Clinical Research Informatics: Supporting the Research Study Lifecycle

S. B. Johnson
Healthcare Policy and Research, Weill Cornell Medicine, New York, USA

Summary

Objectives: The primary goal of this review is to summarize significant developments in the field of Clinical Research Informatics (CRI) over the years 2015-2016. The secondary goal is to contribute to a deeper understanding of CRI as a field, through the development of a strategy for searching and classifying CRI publications.

Methods: A search strategy was developed to query the PubMed database, using medical subject headings to both select and exclude articles, and filtering publications by date and other characteristics. A manual review classified publications using stages in the "research study lifecycle", with key stages that include study definition, participant enrollment, data management, data analysis, and results dissemination.

Results: The search strategy generated 510 publications. The manual classification identified 125 publications as relevant to CRI, which were classified into seven different stages of the research lifecycle, and one additional class that pertained to multiple stages, referring to general infrastructure or standards. Important cross-cutting themes included new applications of electronic media (Internet, social media, mobile devices), standardization of data and procedures, and increased automation through the use of data mining and big data methods.

Conclusions: The review revealed increased interest and support for CRI in large-scale projects across institutions, regionally, nationally, and internationally. A search strategy based on medical subject headings can find many relevant papers, but a large number of non-relevant papers need to be detected using text words which pertain to closely related fields such as computational statistics and clinical informatics. The research lifecycle was useful as a classification scheme by highlighting the relevance to the users of clinical research informatics solutions.

Keywords

Biomedical research; informatics; clinical studies as topic; review literature as topic

Yearb Med Inform 2017:193-200
http://dx.doi.org/10.15265/IY-2017-022
Published online August 18, 2017

1 Introduction

This review seeks to summarize significant developments in the field of clinical research informatics (CRI) for the years 2015-2016. The approach continues the tradition of past reviews of the IMIA Yearbook by focusing on a relatively small number of publications that are representative of current work in clinical research informatics [1-4].

The definition of clinical research used here is based on early work of the Clinical Research Roundtable at the Institute of Medicine [5, 6]. The focus is on scientific studies positioned between two translational "blocks": the translational of basic science into human studies, and the translational of human studies into clinical practice (frequently abbreviated as "T1" and "T2", respectively). CRI can be defined simply as the intersection of the field of clinical research and the field of biomedical informatics. This review attempts to partially formalize the definition of CRI as a formal search strategy, building on the method suggested by Embi [2]. This approach seeks to directly identify a list of publications that illustrate salient areas of investigation in the time period of interest, and to provide a search strategy that may be useful for future reviews.

In addition, this review develops a classification for CRI articles which may be useful for identifying areas of focus within the field. The approach adopts another product of the Clinical Research Roundtable, which subdivides clinical research informatics in terms of the stages of research studies, such as study definition, participant recruitment, data collection, data analysis, and results dissemination [7]. For a single study, these stages form a linear pipeline in which the data produced by one stage is consumed by the next stage. For the clinical research enterprise as a whole, these stages form a research "lifecycle", i.e., a circular flow in which the results of studies serve to stimulate the designs of new studies. This view is reflected in part by the recent conceptual model by Weng and Kahn [4]. One advantage of classifying articles using the research study lifecycle is the user-centric perspective: each stage of the lifecycle is largely defined by the kinds of activities performed by investigators and their staff. Recent qualitative studies confirm the importance of the research lifecycle, and suggest that informatics methods and tools are highly specific to particular stages [8].

The goal of this review is to characterize the field of CRI over the past two years. In addition, the review seeks to apply the research study lifecycle as a method of classifying subfields within CRI.

2 Methods

Table 1 provides an overview of the search strategy employed for this review. The core of the search strategy consists of two conceptual axes: clinical research and informatics. There is no medical subject heading (MeSH) descriptor for clinical research, so the first axis uses *biomedical research*, which includes *human experimentation*, *health services research,* and *comparative effectiveness research*. This term did not retrieve some articles that were searchable by text words such as "clinical trials", "clinical research" and "recruitment", so the axis was extended with MeSH terms *clinical studies as topic*, *patient selection*, and *multicenter studies as topic*, which appear under the *investigative techniques* hierarchy. To shift the focus away from basic science, the search excluded MeSH terms for *genetic research* and *translational medical research*.

The second axis (*informatics*) includes subfields such as *medical informatics*, *nursing informatics,* and *public health informatics*, but excludes *genomics* and *computational biology*. Text searches showed that certain articles were not included under this term, so the axis was extended with *computational methods*, which includes important terms such as *artificial intelligence*, *natural language processing,* and *database management systems*.

All the MeSH terms used to define the conceptual axes were qualified with the "major" subheading, to identify publications that have these terms as their main focus. The terms forming each individual axis were combined with the OR operator, and the two axes were combined with the AND operator, yielding 9,522 citations. The excluded terms were removed with the NOT operator, leaving 8,493 citations.

