EDM Forum

Comparative Effectiveness Research (CER) and Clinical Informatics:
An Annotated Bibliography

by Hamilton Lopez M, Nagda SR, Holve E, Sarkar IN

1. Introduction

Significant national investment in developing key infrastructure to support Comparative Effectiveness Research (CER) necessitates developing techniques for information and knowledge management, such as those pioneered by the Clinical Informatics community. Motivated by the need for resources that specifically track the literature at the intersection of CER and Clinical Informatics, this report provides an annotated bibliography of 132 articles in the peer-reviewed literature.

In particular, this review emphasizes publications that provide insight into the uses of clinical informatics for population-based research. The selected articles address the use of clinical informatics for CER from a historical perspective and consider the development and application of several key projects, platforms, networks, and technologies that are most relevant for CER.

The review also includes information on the structured approach used to identify key papers that are relevant to the emerging body of literature related to developing CER infrastructure with prospective electronic clinical data.

A second report within this series will include annotations based on a review of the grey literature. A literature review, with an analysis of key themes, concepts, and gaps is forthcoming.

2. Methods

Data Collection

The relatively new use of the term CER, as well as the substantial number of concepts that are important to include or reflect on when identifying CER, complicate efforts to develop an automated search for the term. Efforts to identify an appropriate subset of the literature tend to generate extremely small result sets of literature annotated with MeSH terms for "comparative effectiveness research" as a topic (n=375), or extremely large result sets if coded with a range of concepts associated with CER (one to two million citations or more).

To ensure a high-quality review, we employed a curated approach to identify the appropriate subset of the literature. Our efforts utilized a manually chosen set of MeSH-based MEDLINE searches focused on key projects, as well as programs and authors. These efforts were combined with a manual, in-depth review of specific papers that were more effectively located by referencing select publication lists (rather than using MeSH descriptors).[1] For example, in an effort to identify research activities that may have relevance to the EDM Forum (e.g. research based on electronic prospective clinical data), we drew from the following projects (to view the projects' websites, you may select the following links):

- caBIG (cancer Biomedical Informatics Grid)
- DARTNet
- DEcIDE (Developing Evidence to Inform Decisions about Effectiveness)
- HMORN (HMO Research Network)
- iDASH (integrating data for analysis, anonymization, and sharing)
- i2b2 (Informatics for Integrating Biology and the Bedside)
- OMOP (Observational Medical Outcomes Partnership)
- PhysioMIMI (Multi-Modality, Multi-Resource Information Integration environment)
- RedCAP (Research Electronic Data Capture)
- Sentinel Initiative and Mini-Sentinel-
- SHARP Program (Strategic Health IT Advanced Research Projects)
- TRIAD (OSU Clinical and Translational Science Awards)
- VINCI (VA Informatics and Computing Infrastructure)

In total 2,435 papers were identified as potentially relevant for the review. After de-duplicating papers and conducting an initial review of titles and abstracts for relevance, this set was reduced to approximately 400 peer-reviewed articles, in addition to several important reports in the grey literature.

Review

In order to select the most relevant subset for in-depth review, we used a set of exclusion criteria. The following articles were excluded from consideration:

- Articles that that were not explicitly related to clinical informatics;
- Articles that focused solely on clinical outcomes rather than discussing the use of informatics for research; and
- Articles that discussed genomic rather than clinical data.

As a result of this process, a set of 147 peer-reviewed papers was identified for full-text review.

The selected body of work was divided between two reviewers, who each read the full text version of assigned articles. Each reviewer used a standardized abstraction form designed specifically for this project. The form included key information such as the article's primary objective, methods, themes, key quotes, definitions, research findings, author's perspectives/recommendations, and funding sources. To validate findings, the reviewers conducted a second review of a third of the articles from the other reviewer to compare abstracted themes and information. After full-text review, additional articles were excluded based upon the criteria outlined above. In total, 132 articles were identified for the annotated bibliography.

The annotations were then organized into primary themes identified by the reviewers. The themes include: (I) General Overview, (II) Platforms & Projects, (III) Natural Language Processing, (IV) Research Networks, (V) Data use and quality, and (VI) Other: Identifiers and De-Identification; IRBs; Governance; Library of Phenotypes; Metadata; Patient Involvement; Security; Standardized data collection; The Learning Healthcare system and CER.

1 This structured approach resulted in articles selected from the following sources:
- PubMed Searches for specific concepts:
  - A search for clinical informatics and CER with the following string:
    - ((“Informatics”[tiab] OR “Informatics”[MH] OR “Medical Informatics”[MH]) OR (“data mining”[MH] OR “information storage and retrieval”[mh])) AND (“Comparative Effectiveness”[tiab] OR “Comparative Effectiveness Research”[MH]).
  - A keyword search for papers about a “Learning Healthcare System.”
  - Papers produced by research activities that have relevance for the EDM Forum
- Relevant references from the PROSPECT, DRN, and Enhanced Registries studies.
- Relevant articles from the HIT for Actionable Knowledge Annotated Bibliography.
- A subset of papers presented at the 2010 AMIA Symposium.

I. General Overview

Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, Bach PB, Murphy SB. Rapid-learning system for cancer care. J Clin Oncol. 2010 Sep 20; 28(27):4268-74.
Abernethy et al. examine the elements of a rapid-learning system for cancer care by summarizing the 2009 Institute of Medicine's National Cancer Policy Forum "A Foundation for Evidence-Driven Practice: A Rapid-Learning System for Cancer Care." Health IT limitations in creating a learning health care system include selection bias, lag time of linkage, and the challenges of assessing treatment outcomes. The authors argue that Health IT can be used for clinical data, but these efforts need to be coordinated among the groups that are conducting research, and these approaches need to be tested on a national level in order to ensure that the results are useful. The cancer-learning health care system will link research and clinical efforts in the continuously improving treatment of cancer and practice of oncology. This must be balanced by the clinical work that practitioners need to do as their primary function, and the IT requirements should not take away from that.

Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part I. Value Health. 2009 Nov-Dec;12(8):1044-52.
Berger et al. summarize recommendations from the September 2007 ISPOR Good Research Practices for Retrospective Database Analysis Task Force. The Task Force outlined how to ask a good research question, what type of study is most beneficial to answer the question, how to report the data, and how to interpret the results. Because policymakers are often reluctant to support observational data, due to associated bias and confounding, it is important to define a research question that appropriately provides the rationale for conducting observational research. The four primary characteristics proposed for a good research question are: (1) relevance and rationale; (2) specificity; (3) novelty; and (4) feasibility in order for the results to be translated to clinical use.

Corn, M, Rudzinski, KA, Cahn, MA. Bridging the gap in medical informatics and health services research: workshop results and next steps. J Am Med Inform Assoc. 2002;9(2): 140-3.
This article summarizes findings and discussion topics from the January 2000 "Medical Informatics and Health Services Research: Bridging the Gap" workshop cosponsored by AHRQ and the NLM. Attendees included over 100 educators and researchers in medical informatics and health services research. The workshop emphasized the utility of informatics in the conduct of health services research underscored with the importance of translating research findings to the clinical setting. Two identified barriers for interaction between biomedical informatics and health services research were: financial and fostering collaboration. Additionally, there is a lack of professionals in informatics that can progress the necessary work in the medical realm. Next steps include putting emphasis on training programs to cultivate these assets in the field.

Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report--Part II. Value Health. 2009 Nov-Dec;12(8):1053-61.
This is a report by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Good Research Practices for Retrospective Database Analysis Task Force, a multidisciplinary team established in September 2007. The report discusses the measurement of exposure and outcome in secondary data sources. It also examines bias and confounding, which exist in the data collection practices of researchers conducting CER. The Task Force recommends that acknowledging that these issues exist, and determining how to improve the data, will lead to better CER results. Additionally, using causal diagrams and restriction can improve generalizability of the CER results.

D'Avolio LW, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. Am J Med. 2010 Dec;123(12 Suppl 1):e32-7.
D'Avolio et al. provide an overview of the process through which researchers can conduct medical informatics and comparative effectiveness research. The four focus areas that are outlined are: information access, information structure, information analysis, and information interaction. Examples from the VA Healthcare System are offered to describe how these focus areas can be used to leverage HIT for CER research.

D'Avolio LW. Electronic medical records at a crossroads: impetus for change or missed opportunity? JAMA. 2009 Sep 9;302(10):1109-11.
D'Avolio comments upon the electronic medical record (EMR) systems that are in use in the United States following the passage of the American Recovery and Reinvestment Act (ARRA). While ARRA provides investments in EMRs and HIT, there is a lack of focus on using health data to improve quality of care. The author argues that the definition of "meaningful use" of HIT should include requiring ease of access to data and the ability to measure outcomes.

de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam Pract. 2006 Apr;23(2):253-63.
The authors outline the opportunities and challenges related to the utilization of technology to promote health care research. Opportunities include the increasing amount of stored clinical data, improved data quality, and open-sourcing of new technologies for data integration. Challenges include limited sharing of research methods, gleaning meaning from data, and protecting the privacy and security of the data. Areas for improvement in the field include data ownership, privacy, and temporal issues with adopting technology in medicine.

Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc. 2009 May-Jun;16(3):316-27.
Using qualitative methods, Embi and Payne seek to define the domain and scope of Clinical Research Informatics (CRI) as an emerging field. A series of personal and electronic discussions were conducted with AMIA CRI Working Group members. As a result, 13 different themes were identified as being pertinent to the study of CRI, including research planning and conduct; data access, integration, and analysis; regulatory and policy issues; standards; workflow; and CRI innovation and investigation.

Etheredge LM. Creating a high-performance system for comparative effectiveness research. Health Aff (Millwood). 2010 Oct;29(10):1761-7.
According to Etheredge, a greater source of clinical data could lead to better and more comparative effectiveness research. The author argues that this source of data could come from a national clinical research system, built on electronic health records and managed by the Department of Health and Human Services. These efforts would contribute to a national rapid-learning cycle.

Kukafka R, Ancker JS, Chan C, Chelico J, Khan S, Mortoti S, Natarajan K, Presley K, Stephens K. Redesigning electronic health record systems to support public health. J Biomed Inform. 2007 Aug;40(4):398-409.
Kukafka et al. conduct a review of the literature on EHR use for public health. Public health core functions were outlines as assessment, policy development, and assurance. They also list the Certification Commission for Health Information Technology (CCHIT) requirements for EHRs, including: functionality, security, reliability, and interoperability standards. The authors discuss how the research needs in public health should shape the EHR design.

Mandl KD, Lee TH. Integrating medical informatics and health services research: the need for dual training at the clinical health systems and policy levels. J Am Med Inform Assoc. 2002 Mar-Apr;9(2):127-32.
Mandl et al. argue that there must be an integration of health services research and informaticians (through collaborations and dual training) in order to structure and then effectively use information systems and databases to improve health care. However, there are barriers to collaboration between health services researchers and informaticians that include: the lack of researchers trained in both fields; the physical work setting separation between the two fields; and different problem-solving approaches. The authors recommend that a needs assessment is needed to estimate workforce requirements, cross-hybridization of existing training programs and the creation of new programs, and research at the gap.

McCray AT, Scherrer JR, Safran C, Chute CG. Concepts, knowledge, and language in health-care information systems. Methods Inf Med. 1995 Mar;34(1-2):1-4.
McCray et al. outline the goals for health care information systems, as determined by the first meeting of the International Medical Informatics Association Working Group in 1984. This article shows the evolution of the field including its multi-disciplinary roots (1988 conference participants came from the following sectors: linguistics, natural language analysis, medical decision makers, knowledge representation, and computer science). The 1994 conference resulted in two recommendation categories: (1) Mechanisms for sharing research; and (2) Experiments, Testbeds, and Demonstration Projects. Many issues brought forth during these first meetings such as knowledge acquisition, clinical information management, and the incorporation of multiple stakeholders continue to be discussed and debated today.

