AMIA
2015 Summit in CRI
CRI Year-In-Review

Peter J. Embi, MD, MS, FACP, FACMI

Assoc Prof & Vice Chair, Biomedical Informatics
Associate Professor of Medicine & Public Health
Chief Research Information Officer/Assoc Dean for Research Informatics
Director, Biomedical Informatics, CCTS
The Ohio State University

San Francisco, California
March 27, 2015

Disclosures

- Associate Editor, IJMI
- Editorial board, JAMIA
- Co-founder and consultant: Signet Accel LLC
- Consultant to various universities, research organizations

Approach to this presentation

- Mixed approach to article identification:
  - Started with structured approach
    - (akin to ACP “update” sessions)
  - Augment with “what seemed interesting” approach
- Learned a lot from doing this last four years
  - Tracked manuscripts throughout the year
  - Intended to spread work out…

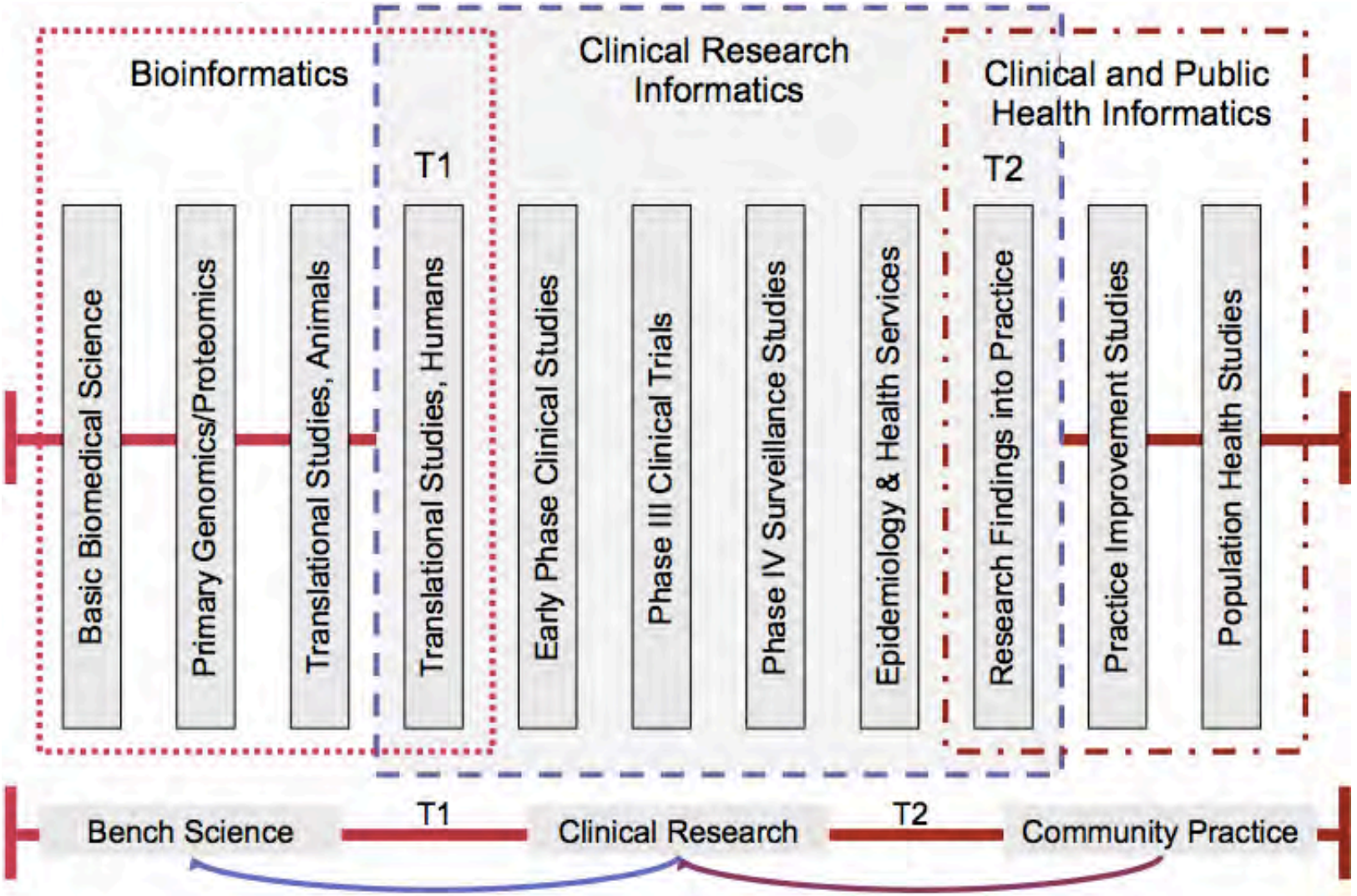
    …still worked down to the wire
- So, what was my approach…

Source of Content for Session

- Literature review:
  - Initial search by MESH terms:
    - "Biomedical Research"[Mesh] AND "Informatics"[Mesh] AND "2014/01/01"[PDat] : "2015/02/01"[Pdat]
    - Resulted in **118** articles
  - Additional articles found via:
    - Recommendations from colleagues
    - Other keyword searches using terms like:
      - Clinical Trials, Clinical Research, Informatics, Translational, Data Warehouse, Research Registries, Recruitment
    - Yielding **308** total, from which…
    - **99** were CRI relevant
  - From those, I've selected **42** representative papers that I'll present here (*briefly*)

Clinical and Translational Research & Informatics: T1, T2, and Areas of Overlap for Informatics

Shaded CRI Region is Main Area of Focus



Embi & Payne, JAMIA 2009

Session caveats

- What this is not…
  - A systematic review of the literature
  - An exhaustive review
- What this is…
  - My best attempt at *briefly* covering *some* of the representative CRI literature from the past year
  - A snap-shot of excellent CRI activity over past year+
  - What I thought was particularly notable

Topics

- Grouped 42 articles into several CRI categories (admittedly, not *all* CRI areas)
  - Data Sharing and Re-Use
  - Methods and Systems in CRI
  - Recruitment and Eligibility
  - Policy & Perspectives
  - Trends in CRI
- In each category, I'll highlight a few key articles and then given a quick "shout out" to some others
- Conclude with notable events from the past year

Apologies up front

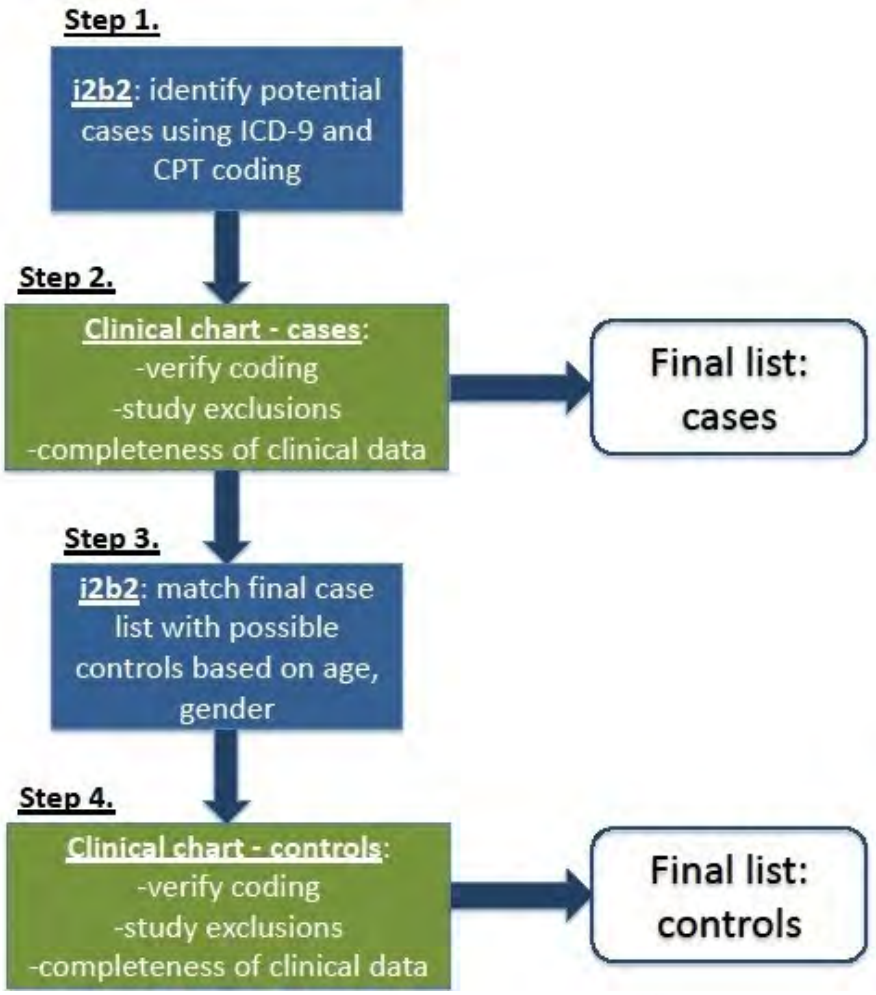
- I’m CERTAIN I’ve missed a lot of great work
- I’m REALLY SORRY about that

Clinical Data Sharing and Re-Use for Research

Use of the i2b2 research query tool to conduct a matched case–control clinical research study: advantages, disadvantages and methodological considerations

(Johnson EK, et al. BMC Medical Research Methodology. 2014)

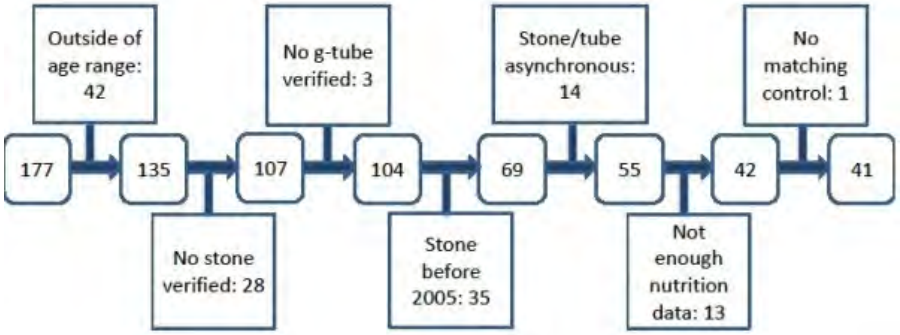
- **Objective:** Describe use of i2b2 research query tool and EMR at Boston Children's Hospital in conducting a case-controlled clinical study
- **Methods:** Analyzed the process of using i2b2 and the EMR together to generate a complete research database for a case–control study that sought to examine risk factors for kidney stones among gastrostomy tube (G-tube) fed children.
- To assemble the database for this study, a multi-step process was followed (Figure 1)



Use of the i2b2 research query tool to conduct a matched case–control clinical research study: advantages, disadvantages and methodological considerations

(Johnson EK, et al. BMC Medical Research Methodology. 2014)

- **Results:** Final case cohort consisted of 41/177 (23%) of potential cases initially identified by i2b2, who were matched with 80/486 (17%) of potential controls. Cases were 10 times more likely to be excluded for inaccurate coding regarding stones vs. inaccurate coding regarding G-tubes. A majority (67%) of cases were excluded due to not meeting clinical inclusion criteria, whereas a majority of control exclusions (72%) occurred due to inadequate clinical data necessary for study completion. Full dataset assembly required complementary information from i2b2 and the EMR.