When limited to the years 2015-2016, this search produced 909 publications. The search was further limited to the English language and availability of abstracts. This restriction was necessary to prepare for the classification of the articles (see description below) which made extensive use of text words. The focus of this review was on original research, so the search excluded publication types for reviews, comments, letters, and clinical trials, which resulted in 510 publications. The clinical trial publication type was excluded to remove articles that describe a specific clinical trial that happens to use some form of informatics, rather than being about the use of informatics to support clinical research more generally.

The citations returned by the search strategy in Table 1 contained a mixture of articles relevant to CRI, and many that were not relevant. Among these publications, those having multiple MeSH terms in both axes were most representative of CRI, while those having a singleton term in one or both axes were least representative, and were frequently not relevant. For example, there are a large number of publications that have *data interpretation, statistical* as the singleton term for the informatics axis. These publications are about the use of computation as part of the statistical methods. Similarly, articles with the singleton term *hospital information systems* or *outcome assessment (health care)* in the clinical research axis were typically studies of informatics in patient care. However, removing all articles with these MeSH terms would have eliminated too many relevant articles.

For this reason, the citations retrieved by the search strategy were manually classified, as shown in Table 2. The upper portion of the table represents the 124 citations that were judged to be relevant to CRI, and the bottom portion represents the 386 citations judged to be non-relevant. As described above, non-relevant publications were identified manually using MeSH terms, but also using text words in the title or abstract. These included articles for which the focus was patient care, statistics, and basic science.

The manual classification also assigned each relevant citation to a class based on the most appropriate stage of the research study life cycle. As with non-relevant citations, MeSH terms were sometimes helpful in defining these classes; however, text words from the abstract were generally the most useful in assigning a stage. The stages of research were suggested by prior work in this area, but were ultimately determined by the data, and ordered by the chronology of activities required to carry out a study: design of study (D), enrollment of participants (E), execution of study (X), management of data (M), use of data (U), communication of results (C), and re-use of publication results (R). A number of publications pertained to many different stages and typically addressed general informatics issues relevant to CRI, such as systems architectures, security, or data standards (A).

Table 1 PubMed search strategy. Column 1 indicates the Boolean operator used to combine terms (AND or NOT). Column 2 specifies the PubMed field being searched: major MeSH term (majr), MeSH term (mesh), date of publication (dp), language (la), and publication type (pt). Column 3 indicates the value used to search within the specified field; terms and publication types are combined with the OR operator. The last column provides the number of articles retrieved for the operation performed in each row.

Operation	Field	Value	Results
	majr	biomedical research OR clinical studies as topic OR patient selection OR multicenter studies as topic	173,008
AND	majr	informatics OR computing methodologies	9,522
NOT	mesh	genetic research OR translational research OR genomics OR computational biology	8,493
AND	dp	2015:2016	909
AND		hasabstract	707
AND	la	eng	658
NOT	pt	review OR clinical trial OR comment OR letter	510

3 Results

The 124 publications that were judged to be relevant to CRI were reviewed. This process revealed additional common themes within each class (research study stage), which are described in the following sections.

3.1 Architectures and Standards

The publications reviewed that pertained to multiple stages of the research lifecycle included 15 articles (12%). These articles typically addressed general approaches to

Table 2 Manual classification of the 510 citations. Column 1 indicates whether the citations are relevant or non-relevant to CRI. Column 2 assigns a class label, with number of citations in column 3, and name of class in column 4. Column 5 provides examples of MeSH terms related to the class, and column 6 shows examples of text words from titles and abstracts.

Relevance	Class	Count	Name	Example MeSH Terms	Example Keywords
Relevant (124)	A	15	Architecture and standards	Computer communication networks, computer systems	Architectures, standards, national
	D	11	Design of study	Randomized controlled trials as topic, computer graphics	Design, protocol, criterion
	E	15	Enrollment of participants	Clinical trials as topic, diagnosis, computer-assisted	Recruitment, eligibility, matching
	X	12	Execution of study	Internet, social media	Workflow, conduct, staff
	M	20	Management of data	Database management systems, information storage and retrieval	Management, database, collection
	U	11	Use of data	Data mining, natural language processing	Mining, big, processing
	C	9	Communication of results	Data curation, MedlinePlus	Dissemination, reporting, public
	R	31	Re-use of publication results analysis	Databases, bibliographic, Medline, pubMed	Evidence, systematic, reproducibility
Non-relevant (386)	H	253	Healthcare	Hospital information systems, outcome assessment (health care)	Care, delivery, therapy
	S	116	Statistics	Data interpretation, statistical, numerical analysis, computer-assisted	Statistical, multivariate, sampling
	B	17	Basic science	Biomedical research, biological ontologies	Biological, laboratory, basic

