (9.8%)	19 (19%)
Did Not Consent	10 (2.3%)	2 (1.7%)

The warehouse yielded significantly fewer "potentially eligible" patients than the Registry (100 vs. 2,033), but higher true positives (31% vs. 6.6%) and unreachable (19% vs. 9.8%) proportions. The Registry generated higher false positives than the Warehouse (81.2% vs. 48%). Nineteen (19%) patients identified from the Warehouse were unreachable, in contrast to 9.8% patients from the Registry who were unreachable.

2. Comparative Recruitment Efficiency

Recruitment for TECOS began in December of 2008. Site 110 at Columbia University was activated on 8/17/09 and enrolled her first participant on the 27th of that month. A necessary prerequisite to site activation was the compilation of a screening log of potential participants. This screening log was compiled using the registry. The research team recruited 1 participant August; 7 in September; and 6 in October. After 11/1/09, site 110 was using the Warehouse query exclusively.

The eligibility E-screening process for TECOS involved a query of either the Registry or the Warehouse during two consecutive 3-month periods, as illustrated by Figure 2. (Of note, this case report focused on the first 5.5 months of the ongoing recruitment period.) Figure 2 shows that **14** enrolled and randomized participants were identified by the registry query, while **30** were randomized using the Warehouse query.

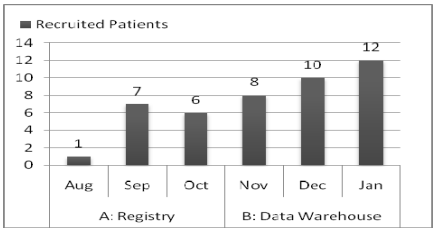

Figure 2. Monthly Recruited Participants Using the Registry (Aug'09-Oct'09) vs. Using the Warehouse (Nov'09-Jan'10).

The number of working days spent on screening and recruiting was 74 on the Registry list, and a total of 59 days spent on the Warehouse list were included in this study. These numbers reflect the differences in the number of holidays between two time periods. The recruitment rate was 1 patient per week using the Registry and 2.5 patients per week using the Warehouse.

When site 110 became active, TECOS had 49 US clinical sites enrolling with a total of 195 participants recruited and 43 worldwide sites enrolling with a total of 70 participants recruited. Site 110 began its enrolling activities almost 9 months after the start of TECOS, when 92 competing sites were active. In less than a year site 110 ranked first in the US and third worldwide in recruitment. We believe that the recruitment strategies used by site 110 enabled them to excel.

Discussion

1. Pros and Cons of Registry and Warehouse

Unlike the Data Warehouse, registries are generally created for quality improvement and clinical care purposes. The Diabetes Registry was seen by the TECOS investigators as a substantial improvement over manual chart review because it consolidated relevant clinical information for patients with diabetes – the focus of TECOS– in a single accessible view. The clinicians' familiarity with the Registry facilitated the initial interaction. However, the time saved in easily accessing a list of "potentially eligible patients" was mitigated by the burden introduced by the high rate of false positives (generated by the lack of exclusion criteria subject to query).

The Warehouse query produced more true positives but the investigators were less confident using the query procedures, including requesting resources and certifying authorization for access. A warehouse query also requires more sophisticated query designs for selecting appropriate data sources (structured vs. unstructured, in-patient vs. out-patient records) for eligibility determination than the simple in-house Registry search. Although in the end, the Warehouse required less work from the research team, the clinician researcher would still naturally prefer the more accessible Registry to the Warehouse. This underscores the importance of building an infrastructure for clinical research on the primary care level that will include the availability of Warehouse resources, the alignment of protocol specifics into Registry creation and the training of research personnel in recruitment strategies.

2. ICD-9 Diagnoses

A total of 16 patients identified using the Warehouse (16%) and 48 (18%) from the Registry were false positives because of inappropriately assigned ICD-9 codes for ischemic vascular diseases by the coding staff. On the manual review, of the unstructured clinical notes, we did not find any evidence for ischemic vascular disease for the patients. The high rate of unreliable ICD-9 codes in this study indicates the need to use narrative (unstructured) clinical data for retrieving an accurate problem list. Inconsistency between structured (e.g., ICD-9) and unstructured data (e.g., notes) in EHR contributes to the number of false positives. Often an ICD code may be assigned to a working diagnosis rather than a confirmed diagnosis. For example, a patient who presents with chest pain and has multiple coronary artery disease (CAD) risk factors might be given, on admission, the diagnosis of CAD. If this diagnosis is not confirmed through diagnostic tests the coding may not get corrected. This diagnosis confirming data remains in text format and is less accessible to data query. Methods that can interrogate structured and unstructured data in EHR would be an important part of future attempts to increase the accuracy of data query from clinical data warehouses.

3. Meaningful Evaluation for E-Screening

The research staff benefited from E-screening and reported a >50% savings in time (100 vs. 437 manually reviewed) mostly attributed to the reduced time reviewing medical records to confirm patient eligibility. Without E-screening, the research team would have had to browse a large volume of patient records and multiple EHRs in our organization for in-patients, outpatients, or different specialty clinics (e.g., WebCIS, Eclipsys, Epic, etc.) to manually aggregate clinical information to determine eligibility.