- **Conclusions:** i2b2 was useful (critical) as a query analysis tool for patient identification in this case–control study. Patient identification via *procedural* coding appeared more accurate compared with *diagnosis* coding. Completion of study required iterative interplay of i2b2 and the EMR to assemble the study cohort. **Caveats for us as we attempt to use such resources as primary sources for such research.**

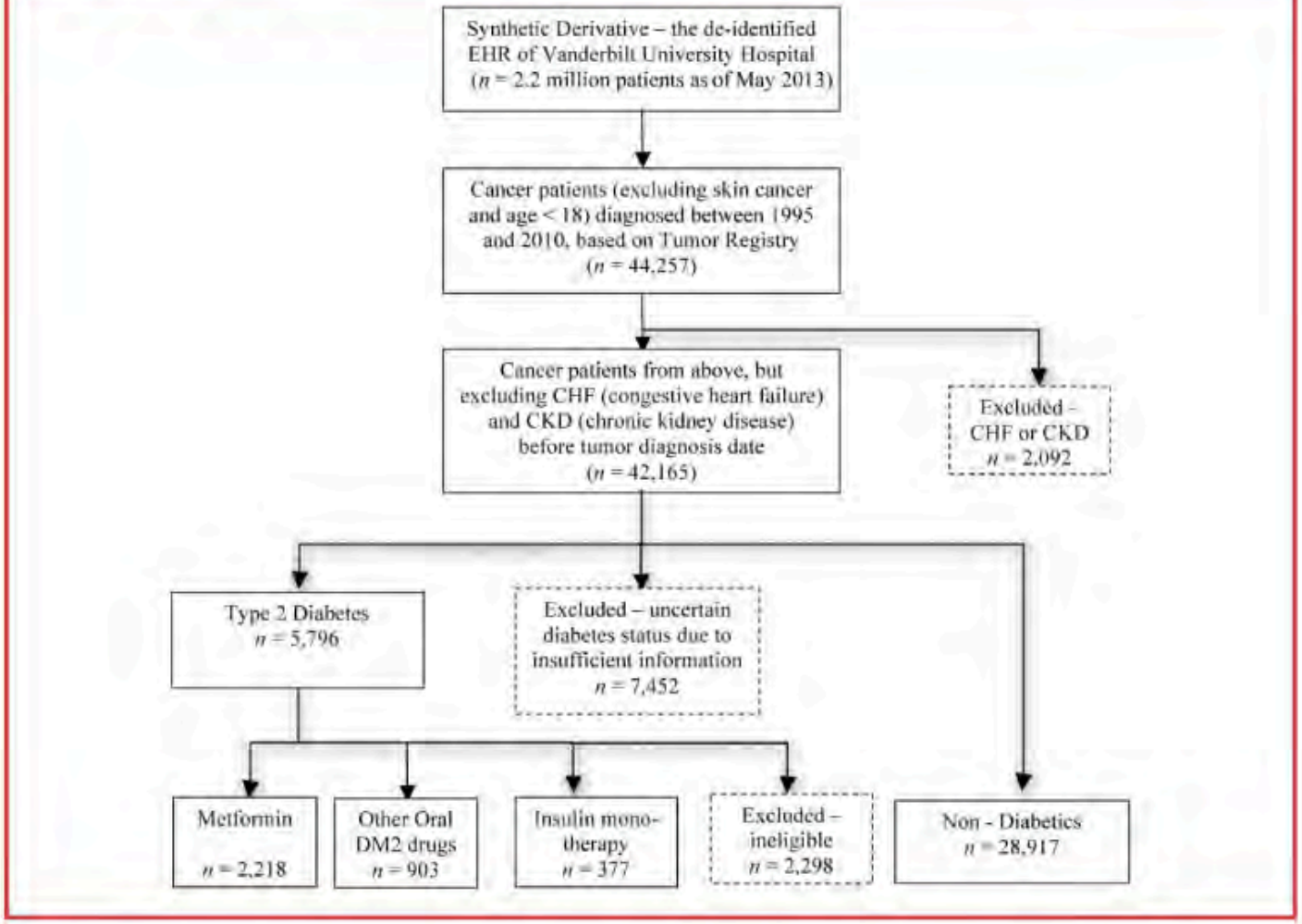
Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality

(Hua Xu, et al. JAMIA. 2014)

- **Objectives** Drug repurposing, which finds new indications for existing drugs, has received great attention recently.
- The goal here – to assess the feasibility of using EHRs and automated informatics methods to efficiently validate a recent drug repurposing association of metformin with reduced cancer mortality.
- **Methods** By linking two large EHRs from Vanderbilt University Medical Center and Mayo Clinic to their tumor registries, they constructed a cohort including 32,415 adults with a cancer diagnosis at Vanderbilt and 79,258 cancer patients at Mayo from 1995 to 2010.
- Using automated informatics methods, we further identified type 2 diabetes patients within the cancer cohort and determined their drug exposure information, as well as other covariates such as smoking status.
- Then estimated hazard ratios (HRs) for all-cause mortality and their associated 95% CIs using stratified Cox proportional hazard models. HRs were estimated according to *metformin exposure*.
  - Adjusted for age at diagnosis, sex, race, body mass index, tobacco use, insulin use, cancer type, and non-cancer Charlson comorbidity index.

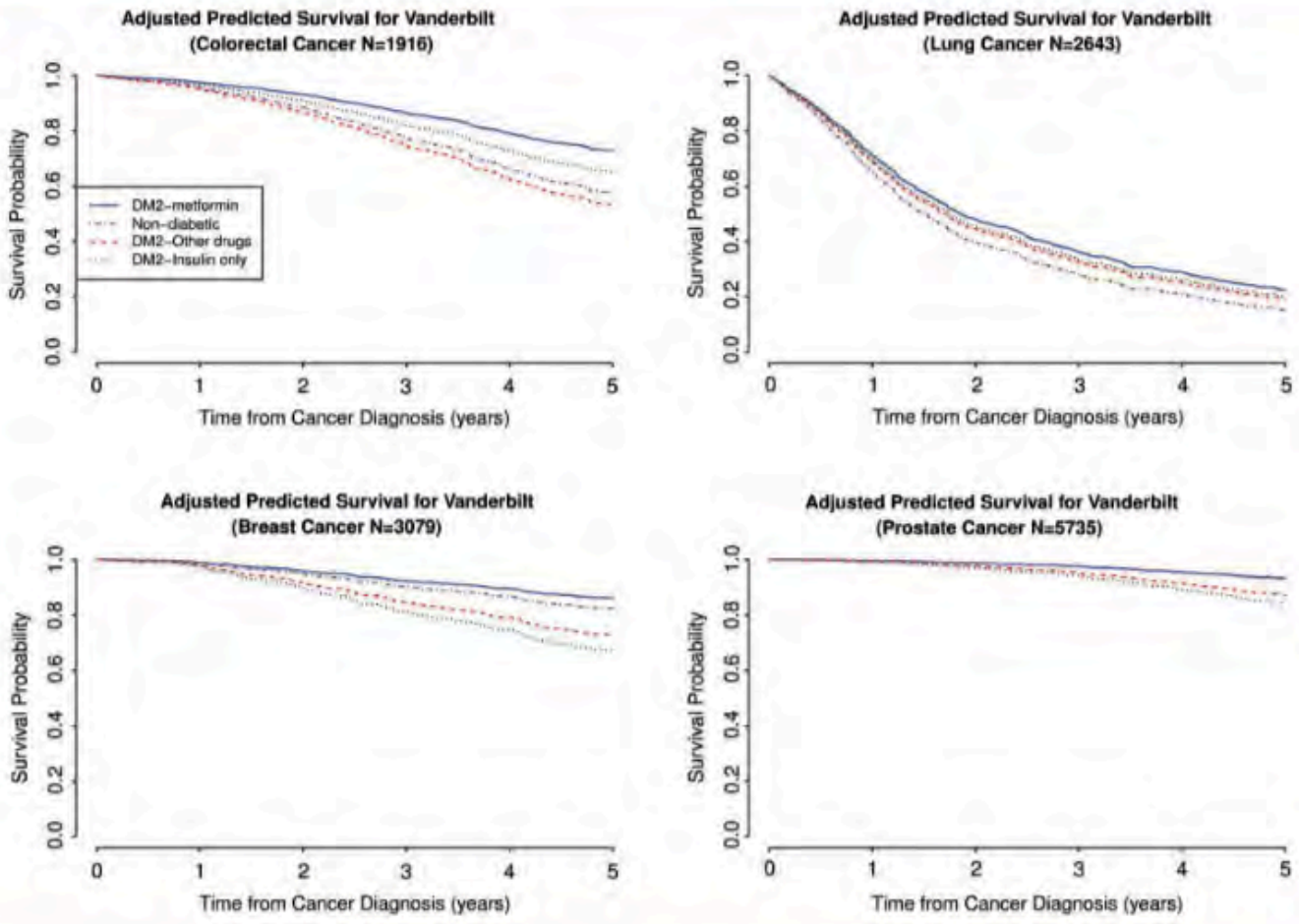
Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality (Hua Xu, et al. JAMIA. 2014)

Figure 1: The study design and data extraction workflow for patients in the Vanderbilt electronic health record (EHR) system from January 1995 to December 2010.



Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality (Hua Xu, et al. JAMIA. 2014)

Figure 6: Adjusted Cox proportional hazards model stratified by tumor type for the Vanderbilt cohort. All models are based on cancer survival in a white smoker, age 58 years, body mass index 27 kg/m^2, and not using insulin. DM2, type 2 diabetes mellitus.



Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality

(Hua Xu, et al. JAMIA. 2014)

- **Results** Among all Vanderbilt cancer patients, metformin was associated with a 22% decrease in overall mortality compared to other oral hypoglycemic medications (HR 0.78; 95% CI 0.69 to 0.88) and with a 39% decrease compared to type 2 diabetes patients on insulin only (HR 0.61; 95% CI 0.50 to 0.73).
- Diabetic patients on metformin also had a 23% improved survival compared with non-diabetic patients (HR 0.77; 95% CI 0.71 to 0.85). These associations were replicated using the Mayo Clinic EHR data. Many site-specific cancers including breast, colorectal, lung, and prostate demonstrated reduced mortality with metformin use in at least one EHR.
- **Conclusions** EHR data suggested that the use of metformin was associated with decreased mortality after a cancer diagnosis compared with diabetic and non-diabetic cancer patients not on metformin, indicating its potential as a chemotherapeutic regimen.
- **A model that EHR data can serve as a source of robust and inexpensive validation studies for drug repurposing signals.**

Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature

(Holmes JH, et al. JAMIA. 2014)

- **Objective**: Review published, peer-reviewed literature on clinical research data warehouse governance in distributed research networks (DRNs).
- **Methods:** Systematic literature review through July 31, 2013 for DRNs in USA only.
- **Results:** 6641 documents retrieved, 39 were included in the final review. Documents were analyzed using a classification framework consisting of 10 facets to identify themes:
  - Data collation
  - Data and process standards
  - Data stewardship
  - Data privacy
  - Query alignment and approval
  - Data use
  - Data security
  - Data retention
  - Data audits
  - User training

Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature (Holmes JH, et al. JAMIA. 2014)

Table 1 List of documents contained in the final corpus for review, with their contributions, classified according to the faceted framework used in this review

	First author	Reference number	Data collation	Data and process standards	Data stewardship	Data privacy	Query alignment and approval	Data use	Data security	Data retention	Data audits	User training
1	Baggs	25	X	X	X		X	X	X			
2	Bailey	26	X	X	X		X					
3	Bloomrosen	29			X							
4	Braff	31				X						
5	Bredfeldt	32				X						
6	Brown	11	X									
7	Brown	28	X		X		X	X		X	X	
8	Curtis	4	X	X	X		X	X	X			
9	Fernandes	13	X	X			X	X	X			
10	Forrow	5	X	X	X	X	X	X	X	X		
11	Fullerton	15				X						
12	Gardner	30				X						
13	Go	21	X	X	X		X	X				
14	Godwin	38										X
15	Greene	44										X
16	Holve	42	X	X	X	X						
17	Kim	36				X						

Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature (Holmes JH, et al. JAMIA. 2014)

Table 1 List of documents contained in the final corpus for review, with their contributions, classified according to the faceted framework used in this review

	First author	Reference number	Data collation	Data and process standards	Data stewardship	Data privacy	Query alignment and approval	Data use	Data security	Data retention	Data audits	User training
18	Lazarus	22	X		X	X	X	X	X			
19	Lopez	41	X	X	X	X		X	X			
20	Magid	23	X	X	X		X	X				
21	Manion	17				X			X		X	
22	Maro	12	X			X	X		X		X	
23	McGarvey	14	X				X				X	
24	McGraw	6	X	X			X	X	X	X		
25	McMurry	7					X					
26	McMurry	33	X		X	X	X	X		X	X	
27	Ohno-Machado	35						X				
28	Parwani	24	X			X	X		X		X	
29	Patel	2	X			X	X		X		X	
30	Platt	8	X					X				
31	Randhawa	18	X	X								
32	Rosenbaum	10		X	X							
33	Shapiro	16	X		X		X				X	
34	Thompson	39										X
35	Toh	1	X		X		X	X	X			
36	Tucci	27	X	X			X					
37	Velentgas	9	X		X				X			
38	Wagner	3	X				X					
39	Willison	37								X		

Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature

(Holmes JH, et al. JAMIA. 2014)

- **Results:** 6641 documents retrieved, 39 were included in the final review. A **peer-reviewed literature on data warehouse governance is emerging, but is still sparse**.
- Even though DRNs growing in importance for research and population health surveillance, understanding of DRN data governance policies and procedures is limited. This is expected to change as more DRN projects disseminate their governance approaches as publicly available toolkits and peer-reviewed publications.
- **Conclusions:** While US-based DRN data warehouse governance publications have increased, more DRN developers and administrators should formalize and publish information about these programs.