McKinney M. 'Huge potential' for EHRs and comparative effectiveness. Hosp Health Netw. 2010 Apr;84(4):41-2.
McKinney provides background information about the recent funding for EHRs and comparative effectiveness and the need for, and direction of, CER research. Current electronic health records have limitations, both with their ability to capture necessary data and capture continuum of care. The new focus on EMRs and comparative effectiveness may have implications for how data are gathered and studied for CER.

Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. Methods Inf Med. 2009;48(1):45-54.
Ohmann et al. advocate for open sharing of information in order to further the field of biomedical informatics. One of the most important tasks is to achieve interoperability between EHR and EDC. One consideration is that the data used for clinical purposes may be different (in content and structure) from the data used for clinical trial recruitment, surveillance, and research. Standards must be created and applied in order for technical interoperability to be achieved.

Pagliari C. Design and evaluation in eHealth: challenges and implications for an interdisciplinary field. J Med Internet Res. 2007 May 27;9(2):e15.
Pagliari describes the interdisciplinary aspect of eHealth systems and services, specifically the collaboration between software developers and health services researchers. Although silos still exist between software development and health service research, the eHealth field has become increasingly interdisciplinary with the introduction of social, economic, and legal sciences, and bioinformatics. Barriers to collaboration include departmental differences in how research problems are addressed, lack of knowledge of different fields and their operational processes, and un-standardized terminologies. However, the author argues that commonalities exist and that eHealth relies on interdisciplinary collaborations.

Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. Towards a data sharing culture: recommendations for leadership from academic health centers. PLoS Med. 2008; 5: e183.
Piwowar et al. present the challenges facing data sharing and make the case for academic health centers' leadership. The authors argue that, in the current climate, barriers to sharing data include technical limitations and the lack of incentives for researchers. Donors, investigators, academic health centers, and industrial sponsors all have financial and intellectual property interests in research data. Nevertheless, there are benefits to sharing. It leads to collaborations and scientific discovers which result in better patient outcomes and reduced research costs. Also, increased visibility of research outputs result in increased publications and more funding opportunities. The authors recommend the encouragement of data sharing contributors and a data sharing infrastructure, the integration of data sharing into education curricula, and a shift in policies to sustain a data sharing environment.

Rubin D, Napel S. Imaging informatics: toward capturing and processing semantic information in radiology images. Yearb Med Inform. 2010; 34-42.
Rubin provides a landscape view of the evolving field of imaging informatics research. Based on a literature review of recent imaging informatics articles, the following three themes emerged: standard terminologies and ontologies for describing images, structured representation of image content, and retrieval of image content for decision support. The author found similarities between imaging and other sub-disciplines of biomedical informatics including the need for standardization, the increase in the amount of available data, and the management of large databases.

Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc. 2007 Jan-Feb;14(1):1-9.
This article summarizes a discussion on the secondary use of health data that occurred at an AMIA meeting expert panel in April 2006, including 36 clinical and non-clinical stakeholders. The consumer perspective highlighted issues related to data security and privacy, personal data ownership, and consent for nonclinical use of health data. The patient safety, quality, and research perspective highlighted issues related to research data, including de-identification, the quality of claims data in the conduct of clinical research, and the development of standards for establishing levels of evidence. The public health perspective discussed how clinical data could be used for public health purposes, including biosurveillance, epidemiology, emergency preparedness, and homeland security.

Sarkar IN. Biomedical informatics and translational medicine. J Transl Med. 2010 Feb 26;8:22.
Sarkar discusses the use of informatics to improve bedside and community medicine, as well as to shape policy through translational research. Between "T1," "T2" and "T3" steps, translation barriers exist that prevent research from being adopted. Sarkar depicts the relationship of the translational medicine and biomedical informatics continuums, and describes how these two fields can synergistically coexist. Clinical informatics themes--including decision support systems, natural language processing, data standards, information retrieval, and electronic health records--are defined in this article in relation to translational medicine.

Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. Ann Intern Med. 2009 Aug 4;151(3):203-5.
In this article, Sox provides an overview of the Comparative Effectiveness Research recommendations presented in 2009 by the Institute of Medicine. The recommendations were developed through a series of public comments, hearing presentations, and a three-step voting process in which the committee defined 100 high-priority topics. Of particular interest is the focus on consumer engagement in the development and implementation of CER, the call to overcome the limitations of observational research, and the importance given to large collections of patients' electronic health records as a tool for CER.

Turisco F, Keogh D, Stubbs C, Glaser J, Crowley Jr WF. Current status of integrating information technologies into the clinical research enterprise within US academic health centers: strategic value and opportunities for investment. J Investig Med. 2005 Dec;53(8):425-33.
Turisco et al. present the findings from a survey of Academic Health Centers' (AHC) vision and capabilities for a comprehensive clinical research IT program. The investigators surveyed 37 AHCs (all members of the Clinical Research Forum). Responses were tabulated (with a respondent rate of 78%), reviewed for findings and outcomes and presented at the 2005 Forum's annual meeting. Some findings include: unanimous support of the importance of a "state-of-the-art" clinical research IT program and its future necessity, zero respondents had in place such a program or the management foundations to build such a program.

Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med. 2009 Sep 1;151(5):359-60.
In this editorial, Weiner and Embi provide an overview of informatics needs, the benefits of a distributed network, and the reasons why multidisciplinary collaboration is crucial. They argue that the necessary technology is there and a viable distributed network must include collaborative efforts between all stakeholders. The quality of electronic patient information has implications for the future of population health, pharmaceutical surveillance, and comparative effectiveness research.

II. Platforms & Projects

Clinical Informatics Platforms

Bellin E, Fletcher DD, Geberer N, Islam S, Srivastava N. Democratizing information creation from health care data for quality improvement, research, and education-the Montefiore Medical Center Experience. Acad Med. 2010 Aug;85(8):1362-8.
This paper describes the development and use of the Clinical Looking Glass (CLG) software at the Montefiore Medical Center at the Albert Einstein College of Medicine. This informatics system recognizes EMRs from any vendor and stores deidentified information in a data repository for subsequent longitudinal data analysis and statistical comparisons. CLG has data sharing capabilities to allow researchers to contribute to a larger knowledge base. Experts in the field can also use CLG's wiki capability to correct and contribute to data. To date, Montefiore's CLG system has been used for comparative effectiveness research of treatments for diabetes.

Buetow KH. An infrastructure for interconnecting research institutions. Drug Discov Today. 2009; 14:605-10.
Buetow describes caBIG, an infrastructure that addresses interoperability, data sharing, and data security issues related to cancer data in the National Cancer Institute cancer centers. The goal of caBIG is the creation of a virtual web of interconnected data, individuals, and organizations. This will redefine how treatment-focused research is conducted resulting in improved patient/participant interaction with biomedical organizations and, ultimately, improved patient outcomes. caBIG uses the GAARDS (Grid Authentication and Authorization with Reliably Distributed Services) security infrastructure. Research is turning quickly from traditional siloed approaches to identifying more innovative ways to collaborate and share data. The caBIG program addresses some of the main barriers to working collaboratively and provides comprehensive solutions to managing data while ensuring data security.

Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc. 2010 Mar-Apr;17(2):131-5.
Chute et al detail the formation and utilization of the Mayo Clinic's Enterprise Data Trust (EDT), a clinical data warehouse used for patient care, administrative, research, and educational purposes. The contents of the EDT are subject oriented, semantically integrated, time variant, and non-volatile. All of the Mayo Clinic services aggregate their data into the EDT using a common vocabulary. The LexGrid platform is utilized to standardize the terminology of the data in the EDT. The EDT system goes beyond EHR capabilities to allow for research and analytics. The Mayo Clinic's Enterprise Data Trust has been successful in utilizing clinical data to improve quality, and is transforming healthcare and research at the Mayo Clinic.

Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc. 2000 May-Jun;7(3):298-303.
In this article, Chute describes the evolution of naming mechanisms in medicine. SNOMED is useful in the naming of various medical concepts. However, its flexibility can cause difficulty by allowing for multiple names for a single concept and preventing the true standardization of the process. This issue is only worsened as multiple competing parallel naming mechanisms were developed, including a naming mechanisms created by the American College of Radiology, and the International Classification of Diseases (ICD).

Chute, C. G., Yang, Y., & Evans, D. A. (1991). Latent Semantic Indexing of medical diagnoses using UMLS semantic structures. Proc Annu Symp Comput Appl Med Care, 185-189.
The authors evaluated the UMLS Metathesaurus as a means to extract medical diagnoses from the medical record. The words and phrases in a medical record can have a lot of variation (number, case, tense, adverbial, or other inflections). This project used a morphology lexicon that reduced variants to a singular noun form. Three concept lists were created in order to test the LSI technique: Tiny-Input (10 matrices), Midi-Input (101 matrices), and Maxi-Input (2,580 matrices). All examples used inquiries for carcinoma of the lung, myocardial infarction, and cerebral ischemia. This latent semantic indexing (LSI) technique is useful and successful in the process of categorization and retrieval of patient data in the electronic medical record.

Cimino JJ. Terminology tools: state of the art and practical lessons. Methods Inf Med. 2001;40(4):298-306.
Cimino conducted a review of the literature on terminology tools, and presents a case from Columbia University/New York Presbyterian Hospital on the development of the Medical Entities Dictionary (MED). The MED was developed specifically for Columbia University, but has been successfully used to browse SNOMED-RT and UMLS. Future directions include creating distributed editing capabilities and to allow for commercial systems to utilize MED.

D'Avolio LW, Bui AA. The Clinical Outcomes Assessment Toolkit: a framework to support automated clinical records-based outcomes assessment and performance measurement research. J Am Med Inform Assoc. 2008 May-Jun;15(3):333-40.
The authors describe the Clinical Outcomes Assessment Toolkit (COAT), which is a framework to conduct clinical outcomes and performance measurement research through the collaboration of UCLA and the Brigham and Women's Hospital. COAT allows developers to create information pipelines for the reuse and analysis of clinical data.

Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med Res Methodol. 2009 Oct 28;9:70.
The authors evaluated the cohort selection tool, implemented in the Integrating Biology and the Bedside (i2b2) hive. The i2b2 hive architecture was used at the University of Utah Healthcare System to determine the generalizability of the i2b2 model to other systems. Data requests were evaluated on several factors: (1) request of counts (2) types of data requested (3) institution-specific requests (4) features of the query (temporal or exclusion criteria, calculated fields, aggregate data) (5) ability to conduct request using i2b2 alone or with little modification, and (6) reasons it was not possible to use i2b2 alone in conducting the data request. Findings: The i2b2 hive was useful in estimating cohort sizes and generating research cohorts from predetermined inclusion and exclusion criteria. The various reasons why the i2b2 hive could not be used to conduct the remaining data requests include temporal issues, exception conditions, calculated fields, and metadata modifications.

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377-81.
Harris et al. present the Research Electronic Data Capture (REDCap), a metadata-driven software toolset and workflow system sponsored by the National Center for Research Resources (NCRR). Piloted at Vanderbilt University, and now being supported by twenty-seven partner institutions, REDCap uses individualized data element requirements to create electronic data capture tools that can be used for a multitude of clinical and translational research purposes.

