Research Data Explorer: Lessons Learned in Design and Development of Context-based Cohort Definition and Selection

Adam Wilcox, PhD[1], David Vawdrey, PhD[2], Chunhua Weng, PhD[2], Mark Velez[2], Suzanne Bakken, RN DNSc[2]

[1]Intermountain Healthcare, Salt Lake City, UT, [2]Columbia University, New York, NY

Abstract

Research Data eXplorer (RedX) was designed to support self-service research data queries and cohort identification from clinical research databases. The primary innovation of RedX was the electronic health record view of patient data, to provide better contextual understanding for non-technical users in building complex data queries. The design of RedX around this need identified multiple functions that would use individual patient views to better understand population-based data, and vice-versa. During development, the more necessary and valuable components of RedX were refined, leading to a functional self-service query and cohort identification tool. However, with the improved capabilities and extensibility of other applications for data querying and navigation, our long-term implementation and dissemination plans have moved towards consolidation and alignment of RedX functions as enhancements in these other initiatives.

Introduction

The last decade has seen an explosion in the amount of clinical data stored in electronic form. Examples of institutions that have improved quality and efficiency of healthcare using health information technology (health IT) have indicated the potential value of electronic health records (EHRs) and electronic health data (1). Government incentives for healthcare adoption of electronic health records have moved EHR adoption to the late majority, with most hospitals and ambulatory providers using EHRs for patients (2,3). Coincident with these trends have been the increased use of large clinical data warehouses for both research and quality improvement. A query of publications in PubMed relating to "data warehouse" or "data warehousing" shows growth in the number of publications in the field growing substantially in the last five years after modest growth before. A similar search for "big data" shows a more dramatic rise (see Figure 1). With this increase in availability of data has come the promise of using the data for research, such as retrospective analyses (4), knowledge discovery (5), cohort identification (6), and phenotype analysis (7,8). Many large research projects have begun efforts to maximize the amount of data available for such studies by linking data across institutions, in population databases (9–11), research networks (12), and research registries (13).

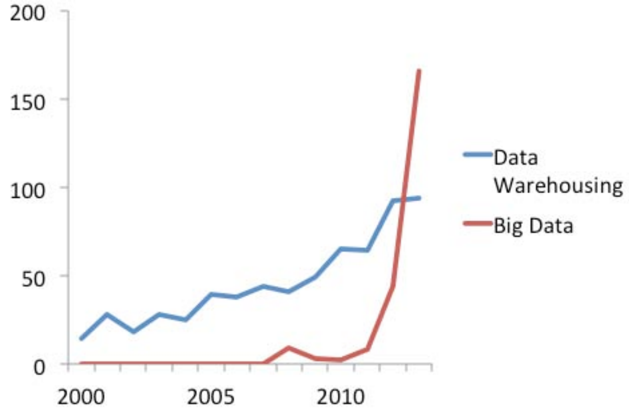


Figure 1: Number of publications per year retrieved from PubMed related to Data Warehousing and Big Data.

With all of these projects, data must be made accessible to researchers. Often this is done through trained data analysts, who receive data requests from researchers and then translate the requests into executable queries against the research databases (14,15). This approach faces two challenges. First, research data mediation is labor-intensive, requiring transfer of clinical knowledge to technical experts to create the queries (16). Projections on mediation-dependent approaches show them to be unsustainable as the number of data requests increases (15,17). Second, it requires the clinical researcher to define the data request in terms of rules, which are an abstraction from the clinical concept that was understood. The definition and clarification of these rules can greatly increase the already unsustainable cost of query mediation. It can also be difficult for researchers to specify the data rules that define a clinical concept, especially when a concept is represented by various disparate data sources (7).