Approaches and costs for sharing clinical research data (Wilhelm EE, et al. JAMA. 2014)

- Brief report analyzing the costs to facilitating access to disease-oriented research databases that promote future research and data sharing. Prototypical example is the established (since 2004) Alzheimer's Disease Neuroimaging Initiative (ADNI).
- Description of the major categories of costs include:
  - (1) infrastructure and administration
  - (2) standardization
  - (3) human resources
  - (4) opportunity costs
- Good discussion of the considerations and implications
  - Most costs borne by those sharing, rather than users
  - Sophistication, standardization adds value but also cost
  - Funding for sharing often inadequate in such research projects, despite increasing expectation and value in sharing
  - Understanding and planning for these costs is key to success
- Sustainability arises as a concern…
  - (topic addressed by other articles in this year's review)

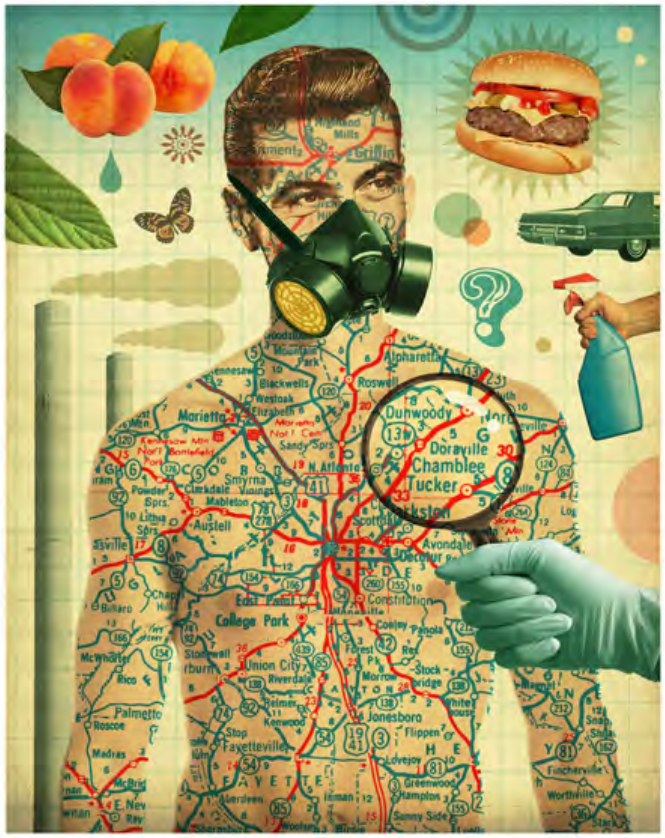
Other notable papers in this (Sharing/Reuse) category:

- **Data collection challenges in community settings: insights from two field studies of patients with chronic disease** (Holden RJ, et al. Quality of Life Research. 2014)
  - Framework of contextual challenges relevant to community-based participatory research and patient-contributed data. Insights for design of CRI solutions
- **Sharing behavioral data through a grid infrastructure using data standards** (Min H, et al. JAMIA. 2014)
  - Developed and refined a method for incorporating measures in NCI grid-enabled measured (GEM) portal for behavioral/social measures (i.e. smoking-related) into cancer data standards registry and repository (caDSR).
  - Created new branch of common data elements (CDE) that extends beyond clinical/biological, and missing terms/concepts for behavioral measures added to NCI thesaurus.
  - *An example of expanding a federated data sharing resource for behavioral science.*

CRI Methods and Systems

Exposome informatics: considerations for the design of future biomedical research information systems (Sanchez FM, et al. JAMIA. 2014.)

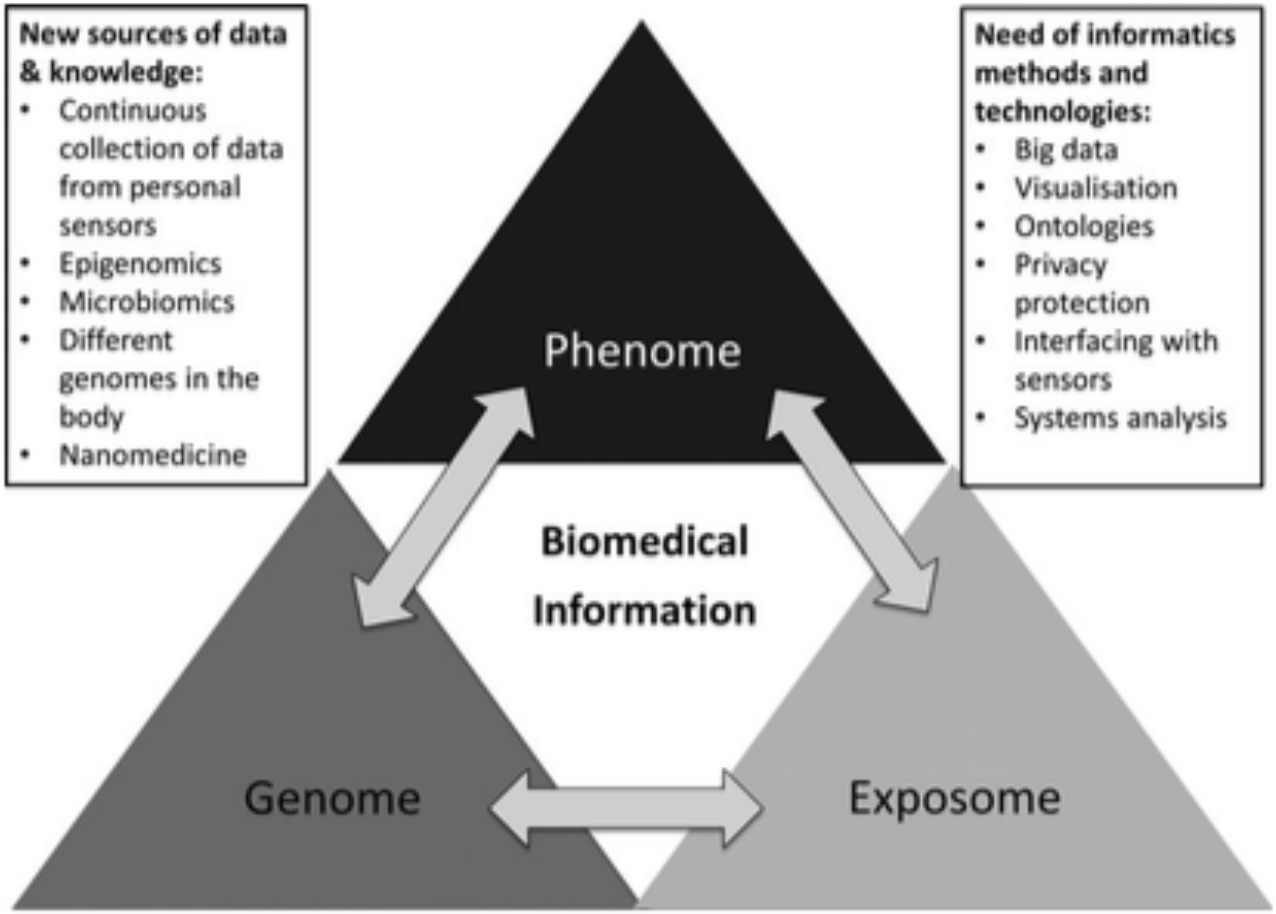
- The environment's contribution to health has been conceptualized as **the exposome**. Biomedical research interest in environmental exposures as a determinant of physio-pathological processes is rising as such data increasingly become available.
- The growth of miniaturized sensing devices now accessible and affordable for individuals to use to monitor a widening range of parameters opens up a new world of research data.
- Biomedical informatics (BMI) must provide a coherent framework for dealing with multi-scale population data including the phenome, the genome, the exposome, and their interconnections.



From:
MICHAEL WARAKSA:
MAPPING THE EXPOSOME
Theispot.com

Exposome informatics: considerations for the design of future biomedical research information systems (Sanchez FM, et al. JAMIA. 2014.)

- New research data types that require changes in informatics methods



Exposome informatics: considerations for the design of future biomedical research information systems (Sanchez FM, et al. JAMIA. 2014.)

- Examples of data of interest:

Table 2 Examples of the data of interest for future information systems

Group	Subgroup	Measure
Exposome	General external	Climate
		Education
		Socio-economical aspects
		Natural and built environment
	Specific external	Noise, humidity, CO, NOx, temperature, O_3, radiation, particulate matter
		Medication, nanomaterials, medical procedures
		Sedentary behaviors, physical activity
		Smoking, diet, sleep, alcohol consumption
		Infectious agents
	Internal	Metabolites, hormones, oxidative stress, inflammation
Phenome	Molecular traits	Gene expression, proteomics
		Lipids, HDL, triglycerides
	Cellular traits	Signaling pathways
		Cell cycle, apoptosis
		Cell migration
	Tissue/organ traits	Organ malformations, morphology, medical imaging
		Blood pressure
	Organismal traits	Body mass index, weight, height
	Disease phenotypes	Pathologies
	Behavior	Stress, mood
	Endophenotypes	Cholesterol, immunoglobulins
Genome	Sequence information	Whole genome, exome
	Genomic variation	Single nucleotide variants (SNPs, mutations, ...), structural variants (CNVs,

Exposome informatics: considerations for the design of future biomedical research information systems (Sanchez FM, et al. JAMIA. 2014.)

- The combination of these more continuous, comprehensive, and personalized data sources requires new research and development approaches to data management, analysis, and visualization.

- Article analyzes the implications of a new paradigm for our discipline that recognizes genome, phenome, **and exposome** data and their intricate interactions as the basis for biomedical research now and for clinical care in the near future.
- Much work to be done here…

The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date

(Cimino JJ, et al. J. Biomedical Informatics. 2014.)

- NIH developed the Biomedical Translational Research Information System (BTRIS) to support researchers' access to translational and clinical data.
- BTRIS includes a data repository, a set of programs for loading data from NIH electronic health records and research data management systems, an ontology for coding the disparate data with a single terminology, and a set of user interface tools that provide access to identified data from individual research studies and data across all studies from which individually identifiable data have been removed.
- This paper reports on unique design elements of the system, progress to date and user experience after five years of development and operation.

The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date

(Cimino JJ, et al. J. Biomedical Informatics. 2014.)

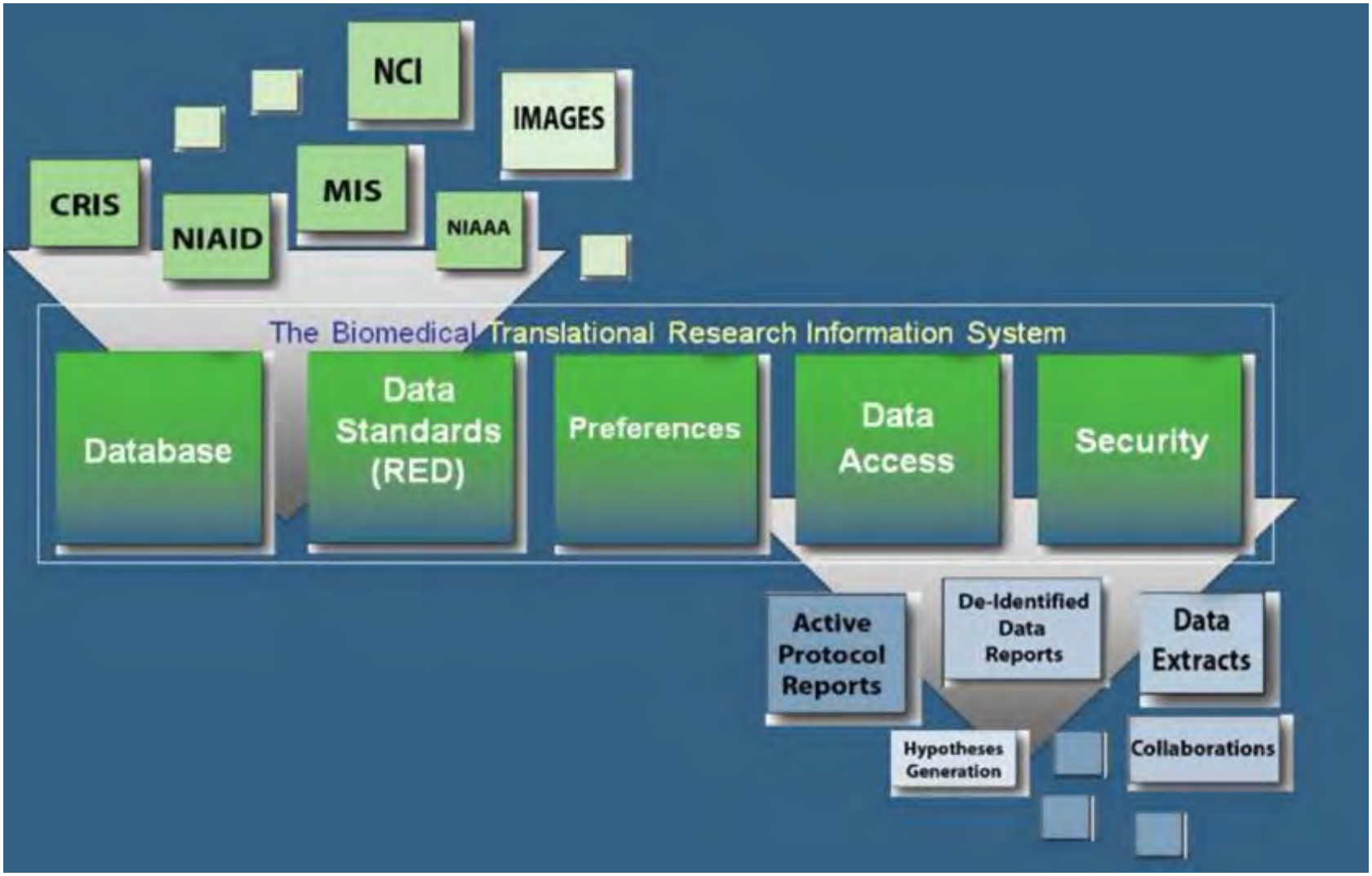
Table 1
NIH clinical research systems.