Hastings S, Oster S, Langella S, Melean C, Borlawsky T, Dhavel R, Payne P. Adoption and Adaptation of caGrid for CTSA. Summit on Translat Bioinforma. 2009 Mar 1;2009:44-8.
The purpose of this article is to describe the use of the caGRID infrastructure as a backbone of the Ohio State University (OSU) CTSA project, TRIAD (Translational Informatics and Data Management Grid). TRIAD is a middleware system that leverages service-oriented architecture, data model management, security infrastructure, grid service creation, and grid service infrastructure.

Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc. 2005 Sep-Oct;12(5):517-29.
Hazlehurst et al. describe the history, influences, and application of MediClass, a knowledge-based system that extracts clinical event information as UMLS concepts from both free-text and coded data. There are three "layers" of the MediClass process architecture that focus on system integration, concept identification, and classification. Lexical processing, concept identification, and concept instantiation are used to extract the data being sought. MediClass is demonstrated by a case of data extraction for smoking cessation discussions in the clinical encounter.

Jiang G, Solbrig HR, Iberson-Hurst D, Kush RD, Chute CG. A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki. AMIA Summits Transl Sci Proc. 2010 Mar 1;2010:11-5.
Jiang et al. present the results of a pilot study that implement CDISC Shared Health and Research Electronic Library (CSHARE) using Semantic MediaWiki to create a harmonized framework for clinical study data elements. There were four roles identified in the harmonization process: Classification, Definition, Value Meanings, and Value Sets. Classification used "semantic keywords" to determine if there was any relation between two or more of the components. Definition of the data elements was another role that was served, and this should be distinguished from classification of the elements. In cases where a value represents a piece of data, the meaning of the code can be mapped to its terminology (e.g. 1 = male, 2 = female). Value sets are simply a group of value meanings for a particular data element.

McConnell P, Dash RC, Chilukuri R, Pietrobon R, Johnson K, Annechiarico R, Cuticchia AJ. The cancer translational research informatics platform. BMC Med Inform Decis Mak. 2008 Dec 24;8:60.
McConnell et al. describe the Cancer Translational Research Informatics Platform (caTRIP) tool, which is designed to allow for intra-institutional aggregated clinical and molecular data sharing from different domains and data sources. The platform was developed on an N-tier architecture, which is based on standardization and security requirements developed by the Cancer Biomedical Informatics Grid (caGIG) initiative, and provides simple to complex query capabilities.

McMurry AJ, Gilbert CA, Reis BY, Chueh HC, Kohane IS, Mandl KD. A self-scaling, distributed information architecture for public health, research, and clinical care. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):527-33.
McMurry et al. describe the development of a scalable architecture to support the National Health Information Network. The desiderata that are included in this infrastructure include: (1) distributed data approach, (2) institutional autonomy, (3) oversight and transparency, (4) variable access to data based on research needs, and (5) self-scaling architecture that allows regional participation to build a national network.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30.
Murphy et al. provide an overview of the Informatics for Integrating Biology and the Bedside (i2b2) software. Through i2b2, investigators have access to de-identified data that is available for queries and aggregated data on the demographics of patient sets (all patient data remains anonymous). This software has the potential to be used for clinical trial recruitments and for providing a source of de-identified and secure patient information that can be used for research purposes (both basis and possibly complex).

Pathak J, Solbrig HR, Buntrock JD, Johnson TM, Chute CG. LexGrid: a framework for representing, storing, and querying biomedical terminologies from simple to sublime. J Am Med Inform Assoc. 2009 May-Jun;16(3):305-15.
Pathak et al. describe the LexGrid framework for representing, storing, and querying biomedical terminologies. LexGrid utilizes terminology and ontology tools to create a standardized format that can be represented uniformly in multiple platforms. Standards must be created and applied in order for technical interoperability to be achieved.

Sharma A, Pan T, Cambazoglu BB, Gurcan M, Kurc T, Saltz J. VirtualPACS - a federating gateway to access remote image data resources over the grid. J Digit Imaging. 2009; 22: 1-10.
Historically, there have not been the mechanisms available to securely and efficiently share images electronically. Instead, image data have been burned on CDs and shared through the mail. Sharma et al. address this limitation by developing a toolkit called VirtualPACS that allows queries and the ability to retrieve and submit functions across a distributed image database. VirtualPACS has three layers: presentation, middleware, and the implementation layer mediation.

Staes CJ, Xu W, LeFevre SD, Price RC, Narus SP, Gundlapalli, Rolfs R, Nangle B, Samore M, Facelli JC. A case for using grid architecture for state public health informatics: the Utah perspective. BMC Med Inform Dec Mak. 2009; 9:32.
This paper discusses the lessons learned from the University of Utah's implementation of a Utah Public Health Informatics Grid (after a one year period and across multiple administrative domains). Staes et al. argue that most public health informatics systems are contained in "data silos" and discovered 79 silo applications and databases within the University prior to the study. Some specifics about the Public Health Informatics Grid implemented here include its open source software that works with legacy applications and the allowance of both distributed and federated databases.

Viangteeravat T, Brooks I, Vuthipadadon S, Huang E, Smith E, Homayouni R, McDonald C. Slim-Prim: an integrated data system for clinical and translational research. BMC Bioinformatics. 2009; 10.
In an effort to support translational science, Viangteeravat et al., developed Slim-Prim an integrated data system. The system is web-based and integrates basic and clinical sciences data from multiple sources. It has controlled access by Project Managers as a way to adhere with HIPAA and includes a metadata design.

Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):624-30
The authors discuss the development of a prototype Shared Health Research Information Network (SHRINE), a Harvard CTSA project that identifies the challenges of creating a federated query tool with the capability to search across multiple Harvard clinical databases. They describe the background of Harvard's work with various query tools, including ClinQuery, W3EMRS, CareWeb, Goldminer, DXtracter, the Research Patient Data Registry, and Informatics for Integrating Biology and the Bedside (i2b2)/i2b2 Clinical Research Chart (open source platform that enables NLP). IRB approval had to be completed for all three institutions separately.

Overview of Clinical Informatics Projects

Aiello Bowles EJ, Tuzzio L, Ritzwoller DP, Williams AE, Ross T, Wagner EH, Neslund-Dudas C, Altschuler A, Quinn V, Hornbrook M, Nekhlyudov L. Accuracy and complexities of using automated clinical data for capturing chemotherapy administrations: implications for future research. Med Care. 2009 Oct;47(10):1091-7.
Aiello Bowles et al. use data aggregated from the HMO Cancer Research Network Virtual Data Warehouse to extract information regarding ovarian cancer chemotherapy treatment. Using Health Care Procedure Coding System (HCPCS) codes, National Drug Codes (NDCs) from pharmacy data, and ICD-9-DM diagnostic codes, the authors extracted data about whether 757 ovarian cancer patients had chemotherapy treatment or not. With just a single code, the sensitivity and specificity of the data extraction was not adequate. The combination of the three codes made this a more accurate method of obtaining chemotherapy data in comparison to only utilizing the tumor registry.

Amin W, Singh H, Pople AK, Winters S, Dhir R, Becich MJ. A decade of experience in the development and implementation of tissue banking informatics tools for intra and inter-institutional translational research. J Pathol Inform. 2010; 1:12.
Amin et al. provide information about the advance tissue banking informatics tools (an Oracle-based organ-specific data mart and a model-driven architecture for biorepositories) developed over the last decade by the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh. The developed tools are interactive with querying capability, utilize an Honest Broker for HIPAA compliance, and support the standardization of clinical annotations.

Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. N Engl J Med. 2011 Feb 10;364(6):498-9.
The FDA has piloted the mini-Sentinel program to conduct drug safety monitoring in a distributed database of over 60 million people. This project advances the efforts already done by, among others, the Vaccine Safety Datalink project, OMOP, PopMedNet, and HMORN. The authors highlight the importance of protecting patient privacy through the use of the distributed data model and the idea of utilizing electronic data to conduct postmarketing pharmaceutical surveillance.

Chen RT, Glasser JW, Rhodes PH, Davis RL, Barlow WE, Thompson RS, Mullooly JP, Black SB, Shinefield HR, Vadheim CM, Marcy SM, Ward JI, Wise RP, Wassilak SG, Hadler SC. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. Pediatrics. 1997 Jun;99(6):765-73.
Previous surveillance systems rely on passive postmarketing reporting, which leads to underreporting and bias. Chen et al. describe the Vaccine Safety Datalink (VSD) Project, which monitors reactions and adverse events related to vaccination. Vaccination records from children ages 0-6 years from four different HMOs were included. There were 549,618 administrations of DTP (diphtheria-tetanus-pertussis) and 310,618 administrations of MMR (measles-mumps-rubella) in the cohort. Analysis of the data and thrity-four possible adverse events showed a 2.1 relative risk of seizures the same day following DTP, and 3.0 for eight to forteen days following MMR. The study showed that the RR of seizures from Haemophilus influenza type b (Hib) between eight to 14 days of follow-up (1.6) was actually attributed to confounding, and was dropped to 1.0 after adjusting for concomitant MMR vaccination. The design, data extraction methods, and analytic capability of VSD have been validated.

Eisen S, Francis J. Transformation of VHA health data into clinically useful information to provide quality veteran care. J Rehabil Res Dev. 2010;47(8):xiii-xv.
Eisen and Francis describe the Veterans Health Administration (VHA) healthcare system, its HIT initiatives and uses, and the next steps being taken to enhance its utilization for clinical and research purposes. The VHA health system is an integrated system with a large amount of data from which to conduct quality improvement research. Findings from this data could be useful for the veteran population and generalizable to the larger population. The VHA will be making improvements to its EMR system by allowing for structured documentation and moving its data to a centralized system for administrative and clinical purposes. VHA is also considering using NLP to extract meaningful data from its records. The vastness of the VHA data systems, and the work that is being done using this information, could greatly impact quality of care as well as hospital management.

Fridsma DB, Evans J, Hastak S, Mead CN.The BRIDG project: a technical report. J Am Med Inform Assoc. 2008; 15:130-7.
Fridsma et al. describe the Biomedical Research Integrated Domain Group (BRIDG) Project, a collaborative effort between HL7, CDISC, caBIG, and the FDA. The BRIDG Project uses Unified Modeling Language (UML) to support the goal of semantic interoperability across applications. The BRIDG Project uses the consensus through harmonization method to unify concepts among the various models. This model has been used by the National Cancer Institute's caBIG to support application development. HL7 and CDISC are utilizing the program to support standards development.

McDonald CJ, Overhage JM, Tierney WM, Dexter PR, Martin DK, Suico JG, Zafar A, Schadow G, Blevins L, Glazener T, Meeks-Johnson J, Lemmon L, Warvel J, Porterfield B, Warvel J, Cassidy P, Lindbergh D, Belsito A, Tucker M, Williams B, Wodniak C. The Regenstrief Medical Record System: a quarter century experience. Int J Med Inform. 1999 Jun;54(3):225-53.
This article describes the development of a centralized clinical database by the Indiana Network for Patient Care (INPC). The database is a means for collecting and sharing data in a local health information infrastructure (LHII) throughout 15 hospitals, public health departments, and Indiana Medicaid and RxHub. This includes most of the Regenstrief Medical Record System, which has a service in place known as DOCS4DOCS (D4D) that alerts clinicians to new reports, generates reports for other providers, and allows clinicians to view reports in the database for clinical and research purposes.