CRI, such as systems architectures, research networks, or data standards. Several of the publications described large-scale efforts to improve the state of CRI through regional, national, or international consortia or funding models. Infrastructure initiatives included interoperable electronic health records, cloud computing, management of big data sources (such as genomics and imaging), collection of patient-reported outcomes, and multi-institution integration for comparative effectiveness research [9-13]. One crucial aspect of systems architectures for CRI is the ability to protect confidentiality of participants; articles in this group covered methods for securely sharing data across sites, detecting protected health information and pseudonymization [14-16]. Efforts related to data standardization included a comparison of data models, processes for data harmonization, federated data sharing, and minimum datasets [17-20]. These papers addressed a wide range of disease areas, including cancer, lung disease, and rare diseases, as well as Down syndrome, heart disease, and diabetes [21-23].

3.2 Improving Study Designs

The first stage of the research study lifecycle involves activities in preparation for conducting a study, such as developing the study protocol. The review included 11 publications (9%) that addressed informatics methods and tools for understanding or improving study designs. One group of these articles provided support for designing various aspects of study design, such as managing confounding factors, comparing placebos, stratification, adaptive designs, group designs, and so-called n of one studies [24-29]. Another group examined broader aspects of protocol design, such as assessing the feasibility of the study, preparing the protocol for the institutional review board (IRB), reducing fraudulent behavior in internet-based studies, managing the protocol across multiple sites, and managing protocols of multiple studies [30-34].

3.3 Enrolling Participants into Studies

Once the study is designed and approved, the next stage of the research study life cycle is concerned with enrolling participants into studies and includes such activities as pre-screening, screening, and consenting.

The review considered 15 papers (12%) which used a variety of strategies to improve recruitment. One group of papers sought to streamline the processes of recruitment by using information retrieval to identify potential participants, speeding electronic chart review, and helping providers to refer patients [35-37]. Standards are a key method in improving the recruitment process, and include obtaining a better understanding of system requirements, using standardized data elements, and applying Semantic Web technologies [38-41].

Sophisticated automated methods are also coming into play to assist with patient matching. Papers reviewed included the use of natural language processing on clinical notes, case-based reasoning, and automated analysis of audiograms [42-45]. The Internet is having an increased impact on patient recruitment. Papers in this group addressed improving patient awareness of available studies, creating online registries and portals, and analyzing the new ethical issues that arise in the use of such systems [46-49].

3.4 Executing Studies

The next stage of the research study lifecycle concerns the execution of the study. The review examined 12 papers (10%) that addressed a variety of factors, such as the roles of staff members performing the work, the workflow required to complete tasks, and the provision of the staff with appropriate training.

Clinical studies can involve large numbers of highly diverse staff members, including investigators, research nurses, coordinators, data managers, and statisticians. The review included papers that focused on understanding the perspectives and requirements of stakeholders, including approaches for engaging them in the research process [8, 50, 51]. In addition, there were articles that discussed the educational needs of research staff for tasks such as record linkage, ethics, and biobanking, which use a range of online and multimedia content to deliver training [52-54].

There are numerous complexities in the workflow of clinical studies, especially with regard to online environments, such as secondary use of electronic health records, social networks, and patient-led research studies [55, 56]. Technologies such as smart phones and the Internet are also providing new opportunities for conducting clinical research which includes tracking patients, assessing compliance, and delivering interventions [57-60].

3.5 Managing Study Data

The next stage of the research lifecycle focuses on data management, which includes tasks for collecting, organizing, and integrating data prior to analysis. The review included 20 articles (16%) that addressed various aspects of data management, including best practices, use of data standards, and integration of multi-media data. Practical guidance for data management addressed a variety of topics, including managing data in the field, improving management workflow, and using databases and registries to structure the data [61-65]. Additional guidance for practice covers using mobile technology for data collection, reusing data from electronic health records, and graphical methods to explore such data collections [66-68].

Data standards continue to be vital for the management of research data. These include the use of semantic open data technologies, open source software, and common data models [69-72]. These standards have important consequences for measuring data quality, source data verification, and data normalization [73-76]. The challenges of additional types of media are giving rise to new approaches for data management, including the use of video, images, physiologic signals, and global positioning system data [77-80].

3.6 Using Study Data

After study data has been collected and prepared, the next stage in the lifecycle is to analyze it. There were 11 papers (9%) captured by this review that addressed data use in clinical studies, which covers important informatics topics such as data mining and the analysis of large datasets. Papers relevant to data mining examined such problems as discovering indications for drugs, extracting instructions from prescriptions, assessing effectiveness, and evaluating patient phenotypes [81-84]. With the explosion of new sources of data for clinical research, methods for the analysis of "big data" are becoming increasingly important. In this review, papers characterized big data in terms of volume, variety, and velocity, with data sources including structured, unstructured, and image data. Problems addressed by big data approaches included drug discovery, health disparities, psychotherapy outcomes, smoking, and nursing research [85-91].