System name	Institute	System type	Developer	Data domains
Medical Information System (MIS)[a]	Clinical Center	EHR	Commercial	Demographics, Laboratory Tests, Blood Bank, Medications, Microbiology, Radiology, Vital Signs, *Clinical Notes, Anatomic Pathology*
Clinical Research Information System (CRIS)	Clinical Center	EHR	Commercial	Demographics, Subject-Study Attribution, Vital Signs, Clinical Documentation, Alerts, Allergies, Observations, Document Images, Medication Administration, Medication Orders, *Other Orders, Admission/Discharge/ Transfer*
Softlab	Clinical Center	Ancillary Department System	Commercial	Clinical Laboratory Tests, Microbiology Tests, Anatomic Pathology, Blood Bank Tests, Blood Bank Products
Softmed	Clinical Center	Medical Records Department System	Commercial	Admission Notes, Discharge Notes, Diagnoses, *Other Dictated Notes*
Vmax[a]	Clinical Center, NHLBI	Ancillary Department System	Commercial	*Pulmonary Function Reports*
Jaeger	Clinical Center, NHLBI	Ancillary Department System	Commercial	*Pulmonary Function Reports*
Pain and Palliative Care System	Clinical Center	Ancillary Department System	Clinical Center	*Pain and Palliative Care Notes*
LinkTools	Clinical Center, NHLBI	Ancillary Department System	Commercial	Electrocardiograms
ProSolv	Clinical Center, NHLBI	Ancillary Department System	Commercial	Echocardiology Reports
RadNet	Clinical Center	Ancillary Department System	Commercial	Radiology Reports
Carestream	Clinical Center	Picture Archiving and Communication System	Commercial	Radiographic Images
Protrak	Clinical Center	Protocol Services Department System	Clinical Center	Studies, Investigators
Clinical Research Information Management System of the NIAID (CRIMSON)	NIAID	CTDMS	NIAID	Study-Subject Attribution, Laboratory Tests, Medications, Patient Problems
Clinical Research Database (CRDB)	NIAAA	CTDMS	NIAAA	Assessments (Surveys)
Labmatrix	NCI	CTDMS	Commercial	Biospecimens
Cancer Central Clinical Database (C3D)	NCI	CTDMS	NCI	Study Attribution, Laboratory Tests, Case Report Forms
Clinical Trials Database (CTDB)	NICHD, NIAAA, NIDDK, Clinical Center	CTDMS	NICHD	Encounter Forms
Labmatrix	NHGRI	CTDMS	Commercial	Biospecimens, Case Report Forms
Varsifter	NHGRI, NIMH	Laboratory Database	NHGRI	Whole Exome Sequences

CTDMS – clinical trials data management system, EHR – electronic health record, NCI – National Cancer Institute, NHGRI – National Human Genome Research Institute, NHLBI – National Heart, Lung and Blood Institute, NIAAA – National Institute on Alcohol Abuse and Alcoholism, NIAID – National Institute of Allergy and Infectious Diseases, NICHD – National Institute of Child Health and Human Development, NIDDK – National Institute of Diabetes and Digestive and Kidney Diseases, NIMH – National Institute of Mental Health.

The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date

(Cimino JJ, et al. J. Biomedical Informatics. 2014.)



The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date (Cimino JJ, et al. J. Biomedical Informatics. 2014.)

- BTRIS filling an important gap in the information infrastructure at NIH
  - Improving research/clinical data access and use
  - Formal modeling of research entities and terminologies are supporting new methods by bringing together data from patient care, clinical research and genomic research into a unified conceptual search environment.
- Lessons and experience (e.g. data model, Research Entities Dictionary, user query interface, and data sharing policies) are relevant beyond NIH to others working to represent disparate clinical and research data to support clinical/translational research.

Other notable papers in this (Methods/Systems) category:

- **A European inventory of common electronic health record data elements for clinical trial feasibility** (Doods J, et al. Trials. 2014)
  - Building upon the 'Electronic Health Records for Clinical Research' (EHR4CR). 75 data elements identified that are frequently used in clinical studies and are available in European EHR systems. While many key data exist, not all information that is frequently used in site feasibility is documented via routine patient care. *Insights and implications for use of EHR data for research.*
- **Development of an online library of patient-reported outcome measures in gastroenterology: the GI-PRO database** (Khanna P, et al. Am. J Gastroenterol. 2014)
  - NIH supported GI-PRO clearinghouse, using protocol from PROMIS. Searchable item database with quality scores attached. 8 “bins” resulted. While many PROs available, many limited by low methodological quality.
- **Ontology-based data integration between clinical and research systems** (Mate S. et al. PLoS One. 2015)
  - Much data in EHRs not linked to standard terminology, even when discrete. Described is an ontology-basd approach to overcome challenges of database-level ETL definitions traditionally used to combine EHR-derived data for reuse. Approach that could be more scalable than current.

Other notable papers in this (Methods/Systems) category:

- **NOR-DMARD data management: implementation of data capture from electronic health records** (Olsen IC, et al. Clinical and Experimental Rheumatology. 2014)
  - Implementation and use of EDC system in all Rheumatology Departments participating in NORwegian Disease-Modifying Anti-Rheumatic Drugs registry led to more patients in registry, lower costs, improved data quality and accessibility.
- **Understanding data requirements of retrospective studies** (Shenvi EC, et al. International Journal of Medical Informatics. 2014)
  - Study of over 100 studies exploring the data elements required for studies and their availability in EHRs. While most frequently used items (e.g. procedures, condition, meds) often available, 49/104 studies had complex criteria that relied in part on data elements that could not be mapped to standard data dictionaries. Informs use of EHRs for such studies and work to be done to improve current state/enable.
- **Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT** (Varghese J, et al. Methods of Information in medicine. 2014)
  - 425 studies at German University analyzed and coded per UMLS. Revealed that relatively few concepts (101/5236) covered 25% of eligibility criteria in studies. Inform concepts that should be made available in EHRs to enable research activities like recruitment.

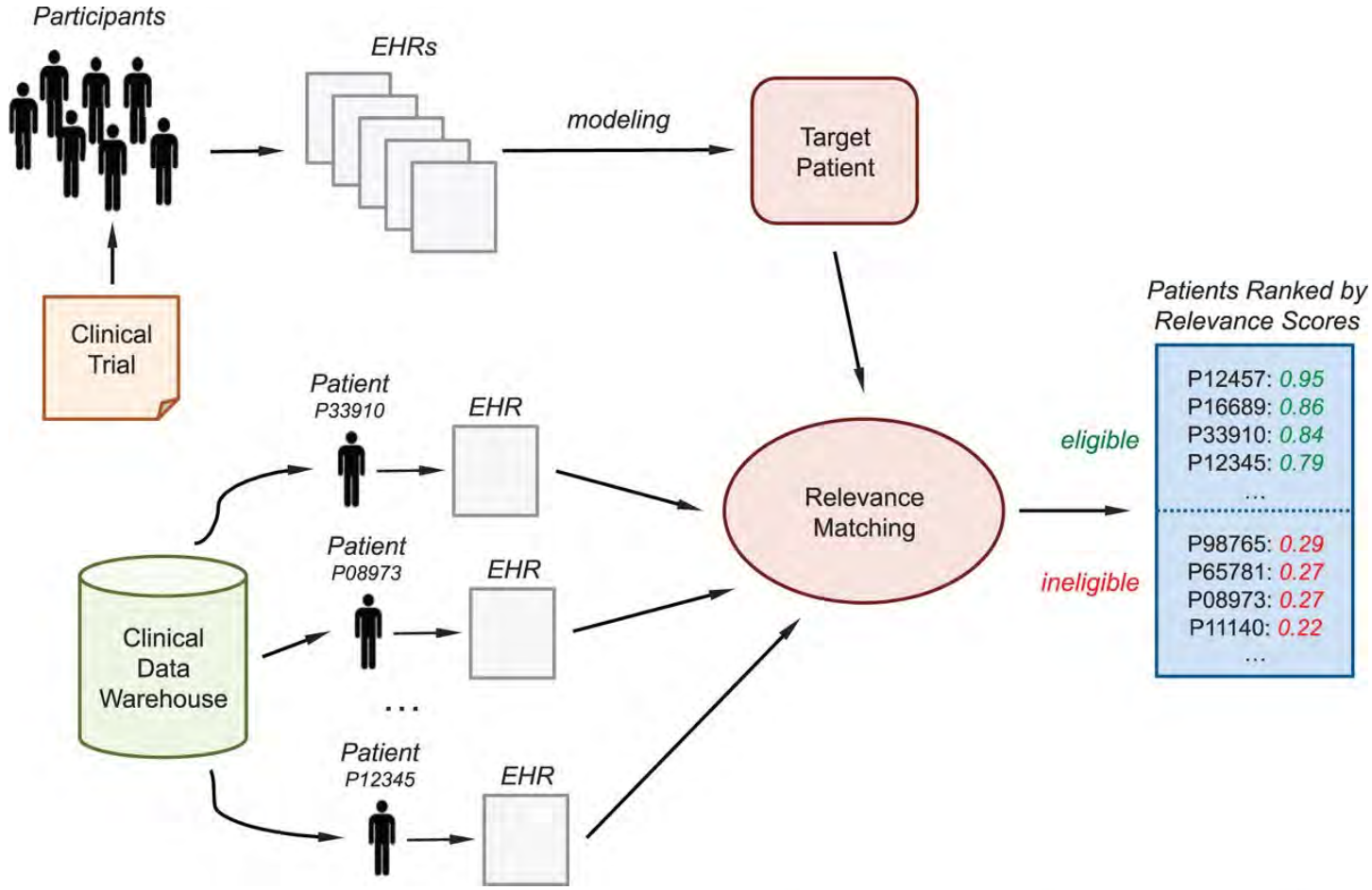
Participant Recruitment and Eligibility

Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials (Miotto R, Weng C. JAMIA. 2015. ePub 2014)

- **OBJECTIVE:** To develop a cost-effective, case-based reasoning framework for clinical research eligibility screening by only reusing the electronic health records (EHRs) of minimal enrolled participants to represent the target patient for each trial under consideration.
- **METHODS:** The EHR data-specifically diagnosis, medications, laboratory results, and clinical notes-of known clinical trial participants were aggregated to profile the "target patient" for a trial, which was used to discover new eligible patients for that trial.
- The EHR data of unseen patients were matched to this "target patient" to determine their relevance to the trial; the higher the relevance, the more likely the patient was eligible.
- Relevance scores were a *weighted linear combination* of cosine similarities computed over individual EHR data types.
- Evaluated using 262 participants of 13 different clinical trials conducted at Columbia University (gold standard). Ran a 2-fold cross validation with half of the participants used for training and the other half used for testing along with other 30,000 patients selected at random from our clinical database. We performed binary classification and ranking experiments.