Morris MJ, Basch EM, Wilding G, Hussain M, Carducci MA, Higano C, Kantoff P, Oh WK, Small EJ, George D, Mathew P, Beer TM, Slovin SF, Ryan C, Logothetis C, Scher HI. Department of the Defense Prostate Cancer Clinical Trials Consortium: a new instrument for prostate cancer clinical research. Clin Genitounrinary Cancer. 2009;7:51-57.
Morris et al. describe the DoD's Prostate Cancer Clinical Trials Consortium including its data management across numerous centers. The Consortium is comprised of 10 prostate cancer research centers. It involves a centralized infrastructure with standardized common data elements, security and regulations, and an online clinical trial management system. However, the data is inputted locally before moving to the central infrastructure.

Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The new Sentinel Network--improving the evidence of medical-product safety. N Engl J Med. 2009 Aug 13;361(7):645-7.
Platt et al. provide an overview of the history and design of the Sentinel Initiative, a FDA distributed network that will conduct post-marketing surveillance through the use of electronic health record systems and medical claims databases. This distributed network is designed to detect and respond quickly to higher-than-expected signals (which may be an indicator of adverse outcomes). Data remains on the local level but data files are standardized and data are summarized and passed on. Associated challenges include the inclusion of multiple data sources, the detection of signals not included in claims data or electronic health records (the two data sources), governance and accountability.

Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010 Aug;19(8):858-68.
Schneeweiss introduces an approach (a basic study design) for conducting expedited signal evaluation in longitudinal healthcare databases, such as the Sentinel System. After a signal has been triggered, the last step (an implementation of pharmacoepidemiolgoic investigations to refute or not the signal) occurs. This developed incident user cohort design seeks to make this portion of the process faster and more efficient.

III. Natural Language Processing

Ambert KH, Cohen AM. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. J Am Med Inform Assoc. 2009 Jul-Aug;16(4):590-5.
Ambert and Cohen describe their entry into the 2008 i2b2 Obesity Challenge Task. The Challenge presented by the i2b2 team was to extract obesity data from the clinical record, along with comorbidities. The authors used a five-step approach to conducting the extraction, which included: preprocessing, tokenization, vectorization, filtering, and classification. The system scored 13th overall in the textual task, and 5th for the intuitive task.

Bramsen P, Deshpande P, Lee YK, Barzilay R. Finding temporal order in discharge summaries. AMIA Annu Symp Proc. 2006:81-5.
Bramsen et al. propose a method to conduct automated temporal analysis of medical discharge summaries. The authors outline the temporal annotation scheme, describe the benefits of using such a system, and then describe their method of conducting this analysis using actual medical discharge summaries. The temporal segmentation algorithm, as performed on 60 summaries, had a recall of 78% and precision of 89% (F-measure 83%). The ordering component had an accuracy of 78.3%, which outperformed the baseline by 6.1%.

Carrell D, Miglioretti D, Smith-Bindman R. Coding free text radiology reports using the Cancer Text Information Extraction System (caTIES). AMIA Annu Symp Proc. 2007; 889.
Carrell et al. describe a method for coding radiology reports using the Cancer Text Information Extraction System (caTIES). caTIES indexes clinical concepts from medical text using NLP algorithms, and identifies negated concepts, as well. The authors selected 700 pelvic ultrasounds from a set of 200,000. These included 600 studies of 273 women who were diagnosed with ovarian cancer, and 100 without ovarian cancer. A radiologist reviewed a sample of 15 reports. Review was ongoing at time of publication. caTIES had a sensitivity of 82% in identifying cases of ovarian cancer, and a specificity of 95%. Future directions for research include algorithm and domain extensions.

Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. Proc AMIA Symp. 1999:42-6.
The authors outline the desiderata for clinical term entry at the workstation. They describe some of the available query services and then describe what elements they would look for in an ideal system for clinicians in the care setting. The desiderata outlined by the authors include the following: (1) Word Normalization, (2) Word Completion, (3) Target Terminology Specifications, (4) Spelling Correction, (5) Lexical Matching, (6) Term Completion, (7) Semantic Locality, (8) Term Composition, and (9) Term Decomposition. While there are already clinical text extractors available for research use, there is a need for clinical care setting modifications. Ease of navigation and user-friendliness is a more important aspect of the program in the clinical workstation.

Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):36-9.
Clark et al. describe a method for data extraction that would allow for classification of smokers from the free text portions of the electronic medical record. This method was created by the Clinical Language Understanding Group at Nuance Communications as part of the i2b2 Challenge in 2006, and has been developed further since then. The authors used a two-step process to extract clinical data, first using a rule-based extraction engine to identify records with any smoking reference, and then using linguistic analysis and machine learning to detect and classify smoking mentions. This project was the winner of the i2b2 Challenge.

Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009 Oct;42(5):760-72.
Demner-Fushman et al. describe how natural language processing (NLP) can be used to improve the clinical decision support systems. This article describes the process of natural language processing, explains how NLP can be used to extract data from the clinical record that can allow for clinical decision-making, and describes the future directions for NLP. The authors describe clinical decision support systems and computerized physician order entry systems, and they outline examples of how NLP can be utilized to improve these functions, leading to better health outcomes.

Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res. 2000 Feb;33(1):1-10.
The authors evaluate the accuracy of manual versus automated NLP coding of neuroradiology reports in the Northern Manhattan Stroke Study (NOMASS). The NOMASS protocol uses a form to evaluate brain images of stroke patients. The forms have 47 separate data fields – 32 for anatomical localization, 16 each for new and old lesions, six for lesion size, and nine for generic characteristics (side of brain, hemorrhagic, edematous, etc.). This study used the MedLEE (Medical Language Extraction and Encoding) system NLP to create a program that could analyze and code neuroradiology reports. Manual coding had an accuracy of 86%, compared to 84% in the automated method (P=0.026).

Farkas R, Szarvas G, Hegedus I, Almási A, Vincze V, Ormándi R, Busa-Fekete R. Semi-automated construction of decision rules to predict morbidities from clinical texts. J Am Med Inform Assoc. 2009 Jul-Aug;16(4):601-5.
Farkas et al. describe an entry for the i2b2 Challenge in Natural Language Processing for Clinical Data, which required the use of automated data extraction to determine obesity and comorbidity status in a data set. The team from the University of Szeged used a rule-based system that capitalizes on a list of key words and concepts to identify important sentences and passages. In the textual model, the authors used a dictionary-lookup-based system to collect a set of terms and abbreviations for the diseases that were being considered. In the intuitive model, the documents categorized as "unmentioned" were evaluated to determine if they could be classified as an intuitive "yes" or "no." They achieved an F-macro score of 76% for the textual model test set and a F-macro score of 67% for the intuitive model test-set.

Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. AMIA Annu Symp Proc. 2008 Nov 6:207-11.
Friedlin et al. provide an example of the use of the Regenstrief Extraction tool (REX) to obtain data about MRSA infections from electronic medical records. They used localization techniques to identify "methicillin" "MRSA" or "aureus" in the text, and REX contextual detection to identify MRSA-related phrases. REX had a success rate for MRSA identification of 99.96% with 74 false positives. The authors identified improper grammar as one of the main hurdles to conducting accurate NLP extraction of data. Structuring and coding messages in a standardized manner (e.g. SNOMED) would decrease the need for more complex detection methods.

Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392-402.
Friedman et al. evaluate MedLEE, a method that uses natural language processing (NLP) techniques to map a clinical document into codes with modifiers. MedLEE was modified to automatically generate codes. Recall and precision based on the UMLS codes were measured as the outcomes. This method performed as well or better than six experts.

Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc. 2008 Nov 6:247-51.
Goryachev et al. describe the use of natural language processing techniques to extract information on family medical history. The technique included pre-processing, family member and diagnosis concept identification, and family member/patient assignment. Pre-processing is done using HITEx and UMLS concepts. 2,000 medical reports were selected from the Partners Research Patient Data Registry (RPDR). From these, 350 sentences were randomly selected, and were manually reviewed by a nurse (one of the authors) for identification, and then reviewed by a physician (not an author) to assess inter-rater reliability. In terms of extracting all diagnoses, this method had 85.12% precision and 86.93% recall. In terms of differentiating family history from patient diagnoses, precision was 96.30% and recall was 92.86%. The assignment of diagnoses to exact family members had a 92.31% precision and recall. The correct extraction of data is very important when using the data for research purposes, but this can be challenging to do with automated NLP. This rule-based algorithm has good accuracy, and can be utilized for many applications, including cancer risk prediction.

Hazlehurst B, Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. Vaccine. 2009 Mar 23;27(14):2077-83.
The authors describe the Vaccine Safety Datalink (VSD) project, which uses NLP techniques to extract vaccine adverse event data from the clinical record. Using the Kaiser Permanente Northwest (KPNW) database of more than 450,000 members, the authors used NLP techniques to extract clinical data about vaccine adverse events, specifically those related to GI adverse events, using MediClass. This large linked database is able to study vaccine adverse effects in approximately 2.5% of the U.S. population. The NLP method based on MediClass is fairly successful in identifying adverse events. Further research and emphasis needs to go into improved coding methods on the physician end, and encouraging better documentation of minor symptoms that may be VAEs.

Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, Winickoff JP, Glasgow R, Palen TE, Rigotti NA. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med. 2005 Dec;29(5):434-9.
Hazlehurst et al. use MediClass, an automated NLP system, to assess clinician ability to evaluate smoking cessation evaluation of patients. MediClass maps phrases from the free text of the clinical record and codes in structured sections to a controlled vocabulary in order to conduct text classification. The authors gathered 1,000 records from 2003-2004 in each of four HMOs. Of these, 125 were selected from each of the sites for review, and were coded by MediClass, and also by trained chart abstractors. Disagreements were assessed and re-run through the system. Then, the remaining 875 records from each site were run through MediClass. This pool was separated into subgroups related to the 5 A's of smoking cessation, and then cut to 500 records. All were coded manually for the presence of the 5 A's. The MediClass system was successful in coding for the 5 A's of smoking cessation in the medical record, and can be useful in determining clinician performance with the evidence-based guidelines for preventive health services.

Himes BE, Dai Y, Kohane IS, Weiss ST, Ramoni MF. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. J Am Med Inform Assoc. 2009 May-Jun;16(3):371-9.
Himes et al. use electronic medical record data from the Partners Healthcare System to identify risk factors for progression to COPD in asthma patients. A predictive model was created with data from 843 asthma cases and 8,506 controls to estimate COPD development rates from a future set of 992 patients (46 cases and 946 controls). The authors used NLP techniques to extract relevant demographic and comorbidity data from the EMR, and using the i2b2 workbench, refined and analyzed this "asthma data mart." The Health Information Text Extraction (HITEx) tool was used to collect smoking status data. The model that was created by the authors using Bayesian networks with the demographic, smoking, and comorbidity history for the initial set of patients was able to predict COPD progression in a future set of patients with 83.3% accuracy.

Honigman B, Lee J, Rothschild J, Light P, Pulling RM, Yu T, Bates DW. Using computerized data to identify adverse drug events in outpatients. J Am Med Inform Assoc. 2001 May-Jun;8(3):254-66.
Honigman et al. evaluate a computer program that extracted adverse drug event information from Brigham and Women's Hospital outpatient clinical records collected from July 1995 to June 1996. The Micromedex M2D2 medical dictionary was used to mine data from a controlled medical vocabulary. Two case studies (warfarin and captopril) were described to show the methodology. It was estimated that there were 864 adverse drug events among 25,056 incidents that were identified by the program (5.5 per 100 patients coming for care). There was an admission rate of 3.4 per 1,000 patients. Sensitivity was 58%, and specificity was 88%. Free text searches in the computerized detection of adverse drug events can be very useful and these programs may have a role in QI efforts.

Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010 Aug;62(8):1120-7.
Liao et al. used EMR data from two large academic medical centers to create an "RA [rheumatoid arthritis] Data Mart" based on diagnosis codes and lab data. The data mart included 29,432 patients. The authors used free-text and coded data to create a classification algorithm. The calculated PPV of RA classification for the complete narrative (using both coded and free-text data) was 94%, compared to the PPV of 88% for coded data alone.

McCormick PJ, Elhadad N, Stetson PD. Use of semantic features to classify patient smoking status. AMIA Annu Symp Proc. 2008 Nov 6:450-4.
McCormick et al. present their submission to the i2b2 NLP Challenge smoking classification task. The team looked at semantic features only, as opposed to other research that usually focuses on a combination of semantic and medical knowledge or lexical techniques. They used the MedLEE natural language processor to identify semantic features. Their supervised MedLEE classifier placed second in both micro-F1 and macro-F1 measures.

Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10:70.
Meystre et al. conducted a literature review of research that studies the automated de-identification of narrative text documents from electronic health records. Study results include: while all the automated systems address some of the PHIs, none target all 18; most research only look at the de-identification of certain types of documents instead of evaluating the systems' success with a mixture of report types; and two methodologies were identified for de-identification (pattern matching and machine learning) or a combination of the two.

Mishra NK, Cummo DM, Arnzen JJ, Bonander J. A rule-based approach for identifying obesity and its comorbidities in medical discharge summaries. J Am Med Inform Assoc. 2009 Jul-Aug;16(4):576-9.
Mishra et al. participated in the i2b2 Obesity Challenge. The authors used a rule-based approach and searched for occurrences of morbidity-related keywords. For the 16 morbidities involved within the challenge, this approach achieved macro F-measures scores above 0.8 for 12 of the morbidities and above 0.9 for five of the morbidities.

Morrison FP, Li L, Lai AM, Hripcsak G. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J Am Med Inform Assoc. 2009 Jan-Feb;16(1):37-9.
Morrison et al. summarize research on the MedLEE NLP system's ability to de-identify PHIs from medical clinical narrative notes. For this test, an unmodified MedLEE NLP system evaluated electronic outpatient internal medicine practitioners' clinical follow-up notes. 26 PHIs escaped detection mostly due to misclassification of medical terms and numbers. The authors contend that several methods should be used to de-identify data (as opposed to merely improving existing systems) and propose pairing a traditional de-identification system with MedLEE.

Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care. 2007 Jun;13(6 Part 1):281-8.
Pakhomov et al. evaluate the use of NLP and predictive modeling methods to identify heart failure (HF) patients in the EMR of the Mayo Clinic. In the pilot study, the authors use NLP methodology that searches for key disease-related non-negated vocabulary and their synonyms. These codes were manually entered. This protocol was periodically checked against billing codes to determine if there were any omissions. A predictive model algorithm was also used, taking statistical methods to determine if a patient is HF positive or negative. NLP had a sensitivity of 81.6%, a specificity of 97.8%, and a PPV of 49.3%. Predictive modeling had a sensitivity of 56.0%, a specificity of 96.0%, and a PPV of 82.2%

Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13.
Savova et al. describe the clinical Text Analysis and Knowledge Extraction System (cTAKES), an open-source natural language processing system developed at the Mayo Clinic. The cTAKES system uses rule-based and machine learning to extract clinical data for research purposes. The authors evaluated cTAKES for accuracy on several measures, and determined that there is good speed, but the Named Entity Recognition (NER) was unable to perform well with complex queries.

Sibanda T, He T, Szolovits P, Uzuner O. Syntactically-informed semantic category recognition in discharge summaries. AMIA Annu Symp Proc. 2006:714-8.
Sibanda et al. show the results of a statistical recognizer used for natural language processing purposes. They rely on information from the UMLS plus they supplement the system with lexical and syntactic context and incorporate a statistical semantic category recognizer to identify eight semantic categories in 48 discharge summaries (diseases, signs and symptoms, treatments, diagnostic tests, results, dosage information, abusive substances, and medical practitioners). The statistical recognizer performed better in comparison to baseline for all categories. Weakness was due to vocabulary misclassification.

Solt I, Tikk D, Gál V, Kardkovács ZT. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. J Am Med Inform Assoc. 2009 Jul-Aug;16(4):580-4.
Solt et al. present results from their submission to the i2b2 Obesity Challenge: a classification technique. The task required the assignment of semantic labels to the diseases: present, absent, questionable, or unmentioned. The authors used a context-aware rule-based semantic classification technique. Unique to this submission, the authors preprocessed the text, created rules for the textual and intuitive subtasks, and developed a dictionary with disease names and alternatives. Challenge results: second place at the textual and first place at the intuitive subtasks.

Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):550-63.
This paper summarizes the findings from the de-identification submissions (part of the i2b2 Smoking Challenge). The authors provide an overall summary of all the submission's systems and their results. There were submissions from 18 teams for two challenges: de-identification of medical records (seven) and determining smoking status from medical records (11). The best systems within the competition were statistical learning systems that utilized rule templates as features. Next, in descending order of performance: hybrid systems of rules and machine learning, pure machine learning, and pure rule-based systems.

Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009 Jul-Aug;16(4):561-70.
This article provides an overview of the i2b2: Obesity NLP Challenge including the methods used to judge submissions and summaries of the systems and their results. 30 teams participated in the Obesity Challenge (with each permitted three runs) for a total of 136 submissions. Each submission was evaluated using micro-and macro-averaged precision, recal, and F-measure. The machine learning approaches used were top performers in the intuitive test; less so for textual test.

Uzuner O, Mailoa J, Ryan R, Sibanda T. Semantic relations for problem-oriented medical records. Artif Intell Med. 2010 Oct;50(2):63-73.
Uzuner et al. developed a semantic relation (SR) classifier that identifies "fine-grained" relations based on a review of data within one sentence at a time and a pair of concepts at a time (candidate pair). The authors sought to determine the relation of each medical problem with other possible medical problems, treatments, and tests mentioned in the same sentence. The authors applied their classifier to two corpora, with a total of nearly 200 medical discharge summaries and over 25 thousand sentences. The SR classifier outperformed two baseline feature sets (with significant gains over baselines in present disease-treatment, present symptom-treatment, and disease-test relation types and failure to gain significantly over baselines with possible disease-treatment, possible symptom-treatment, and disease-symptom relation types.

Ware H, Mullett CJ, Jagannathan V. Natural language processing framework to assess clinical conditions. J Am Med Inform Assoc. 2009 Jul-Aug;16(4):585-9.
Ware et al. (in response to the i2b2 NLP Obesity Challenge) developed a NLP framework for the inclusion or exclusion of 16 patient conditions from discharge summaries. Findings could either be "textual" (directly worded within the summaries) or "intuitive" (not directly pulled from the text but inferred based on other indicators). Working from the released 700 de-identified discharge summaries, the authors preprocessed the text, detected a concept match, studied its surrounding neighborhood, and applied a variety of rules for the intuitive component. The authors finished fourth in the intuitive scoring and third in the textual scoring in the NLP challenge contest.

Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, Yeh A, Hitzeman J, Hirschman L. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):564-73.
Wellner et al. describe one de-identification process that was submitted to the AMIA Challenges in Natural Language Processing for Clinical Data. The investigators slightly adjusted two toolkits: Carafe and Ling Pipe and labeled the entire text as a single sentence for both systems. With their modest modifications, the highest F-measure was obtained with Carafe-ALL_PP (which was the original Carafe toolkit tuned for high recall coupled with a manual post-processing approach). The authors stated goals for a de-identification system include high performance, rapid retargeting (which they demonstrated in their slight modification to out-of-the-box toolkits), adjustability, introspectivity, and interactivity.

Wilcox AB, Vawdrey DK, Chen YH, Forman B, Hripcsak G. The evolving use of a clinical data repository: facilitating data access within an electronic medical record. AMIA Annu Symp Proc. 2009 Nov 14;2009:701-5.
Wilcox et al. conduct a study to determine how expert-written rules could improve the text classification aspect of natural language processing. Using records of 200 chest radiographs, five types of rule-based, instance-based, and probabilistic learning algorithms were applied in various combinations and compared with classification done using expert knowledge. Data was represented using the MedLEE NLP system. Extraction/selection algorithms included *predictive, medical*, *negated*, and *med+neg*. The predictive model performed significantly worse than the others, and the med+neg performed significantly better. The expert rules performed better than any of the algorithms.

IV. Research Networks

Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010 Jun;48(6 Suppl):S45-51.
Brown et al. describe a pilot distributed data network that they have created with five participating data holders. They argue that the distributed network is preferable over a centralized network because it performed as well as a centralized network in terms of using a central portal while allowing local control of data.

Diamond CC, Mostashari F, Shirky C. Collecting and sharing data for population health: a new paradigm. Health Aff (Millwood). 2009 Mar-Apr;28(2):454-66.
The authors advocate for the use of distributed data networks to conduct public health surveillance, as it reduces the issues related to aggregated models. Issues related to incomplete participation of public health institutions in sharing data; "dirty data" that is not complete, consistent, or correct; and time lags between data collection and analysis and between quality reporting and improvement.

Hynes DM, Perrin RA, Rappaport S, Stevens JM, Demakis JG. Informatics resources to support health care quality improvement in the veterans health administration. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):344-50.
Hynes et al. provide an example of a centralized network through their study of the VHA Quality Enhancement Research Initiative (QUERI) and its interaction with the Veterans Health Information Systems and Technology Architecture (VistA), the VHA clinical information system. The VistA system incorporates and standardizes data from multiple VA sources. The VHA QUERI is a multidisciplinary quality improvement initiative that oversees the organization's health care systems. The challenges associated with using the data from the VistA for research purposes include: the requirement of data extraction from multiple facilities, IRB authorization; the incorporation of data stored within other non-VA sites, and extracting data from narrative text.

Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. BMC Public Health. 2006 Sep 19;6:235.
Lazarus et al. provide information about the design and implementation of the National Bioterrorism Syndromic Surveillance Demonstration Program (NDP)'s distributed network. Currently operating in five states, the network provides surveillance and alerts to public health agencies. Raw data is stored and controlled by local providers while the aggregate data is transferred (using the CDC developed Public Health Information Network Messaging System) to datacenters for statistical processing and signal detection. In order to accrue a large amount of data, the network focuses on large ambulatory group practices with substantial technical capabilities using electronic medical record systems. The distributed network, in comparison to the centralized model, decreases the risk of the release of PHI.

Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, Brown JS. Design of a national distributed health data network. Ann Intern Med. 2009 Sep 1;151(5):341-4.
Maro et al. argue that distributed networks, because they allow their data holders to store and control their own data, alleviate many security, proprietary, legal, and privacy concerns. The authors also argue for a single national distributed database that could be used for multiple purposes by multiple end users. With a national distributed network, all data would remain local but adhere to a common data model. Its success would depend upon stakeholder commitment, software development, governance, and sustained funding. The authors recommend starting with a small number of users and data sources and see the federal government taking the lead with regards to initial funding with subsequent contributions coming from the system's users.