3.7 Communicating Study Results

This stage of the research lifecycle focuses on the dissemination of the results of a study, through publication, data sharing, and communication directed at specific stakeholders. This review captured 9 papers (7%) that described methods for dissemination through a variety of electronic media. One theme in this group included standards for dissemination including open databases for sharing study-related materials, standards for reporting results and sharing study documentation, and standards for registering studies and complying with regulatory requirements [92-95]. Another group examined how to assess the availability of study results, readability of study descriptions, and potential impact of a study using bibiliometrics [96-98]. Two papers described informatics methods for the dissemination of research to broader audiences, seeking to improve public understanding, government support for research, and policy makers awareness [99, 100].

3.8 Analyzing Study Publications

The final stage of the research lifecycle uses the results generated by multiple studies. This stage completes the cycle by using the results of prior studies to inform the design of new studies. This was the largest group of papers examined by the review, containing 31 papers (25%). One group included articles that conduct systematic reviews using a variety of online databases. These covered a wide variety of therapies, including medications, brachytherapy, exercise, nutrition, and ophthalmological treatments [101-108]. Another group analyzed various trends in publications, such as how researchers access the literature, extent to which studies comply with registration and reporting requirements, reasons for study termination, transparency regarding sponsorship and conflicts of interest, and inclusion of patient-reported outcomes [109, 110, 111-115].

The next group used more advanced methods to mine various aspects of published studies. These included a wide range of goals, including predicting regulatory approval, assessing bias, extracting data from study text and figures, selecting articles for systematic review, identifying available evidence for a topic, extracting characteristics of study participants, and detecting articles that describe the same study [116-125].

The last group of papers sought to directly re-use the results generated from prior studies. Goals pursued by this group included improving decision-making in future trials, controlling access to shared study data, imputation of missing data, pooling data to improve analysis, and reproducing results [126-131].

4 Discussion

4.1 Developments in the CRI field

The articles targeted by this review demonstrate the vital role that informatics has come to play in clinical research. At an early time in the field, research in CRI was often limited to single institutions, or even individual informatics investigators. The current state of the field illustrates the importance of informatics, with support for large-scale projects across

multiple institutions, either regionally, or nationally, and internationally. This is particularly the case for articles describing general methods such as architectures, research networks, and standards, but also in numerous articles addressing more specific methods, such as recruitment.

The reviewed articles also to some extent reflect the evolution of informatics as a field. The papers are dominated by "classic" informatics approaches that pertain to standards such as terminology, data models, and interoperability. We also see the growth of these approaches into more "modern" methods which include open source, semantic Web, and data sharing. There is also increasing awareness of the importance of human factors in clinical research, including workflow, stakeholder engagement, and training. Finally, we see considerable discussion of "hot" topics, such as big data, data mining, and text mining.

These developments are paralleled, for the most part, by applications of different kinds of online technologies and media. While not new in itself, the Internet is providing enormous opportunities for new applications in clinical research, including promoting awareness, recruiting participants, engaging stakeholders, delivering interventions, and disseminating results. In particular, we see many novel applications in the use of social media and mobile technologies.

4.2 Methods to Define the CRI Field

It continues to be challenging to provide a formal definition for clinical research informatics. This review adopted a standard query strategy that combined two conceptual axes using MeSH terms. Excluded MeSH terms were largely successful at removing basic science papers from the collection (such as computational biology), but were not effective in removing very closely related fields such as computational statistics or clinical informatics. Excluding MeSH terms associated with these fields in the query would have removed too many papers relevant to CRI.

The manual review demonstrates that these terms do provide strong evidence for non-relevance, which can be strengthened with the use of particular text words. One possible way forward is to employ a simple classification algorithm (such as k-means) using a combination of MeSH terms and text words. The preliminary results here suggest that this approach could be highly effective.

For articles relevant to CRI, the research lifecycle proved to be a useful approach for classification, particularly because this viewpoint considers how investigators will use the technology, in contrast to classifying by type of informatics methods used (such as data mining). An added benefit of this approach is that the resulting classification suggests how individual tools and methods might be combined to form larger portions of the research "pipeline", with the results of one tool feeding into the next. One challenge for this approach is that the lifecycle is a continuum of activities, and so there are different ways to partition the stages as discrete intervals. For example, the stages E (enrollment) and X (execution) might be combined, as could stages for data management (M) and use (U). Further work is required to standardize the stages of research for classification purposes.