Tools have been developed to bypass the query mediation process, by allowing researchers to create queries to the database directly (18). Integrating Informatics and Biology at the Bedside (i2b2) is such a tool with a simplified user interface for selecting clinical concepts, defining constraints for the concepts, and creating queries as Boolean combinations of element definitions (19). It has been disseminated broadly, to over a hundred clinical research institutions. The Observational Health Data Sciences and Informatics (OHDSI) program has also created tools for querying electronic health data for research (20). OHDSI is based on the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), and has successfully distributed data analytic tools to many of its nearly 100 named collaborators. Both these projects have created robust tools and effective distribution models. They have also leveraged the efforts of collaborators to expand functionality through user group development. They are likely the most successfully distributed data infrastructure applications in the clinical research informatics field. However, while i2b2 and OHDSI/OMOP efforts have been effective at making self-service query development possible, and thus improving the sustainability of data access to research, they have not fully addressed the challenge of reducing the abstraction of concepts to rules by the clinical researcher, or the challenge of identifying to what extent different data concepts are actually in the database to represent the clinical concept.

In this paper, we describe the design and development efforts of Research Data eXplorer (RedX), a cohort identification and selection tool developed at Columbia University to support clinical researchers. We also describe the current dissemination and enhancement approach for RedX. Our goal is to both present an effective design to address two unmet challenges in query development and cohort identification with data from electronic health records, and to recommend based on our experience appropriate methods for the sustainability of such tools.

Methods

RedX was an informatics component of the Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) at Columbia University. WICER was funded by the Agency for Healthcare Research and Policy to improve the capability for prospective comparative effectiveness research. RedX was a critical component of this goal, because it could provide methods of improved cohort identification for both prospective and retrospective studies.

RedX was designed based on years of experience in observing and managing the query mediation process for research data requests by Columbia researchers from the clinical data warehouse. A formal request process was developed where researchers would submit a request for data to support a research studies. The request would include a brief description of the data needed, contact information for both the submitter and sponsor of the request, and information on the purpose of the research study and the need for the specific data, to support data governance and security requirements for both the institutional review board and NewYork Presbyterian Hospital, where the clinical data was collected. Once a request was approved, a data analyst within the clinical data warehouse group would contact the researcher requesting the data, and the query mediation process would begin.

Data analysts learned by experience that the most effective method for communicating with clinical researchers about which data were needed was to identify patients from the researcher's panel of patients, and then review those patients' electronic health record to identify data that represented the patient meeting the inclusion criteria. That is, the requesters defined their data request based on patient data examples. When multiple different data sources were used as possible inclusion criteria, different patient examples would be identified for the different criteria. This "query by example" approach had the benefit of allowing clinical researchers to communicate the patient characteristics in the context of data navigation that they were most accustomed, either from providing care or performing electronic chart reviews. The data analysts were able to identify the data elements from the EHR data directly.

The RedX design mimicked this "query by example" approach directly. RedX was designed as a query interface integrated with the EHR view. We cloned the existing user interface from the EHR implemented at CUMC (21), and mapped the EHR functions to a de-identified research data set. Within this EHR view, users could then browse data for an individual patient in the same way they would review data in the EHR. No translation or data abstraction was necessary for the user. When reviewing specific data element values, such as problems on the problem list or laboratory result values, users could select the data value and directly create a query to identify other patients with similar data. We designed the RedX "query by example" button after the "infobutton" previously developed within the CUMC EHR (22).

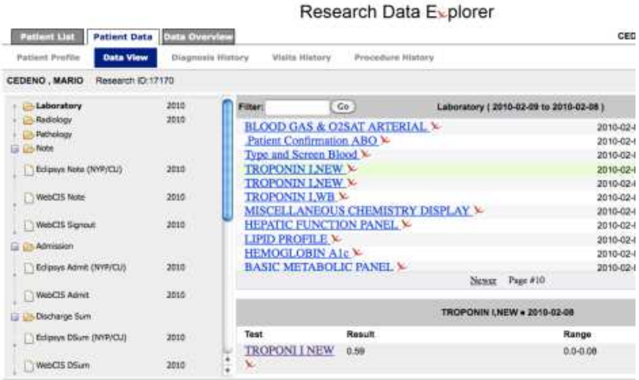
Queries based on individual data elements would necessarily be simple, so we designed a query interface to iteratively combine complicated queries from simple components. We modeled this design after the PubMed query combination tool (23), considering it both an effective design and one which clinical researchers would likely be familiar. Similar Boolean query combination tools exist in other health data query applications such as i2b2 (24). A value of iterative query combinations is that users can see which query components are most restrictive and permissive in defining the cohort. To support the context-based data navigation, we also designed RedX to allow navigation to the EHR-view from any patient cohort list within the system. Cohort groups created by any query could be viewed as a list, and selecting the patient from the list would show the individual EHR view.