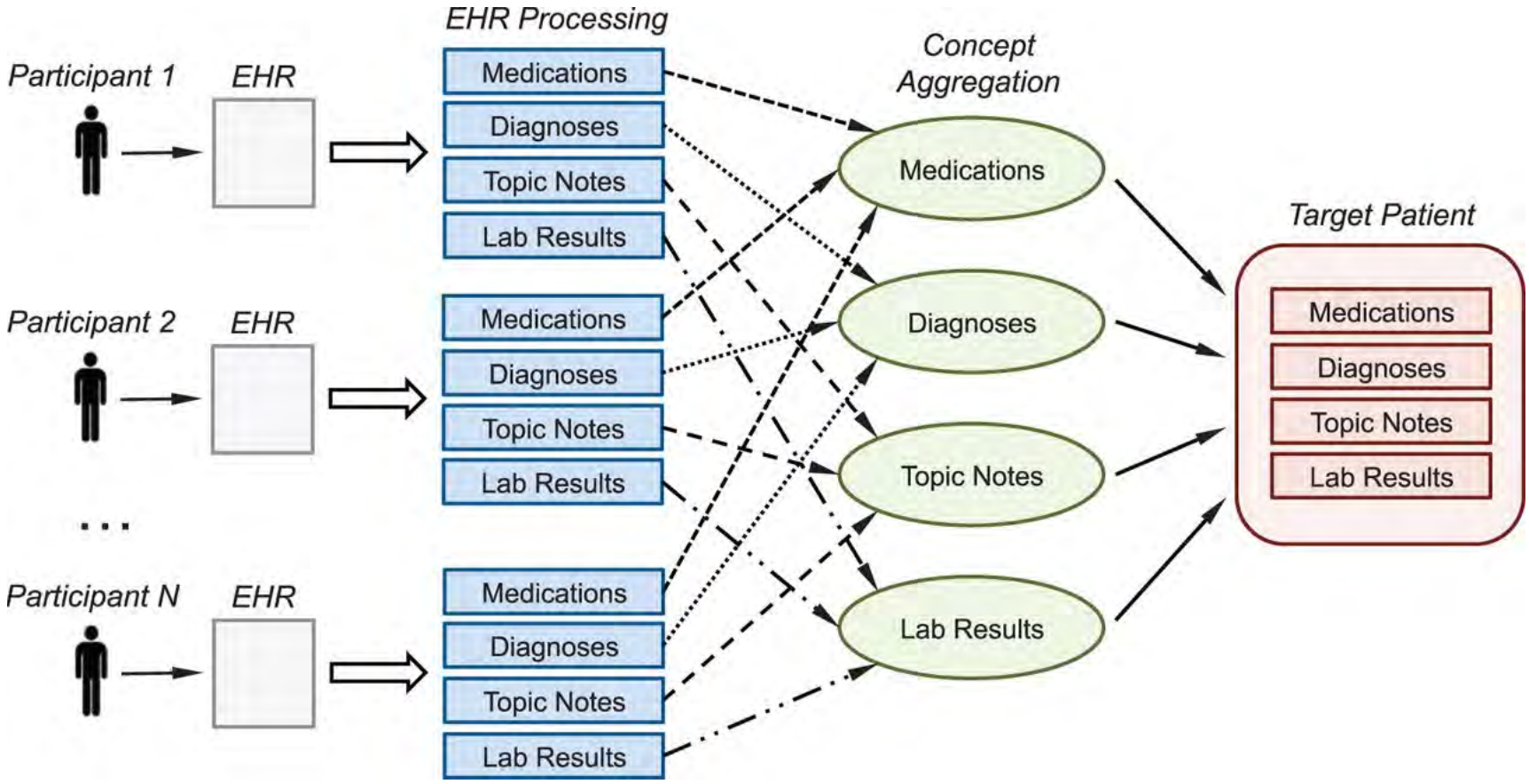
Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials (Miotto R, Weng C. JAMIA. 2015. ePub 2014)

- Figure 1: Overview of the "case-based reasoning" framework to discover eligible patients for a clinical trial through the "target patient," a representation of the trial derived from the EHR data of a minimal sample of participants.



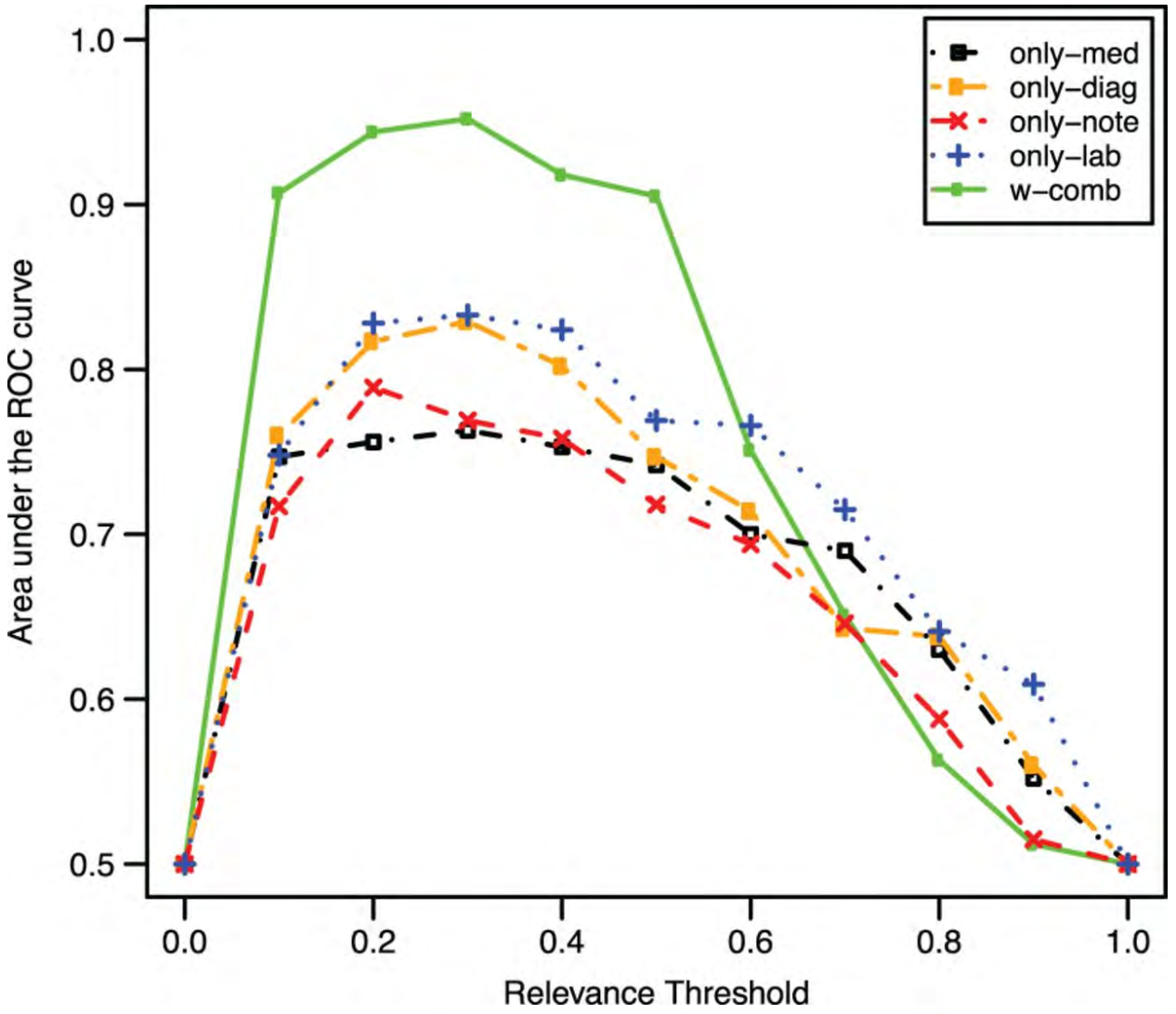
Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials (Miotto R, Weng C. JAMIA. 2015. ePub 2014)

- **Figure 2:** Overview of the process to derive the clinical trial's "target patient" by modeling the EHR data of minimal enrolled participants.



Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials (Miotto R, Weng C. JAMIA. 2015. ePub 2014)

- Figure 3: Classification results in terms of the area under the ROC curve averaged over both the evaluation folds.
- A patient was considered eligible if its relevance score with the corresponding "target patient" was over a threshold (ranged between 0 and 1), ineligible otherwise.



Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials (Miotto R, Weng C. JAMIA. 2015. ePub 2014)

- **RESULTS:** The overall area under the ROC curve for classification was 0.95, enabling the highlight of eligible patients with good precision. Ranking showed satisfactory results especially at the top of the recommended list, with each trial having at least one eligible patient in the top five positions.

- **CONCLUSIONS:** This relevance-based method can potentially be used to identify eligible patients for clinical trials by processing patient EHR data alone without parsing free-text eligibility criteria, and shows promise of efficient "case-based reasoning" modeled only on minimal trial participants.
- *Innovative approach that adds to our toolbox of approaches for recruitment – sorely needed.*

ClinicalTrials.gov as a data source for semi-automated point-of-care trial eligibility screening (Pfiffner PB, et al. PLoS One. 2014.)

- **BACKGROUND:** Implementing semi-automated processes to efficiently match patients to clinical trials at the point of care requires both detailed patient data and authoritative information about open studies.

- **OBJECTIVE:** To evaluate the utility of the ClinicalTrials.gov registry as a data source for semi-automated trial eligibility screening.

- **METHODS:** Eligibility criteria and metadata for 437 trials open for recruitment in four different clinical domains were identified in ClinicalTrials.gov. Trials were evaluated for up to date recruitment status and eligibility criteria were evaluated for obstacles to automated interpretation. Finally, phone or email outreach to coordinators at a subset of the trials was made to assess the accuracy of contact details and recruitment status.

ClinicalTrials.gov as a data source for semi-automated point-of-care trial eligibility screening (Pfiffner PB, et al. PLoS One. 2014.)

- **RESULTS:** 24% (104 of 437) of trials declaring on open recruitment status list a study completion date in the past, indicating out of date records. Substantial barriers to automated eligibility interpretation in free form text are present in 81% to up to 94% of all trials. Were unable to contact coordinators at 31% (45 of 146) of the trials in the subset, either by phone or by email. Only 53% (74 of 146) would confirm that they were still recruiting patients.
- **CONCLUSION:** Because ClinicalTrials.gov has entries on most US and many international trials, the registry could be repurposed as a comprehensive trial matching data source. Semi-automated point of care recruitment would be facilitated by matching the registry's eligibility criteria against clinical data from electronic health records. But the current entries fall short. Ultimately, improved techniques in natural language processing will facilitate semi-automated complex matching.
- As immediate next steps, recommend augmenting ClinicalTrials.gov data entry forms to capture key eligibility criteria in a simple, structured format.

Effectiveness of a community research registry to recruit minority and underserved adults for health research (Bishop WP, et al. Clin Transl Sci. epub 2014.)

- **BACKGROUND:** Recruiting minorities and underserved populations into population-based studies is a long standing challenge. This study examined the feasibility of recruiting adults from a community research registry.
- **METHODS:** Ethnically diverse, bilingual staff attended health fairs, inviting adults to join a registry. Examined rates of successful contact, scheduling, and participation for studies that used the registry.
- **RESULTS:** Five studies queried 6,886 research registry members (48% Hispanic and 38% black) and attempted to contact 2,301 potentially eligible participants; eligibility criteria varied across studies.
- Successfully contacted 1,130 members, 51.9% were scheduled to participate and of those, 60.8% completed their study appointment. Non-Hispanic whites were less likely than Hispanics to be interested, but among those scheduling an appointment, participation did not differ by race/ethnicity.
- **CONCLUSION:** Community research registries are a feasible and efficient method for recruiting minority and underserved adults and may address disparities in access to and participation in health research.

Other notable papers in this (Recruitment) category:

- **Accrual and recruitment practices at Clinical and Translational Science Award (CTSA) institutions: a call for expectations, expertise, and evaluation** (Kost RG, et al. Acad Med. 2014.)
  - Survey of 44 CTSAs revealed insights about common recruitment practices, services offered (or not by most), etc. Many CTSAs lack institutional frameworks to support study accrual. Recommendations made for improvements.
- **Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department** (Ni Y, et al. JAMIA. 2015. ePub 2014)
  - Approach to automated eligibility screening for clinical trials in an urban tertiary care pediatric emergency department (ED). Tests involving natural language processing (NLP), information extraction (IE), and machine learning (ML) techniques on real-world clinical data and trials yielded significant improvements in efficiency on retrospective data.

Other notable papers in this (recruitment) category:

- **Design and multicentric implementation of a generic software architecture for patient recruitment systems re-using existing HIS tools and routine patient data** (Trinczek B. Applied Clinical Informatics. 2014)
  - Definition and design of patient recruitment system, identified 24 commonly reported/requested features, 13 considered required, and implementation options for common hospital IS environments (Germany) indicate feasibility of approach.
- **Developing a new model for patient recruitment in mental health services: a cohort study using Electronic Health Records** (Callard F, et al. BMJ Open. 2014)
  - To address ethical, legal, technical issues, model for universal “consent to contact” “registry” developed in UK NHS mental health system. Enables patient participation and autonomy, and allows researchers to contact patients for future studies.
- **Feasibility platform for stroke studies: an online tool to improve eligibility criteria for clinical trials** (Minnerup J. Stroke. 2015. ePub 2014)
  - Development of feasibility platform for stroke studies (FePASS) to estimate proportions of eligible patients for acute stroke studies. Applied FePASS to 4 recent stroke studies. Proportion of eligible patients found to range 2.1-11.3%, and slight variations in inclusion criteria could have substantial increases in proportions. Tools is open access online resource.

Other notable papers in this (recruitment) category:

- **Completion and Publication Rates of Randomized Controlled Trials in Surgery: An Empirical Study** (Rosenthal R. Ann. Surg. 2014.)
  - Of 836 RCTs, discontinuation (43% surgical, 27% medical), mostly commonly to poor recruitment (18%); non-publication (44%) rates high. Need for CRI solutions.
- **Prevalence, characteristics, and publication of discontinued randomized trials** (Kasenda B, et al. JAMA. 2014)
  - 1017 RCTs from Switzerland, Germany, Canada analyzed. 25% discontinued, only 38% of those reported to ethics committees. Most common reason was poor recruitment. Factors associated with success and failure reported.