The classification that emerged from the manual review process was relatively uniform across the stages of the research lifecycle. Some stages were expected to have ample numbers of papers, such as participant enrollment, data management, and data use. It was not expected to find so many articles on study definition, study execution, or results communication. It was also not anticipated that the results re-use stage would have the largest number of papers. One issue here is when a paper about systematic review can be considered "informatics"; future strategies for CRI may wish to exclude these papers. The last group of papers in this stage focused on the actual reuse of data generated by studies, in contrast to a simple synthesis of the literature. These papers address the long-standing problem of lack of reproducibility in scientific findings, which may ultimately be considered as a separate stage from literature review.

5 Conclusions

This review revealed increased interest and support for CRI in large-scale projects across institutions, regionally, nationally, and internationally. The most important influence of informatics on CRI is in the use of standards for data, software, and best practices. The publications address an increasing richness of data sources, variety of media, and explosion of new computational methods for data mining and big data.

A search strategy using two conceptual axes helped to identify publications relevant to CRI, but significant manual effort was required to remove non-relevant papers. Additional work was required to sharpen the boundaries between CRI and closely related fields such as computational statistics and patient care informatics. The research study lifecycle was useful in classifying relevant CRI publications, and helped to focus attention on how a CRI method or tool could benefit end users, including clinical investigators and their staff members. In addition, the life cycle suggests how these individual CRI initiatives might be combined into a larger "pipeline" that supports the clinical research enterprise.

References

1. Dugas M. Clinical Research Informatics: Recent Advances and Future Directions. Yearb Med Inform 2015;10:174-7.
2. Embi PJ. Clinical Research Informatics: Survey of Recent Advances and Trends in a Maturing Field. Yearb Med Inform 2013;8:178-84.
3. Richesson RL, Horvath MM, Rusincovitch SA. Clinical Research Informatics and Electronic Health Record Data. Yearb Med Inform 2014;9:215-23.
4. Weng C, Kahn MG. Clinical Research Informatics for Big Data and Precision Medicine. Yearb Med Inform 2016(1):211-8.
5. Crowley WF, Jr., Sherwood L, Salber P, Scheinberg D, Slavkin H, Tilson H, et al. Clinical Research in the United States at a Crossroads: Proposal for a Novel Public-Private Partnership to Establish a National Clinical Research Enterprise. JAMA 2004;291(9):1120-6.
6. Sung NS, Crowley WF, Jr., Genel M, Salber P, Sandy L, Sherwood LM, et al. Central Challenges Facing the National Clinical Research Enterprise. JAMA 2003;289(10):1278-87.
7. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the Translational Barriers: the Value of Integrating Biomedical Informatics and Translational Research. J Investig Med 2005;53(4):192-200.
8. Johnson SB, Farach FJ, Pelphrey K, Rozenblit L. Data Management in Clinical Research: Synthesizing Stakeholder Perspectives. J Biomed Inform 2016;60:286-93.
9. De Moor G, Sundgren M, Kalra D, Schmidt A,