McDonald, CJ, Dexter, P, Schadow G, Chueh HC, Abernathy G, Hook J, Blevins L, Overhage M, Berman JJ. SPIN query tools for de-identified research on a humongous database. AMIA Annu Symp Proc. 2005; 515-9.
McDonald et al. describe the query tool (its content and capabilities) developed for the Indianapolis/Regenstrief SPIN note (part of the Indiana Network for Patient care (INPC)). The INPC is a centralized network. However, the SPIN/INPC is a federated network (and an extension of the Regenstrief Medical Record system (RMRS)) that includes standardized data collected from a number of area hospitals and health institutions. The query tool allows users to retrieve de-identified information by defining the cohort, developing datasets, and applying statistical analysis plans.

Pace WD, Cifuentes M, Valuck RJ, Staton EW, Brandt EC, West DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009 Sep 1;151(5):338-40.
Pace et al. give an overview of the Distributed Ambulatory Research in Therapeutics Network (DARTNet). DARTNet is an AHRQ-funded federated database that houses clinical data from eight different organizations comprising over 400,000 patients. Data can be accessed through a database query. The DARTNet system currently works with five brands of ambulatory EHRs, but is brand agnostic. It can prompt clinicians to obtain specific information on an encounter. It also has capabilities for community learning. Challenges included: locating particular types of data, standardizing data from separate practices, EMRs lacking in reasonable range checks. To improve its ability to examine episodes of care, DARTNet will include billing data. The ongoing use of extracted clinical data serves a continuous quality control purpose. DARTNet is capable of bidirectional communication with practices that use EHRs

Platt R, Davis R, Finkelstein J, Go AS, Gurwitz JH, Roblin D, Soumerai S, Ross-Degnan D, Andrade S, Goodman MJ, Martinson B, Raebel MA, Smith D, Ulcickas-Yood M, Chan KA. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiol Drug Saf. 2001 Aug-Sep;10(5):373-7.
Platt et al. describe the HMO Research Network (HMORN) Center for Education and Research on Therapeutics (CERT). This multicenter research effort is done on a distributed data model, so as to control costs and maintain data security. National Drug Codes (NDCs) are used to conduct research on dispensing records, answering questions related to therapeutic benefit, drug safety, physician and patient behavioral changes, personalized medicine and costs.

Saver B. One system for electronic health records. Health Aff (Millwood). 2010 Jun;29(6):1273.
Saver makes the argument that the nation should adopt one centralized electronic health record system. Distributed networks are too expensive due competing software, implementation, maintenance, training and retraining. Instead, the author advocates that the nation should follow the Veteran Health Administration's lead and implement a centralized network. Such a network would ensure standardization and could become a secure source from which to conduct quality improvement, surveillance, and research.

V. Data use and quality

Atkins D. Connecting research and patient care: lessons from the VA's Quality Enhancement Research Initiative. J Gen Intern Med. 2010 Jan;25 Suppl 1:1-2.
Atkins gives an overview and lessons learned from the VA's Quality Enhancement Research Initiative (QUERI) Program to improve quality of care for priority conditions within the Veterans Health Administration. The author relates that, to date, the lessons derived from the QUERI Program include the slow pace of change and innovation adoption, the lack of focus on T2 translational research, the mutual learning that can occur within and outside of the VA for quality improvement research, and the need to adapt to changing clinical research priorities.

Aronsky D, Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index. J Am Med Inform Assoc. 2000 Jan-Feb;7(1):55-65.
Aronsky et al. describe a study that was conducted in an emergency department (ED) setting to determine if clinical data in the electronic record can drive complex point-of-care guidelines for community acquired pneumonia. This study was conducted on the 241 patients that arrived in the ED with a primary discharge diagnosis of community acquired pneumonia over the period from June 1995 to November 1996. Each patient in the study was evaluated for a Pneumonia Severity Index (PSI) score using a self-coding data sheet from information in the clinical record, the HELP system.

There are 20 variables used in the PSI guidelines. Of these, 12 are always documented in the HELP system, seven were available only for a subset of patients, and one was not typically used. The data availability for the PSI elements was strong enough to create a 0.98 concordance rate with actual PSI scores. The most frequent error was related to charting, particularly in the free-text portions. The level of data quality in HELP can support the PSI as a clinical decision support tool.

Bates DW, Pappius E, Kuperman GJ, Sittig D, Burstin H, Fairchild D, Brennan TA, Teich JM. Using information systems to measure and improve quality. Int J Med Inform. 1999 Feb-Mar;53(2-3):115-24.
Bates et al. describe the quality improvement measures developed by Brigham and Women's Hospital. The hospital has moved to measure as many aspects of its performance by information systems as possible. They use electronic medical record data rather than billing data for a more thorough, accurate picture of the events of the hospital. They implemented CDSS and CPOE and evaluated the initial results of these programs. BWH is currently building a data warehouse, whose focus will be on data analysis of retrospective data. In contrast, there is a data repository in the hospital that is used for clinical purposes for the individual patient. The data warehouse will integrate data collection into routine clinical care. The system should soon be able to evaluate the Maryland Hospital Association measures for hospitals and the HEDIS outpatient criteria. Long-term goals include measurement of quality and utilization outcomes. Although not statistically significant, the implementation of CDSS for lab testing reduced the number of tests ordered by 4.5% and reduced the total charges for these tests by 4.2%. Another study of the CDSS showed that 70% of tests were canceled with a reminder system in place.

Brennan PF, Stead WW. Assessing data quality: from concordance, through correctness and completeness, to valid manipulatable representations. J Am Med Inform Assoc. 2000 Jan-Feb;7(1):106-7.
Brennan and Stead conduct a literature review of approaches (as of 2000) to assessing data quality. The authors first examined the idea of concordance as a measure of data quality. Stein measured concordance within record systems, while Aronsky looked at this measure across two different record systems. In the Hogan and Wagner article, the authors take data quality a step further by creating a gold standard of information in the clinical record (using the patient and provider as sources of information). This method takes into account data correctness and completeness in the clinical record. They conclude that higher levels of concordance in the medical records (within a record or between a record and its reference standard) lead to better and more consistent clinical recommendations.

Brown JS, Moore KM, Braun MM, Ziyadeh N, Chan KA, Lee GM, Kulldorff M, Walker AM, Platt R. Active influenza vaccine safety surveillance: potential within a healthcare claims environment. Med Care. 2009 Dec;47(12):1251-7.
Brown et al. utilize a claims database to conduct a study on influenza vaccine safety. The study was conducted using the claims data warehouse of a large, multi-state health insurer, which included approximately 13.5 million members at the time the research was conducted in 2006. Retrospective data was examined, including a final data extraction six months after the last study date. 10 potential vaccine-related adverse events (AEs) were evaluated in the study, as occurring within a 60-day window follow vaccine administration, in the inpatient or outpatient setting. In 2005-2006, there were 4.9 AEs per 1,000 influenza vaccinations (3,417 total). In 2006-2007, there were 5.3 AEs per 1,000 influenza vaccinations (4,627 total). There was a slight lag in data accrual, with 90% of data being received by the end of 1 month following the vaccination.

Bu D, Pan E, Walker J, Adler-Milstein J, Kendrick D, Hook JM, Cusack CM, Bates DW, Middleton B. Benefits of information technology-enabled diabetes management. Diabetes Care. 2007 May;30(5):1137-42.
Bu et al. describe how IT-enabled diabetes management (ITDM) programs can affect diabetes care outcomes. They also look at the cost of providing diabetes treatment through their creation of a computer simulation model that examines the impact of technologies used by providers, patients, and payers, as well as an integrated diabetes management system. Payer systems use claims data to compare the care that patients receive to the care that is recommended. This can allow payers to provide feedback to patients and providers about care management, and can also lead to behavior change programs, which leads to cost savings and improved outcomes.

Chan KA. Development of a Multipurpose Dataset to Evaluate Potential Medication Errors in Ambulatory Settings. In: Henriksen K, Battles JB, Marks ES, Lewin DI, editors. Advances in Patient Safety: From Research to Implementation (Volume 2: Concepts and Methodology). Rockville (MD): Agency for Healthcare Research and Quality (US); 2005 Feb.
Chan describes the data development process and study subject demographics of a study conducted by the HMORN CERT on the frequency of potential outpatient medication errors. The data that was extracted for each of the 2,020,037 study participants include factors related to demographics, membership in one of 10 HMOs, medications dispensed and health care utilization. The method that the researchers chose to de-identify data was based on the idea of having the "minimum necessary data" to conduct the research. Each study conducted used individual SAS codes that were distributed to each of the participating HMOs. This allowed consistent implementation across all of the HMOs. This also allowed for manual review of the medical records. Abstracted information was entered into forms with the Study ID, into databases at the Data Coordinating Center. While there are limitations of using the HMO records to extract prescribing data, it is a useful means of gathering large amounts of patient data. This data set will continue gathering records for five years in order to conduct research on this data.

El-Ghatta SB, Cladé T, Snyder JC. Integrating clinical trial imaging data resources using service-oriented architecture and grid computing. Neuroinformatics. 2010.
El-Ghatta et al. provide an overview of the challenges currently associated with using imaging in clinical trials, including the unresolved issues of protecting patients' privacy, ensuring image readability, maintaining unchanged source imaging, and securely conducting data transfers. The authors provide a series of recommendations specifically focused on a need for standards and tools (for de-identification, QC profiles, data exchange services, and interoperability). They also recommend the use of "middleware" to facilitate interface between services and recommend caGrid.

Elkin PL, Mohr DN, Tuttle MS, Cole WG, Atkin GE, Keck K, Fisk TB, Kaihoi BH, Lee KE, Higgins MC, Suermondt HJ, Olson N, Claus PL, Carpenter PC, Chute CG. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. Proc AMIA Annu Fall Symp. 1997:500-4.
Elkin et al. evaluate a problem list entry tool and a clinically derived problem list vocabulary developed at the Mayo Clinic. The Clinical Notes module of the Mayo system creates a clinical note where clinicians add diagnosis data to the Impression/Report/Plan section. The authors conducted a Usability study, which allowed users to test the system, and videotaped their experience with the product. Users are encouraged to talk aloud about their thoughts, behaviors, and actions. There were eight clinicians who participated in the Usability study. Each clinician participated in 2 test scenarios, followed by nine additional free-form scenarios with various diagnoses from CHF to acute cystitis. This data has research utility, but has also been used to create a clinical lexicon. 91.1% of clinicians found acceptable clinical diagnoses using the Mayo vocabulary navigation system, and there was an acceptable response time in 92.5% of cases. 71.4% of participants felt that the number of terms presented was just right, while 12.9% felt there were too many, and 15.7% felt there were too few. The presentation of Related Terms was deemed to be helpful by 87.5% of the clinicians. 100% of the clinicians found it useful to enter abbreviations and word fragments, but only 70.7% of the time.

Grant A, Moshyk A, Diab H, Caron P, de Lorenzi F, Bisson G, Menard L, Lefebvre R, Gauthier P, Grondin R, Desautels M. Integrating feedback from a clinical data warehouse into practice organisation. Int J Med Inform. 2006 Mar-Apr;75(3-4):232-9.
Grant et al. describe the CIRESSS (Centre informatisé de recherche évaluative en soins et systèmes de la santé) Program at the Sherbrooke University in Canada, that maintains an EHR (ARIANE) and a clinical data warehouse (CDW). The CDW uses SNOMED and ICD-9 and -10 codes to automatically encode clinical data into the warehouse. CIRESSS communicates with clinical teams to determine which data sets would be useful to have in the practice setting that could provide useful feedback for quality improvement measures. They created a dashboard tool for the clinics that contains no identifiable data and is not used for patient care. They used the dashboard in the emergency department and in the clinical biochemistry department. The emergency department prototype allows an automated analysis of patient occupancy. The biochemistry department prototype allows for an automated analysis of quality assurance. The uses of the clinical data warehouse are still being explored, but it is expected to have an impact on quality assurance analysis of clinical practices.

Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part III. Value Health. 2009 Nov-Dec;12(8):1062-73.
In 2007, a multidisciplinary, multi-national Task Force was commissioned by the International Society for Pharmacoeconomics and Outcomes Research to recommend Good Research Practices for Designing and Analyzing Retrospective Databases. Johnson et al. present the recommendations that came from the Task Force's review of the relationship of statistical analysis on causal inference as it relates to comparative treatment effects. Recommendations address a variety of topics including identifying potential confounding factors, dealing with missing data, the reporting and assessment of performance measures, the inclusion of factors related to outcome or treatment selection, the reporting of instrument strength, and the importance of the use of correct statistical analysis techniques for improved validity.

Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. Int J Med Inform. 1998 Mar;49(1):117-22.
Quantin et al discuss a computerized record hash tag coding and linkage procedure used to chain medical information while maintaining anonymization for epidemiological follow-up. This method was applied to two different files: a Burgundian registry of digestive tumors (1,570 cases of colorectal and pancreatic tumors diagnosed between 1990 and 1995) and discharges from a Dijon public teaching hospital (334,848 discharges from 1990 to 1996). Automated linkages were manually checked to evaluate for false positives. False negatives were determined by manual review. Linkage protocols take into account different weights for various demographic variables in order to ensure greater reliability that the linkages are between records for the same individual. The method of anonymized record linkages had 100% specificity and 95% sensitivity.

Soto CM, Kleinman KP, Simon SR. Quality and correlates of medical record documentation in the ambulatory care setting. BMC Health Serv Res. 2002 Dec 10;2(1):22.
Soto et al. reviewed EMR data from 834 patients receiving care from 167 clinicians to determine how medical documentation was conducted on five measures, including smoking history, medications, drug allergies, compliance with screening guidelines, and immunizations. Certain elements of the clinical record have a better documentation rate than others. This also varies with the characteristic of the physician, including specialty and gender. There may be implications of this data to create an intervention for clinicians on how to improve documentation into the medical record so that data quality and completeness could be improved for clinical and research purposes.

Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. J Am Med Inform Assoc. 2000 Jan-Feb;7(1):42-54.
Stein et al. conducted a study of a clinical data repository (CDR) to determine if coded versus free text data fields would result in confirmatory, complementary, or conflicting information. Coded data fields can be limited in scope for clinician data entry, and thus free text fields are frequently used in the health record. Coded data are easier to use for data storage and retrieval; it is more difficult to automatically extract meaningful data from free text. Errors in accuracy of data (missing data, data that is incorrectly entered, patient neglect in reporting data) have previously been studied, and this study goes beyond that to look into data concordance. Contradictory data were present in 5-8% of the records examined.

Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. BMJ. 2009 Jan 27;338:b81. doi: 10.1136/bmj.b81.
Tannen et al. present the findings from a study exploring whether observational studies using electronic medical databases can provide results comparable to those from randomized controlled trials. The investigators compared data from the UK GRPD (the UK General Practice Research Database that contains comprehensive medical information for 5.7% of the UK population from 1990-2000) using studies designed to replicate (as much as possible) the design of previous random controlled trials with six published randomized control trials. The results were conducted using PERR analysis and standard biostatistical methods. There were no significant differences between the two methods in nine of the 17 outcome comparisons. In eight studies, Cox adjusted hazard ratios in the database differed significantly from the randomized trials. The authors suggest that by using the PERR technique, the level of confounding lessens with observational studies.

Wilcox A, Natarajan K, Weng C. Using Personal Health Records for Automated Clinical Trials Recruitment: the ePaIRing Model. Summit on Translat Bioinforma. 2009 Mar 1;2009:136-40.
Wilcox et al. have developed a model for patient recruitment through the use of electronic health records. Using grounded theory analysis, the authors collected data from clinical trial participants (principal investigators, patients, physicians, study coordinators, and study sponsors) and identified three levels of information flow of consent and criteria matching between participants: traditional (PI and patient), brokered (physician driven), and information-enabled (requires use and acceptance of personal health records to identify potential participants). The authors argue that the information-enabled model could increase recruitment opportunities although there are a number of security and regulatory considerations that need to be addressed.

VI. Other

Identifiers and De-identification

Black N. Secondary use of personal data for health and health services research: why identifiable data are essential. J Health Serv Res Policy. 2003 Jul;8 Suppl 1:S1:36-40.
Black argues that there are research purposes for which data identification is necessary, including: creating linkages within a database, creating linkages between databases, ensuring meaningful comparisons that take into account confounding factors, ensuring completeness of recruitment, taking into account social and economic patient characteristics, and assessing the applicability and generalizability of research results. The author recommends the use of identifiable patient data in certain instances for the improvement of quality of care and research purposes. He notes that privacy can still be maintained by protecting access and a detailed informed consent process.

Carpenter PC, Chute CG. The Universal Patient Identifier: a discussion and proposal. Proc Annu Symp Comput Appl Med Care. 1993:49-53.
This article describes some of the benefits and pitfalls of using the Social Security Number as an identifier in the medical record, and proposes a new Unique Patient Identifier (UPI) System. The Social Security Number has become a de facto identifier in the medical record, but because it is tied to other non-clinical uses, and cannot be verified or validated, there are security and confidentiality issues associated with its use in medicine. The UPI proposal is based on immutable properties of an individual. Personal (e.g. changing name with marriage) and political factors (change of country due to boundary reassignments) will not affect this system. It takes into account several factors that would ensure that each person would have a unique combination of digits. It could be used easily by patients and providers. This new UPI number would be a more secure and precise way of identifying patients in the medical record. However, there could still be misuse of these numbers. Future directions in patient identification could include fingerprint recognition or other automated system that would be unique to each individual.

El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. J Med Internet Res. 2006 Nov 21;8(4):e28.
El Emam et al. distinguishes between *identifying variables* that can be used to directly identify an individual, *quasi-identifiers* that can play a role in indirect reidentification, and *nonidentifying variables* that cannot be used for direct reidentification. They performed a simulation with non-clinical data sets to determine if re-identification could be achieved. They found that gender, region and year of birth individually and the combination of region and gender are all relatively stable whereas home or practice postal code, date of birth, initials, and the combination gender and year of birth are unstable.

Fefferman NH, O'Neil EA, Naumova EN. Confidentiality and confidence: is data aggregation a means to achieve both? J Public Health Policy. 2005 Dec;26(4):430-49.
Fefferman et. al. conduct a hypothetical study of data aggregation methods that shows that public health research is not being well-served through this technique as a means to de-identify data. Rather, it causes a loss of critical information (such as temporal and spatial epidemiologic data) that does not necessarily provide an increased level of security for patients. Instead, they recommend using cryptographic methods that anonymize patient identifiers in order to improve disease tracking.

IRBs

Dokholyan RS, Muhlbaier LH, Falletta JM, Jacobs JP, Shahian D, Haan CK, Peterson ED. Regulatory and ethical considerations for linking clinical and administrative databases. Am Heart J. 2009 Jun;157(6):971-82.

Dokholyan et al. describe the benefits and downfalls of the use of claims data, registry data, and a linked data set combining the two for research purposes. Claims databases are comprised of billing information, and they have a large set of data, are comprehensive in the information regarding treatment of a patient, and include the patient identifiers that could allow the creation of a longitudinal record. The disadvantage of the claims database is the limited amount of clinical data that can be used for QI or research. They discuss the Privacy Rule and the Common Rule and outline when IRB approval can be expedited and informed consent can be waived. Also discussed are the ethical considerations around patient protection, and if the rules are going too far. Even if patient identifiers are not removed, there may not be any reason to prevent the disclosure of data, even without informed consent.

Goldstein MM. Health information technology and the idea of informed consent. J Law Med Ethics. 2010 Spring;38(1):27-35.

Goldstein differentiates between "autonomous authorization" by a patient of an intervention by a clinician and "rule-based consent" which is achieved through following certain rules, policies, and practices. With the advancement in the use of HIT for research purposes, she argues that more than simply a signature should be required for truly informed consent. Additionally, this system puts the onus on patients to protect their own health information, and there should be required more stringent policies to ensure that researchers protect data.

Governance

Nazi KM, Hogan TP, Wagner TH, McInnes DK, Smith BM, Haggstrom D, Chumbler NR, Gifford AL, Charters KG, Saleem JJ, Weingardt KR, Fischetti LF, Weaver FM. Embracing a health services research perspective on personal health records: lessons learned from the VA My HealtheVet system. J Gen Intern Med. 2010 Jan;25 Suppl 1:62-7.

This article discusses the personal health record (PHR) using the VA's My HealtheVet program as an example. The My HealtheVet program includes three layers of functionality and access. The first layer is available to any user without any authentication necessary. The second layer is a password-protected section, in which veterans can create an individualized account. The third layer involves in-person identity authentication which allows access to a greater amount of health data.

Library of Phenotypes

Chute CG. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. Proc AMIA Symp. 2002:165-9.

Chute introduces a new method for patient data retrieval. The author discusses the currently available strategies for patient phenotype recognition for study cohort retrieval. He then outlines a new approach, which addresses the challenges with traditional systems. Automatic encoding of the literature into MeSH codes has been only moderately successful, and it has not replaced human assignment. Classifying the data is very macro and too granular, and often does not get into sufficient detail for research purposes. It is also too costly, and requires too much effort, to perform human classification. Using nomenclature to encode data fields is more detailed than data classification, and is tighter in scope (shorter phrases versus the entire document), but requires algorithmic methods that are not readily available. The steps for the proposed new approach include: (1) code the question, (2) index and normalize the text (pre-processing), and (3) invoke thesauri (accommodates exact word matches and lexical variants). By invoking a statistical machine learning technique in order to code the medical text, the author aims to address these concerns and improve the precision and recall of this process.

Metadata

Carvalho ECA de, Batilana AP, Simkins J, Martins H, Shah J, Rajgor D, Shah A, Rockart S, Pietrobon R. Application description and policy model in collaborative environment for sharing of information on epidemiological and clinical research data sets. PLoS ONE. 2010; 5(2):e9314.

The authors maintain that investigators have little incentive to share data due to funding and publication practices. Consequently, this leads to data missharing within open data networks. In response, the authors' developed a database that provides information about epidemiological and clinical research data sets, thus allowing researchers to follow-up with one another for additional information and leading to real partnerships, and future publishing and funding opportunities. The authors also found that a combination of open data sets plus databases about data sets lead to maximum collaboration (as opposed to no sharing policies or one without the other).

Smedley D, Schofield P, Chen CK, Aidinis V, Ainali C, Bard J, Balling R, Birney E, Blake A, Bongcam-Rudloff E, Brookes AJ, Cesareni G, Chandras C, Eppig J, Flicek P, Gkoutos G, Greenaway S, Gruenberger M, Hériché JK, Lyall A, Mallon AM, Muddyman D, Reisinger F, Ringwald M, Rosenthal N, Schughart K, Swertz M, Thorisson GA, Zouberakis M, Hancock JM. Finding and sharing: new approaches to registries of databases and services for the biomedical sciences. Database. 2010; baq 014.