Another component of the RedX design was the view of data distributions for individual data elements. While end users would likely be familiar with data elements for individual patients, they may not understand the overall distribution of values for that element. For discrete data elements, they also may not understand the possible or most common selected elements. We thus felt it was important to provide a population-context view of the data when creating queries based on the example of an individual patient. We identified three characteristics of data elements that would vary, where a view of the variance would help users understand how the data were typically represented. The three characteristics of the data elements were frequency, value, and time. Three two-dimensional views were designed for each characteristic pairing, and each pairing was designed to help users understand the use of the data element. The first pairing was frequency vs. value, or a standard histogram. This view helped understand the most common values, and how different cutoff value could affect the population selected. The second pairing was frequency vs. time. This view would indicate how the use of a data element could change over time. It was thought as most useful in instances where the data element was used differently over time, either because it was recently adopted or replaced by a different data element. The third pairing was value vs. time. This view would indicate how the mean, variance and range of a data element would change over time. It was considered useful for identifying data elements that may have changing normal ranges. Both the first and second pairings could be applied to continuous and discrete data, while the third pairing would only be applied to continuous data.
Since end users would likely be less-experienced with the data

Two other components of the RedX design were intended to help the researcher understand a subgroup of a population, to verify their inclusion in a cohort. Since clinician-researchers likely best understood their own patients and how they should be included in a cohort, we designed RedX to query cohorts from a de-identified database and an existing data warehouse. To prevent the release of protected health information, we determined that clinician-researchers would only be allowed to view real patient data for their own patients – all other patient data would only be accessible in de-identified form. Second, we designed views to compare data element values among different populations. This would allow researchers to identify other data elements that could be used to increase the sensitivity or specificity of cohort selection criteria.

Results

We developed RedX according to the design, and successfully completed five major functions as specified in the design. First, we created the EHR view of de-identified research data that allowed users to browse the data in the context of an individual patient. We also created a "query by example" button that allowed query development based on data elements within the patient record (Figure 2). Third, we created patient cohort lists, where the results of a query could be viewed as a list of patients in that cohort, allowing a user to navigate directly to that patient record (Figure 3). Fourth, we created a query combination tool, allowing for Boolean combinations of simpler element values. Finally, we created data distribution views, specifically for frequency vs. value, for both continuous and discrete data (Figure 4).



CEDENO , MARIO Research ID:17170

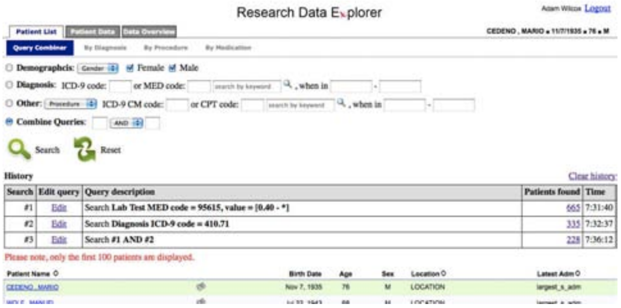
Laboratory (2010-02-09 to 2010-02-08)