Dugas M, Claerhout B, et al. Using Electronic Health Records for Clinical Research: the Case of the EHR4CR Project. J Biomed Inform 2015;53:162-73.
10. Ohmann C, Canham S, Danielyan E, Robertshaw S, Legre Y, Clivio L, et al. 'Cloud computing' and Clinical Trials: Report from an ECRIN Workshop. Trials 2015;16:318.
11. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making Sense of Big Data in Health Research: Towards an EU Action Plan. Genome Med 2016;8(1):71.
12. Dymek C, Gingold J, Shanbhag A, Fridsma D, Yong PL. A National Data Infrastructure for Patient-centered Outcomes Research. J Comp Eff Res 2015;4(1):75-87.
13. Hazlehurst BL, Kurtz SE, Masica A, Stevens VJ, McBurnie MA, Puro JE, et al. CER Hub: An Informatics Platform for Conducting Comparative Effectiveness Research using Multi-Institutional, Heterogeneous, Electronic Clinical Data. Int J Med Inform 2015;84(10):763-73.
14. Budin-Ljosne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ, et al. DataSHIELD: an Ethically Robust Solution to Multiple-Site Individual-level Data Analysis. Public Health Genomics 2015;18(2):87-96.
15. Lautenschlager R, Kohlmayer F, Prasser F, Kuhn KA. A Generic Solution for Web-based Management of Pseudonymized Data. BMC Med Inform Decis Mak 2015;15:100.
16. Redd A, Pickard S, Meystre S, Scehnet J, Bolton D, Heavirland J, et al. Evaluation of PHI Hunter in Natural Language Processing Research. Perspect Health Inf Manag 2015;12:1f.
17. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of Medical Data Models: Information Infrastructure for Medical Research and Healthcare. Database (Oxford) 2016;2016.
18. Meeker D, Jiang X, Matheny ME, Farcas C, D'Arcy M, Pearlman L, et al. A System to Build Distributed Multivariate Models and Manage Disparate Data Sharing Policies: Implementation in the Scalable National Network for Effectiveness Research. J Am Med Inform Assoc 2015;22(6):1187-95.
19. Siesling S, Louwman WJ, Kwast A, van den Hurk C, O'Callaghan M, Rosso S, et al. Uses of Cancer Registries for Public Health and Clinical Research in Europe: Results of the European Network of Cancer Registries Survey among 161 Population-based Cancer Registries during 2010-2012. Eur J Cancer 2015;51(9):1039-49.
20. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A Methodology for a Minimum Data Set for Rare Diseases to Support National Centers of Excellence for Healthcare and Research. J Am Med Inform Assoc 2015;22(1):76-85.
21. Lavigne J, Sharr C, Ozonoff A, Prock LA, Baumer N, Brasington C, et al. National Down Syndrome Patient Database: Insights from the Development of a Multi-center Registry Study. Am J Med Genet A 2015;167A(11):2520-6.
22. Rasooly RS, Akolkar B, Spain LM, Guill MH, Del Vecchio CT, Carroll LE. The National Institute of Diabetes and Digestive and Kidney Diseases Central Repositories: a Valuable Resource for Nephrology Research. Clin J Am Soc Nephrol 2015;10(4):710-5.
23. Antman EM, Benjamin EJ, Harrington RA, Houser SR, Peterson ED, Bauman MA, et al. Acquisition, Analysis, and Sharing of Data in 2015 and Beyond: A Survey of the Landscape: A Conference Report From the American Heart Association Data Summit 2015. J Am Heart Assoc 2015;4(11).
24. Neugebauer R, Schmittdiel JA, Zhu Z, Rassen JA, Seeger JD, Schneeweiss S. High-dimensional Propensity Score Algorithm in Comparative Effectiveness Research with Time-varying Interventions. Stat Med 2015;34(5):753-81.
25. Goldenholz DM, Moss R, Scott J, Auh S, Theodore WH. Confusing Placebo Effect with Natural History in Epilepsy: A Big Data Approach. Ann Neurol 2015;78(3):329-36.
26. Hesse K, MacIsaac RL, Abdul-Rahim AH, Lyden PD, Bluhmki E, Lees KR, et al. Online Tool to Improve Stratification of Adverse Events in Stroke Clinical Trials. Stroke 2016;47(3):882-5.
27. Sugitani T, Bretz F, Maurer W. A Simple and Flexible Graphical Approach for Adaptive Group-Sequential Clinical Trials. J Biopharm Stat 2016;26(2):202-16.
28. Pouwels KB, Mulder B, Hak E. Moderate Concordance was found between Case-only and Parallel Group Designs in Systematic Comparison. J Clin Epidemiol 2016;71:18-24.
29. Lehrach H. Virtual Clinical Trials, an Essential Step in Increasing the Effectiveness of the Drug Development Process. Public Health Genomics 2015;18(6):366-71.