Smedley et al. describe recent efforts to provide metadata about databases, specifically the Database Description Framework. There are many different databases for research and it is difficult for biologists and bio-informaticians to make informed decisions about which ones are of high quality and appropriate for their studies. The authors identify two approaches for advertising data: databases that provide metadata of databases and web service registries that provide information about services available at websites. The Database Description Framework presents the metadata in a standardized format and has querying capability.

Vardaki M, Papageorgiou H, Pentaris F. A statistical metadata model for clinical trials data management. Computer Methods and Programs in Biomedicine. 2009; 95:129-45.
There is growing interest in the concurrent use of datasets from different trials and time periods. However, statisticians performing the analysis may be unaware of the specific processes assigned to each dataset. Vardaki et al. present a data and metadata model (a metadata-enabled statistical software system) for clinical study systems that describe the sequence of processes within trials and looks for differences between datasets that might negatively affect a merge. The authors argue that this process-oriented model, which can process data and metadata partially automated, can reduce human error and provide the mechanism to better track the operation of datasets.

Patient Involvement

Abernethy AP, Wheeler JL, Zafar SY. Management of gastrointestinal symptoms in advanced cancer patients: the rapid learning cancer clinic model. Curr Opin Support Palliat Care. 2010;4(1):36-45.
Abernethy et al. present how ePRO data collection technology is used for gastrointestinal symptom monitoring and management with the Duke University Medical Center's GI Oncology clinics. The authors articulate the benefits of rapid learning healthcare and outline how this system works to achieve its implementation. GI patients report their symptoms through the use of ePRO data collection technology (in the form of e/Tablets). Results are immediately read by the physicians and incorporated into their clinical care. This system has been tested for feasibility, acceptability, validity, and clinical utility supports. Partnerships with drug developers, clinical trialists, and biostatisticians spur further research and interventions. More information about the implications for research, beyond the clinical care setting, is needed.

Tripathi M, Delano D, Lund B, Rudolph L. Engaging patients for health information exchange. Health Aff (Millwood). 2009 Mar-Apr;28(2):435-43.
Tripathi et al. provide information, lessons learned, and recommendations regarding the Massachusetts eHealth Collaborative (MAeHC), a health information exchange that has successfully incorporated consumer/patient involvement in its design and implementation. Three pilot projects were launched by MAeHC in three different Massachusetts communities encompassing 597 ambulatory care providers in 142 practices. The goal of the pilot projects was to implement EHRs and then connect all three sites electronically and centrally for data sharing and collaboration. In order to engage consumers/patients, investigators worked closely with a leading patient advocacy organization concerned with privacy and security issues; adopted the opt-in policy; negotiated with patients/consumers about what data could be shared; enlisted outside help for information dissemination and marketing; and held focus groups to discuss the policies, pilot projects, and marketing materials. As a result, there was 90% opt in.

Security

Krishna R, Kelleher K, Stahlberg E. Patient confidentiality in the research use of clinical medical databases. Am J Public Health. 2007 Apr;97(4):654-8. Epub 2007 Feb 28.
Krishna et al. discuss the legal issues surrounding patient confidentiality and institutional liability in the use of medical records for comparative effectiveness research. There are issues with the conduct of research using clinical data. Physical safety of the data and the proper de-identification of the data is often assumed and depended upon, and this can be problematic when it is not confirmed and protected. Loss of hardware and free access to too many data users can cause breaches in security. The authors have developed a framework for data security that includes data exclusion, data transformation, and data encryption. They also suggest a vocabulary through which these security measures can be noted in research papers.

Langella S, Hastings S, Oster S, Pan T, Sharma A, Permar J, Ervin D, Cambazoglu BB, Kurc T, Saltz J. Sharing data and analytical resources securely in a biomedical research grid environment. J Am Med Inform Assoc. 2008; 15:363-73.
Langella et al. developed software, the Grid Authentication and Authorization with Reliably Distributed Services (GAARDS), in response to security concerns surrounding caBIG. GAARD is designed to support three main components in a federated environment: authentication (through Dorian a management service that allows users to access the GRID through their institution's log-in); authorization (through Grid Trust Service where Grid-level group membership dictates the authorization to perform certain tasks), and trust fabric (through Grid Grouper, a mechanism to integrate multiple certificate authorities and Certificate Revocation Lists from different institutions).

Manion FJ, Robbins RJ, Weems WA, Crowley RS. Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study. BMC Med Inform Decis Mak. 2009 Jun 15;9:31.
Manion et al. conducted 19 interviews with key regulatory personnel at six cancer centers involved with data sharing through caBIG. Interviewers manually coded the interviews, quantified those answers with single responses, and used qualitative analysis to identify themes. Themes included security, governance, IRB compliance, and authentication and authorization. As a result of this research, the authors recommend the following: construct a separate legal entity for governance; develop consensus on foreign and commercial partnerships; define risk models and risk management processes for data within the Federation; develop specific technical infrastructure to support the credentialing process in the regulated environment; study the feasibility of creating a federated honest broker system; identity provisioning and authorization of users; develop or acquire acceptable HIPAA and research ethics training modules for the entire federated community; and establish a central auditing authority.

McGraw D, Dempsey JX, Harris L, Goldman J. Privacy as an enabler, not an impediment: building trust into health information exchange. Health Aff (Millwood). 2009 Mar-Apr;28(2):416-27.
McGraw et al. discuss the policy implications of implementing privacy and security measures into health information technology systems. HIPAA rules are useful in protecting health information in a clinical setting, but do not apply in other important situations, including the use of private personal health record services not affiliated with health care institutions. Their recommendations include: the creation of laws to regulate the use of clinical information for personal health records, the limitation of secondary use included marketing from health care products and services organization, clarification of consent in the health information exchange, better enforcement of pre-existing laws designed to protect patient privacy, and the regulation of costs associated with providing patients with their medical records.

Standardized data collection

Deitzer JR, Payne PR, Starren JB. Coverage of clinical trials in existing ontologies. AMIA Annu Symp Proc. 2006; 903.
Deitzer et al. present the findings from a study that evaluated two ontologies and their domain coverage: SNOMED CT and the NCI Thesaurus. In the first phase, manual abstraction of tasks and events from 20 clinical trial protocol documents was conducted. The resulting set of tasks and events were mapped to the UMLS by two researchers. In the second phase, the source terminologies were evaluated. Of the 102 unique concepts abstracted, 84.3% were mapped to the UMLS concepts. Neither of the ontologies provided exact matches for most of the evaluations (the UMLS matched exactly two-thirds of abstracted tasks).
Kush RD, Helton E, Rockhold FW, Hardison CD. Electronic health records, medical research, and the Tower of Babel. N Engl J Med. 2008 Apr 17;358(16):1738-40.
Kush et al. discuss the challenges to standardization and also provide an overview of the current initiatives to standardize medical and health information for research. The authors make the case for common vocabularies as a means to compare data across institutions. Nevertheless, there are multiple barriers to uniformed standardization of information. Most notably, standardization efforts come after the development and implementation of many electronic databases and the transition to a new standardized system can be challenging. The authors call for multidisciplinary support, both with regards to funding and resources, in order to actualize standardization efforts.

Los RK, van Ginneken AM, van der Lei J. OpenSDE: a strategy for expressive and flexible structured data entry. Int J Med Inform. 2005 Jul;74(6):481-90.
Los et al. describe OpenSDE, an open source application for collecting standardized information in clinical narratives. Arguing that current technologies are not detailed enough for use by clinicians, but that standardization is essential for research purposes, this application allows data to be entered using predefined concepts or domain models. Unique to this program is its goal of usability by both the clinical and research communities. At the time of this article's publication, OpenSDE was being used in numerous academic hospitals for clinical care and being pilot tested for its research capabilities.

Mohanty SK, Mistry AT, Amin W, Parwani AV, Pople AK, Schmandt L, Winters SB, Milliken E, Kim P, Whelan NB, Farhat G, Melamed J, Taioli E, Dihr R, Pass HI, Becich MJ. The development and deployment of Common Data Elements for tissue banks for translational research in cancer - an emerging standard based approach for the Mesothelioma Virtual Tissue Bank. BMC Cancer. 2008; 8: 91.
Mohanty et al. explain the process associated with developing common data elements (CDE), based on controlled vocabulary, ontology, and semantic modeling methods, for the Mesothelioma Virtual Tissue Bank. Combining and formalizing standards from other networks and initiatives, the MVB developed CDE, metadata for the CDEs and a database with querying capability. The goal is to increase interoperability between institutions.

Pathak J, Peters L, Chute CG, Bodenreider O. Comparing and evaluating terminology services application programming interfaces: RxNav, UMLSKS and LexBIG. JAMIA. 2010; 17(6):714-9.
Using qualitative methods, Pathak et al. evaluate three application programming interfaces (API) (RxNav, UMLSKS and LexBIG) and their ability to retrieve information from RxNorm, a biomedical terminology. 100 test values were selected for each simple function and queries were conducted for drug entities by name, code, national drug codes for a drug, properties of a drug concept, proprietary information about a drug concept, related drugs by relationship, and related drugs by type of drug entity. Overall, the three API produced similar results in information retrieval. Most performance differences were due to dataset alignment and content loading.

Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. Common data element (CDE) management and deployment in clinical trials. AMIA Annu Symp Proc. 2003;1048.
Warzel et al. describe the use of common data elements (CDEs) in clinical research, barriers to their standardization, and highlight the Cancer Data Standards Repository (caDSR) which is part of the NCI caCORE. Historically, CDEs have been developed anew for each research project. There was no one source for CDEs or a tool in which to perform queries and modifications of them. caDSR manages the adherence to ISO/IEC 11179 metadata standards and, through the CDE Browser allows for query and download capabilities of existing CDEs. Investigators can be matched to existing CDEs through the CDE Compliance Review Tool and new CDEs can be developed, in adherence to standard vocabulary, through the CDE Curation Tool.

Weng C, Gennari JH, Fridsma DB. User-centered semantic harmonization: a case study. J Biomed Inform. 2007; 40:353-64.
Weng et al describe BRIDG (Biomedical Research Integrated Domain Group), whose mission is to harmonize the semantics from available clinical trials information models into a shared model and to explore a methodology for user-centered semantic harmonization. The paper gives an overview of BRIDG's successful experience supporting community-based domain analysis. The authors also address technical and social challenges based on the BRIDG experience, as well as highlights technical gaps to address for the future. Some of the challenges of BRIDG's mission include: difficulty prioritizing relevant source models; the abstract nature of the system that makes it difficult to be used directly by many application-oriented users; providing inadequate application development support; being disconnected from realistic application models; the utilization of UML which is not a satisfying knowledge representation language for constructing a shared domain reference; and the decreased efficiency of the stewardship mechanism that BRIDG uses.

The Learning Healthcare system and CER

Selker HP, Strom BL, Ford DE, Meltzer DO, Pauker SG, Pincus HA, Rich EC, Tompkins C, Whitlock EP. White paper on CTSA consortium role in facilitating comparative effectiveness research: September 23, 2009 CTSA consortium strategic goal committee on comparative effectiveness research. Clin Transl Sci. 2010 Feb;3(1):29-37.
The CTSA Consortium Strategic Goal Committee on Comparative Effectiveness Research was convened in 2009 to outline the ways in which the CTSA Consortium could utilize CER methods within the National Institutes of Health. The committee cited two definitions for CER given by the Institute of Medicine (IOM) and the Federal Coordinating Council for CER (FCC-CER). The Committee concluded that the focus of their CER efforts should be on creating studies with generalizable results that are relevant to the community, utilizing HIT and EHR data to conduct more efficient research, and training students and scientists in the utilization of CER.