BLOOD GAS & O2SAT ARTERIAL	2010-02-
Patient Confirmation ABO	2010-02-
Type and Screen Blood	2010-02-
TROPONIN I,NEW	2010-02-
TROPONIN I,NEW	2010-02-
TROPONIN I,WB	2010-02-
MISCELLANEOUS CHEMISTRY DISPLAY	2010-02-
HEPATIC FUNCTION PANEL	2010-02-
LIPID PROFILE	2010-02-
HEMOGLOBIN A1c	2010-02-
BASIC METABOLIC PANEL	2010-02-

TROPONIN I,NEW • 2010-02-08

Test	Result	Range
TROPONI I NEW	0.59	0.0-0.08

Figure 2: RedX single patient view. This view is similar to the EHR view. Individual data elements have a red "X" next to them. Selecting the red X will create a query with adjustable parameters based on that data element.



CEDENO , MARIO • 11/7/1935 • 76 • M

○ Demographcis: Gender ☑ Female ☑ Male

○ Diagnosis: ICD-9 code: or MED code: search by keyword , when in -

○ Other: Procedure ICD-9 CM code: or CPT code: search by keyword , when in -

● Combine Queries: AND

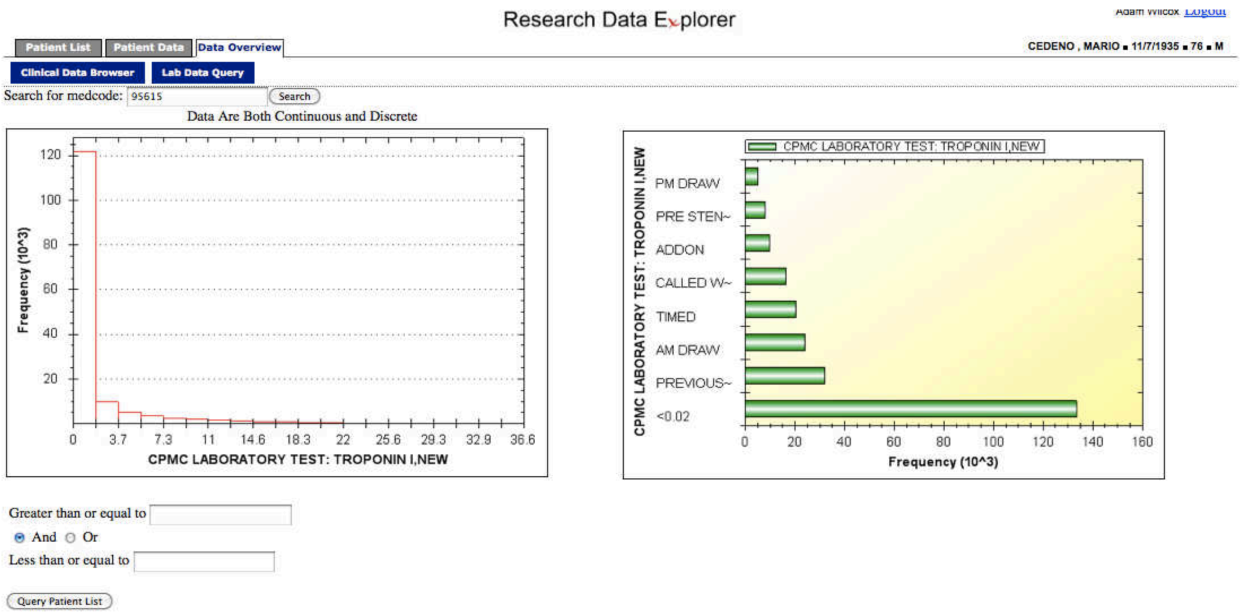
History

Search	Edit query	Query description	Patients found	Time
#1	Edit	Search **Lab Test MED code = 95615, value = [0.40 - *]**	665	7:31:40
#2	Edit	Search **Diagnosis ICD-9 code = 410.71**	335	7:32:37
#3	Edit	Search **#1 AND #2**	228	7:36:12

Please note, only the first 100 patients are displayed.