There is no gold standard for evaluating E-screening. Following Friedman's suggestion that decision support should focus on augmenting the productivity of a user[7], we suggest that the balance between *sensitivity and specificity* for E-screening is study-specific. If the "potentially eligible" population is very small, the goal of E-screening should be to minimize the "false negative" rate by reliably excluding cases that do not need further manual review[2]. In contrast, if the "potentially eligible" population is very large, such as the diabetes population for TECOS, the priority of E-screening should be to minimize false positives so that the "potentially eligible" cases recommended by the E-screening query actually merits manual review. We feel that this user-centered, protocol-specific paradigm should be used to guide E-screening queries, whether to a registry or data warehouse.

4. Practical Recruitment Considerations

In designing our E-screening query for the Warehouse we used a rough condition ("*patient seen in the past 12 months*") to filter likely "reachable patients" and used a list of clinicians within the ACN Network to narrow the screening scope. These measures increased our likelihood of obtaining cases that could be contacted.

5. Toward Proactive Registries: ELiXR

In this case study, both the Data Warehouse and the Registry improved recruitment efficiency for TECOS. The Warehouse reduced manual review by nearly 20 times that of the Registry (100 vs. 2033 cases that needed manual review). On 5/12/10, site 110 reached its recruitment goal of 60 participants and stopped recruiting. Using the same protocol, in the same time period and with identical study start-up procedures, site 110 was catapulted, by use of the Warehouse, to the rank of top recruiter among the 64 United States sites and third among 332 sites worldwide.

Compared with the Warehouse, the Registry was more efficient in having updated disease-specific markers such as A1C, consistent with the purpose of creating a disease management database but introduced too many false positives. However, the use of the Registry was ad hoc, which means that the researchers have no control on the design of the Registry or its variable section. We therefore envision that a proactive approach to using registries can integrate the advantages of both the Warehosue

and the Registry and further improve efficiency. Therefore, a good protocol-specific query interface linked to a rich clinical data warehosue is a key to successful screening and recruitment for researchers.

We propose to use a data warehouse to dynamically generating research registries to improve the accuracy of clinical trial electronic screening. Our design called EliXR (Eligibility Criteria Extraction and Representation) is illustrated by Figure 3. Our design consists of three steps: (1) extracting eligibility criteria from a research protocol; (2) extracting corresponding data representations for the eligibility criteria as available from the a data warehouse and collecting or linking additional data needed for E-Screening from complementary sources (e.g., public health questionnaires, family histories, or other specialized research databases); and (3) developing a protocol-aware research registry including comprehensive data variables with "active links" to miscellenous data sources and can receive regular updates for all the variables required for eligibility determination.

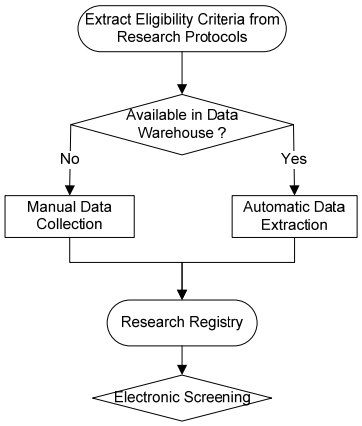

Figure 3. A Dynamic Protocol-Specific Screening Tool: EliXR

ELIXR registries can be designed once but used repeatedly throughout the multi-year recruitment process for a clinical trial study. Its design underscores the importance of linking a clinical data warehouse with disconnected research databases or enriching it with data variables that are generally not captured in EHR. We can even use natural language processing methods to support criteria extraction or variable extraction steps. Data interoperability and data reconciliation from miscellenous sources will be related research issues that need further experiments.

Conclusion

Electronic screening using digital data sources such as clinical registries and clinical data warehouses can both improve clinical research recruitment efficiency. To combine the advantages of both technologies, we proposed the design of EliXR to generate protocol-aware research registries to facilitate electronic screening for clinical trial recruitment. We will investigate the effectiveness of our proposed method for research registry development in our future work.

Acknowledgment

This research was funded by NLM grant R01 LM009886 and CTSA award UL1 RR024156. Its contents are solely the responsibility of the authors and do not represent the official view of NIH. We thank the reviewers for their valuable comments and the exceptional clinical research coordinator, Sabrina Durant, MD for her hard work.

Reference

1. NCRR Director's comment on Recruitment. http://www.nih.gov/news/health/nov2009/ncrr-10.htm. Accessed 3/7/2010.
2. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic Screening Improves Efficiency in Clinical Trial Recruitment. *Journal of the American Medical Informatics Association.* November 2009 2009;16(6):869-873.
3. TECOS. http://www.tecos-study.org/. Accessed March 2, 2010.
4. NewYork Presbyterian Hospital Ambulatory Care Network (ACN). http://nyp.org/services/amb-care-network.html. Accessed 07/03/2010.
5. Johnson S. Generic Data Modeling for Clinical Repositories. *Journal of the American Medical Informatics Association.* 1996;3(5):328–339.
6. Medical Entities Dictionary. http://med.dmi.columbia.edu/. Accessed March 5, 2010.
7. Friedman CP. A Fundamental Theorem of Biomedical Informatics. *Journal of the American Medical Informatics Association.* March 2009 2009;16(2):169-170.