30. Soto-Rey I, Trinczek B, Amo JI, Bauselas J, Dugas M, Fritz F. Web-based Multi-site Feasibility Questionnaire Tool. Stud Health Technol Inform 2015;212:88-93.
31. Kury FS, Cimino JJ. Identifying Repetitive Institutional Review Board Stipulations by Natural Language Processing and Network Analysis. Stud Health Technol Inform 2015;216:579-83.
32. Teitcher JE, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, Preventing, and Responding to "fraudsters" in Internet Research: Ethics and Tradeoffs. J Law Med Ethics 2015;43(1):116-33.
33. Klungel OH, Kurz X, de Groot MC, Schlienger RG, Tcherny-Lessenot S, Grimaldi L, et al. Multi-centre, Multi-database Studies with Common Protocols: Lessons Learnt from the IMI PROTECT Project. Pharmacoepidemiol Drug Saf 2016;25 Suppl 1:156-65.
34. He Z, Carini S, Sim I, Weng C. Visual Aggregate Analysis of Eligibility Features of Clinical Trials. J Biomed Inform 2015;54:241-55.
35. Patrao DF, Oleynik M, Massicano F, Morassi Sasso A. Recruit--An Ontology Based Information Retrieval System for Clinical Trials Recruitment. Stud Health Technol Inform 2015;216:534-8.
36. Shivade C, Hebert C, Lopetegui M, de Marneffe MC, Fosler-Lussier E, Lai AM. Textual Inference for Eligibility Criteria Resolution in Clinical Trials. J Biomed Inform 2015;58 Suppl:S211-8.
37. Afrin LB, Oates JC, Kamen DL. Improving Clinical Trial Accrual by Streamlining the Referral Process. Int J Med Inform 2015;84(1):15-23.
38. Schreiweis B, Bergh B. Requirements for a Patient Recruitment System. Stud Health Technol Inform 2015;210:521-5.
39. Doods J, Lafitte C, Ulliac-Sagnes N, Proeve J, Botteri F, Walls R, et al. A European Inventory of Data Elements for Patient Recruitment. Stud Health Technol Inform 2015;210:506-10.
40. Cuggia M, Campillo-Gimenez B, Bouzille G, Besana P, Jouini W, Dufour JC, et al. Automatic Selection of Clinical Trials Based on A Semantic Web Approach. Stud Health Technol Inform 2015;216:564-8.
41. Schreiweis B, Bergh B. Applicability of different types of Patient Records for Patient Recruitment Systems. Stud Health Technol Inform 2015;216:884.
42. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated Clinical Trial Eligibility Prescreening: Increasing the Efficiency of Patient Identification for Clinical Trials in the Emergency Department. J Am Med Inform Assoc 2015;22(1):166-78.
43. Rahne T, Buthut F, Plossl S, Plontke SK. A Software Tool for Puretone Audiometry. Classification of Audiograms for Inclusion of Patients in Clinical Trials. English version. HNO 2016;64 Suppl 1:S1-6.
44. Miotto R, Weng C. Case-based Reasoning using Electronic Health Records Efficiently identifies Eligible Patients for Clinical Trials. J Am Med Inform Assoc 2015;22(e1):e141-50.
45. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the Efficiency of Trial-patient Matching: Automated Clinical Trial Eligibility Pre-screening for Pediatric Oncology Patients. BMC Med Inform Decis Mak 2015;15:28.
46. Abel GA, Cronin AM, Earles K, Gray SW. Accessibility and Quality of Online Cancer-Related Clinical Trial Information for Naive Searchers. Cancer Epidemiol Biomarkers Prev 2015;24(10):1629-31.
47. Rimel BJ, Lester J, Sabacan L, Park D, Bresee C, Dang C, et al. A Novel Clinical Trial Recruitment Strategy for Women's Cancer. Gynecol Oncol 2015;138(2):445-8.
48. Rocker C, Cappelletti L, Marshall C, Meunier CC, Brooks DW, Sherer T, et al. Use of an Online Portal to Facilitate Clinical Trial Recruitment: a Preliminary Analysis of Fox Trial Finder. J Parkinsons Dis 2015;5(1):55-66.
49. Refolo P, Sacchini D, Minacori R, Daloiso V, Spagnolo AG. E-recruitment Based Clinical Research: Notes for Research Ethics Committees/Institutional Review Boards. Eur Rev Med Pharmacol Sci 2015;19(5):800-4.
50. Khatri C, Chapman SJ, Glasbey J, Kelly M, Nepogodiev D, Bhangu A, et al. Social Media and Internet driven Study Recruitment: Evaluating a New Model for Promoting Collaborator Engagement and Participation. PLoS One 2015;10(3):e0118899.
51. Jackson D, Waine ML, Hutchinson M. Blogs as a Way to Elicit Feedback on Research and Engage Stakeholders. Nurse Res 2015;22(3):41-7.
52. Tan KM, Flack FS, Bear NL, Allen JA. An Evaluation of a Data Linkage Training Workshop for Research Ethics Committees. BMC Med Ethics 2015;16:13.
53. Haugen M, Gasber E, Leonard M, Landier W. Harnessing Technology to Enhance Delivery of