Patient Name	Birth Date	Age	Sex	Location	Latest Adm
CEDENO, MARIO	Nov 7, 1935	76	M	LOCATION	largest_s_adm

Figure 3: RedX ad hoc query builder, query combiner, and patient list view. Queries can be created according to data categories, or as combinations of other queries. The query results are given as a number, and patients in the result set are listed. Selecting a patient in the result set navigates to that patient EHR view.



Research Data Explorer

Adam Wilcox Logout

CEDENO , MARIO ▪ 11/7/1935 ▪ 76 ▪ M

Search for medcode: 95615

Data Are Both Continuous and Discrete

Greater than or equal to

● And ○ Or

Less than or equal to

Figure 4: RedX data distribution view for Troponin test. Data elements could be either continuous, discrete or both. The list of common discrete elements shows how a user would need to include both discrete and continuous data for some defined queries.

During RedX development, we identified two functions needed for its effective use that were not part of the original design. First, we identified that while query-by-example was useful for refining queries, it was less effective as a starting point to create initial queries. This was because the population of relevant example patients first needed to be defined by at least some simple inclusion and exclusion criteria. Therefore, we developed a query builder for ad hoc queries, to get users to at least define a sub-population from which to narrow the inclusion criteria (Figure 3). A second function was identified from the need to choose more generic or inclusive concepts, while understanding how that would affect the cohort size. For this, we adapted an earlier-designed Clinical Database Browser (25), that displayed data within a database in the context of a semantic hierarchy.

Other functions were not developed during the project. This change in development priorities was common among data infrastructure projects at the time, when it was common to underestimate the work required to navigate new institutional rules of data governance after the HITECH act (26). Due to the data governance delay, implementation priorities for the project became more important and some development was reduced. Specifically, we did not develop the additional data distributions comparing frequency or value to time. We found frequency vs. time was solved more by querying vocabulary classes, and value vs. time was not understood or requested by users. The "My Patients" view and comparisons between different cohorts was not developed. For the "My Patients" view, it was too difficult to get a design approved by the institutional data security office that would mix views between identified and de-identified data. The comparisons between cohorts was also seen as less useful to users. In addition, other applications such as i2b2 built similar and more effective solutions during our development period.

Discussion and Conclusion

Our design of RedX focused on providing a patient-context view for clinicians. This design was seen as critical and differentiating in RedX supporting self-service queries at Columbia University. Usability studies and design reviews showed this view combined with the clinical database browser view of data density were most important in helping clinician researchers understand the data behind the system. Resource pressures were effective at identifying other components of the design that were less necessary, and were removed during development.

While the core RedX innovations were critical to its success at CUMC, the development of other freely-available and extensible tools has affected the long-term implementation plan. Rather than supporting a separate tool, we are now adapting the innovative components of RedX to a more extensible data model (OMOP CDM), and reviewing how the components could be built as an i2b2 plugin. At the time RedX was initially designed and developed, these other tools were not as robust. But they have since reached a point that we have determined contributing to those efforts will best support the RedX design principles and discoveries. We feel this represents an important stage in the consolidation of clinical research informatics tools.

Acknowledgment

This work was supported in part by R01 HS019853 from AHRQ, Title: Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) and R01 HS022961 from AHRQ, Title: WICER 4 U.