CRI Trends (1): Clinical Research Networks & LHS

- **Launching PCORnet, a national patient-centered clinical research network** (Fleurence RL, et al. JAMIA. 2014)
  - Describes launch of the PCORI funded network of 11 Clinical Data Research Networks (CDRNs) and 18 Patient-Powered Research Networks (PPRNs)
  - CDRNs focused on implementing systems to support multi-institutional research and enable rapid learning.
- **PaTH: towards a learning health system in the Mid-Atlantic region** (Amin W, et al. JAMIA. 2014)
  - The PaTH (University of Pittsburgh/UPMC, Penn State College of Medicine, Temple University Hospital, and Johns Hopkins University) CDRN
- **The ADVANCE network: accelerating data value across a national community health center network** (DeVoe JE, et al. JAMIA. 2014)
  - ADVANCE (Accelerating Data Value Across a National Community Health Center Network) CDRN is led by the OCHIN; integrates outpatient EHR data for over one million federally qualified health center patients, and integrates hospital, health plan, and community data for these patients, under-represented in studies.
- **PEDSnet: a National Pediatric Learning Health System** (Forrest CB. JAMIA. 2014)
  - CDRN including 8 AMCs, 2 disease-specific pediatric networks, 2 national partners to form the National Pediatric Learning Health System (NPLHS)

CRI Trends (1): Clinical Research Networks & LHS

- **CAPriCORN: Chicago Area Patient-Centered Outcomes Research Network** (Kho AN, et al. JAMIA. 2014)
  - The Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN); collaboration across private, county, and state hospitals and health systems, a consortium of Federally Qualified Health Centers, and 2 VAMCs.
- **The Greater Plains Collaborative: a PCORnet Clinical Research Data Network** (Waitman LR, et al. JAMIA. 2014.)
  - The GPC is composed of 10 leading medical centers (University of Kansas Medical Center, Children's Mercy Hospital, University of Iowa Healthcare, the University of Wisconsin-Madison, the Medical College of Wisconsin and Marshfield Clinic, the University of Minnesota Academic Health Center, the University of Nebraska Medical Center, the University of Texas Health Sciences Center at San Antonio, and the University of Texas Southwestern Medical Center). Over 10 million patients represented.
- **Developing a data infrastructure for a learning health system: the PORTAL network** (McGlynn EA, et al. JAMIA. 2014.)
  - The Kaiser Permanente & Strategic Partners Patient Outcomes Research To Advance Learning (PORTAL) network engages four healthcare delivery systems (Kaiser Permanente, Group Health Cooperative, HealthPartners, and Denver Health) and their affiliated research centers to create a new network.

CRI Trends (2): Pragmatic Clinical Trials

- **The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials** (van Staa TP, et al. Health Technol Assess. 2014)
  - 17 English and Scottish practies contributing EHR data to a common research database participated. Findings: EHR point-of-care trials are feasible, although the recruitment of clinicians is a major challenge owing to the complexity of trial approvals. These trials will provide substantial evidence on clinical effectiveness only if trial interventions and participating clinicians and patients are typical of usual clinical care and trials are simple to initiate and conduct. Administrative burdens and disruptions must be reduced to encourage participation.
- **Informed Consent for Pragmatic Trials — The Integrated Consent Model** (Kim SYH, Miller FG. NEJM. 2014)
  - Given goal to integrate clinical care and research (as in LHS), proposal is to include documentation of informed consent through routine documentation of discussion with patient. Once decision made to “enroll” in trial, physician performs as regular practice, and documents that there was agreement (eg. “We discussed the rationale, the risks and benefits of both options,” and so on) , and that a treatment (A or B) was chosen — including the process of random selection.
- **Ethics and Regulatory Complexities for Pragmatic Clinical Trials** (Sugarman J, Califf RM. JAMA. 2014)

CRI Policy & Perspectives:

- **A new initiative on precision medicine**

  (Collins FS, Varmus H. NEJM. 2015)

  - Following up on President's call for precision medicine initiative
  - Sure to have a major impact on and rely upon CRI and TBI activities

The NEW ENGLAND JOURNAL of MEDICINE

Perspective

FEBRUARY 26, 2015

A New Initiative on Precision Medicine

Francis S. Collins, M.D., Ph.D., and Harold Varmus, M.D.

"Tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes — and to give all of us access to the personalized information we need to keep ourselves and our families healthier."

— President Barack Obama, State of the Union Address, January 20, 2015

President Obama has long expressed a strong conviction that science offers great potential for improving health. Now, the President has announced a research initiative that aims to accelerate progress toward a new era of precision medicine (www.whitehouse.gov/precisionmedicine). We believe that the time is right for this visionary initiative, and the National Institutes of Health (NIH) and other partners will work to achieve this vision.

The concept of precision medicine — prevention and treatment strategies that take individual variability into account — is not new[1]; blood typing, for instance, has been used to guide blood transfusions for more than a century. But the prospect of applying this concept broadly has been dramatically improved by the recent development of large-scale biologic databases (such as the human genome sequence), powerful methods for characterizing patients (such as proteomics, metabolomics, genomics, diverse cellular assays, and even mobile health technology), and computational tools for analyzing large sets of data. What is needed now is a broad research program to encourage creative approaches to precision medicine, test them rigorously, and ultimately use them to build the evidence base needed to guide clinical practice.

The proposed initiative has two main components: a near-term focus on cancers and a longer-term aim to generate knowledge applicable to the whole range of health and disease. Both components are now within our reach because of advances in basic research, including molecular biology, genomics, and bioinformatics. Furthermore, the initiative taps into converging trends of increased connectivity, through social media and mobile devices, and Americans' growing desire to be active partners in medical research.

Oncology is the clear choice for enhancing the near-term impact of precision medicine. Can-

CRI Policy & Perspectives:

- **NIH's Big Data to Knowledge (BD2K) initiatives and the advancement of biomedical informatics** (Ohno-Machado L. JAMIA. 2014)
  - Editorial lead-in to special issue dedicated to informatics solutions focused on Big Data
  - Published in March 2014
  - Many excellent articles addressing issues such as:
    - Technical and policy infrastructure
    - Data processing and organization
    - Knowledge generation

Highlights

NIH's *Big Data to Knowledge* initiative and the advancement of biomedical informatics

doi:10.1136/amiajnl-2014-002666

Lucila Ohno-Machado, *Editor-in-Chief*

Two influential reports on data and computation from advisory committees to the NIH leadership resulted in important initiatives: (1) the report from the Working Group on Biomedical Computing for the Advisory Committee to the Director (ACD) of the National Institutes of Health (NIH) in 1999[i] led to the *Biomedical Information Science and Technology Initiative* (BISTI), and (2) the more recent report from the Working Group on Data and Informatics for the ACD in 2012[ii] led to the *Big Data to Knowledge* (BD2K) initiative.[iii] Several AMIA members participated in this working group. Both reports recommended strong support for data science and computation in biomedical sciences and healthcare.

The BD2K initiative was launched at NIH in 2013 through the development of several focused workshops, calls for proposals for centers of excellence, for a data discovery index, for training programs, and through the creation of the new position of Associate Director of Data Sciences, reporting directly to the NIH director. According to Francis Collins, the charge is to *"lead an NIH-wide priority initiative to take better advantage of the exponential growth of biomedical research datasets, which is an area of critical importance to biomedical research. 'The era of 'Big Data' has arrived, and it is vital that the NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information."* The first NIH Director of Data Sciences is a member of AMIA and an elected fellow of the American College of Medical Informatics. Phillip Bourne, former Editor-in-Chief of PLoS Computational Biology, developer of the Protein Data Bank, and Professor of Pharmacology at the University California San Diego, introduces this special *Big Data* focus issue of *JAMIA* with an editorial describing his vision of a 'Digital Enterprise'.

Bourne's vision reflects a new reality in which informatics has moved from the periphery of the healthcare and biomedical research enterprise to the center of action. *JAMIA* has been documenting solutions to data acquisition, management, and knowledge generation, and will be a premier venue for reporting on BD2K and related activities. In this first special issue of *JAMIA* on *Big Data*, we present creative solutions to challenges in data acquisition, organization, and analysis, with a particular emphasis on electronic health record data.

1. *Technical and policy infrastructure for data acquisition, efficient storage, and management.* Articles by LeDuc *et al* (*See page 195*), White *et al* (*See page 379*),[iv] and Sahoo *et al* (*See page 263*) focus on technical infrastructure for research data, and articles by Bloomrosen *et al* (*See page 204*) and Akagu *et al* (*See page 374*) focus on secondary use of healthcare data, from a sociotechnical perspective, which includes privacy concerns. Goldwater *et al* (*See page 280*) describes how the acquisition of electronic health data can be feasible in institutions that compose the U.S. federal safety net. EHRs are often distributed, so techniques for record linkage are important to integrate the data – Kum *et al* (*See page 212*) and Rajasekaran *et al* (*See page 252*) describe algorithms for this task. Additionally, new data modalities are increasingly augmenting the EHR, and some of these data can challenge current organizational structures and storage capabilities. Tenenbaum *et al* (*See page 200*)[v] discusses standards for 'omics' data, and Li *et al* (*See page 363*) describes a novel algorithm for genomic data compression.
2. *Data processing and organization.* EHR phenotyping, which was the focus of *JAMIA*'s December 2013 issue, refers to data processing for accurate characterization of disease status and health conditions using data collected in the process of care. A review by Shivade *et al* (*See page 221*) provides the context for EHR phenotyping, and articles by Tate *et al* (*See page 292*) and Melton *et al* (*See page 299*) describe algorithms for EHR phenotyping based on structured data and resources for structured data derivation from clinical notes. Rosenman *et al* (*See page 345*) focuses on database queries on hospitalizations for acute congestive hearth failure, while Dentler *et al* (*See page 285*) and Perotte *et al* (*See page 231*) focus on quality measures and diagnosis code assignment on EHRs.
3. *Knowledge generation.* The articles by Iyer *et al* (*See page 353*), Friedman *et al* (*See page 308*), and Liu *et al* (*See page 245*) focus on detecting drug-related adverse events based on EHRs and related sources. Huston *et al* (*See page 238*) uses data mining techniques to suggest drug repurposing. Data mining techniques for predicting hospital readmissions, early detection of neonatal sepsis, outcome of septic patients, and changes in hypertension control are presented in articles by He *et al* (*See page 272*), Mani *et al* (*See page 326*), Taglopoulos *et al* (*See page 315*), and Sun *et al* (*See page 337*), respectively.

Big Data is a big deal in biomedical research and healthcare. I hope our readers enjoy this special issue and continue to submit the products of *Big Data* initiatives such as BD2K for dissemination through *JAMIA*. In the highly diverse biomedical informatics community, professionals with expertise in library science, statistics, management, computer science, software engineering, natural language processing, and implementation science focus on biomedical and healthcare data. Our community is uniquely positioned to translate these data into actionable knowledge to promote health and advance science.

[i] http://www.bisti.nih.gov/library/june_1999_Rpt.asp (accessed 19 Jan 2014).
[ii] http://acd.od.nih.gov/Data%20and%20Informatics%20Working%20Group%20Report.pdf (accessed 19 Jan 2014).
[iii] http://http://bd2k.nih.gov/ (accessed 24 Jan 2014).
[iv] Watch journal club webinars by White and Italia, and by Tenenbaum and Haendel at http://dbmi.ucsd.edu/display/DBMI/UCSD-iDASH+Journal+Club

CRI Policy & Perspectives:

- **Finding the missing link for big biomedical data** (Weber G, et al. JAMA 2014)
  - Brief piece on key issues facing us as we try to leverage “big data”
  - Data sources are organized along different dimensions of “bigness” to help orient to the issues at play….
  - Helpful framing for so many

Opinion

VIEWPOINT

Finding the Missing Link for Big Biomedical Data

Griffin M. Weber, MD, PhD
Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, and Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts.

Kenneth D. Mandl, MD, MPH
Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, and Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts.

Isaac S. Kohane, MD, PhD
Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, and Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts.

It has been argued that big data will enable efficiencies and accountability in health care.[1,2] However, to date, other industries have been far more successful at obtaining value from large-scale integration and analysis of heterogeneous data sources. What these industries have figured out is that big data becomes transformative when disparate data sets can be linked at the individual person level. In contrast, big biomedical data are scattered across institutions and intentionally isolated to protect patient privacy. Both technical and social challenges to linking these data must be addressed before big biomedical data can have their full influence on health care. It is this linkage challenge that we address in this Viewpoint.

Political campaigns, government, and businesses use big data to learn everything possible about their constituents or customers, and then apply advanced computation to hone strategy. The 2012 Obama campaign identified, approached, and influenced swing voters using data fused from Facebook, census, voter lists, and active outreach. The National Security Agency employs massive data on individuals from phone and Internet companies to identify terrorists. Google personalizes search results with the user's web history and geographic context. In all these examples, the key has been to go beyond aggregate data and link information to individual people. Knowing that there are many swing voters in a zip code is helpful, but contacting those specific individuals may help to win an election.