References

1. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006 May 16;144(10):742–52.
2. Adler-Milstein J, DesRoches CM, Furukawa MF, Worzala C, Charles D, Kralovec P, et al. More Than Half of US Hospitals Have At Least A Basic EHR, But Stage 2 Criteria Remain Challenging For Most. Health Aff Proj Hope. 2014 Sep 1;33(9):1664–71.
3. Krist AH, Beasley JW, Crosson JC, Kibbe DC, Klinkman MS, Lehmann CU, et al. Electronic health record functionality needed to better support primary care. J Am Med Inform Assoc JAMIA. 2014 Sep;21(5):764–71.
4. Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology. 2002 Jul;224(1):157–63.
5. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012 Mar 14;4(125):125ra31.
6. Weng C, Batres C, Borda T, Weiskopf NG, Wilcox AB, Bigger JT, et al. A real-time screening alert improves patient recruitment efficiency. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2011;2011:1489–98.
7. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc JAMIA. 2013 Sep 11;
8. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc JAMIA. 2013 Jan 1;20(1):117–21.
9. Bakken S. Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER). 2010-2013 AHRQ ARRA Infrastruct Proj [Internet]. 2013 Jan 1; Available from: http://repository.academyhealth.org/ahrq_arra_cergrants/6
10. Wilcox A, Fort D, Bakken S. Creating a Next-generation Research Informatics Infrastructure: WICER Lessons for Data Integration. AMIA 2014 Joint Summits on Translational Science. San Francisco, CA; 2014.
11. Lee YJ, Boden-Albala B, Larson E, Wilcox A, Bakken S. Online health information seeking behaviors of Hispanics in New York City: a community-based cross-sectional study. J Med Internet Res. 2014;16(7):e176.
12. Ohno-Machado L, Agha Z, Bell DS, Dahm L, Day ME, Doctor JN, et al. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. J Am Med Inform Assoc JAMIA. 2014 Aug;21(4):621–6.
13. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care. 2012 Jul;50 Suppl:S60–67.
14. Natarajan K, Wilcox A, Sobhani N, Boyer A. Analyzing Requests for Clinical Data for Self-Service Penetration. AMIA Annu Symp Proc [Internet]. [cited 2014 Sep 26];2014. Available from: http://knowledge.amia.org/amia-55142-a2013e-1.580047/t-06-1.582200/f-006-1.582201/a-373-1.582875/a-374-1.582872?qr=1
15. Wilcox A, Yoon S, Boden-Albala B, Bigger JT, Feldman PH, Weng C, et al. Developing a Framework for Sustaining Multi-institutional Interdisciplinary Community Participatory Comparative Effectiveness Research. AMIA 2013 Joint Summits on Translational Science. San Francisco, CA; 2013.
16. Hruby GW, Boland MR, Cimino JJ, Gao J, Wilcox AB, Hirschberg J, et al. Characterization of the biomedical query mediation process. AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci. 2013;2013:89–93.
17. Wilcox A, Randhawa G, Embi P, Cao H, Kuperman G. Sustainability Considerations for Health Research and Analytic Data Infrastructures. EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 2014 Sep 17;2(2). Available from: http://repository.academyhealth.org/egems/vol2/iss2/8
18. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care. 2012 Jul;50 Suppl:S49–59.
19. Murphy S, Wilcox A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2). EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 2014 Sep 11;2(2). Available from: http://repository.academyhealth.org/egems/vol2/iss2/7
20. Weng C, Li Y, Ryan P, Zhang Y, Liu F, Gao J, et al. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. Appl Clin Inform. 2014;5(2):463–79.
21. Wilcox AB, Vawdrey DK, Chen Y-H, Forman B, Hripcsak G. The evolving use of a clinical data repository: facilitating data access within an electronic medical record. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2009;2009:701–5.
22. Cimino JJ. Infobuttons: anticipatory passive decision support. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2008;1203–4.
23. Motschall E, Falck-Ytter Y. Searching the MEDLINE literature database through PubMed: a short guide. Onkologie. 2005 Oct;28(10):517–22.
24. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc JAMIA. 2010 Apr;17(2):124–30.
25. Wilcox A, Hripcsak G, Chen C. Creating an environment for linking knowledge-based systems to a clinical database: a suite of tools. Proc Conf Am Med Inform Assoc AMIA Annu Fall Symp AMIA Fall Symp. 1997;303–7.
26. McGraw D, Leiter A. Pathways to Success for Multi-Site Clinical Data Research. EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 2013 Sep 19;1(1). Available from: http://repository.academyhealth.org/egems/vol1/iss1/13