Linking big data will enable physicians and researchers to test new hypotheses and identify areas of possible intervention. For example, do grocery shopping patterns obtained from stores in various areas predict rates of obesity and type 2 diabetes in public health databases? Does level of exercise recorded by home monitoring devices correlate with response rates of cholesterol-lowering drugs, as measured by continued refills at the pharmacy? Does increased physical distance from patients' homes to hospitals and pharmacies affect utilization of health care and result in distinct patterns in claims data? To what extent do patients' Facebook friends influence lifestyle choices and compliance with medical treatments? It is unknown whether these types of correlative inferences will really be found in big data and how physicians would use that information. However, being able to link data at the patient level is a prerequisite to exploring the possibilities.[2]

The first challenge in using big biomedical data effectively is to identify what the potential sources of health care information are and to determine the value of linking these together. The Figure presents a potential way of approaching this problem by organizing data sets along different dimensions of "bigness." Although some big data, such as electronic health records (EHRs), provide depth by including multiple types of data (eg, images, notes, etc) about individual patient encounters, others, such as claims data provide longitudinality—a view of a patient's medical history over an extended period for a narrow range of categories. Linking data adds value when they help fill in the gaps. With this in mind, it becomes easier to see how nontraditional sources of biomedical data outside of the health care system fit into the picture. Social media, credit card purchases, census records, and numerous other types of data, despite varying degrees of quality, can help assemble a holistic view of a patient, and, in particular, shed light on social and environmental factors that may be influencing health.

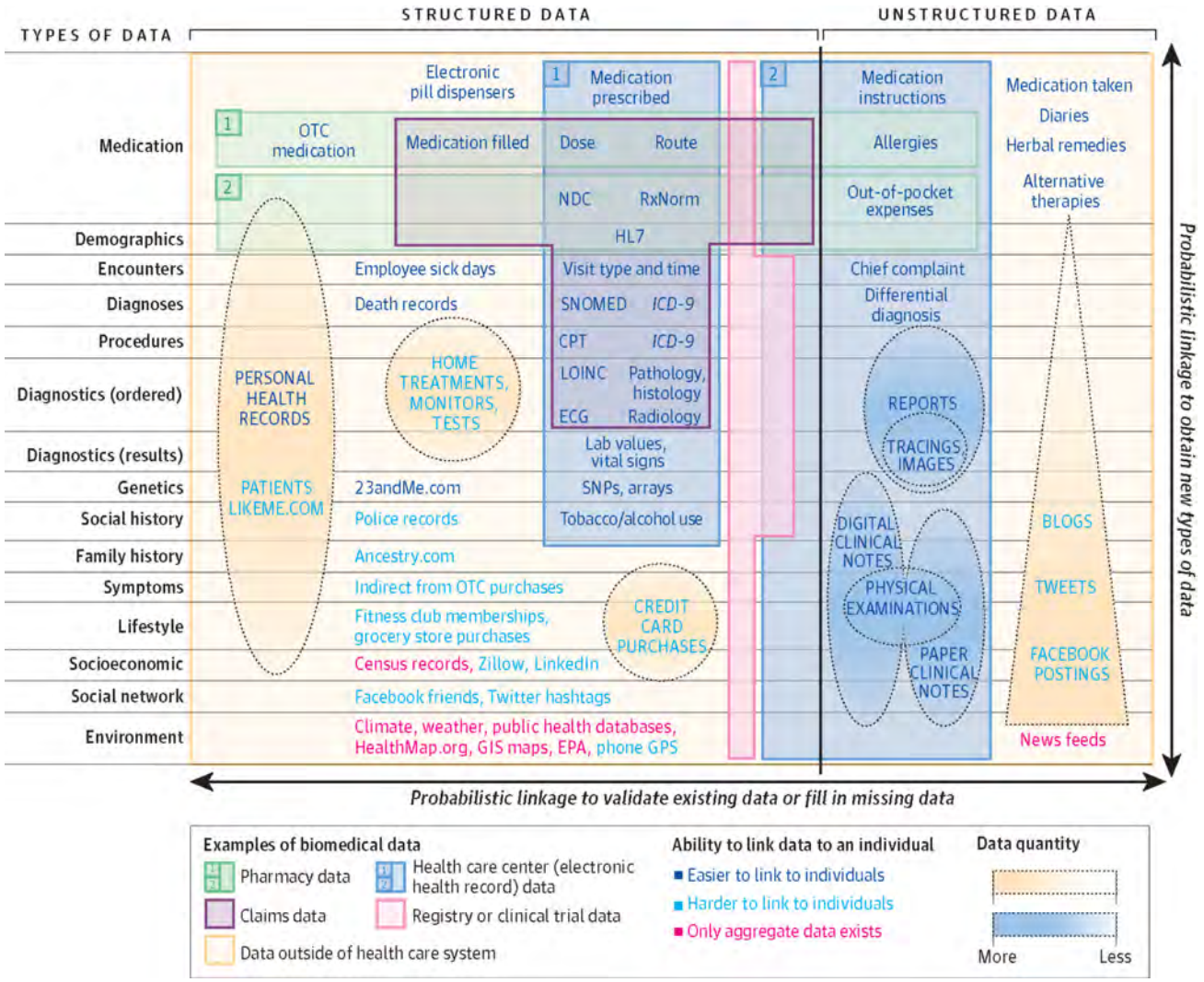
The lack of a national unique patient identifier (UPI) in the United States introduces another technical obstacle in linking big biomedical data. However, driven by the absence of a UPI to enable precise linkage, hospitals and clinics have developed sophisticated probabilistic linkage algorithms based on other information, such as demographics.[2] Although 2 different patients may share the same name, age, zip code, or other characteristics, by requiring enough variables to match, hospitals and clinics are able to reduce the risk of linkage errors to an acceptable level. An advantage of probabilistic linkage is that the same techniques used to match patients across different EHRs can be extended to data sources outside of health care. However, as indicated in the Figure, some data have fewer variables available for linkage, either because they do not contain the information or because policies restrict their use. These data may still be linked to patients, but the likelihood of errors is greater. Uncertainty due to possible linkage errors may be balanced by the advantages of having access to data about millions of patients. Therefore, future tools that use big data for health research or clinical decision making will need to use statistical techniques that correctly model these trade-offs.

Privacy and security concerns present a social challenge in linking big biomedical data. As more data are linked, they become increasingly more difficult to deidentify.[4,5] The consequences of this in health care, particularly for mental health records and genetic markers, have been extensively studied and discussed.[6,7] However, given that data linkage is already happening in other industries and is increasingly being thought of as an informational asset for health care delivery, monitoring, and marketing, it would behoove the medical establishment to guide societal and legislative standards in this regard. One constructive response would be to regulate what is legal and ethical, to ensure that benefits outweigh risks, and to include patients in the decision-making process.[8] An alternative approach would simply be to put the onus entirely on the patients and give them control over their data. However, as has been seen for far less private data, individuals are likely to share

Corresponding Author: Griffin M. Weber, MD, PhD, Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck St, Room 316A, Boston, MA 02115 (weber@hms.harvard.edu).

CRI Policy & Perspectives:

- **Finding the missing link for big biomedical data** (Weber et al)



CRI Policy & Perspectives:

- **Clinical Research Informatics and Electronic Health Record Data** (Richesson R, et al. IMIA Yearbook of Medical Informatics. 2014)
- Survey of issues related to EHR data reuse for research in context of LHS. Identification of major challenges related to data quality, completeness, and provenance.
- Discussion of issues including: integrating data from heterogeneous sources, guidelines (including explicit phenotype definitions) for using these data in both pragmatic clinical trials and observational investigations, strong data governance to better understand and control quality of enterprise data, and promotion of national standards for representing and using clinical data.

Clinical Research Informatics and Electronic Health Record Data

R. L. Richesson[1], M. M. Horvath[2], S. A. Rusincovitch[3]
[1] Duke University School of Nursing, Durham, NC, USA
[2] Health Intelligence and Research Services, Duke Health Technology Solutions, Durham, NC, USA
[3] Duke Translational Medicine Institute, Duke University, Durham, NC, USA

Summary
Objectives: The goal of this survey is to discuss the impact of the growing availability of electronic health record (EHR) data on the evolving field of Clinical Research Informatics (CRI), which is the union of biomedical research and informatics.
Results: Major challenges for the use of EHR-derived data for research include the lack of standard methods for ensuring that data quality, completeness, and provenance are sufficient to assess the appropriateness of its use for research. Areas that need continued emphasis include methods for integrating data from heterogeneous sources, guidelines (including explicit phenotype definitions) for using these data in both pragmatic clinical trials and observational investigations, strong data governance to better understand and control quality of enterprise data, and promotion of national standards for representing and using clinical data.
Conclusions: The use of EHR data has become a priority in CRI. Awareness of underlying clinical data collection processes will be essential in order to leverage these data for clinical research and patient care, and will require multi-disciplinary teams representing clinical research, informatics, and healthcare operations. Considerations for the use of EHR data provide a starting point for practical applications and a CRI research agenda, which will be facilitated by CRI's key role in the infrastructure of a learning healthcare system.

Keywords
Biomedical research, electronic health records, data collection, research design

Yearb Med Inform 2014:215-23
http://dx.doi.org/10.15265/IY-2014-0009
Published online August 15, 2014

Introduction

The use of data derived from electronic health records (EHRs) for research and discovery is a growing area of investigation in clinical research informatics (CRI), defined as the intersection of research and biomedical informatics [1]. CRI has matured in recent years to be a prominent and active informatics sub-discipline [1, 2]. CRI develops tools and methods to support researchers in study design, recruitment, and data collection, acquisition (including from EHR sources), and analysis [1]. To complement the "Big Data" theme of the IMIA 2014 Yearbook, this summary explores the impact of increasing volumes of EHR data on the field of CRI.

There is tremendous potential for leveraging electronic clinical data to solve complex problems in medicine [3]. The impact on the CRI domain is exemplified by a growing number of publications related to the use of EHRs, including medical record systems, algorithms and methods [4]. The analysis of existing clinical, environmental, and genomic data for predicting diseases and health outcomes is growing [5-7]. The regulatory and ethical challenges for using EHR data for research – though complex – are being addressed [8, 9]. Research use of EHR data is inherent to the vision of the learning healthcare system [10]. In this context, CRI will play a central role bridging different perspectives from research and healthcare operations, particularly as they relate to new demonstrations of interventional clinical trials embedded within healthcare systems [11]. The more immediate uses of EHR data are for observational research (i.e., investigations that observe and explore patient phenomena related to the "natural" – rather than researcher controlled – assignment of interventions), because these designs have less inherent risk and disruption to clinical workflows than do interventional trials.

Definitions

Clinical research is the science that supports the evaluation of safety and effectiveness of therapeutics (medications and devices), diagnostic tools, and treatment regimens. Clinical research includes a variety of study designs and methods to support patient-oriented research (i.e., conducted with human subjects or their biospecimens), clinical trials, outcomes research, epidemiologic and behavioral studies, and health services research [12]. Clinical research informatics, then, is the branch of informatics that supports all these research activities, particularly the collection, management, and analysis of data for varied types of studies. Research approaches can be characterized broadly as either interventional (or experimental trials, where the researcher assigns treatments) or observational (where treatments are not assigned by the researcher). To date, CRI has focused largely on the support of interventional trials, but there is momentum around observational research and clinical data mining [6], both of which are particularly relevant to this IMIA Yearbook theme of "Big Data". We defer to other issue authors for precise definitions of the term "Big Data," but premise this discussion on the assumption that the large amounts of clinical and administrative data from institutional repositories and EHR sys-

CRI Policy & Perspectives:

- **Clinical Research Informatics and Electronic Health Record Data** (Richesson R, et al. IMIA Yearbook of Medical Informatics. 2014)

Figure 1: Central role of CRI in a LHS

CRI Policy & Perspectives:

- **Building electronic data infrastructure for comparative effectiveness research: accomplishments, lessons learned and future steps** (Randhawa GS. J. Comp Eff Res. 2014)
- Overview of ~$100million ARRA investment in 12 projects managed by AHRQ to build an electronic clinical data infrastructure that connects research with healthcare delivery.
- Achievements and lessons learned form foundation for initiatives like PCORnet and serve as guides for infrastructure development for an efficient, scalable, and sustainable LHS

SPECIAL FOCUS | The ARRA investment in CER

Opinion

For reprint orders, please contact: reprints@futuremedicine.com

Building electronic data infrastructure for comparative effectiveness research: accomplishments, lessons learned and future steps

Journal of Comparative Effectiveness Research

"The collective infrastructure assets and experiences of these American Recovery and Reinvestment Act-funded projects have laid the foundation to build sustainable learning health systems..."

There are large gaps in our knowledge on the potential impact of diagnostics and therapeutics on outcomes of patients treated in the real world. Comparative effectiveness research aims to fill these gaps to maximize effectiveness of these interventions. Health information technology has the potential to dramatically improve the practice of medicine and of research. This is an overview of about US$100 million of American Recovery and Reinvestment Act investment in 12 projects managed by the Agency for Healthcare Research and Quality to build an electronic clinical data infrastructure that connects research with healthcare delivery. The achievements and lessons learned from these projects provided a foundation for the National Patient-Centered Clinical Research Network (PCORnet) and will help to guide future infrastructure development needed to build an efficient, scalable and sustainable learning health system.

Keywords: American Recovery Reinvestment Act • clinical informatics • comparative effectiveness research • data infrastructure • distributed research • learning health system • quality improvement • registry • sustainability

Gurvaneet S Randhawa
Center for Evidence & Practice Improvement, Agency for Healthcare Research & Quality, Rockville, MD 20850, USA
Tel.: +1 301 427 1619
Fax: +1 301 427 1639
Gurvaneet.Randhawa@ahrq.hhs.gov

Background

A sustained investment in comparative effectiveness research (CER) is needed to bridge the large knowledge gaps on the outcomes of diagnostics and therapeutics on patients treated in the real world. These gaps have been documented in numerous systematic literature reviews [1]. Several factors have contributed to creating these knowledge gaps, including a lack of alignment in the priorities and incentives of researchers and clinicians, lack of ready access to data, varying quality of data created during clinical care, paucity of tools and methods to analyze 'big data', and lack of a health information infrastructure that seamlessly and accurately captures a patient's experience across multiple healthcare delivery sites. The Agency for Healthcare Research and Quality (AHRQ) invested about US$100 million of American Recovery and Reinvestment Act (ARRA) funds to build electronic clinical data systems for collecting patient-centered data that can be used for research, quality improvement and clinical care in order to address these issues.

The investment had two main goals: create an electronic clinical data infrastructure for conduct of CER in diverse diseases, populations, and care delivery sites; and connect the research and healthcare delivery infrastructures to improve efficiency of research and the quality of care. Three programs (consisting of 11 grants) were started to build the infrastructure: Prospective Outcome Systems using Patient-specific Electronic data to Compare Test and therapies (PROSPECT); enhanced registries for Quality Improvement (QI) and CER; and Scalable Distributed Research Networks (DRNs). An additional investment was made to create the Electronic Data Methods (EDM) Forum to foster interdisciplinary collaborations and advance the methods in clinical informatics, analytics, governance and

10.2217/CER.14.73 © 2014 Future Medicine Ltd *J. Comp. Eff. Res.* (2014) 3(5), 567–572 ISSN 2042-6305

CRI Policy & Perspectives:

- **Sustainability Considerations for Health Research and Analytic Data Infrastructures** (Wilcox A, et al. eGEMs. 2014.)
- Framework including factors relevant to developing a sustainability strategy:
  - Assets, expansion, complexity, and stakeholders.
- Each factor is described, with examples of how it is applied.
- These observations are presented as lessons learned, to be applied to other sustainability efforts.

eGEMs

Wilcox et al.: Sustaining Health Research Data Infrastructures

Sustainability Considerations for Health Research and Analytic Data Infrastructures

Adam Wilcox, PhD;[a] Gurvaneet Randhawa, MD, MPH;[a] Peter Embi, MD, MS, FACP, FACMI;[a] Hui Cao, MD, PhD, MS;[b] Gilad J. Kuperman, MD, PhD, FACMI[c]

Abstract

Introduction: The United States has made recent large investments in creating data infrastructures to support the important goals of patient-centered outcomes research (PCOR) and comparative effectiveness research (CER), with still more investment planned. These initial investments, while critical to the creation of the infrastructures, are not expected to sustain them much beyond the initial development. To provide the maximum benefit, the infrastructures need to be sustained through innovative financing models while providing value to PCOR and CER researchers.

Sustainability Factors: Based on our experience with creating flexible sustainability strategies (i.e., strategies that are adaptive to the different characteristics and opportunities of a resource or infrastructure), we define specific factors that are important considerations in developing a sustainability strategy. These factors include assets, expansion, complexity, and stakeholders. Each factor is described, with examples of how it is applied. These factors are dimensions of variation in different resources, to which a sustainability strategy should adapt.

Summary Observations: We also identify specific important considerations for maintaining an infrastructure, so that the long-term intended benefits can be realized. These observations are presented as lessons learned, to be applied to other sustainability efforts. We define the lessons learned, relating them to the defined sustainability factors as interactions between factors.

Conclusion and Next Steps: Using perspectives and experiences from a diverse group of experts, we define broad characteristics of sustainability strategies and important observations, which can vary for different projects. Other descriptions of adaptive, flexible, and successful models of collaboration between stakeholders and data infrastructures can expand this framework by identifying other factors for sustainability, and give more concrete directions on how sustainability can be best achieved.

Introduction

The increasing prevalence of electronic health record (EHR) systems and other health-related data sources is enabling the development of data infrastructures that can more efficiently support research needs in health care, including patient-centered outcomes research (PCOR) and comparative effectiveness research (CER), as well as a wide variety of biomedical research and health delivery-related operational questions. PCOR and CER are seen as emerging approaches to efficiently identify ways to improve health care delivery.[1] To address challenges to the use of health-related data for research, the Agency for Healthcare Research and Quality (AHRQ) has funded several large projects to develop an infrastructure to both support and demonstrate the value of CER and PCOR,[2] as well as to increase the understanding of issues related to creating and using the data infrastructures. The Electronic Data Methods (EDM) Forum, also an AHRQ-funded program, has collaborated with the infrastructure projects to facilitate cross-project learning and to synthesize and disseminate lessons learned.[3–5]

Among the most important issues that have been recognized and considered by the AHRQ-funded projects and the EDM Forum is how to sustain the research data infrastructure beyond the initial investment. The investment in the large infrastructure projects and the EDM Forum was funded through the American Reinvestment and Recover Act of 2009.[3,5] This funding was expected to be a one-time instance, lasting only three years during the initial development stage, though the projects themselves were expected to last beyond the development funding. As a result, efforts to define challenges and solutions to data infrastructure sustainability have been substantial over the course of the projects. In this paper, we have convened a broad group of experts who have participated in these defining efforts. The experts include a principal investigator of an AHRQ-funded infrastructure project, the program officer for the group of projects, a clinical research informatics expert actively applying a sustainability strategy for an existing institutional infrastructure, an informatics researcher working in the pharmaceutical industry, and the leader of a sustained health information

[a]Intermountain Healthcare, [b]Agency for Healthcare Research & Quality, [c]The Ohio State University, [d]Astrazeneca Pharmaceuticals, [e]NewYork-Presbyterian Hospital

CRI Policy & Perspectives:

- eGEMs special issue focused on sustainability for healthcare data

eGEMs

Wilcox and Holve: Sustainability Issue: Message from the Editors

Sustaining the Effective Use of Health Care Data: A Message from the Editors

Adam Wilcox, PhD,[1] Erin Holve, PhD, MPH, MPP[2]

Abstract

Introduction: Over the past decade, several initiatives have funded large projects to develop clinical research data infrastructures totaling several hundred million dollars. While most of this funding has ended or is expected to end soon, the projects themselves must struggle to continue operations beyond the initial funding. Examples of sustained research-data infrastructures are lacking, and recommended approaches to improve sustainability of developing infrastructures are even rarer. Early on, the Electronic Data Methods (EDM) Forum—and the Agency for Healthcare Research and Quality (AHRQ) as its sponsor—recognized the need to study strategies for sustainability.

Themes: Three prominent themes relating to sustainability arise among the articles in this special issue: the importance of project maturity, commercialization activities, and stakeholder support. Maturity was relevant to all the papers since a project's maturity directly influences the opportunities that are available, while commercialization and stakeholder support emerged from comparisons among subsets of articles.

Next Steps: The papers in this issue create a useful initial set of case studies to help in understanding sustainability issues for data infrastructures needed for research and QI. Each paper includes important lessons learned from the authors' experience with the different projects that should resonate with the broader fields of clinical research and clinical research informatics. There is an ongoing need for greater understanding of sustainability beyond what this issue provides. As more case studies of sustainability are accumulated, it is expected even more important themes will emerge from qualitative reviews that can eventually be demonstrated quantitatively.

Introduction

Over the past decade, several initiatives have funded large projects to develop clinical research data infrastructures totaling several hundred million dollars.[1-6] Other substantial investments have occurred with health information technology (HIT) and clinical data exchange projects, with characteristics similar to infrastructure for clinical research data.[7] Initial funding was critical for each project to build collaborations among institutions contributing data, create agreements and methods to actually accumulate the data, and develop and implement software to manage and use the data. While most of this funding has ended or is expected to end soon, the projects themselves must struggle to continue operations beyond the initial funding.

These recent time-limited investments have emphasized the importance of considering sustainability early in the process of *developing* data infrastructure, not only to continue to support operations but also to create an appropriate return on the public funding investments that have been made. In some cases, sustainability plans were not implemented successfully, and those projects failed to survive beyond the initial investment.[8] In other instances, some additional investment was awarded to ensure progress according to a defined sustainability plan. Projects lacking sustainability awareness were less likely to be selected for additional funding from other sources, as subsequent awards required sustainability plans.[9]

Taken together, these examples highlight the need to better understand approaches to achieve sustainability for large clinical data projects. Examples of sustained research-data infrastructures are lacking, and recommended approaches to improve sustainability of developing infrastructures are even rarer. One reason for this gap in knowledge is that, early on, projects often focus substantial effort on navigating data governance and use among changing regulations and incentives during the infrastructure development. As a result, considerations of long-term sustainability were comparatively less urgent.[10] Toward the end of projects, teams may have little time and few resources to discover and develop effective sustainability strategies. Consequently, successful case studies and lessons learned in order to achieve sustainability are often not published or shared—but are desperately needed.

[1]Intermountain Healthcare, [2]Academy Health

Informatics

PDF Sustainability Through Technology Licensing and Commercialization: Lessons Learned from the TRIAD Project
Philip R.O. Payne
Jul 2014, Vol. 2, Iss. 2

PDF Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2)
Shawn N. Murphy and Adam Wilcox
Sep 2014, Vol. 2, Iss. 2

Learning Health System

PDF Sustaining Research Networks: the Twenty-Year Experience of the HMO Research Network
John F. Steiner, Andrea R. Paolino, Ella E. Thompson, and Eric B. Larson
Jun 2014, Vol. 2, Iss. 2

PDF Achieving and Sustaining Automated Health Data Linkages for Learning Systems: Barriers and Solutions
Erik G. Van Eaton, Allison B. Devlin, Emily Beth Devine, David R. Flum, and Peter Tarczy-Hornoch
Jul 2014, Vol. 2, Iss. 2

PDF State Synergies and Disease Surveillance: Creating an Electronic Health Data Communication Model for Cancer Reporting and Comparative Effectiveness Research in Kentucky
Christopher Reams, Mallory Powell, and Rob Edwards
Aug 2014, Vol. 2, Iss. 2

PDF The National Anesthesia Clinical Outcomes Registry: A Sustainable Model for the Information Age?
Richard Dutton
Aug 2014, Vol. 2, Iss. 2

PDF The DARTNet Institute: Seeking a Sustainable Support Mechanism for Electronic Data Enabled Research Networks
Wilson D. Pace, Chet Fox, Turner White, Deborah Graham, Lisa M. Schilling, and David R. West PhD
Sep 2014, Vol. 2, Iss. 2

PDF Sustainability Considerations for Health Research and Analytic Data Infrastructures
Adam Wilcox, Gurvaneet Randhawa, Peter Embi, Hui Cao, and Gil Kuperman
Sep 2014, Vol. 2, Iss. 2

Notable CRI-Related Events

Notable CRI-Related Events

- President's Precision Medicine Initiative
- NIH Big Data initiative
- FDA: Sentinel Project to follow pilot mini-Sentinel
- NLM changes
  - Retirement of Don Lindberg
  - RFI to inform Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM)
- PCORnet phase 1 nearly done…
  - PCORI funding/initiatives (CDRN #2) **(keep writing!)**
- Special issues related to CRI
  - JAMIA, eGEMs, JBI…

Special Journal Issues dedicated to CRI Topics

- Completion of first ever special issue focused on CRI methods
  - Some highlighted this year, some last
  - All worth a read
  - Topics include:
    - Secure, reliable, and reproducible capture, collection, and re-use of data for research purposes.
    - Standards for consistently modeling and devising informatics solutions to research information needs.
    - Approaches to accelerate science via informatics while balancing the very real and critical regulatory and ethical issues inherent in such activities.
    - Better understanding, visualization, and facilitation of research collaboration networks and team science via informatics approaches.
    - Improved user experience and CRI solution development through better understanding of research users' information needs.
    - The integration of clinical practice and research activities in order to leverage our health system for evidence generation.

Embi & Payne, JBI, 2014

One more notable event:
Announcement of Apple Research Kit

- Great potential for new research across populations
- Novel consent procedures, innovation opportunities
- Nearly 40K people signed up to participate in first 72 hours with just one of the apps (Parkinson's study)

In Summary…

- Maturing informatics approaches in CRI – accelerating
  - Much more activity than in years past
  - I'm sure it will only continue!
- CRI infrastructure also maturing and beginning to drive science
- Multiple groups/initiatives converging on common needs to advance the field
- CRI initiatives and investments beginning to realize the vision of the "*learning health system*"

- Exciting time to be in CRI!

Thanks!

Special thanks to those who suggested articles/ events to highlight, particularly:

- Philip Payne
- Rachel Richesson
- Adam Wilcox
- Chunhua Weng
- Erin Holve

Thanks!

Peter.Embi@osumc.edu

Slides will be linked to on

http://www.embi.net/ (click on “Informatics”)