Clinical Trials Education for Nurses: a Report from the Children's Oncology Group. J Pediatr Oncol Nurs 2015;32(2):96-102.
54. Sehovic I, Gwede CK, Meade CD, Sodeke S, Pentz R, Quinn GP. A Web-Based Platform for Educating Researchers About Bioethics and Biobanking. J Cancer Educ 2016;31(2):397-404.
55. Lamas E, Salinas R, Vuillaume D. A New Challenge to Research Ethics: Patients-Led Research (PLR) and the Role of Internet Based Social Networks. Stud Health Technol Inform 2016;221:36-40.
56. Bouzille G, Sylvestre E, Campillo-Gimenez B, Renault E, Ledieu T, Delamarre D, et al. An Integrated Workflow For Secondary Use of Patient Data for Clinical Research. Stud Health Technol Inform 2015;216:913.
57. Mitchell SG, Schwartz RP, Alvanzo AA, Weisman MS, Kyle TL, Turrigiano EM, et al. The Use of Technology in Participant Tracking and Study Retention: Lessons Learned From a Clinical Trials Network Study. Subst Abus 2015;36(4):420-6.
58. Geller NL, Kim DY, Tian X. Smart Technology in Lung Disease Clinical Trials. Chest 2016;149(1):22-6.
59. Akmatov MK, Rubsamen N, Schultze A, Kemmling Y, Obi N, Gunther K, et al. Diverse Recruitment Strategies Result in Different Participation Percentages in a Web-based Study, but in Similar Compliance. Int J Public Health 2015;60(8):937-43.
60. Stephens TM, Gunther ME. Twitter, Millennials, and Nursing Education Research. Nurs Educ Perspect 2016;37(1):23-7.
61. Park JY, Kim DR, Haldar B, Mallick AH, Kim SA, Dey A, et al. Use of the Data System for Field Management of a Clinical Study Conducted in Kolkata, India. BMC Res Notes 2016;9:20.
62. Lee H, Chapiro J, Schernthaner R, Duran R, Wang Z, Gorodetski B, et al. How I do it: a Practical Database Management System to Assist Clinical Research Teams with Data Collection, Organization, and Reporting. Acad Radiol 2015;22(4):527-33.
63. Long E, Huang B, Wang L, Lin X, Lin H. Construction of Databases: Advances and Significance in Clinical Research. Eye Sci 2015;30(4):184-9.
64. Adkinson JM, Casale MT, Kim JY, Khavanin N, Gutowski KA, Gosain AK. So You Have a Research Idea: A Survey of Databases Available for Plastic Surgery Research. Plast Reconstr Surg 2016;137(2):680-9.
65. Baili P, Torresani M, Agresti R, Rosito G, Daidone MG, Veneroni S, et al. A Breast Cancer Clinical Registry in an Italian Comprehensive Cancer Center: an Instrument for Descriptive, Clinical, and Experimental Research. Tumori 2015;101(4):440-6.
66. Kaka H, Ayearst R, Tran M, Touma Z, Bagovich M, Vinik O, et al. Developing an Ipad(R) Application for Data Collection in a Rheumatology Research Clinic. Int J Technol Assess Health Care 2015;31(1-2):99-102.
67. Huang CW, Lu R, Iqbal U, Lin SH, Nguyen PA, Yang HC, et al. A richly Interactive Exploratory Data Analysis and Visualization Tool using Electronic Medical Records. BMC Med Inform Decis Mak 2015;15:92.
68. Mayer MA, Furlong LI, Torre P, Planas I, Cots F, Izquierdo E, et al. Reuse of EHRs to Support Clinical Research in a Hospital of Reference. Stud Health Technol Inform 2015;210:224-6.
69. Legaz-Garcia Mdel C, Minarro-Gimenez JA, Menarguez-Tortosa M, Fernandez-Breis JT. Lessons Learned in the Generation of Biomedical Research Datasets using Semantic Open Data technologies. Stud Health Technol Inform 2015;210:165-9.
70. Tuti T, Bitok M, Paton C, Makone B, Malla L, Muinga N, et al. Innovating to Enhance Clinical Data Management using Non-commercial and Open Source Solutions across a Multi-center Network supporting Inpatient Pediatric Care and Research in Kenya. J Am Med Inform Assoc 2016;23(1):184-92.
71. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and Utility of Applications of the Common Data Model to Multiple, Disparate Observational Health Databases. J Am Med Inform Assoc 2015;22(3):553-64.
72. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-8.
73. Houston L, Probst Y, Humphries A. Measuring Data Quality Through a Source Data Verification Audit in a Clinical Research Setting. Stud Health Technol Inform 2015;214:107-13.
74. Elkhenini HF, Davis KJ, Stein ND, New JP, Delderfield MR, Gibson M, et al. Using an Electronic Medical Record (EMR) to Conduct Clinical Trials: Salford Lung Study feasibility. BMC Med Inform Decis Mak 2015;15:8.
75. Yoon D, Schuemie MJ, Kim JH, Kim DK, Park MY, Ahn EK, et al. A Normalization Method for Combination of Laboratory Test Results from Different Electronic Healthcare Databases in a Distributed Research Network. Pharmacoepidemiol Drug Saf 2016;25(3):307-16.
76. Miller TP, Troxel AB, Li Y, Huang YS, Alonzo TA, Gerbing RB, et al. Comparison of Administrative/Billing Data to Expected Protocol-mandated Chemotherapy Exposure in Children with Acute Myeloid Leukemia: a Report from the Children's Oncology Group. Pediatr Blood Cancer 2015;62(7):1184-9.
77. Yan L, Hicks M, Winslow K, Comella C, Ludlow C, Jinnah HA, et al. Secured Web-based Video Repository for Multicenter Studies. Parkinsonism Relat Disord 2015;21(4):366-71.
78. McGregor C, Heath J, Choi Y. Streaming Physiological Data: General Public Perceptions of Secondary Use and Application to Research in Neonatal Intensive Care. Stud Health Technol Inform 2015;216:453-7.
79. Deserno TM, Deserno V, Haak D, Kabino K. Digital Imaging and Electronic Data Capture in Multi-Center Clinical Trials. Stud Health Technol Inform 2015;216:930.
80. Jankowska MM, Schipperijn J, Kerr J. A Framework for Using GPS Data in Physical Activity and Sedentary Behavior Studies. Exerc Sport Sci Rev 2015;43(1):48-56.
81. Jang D, Lee S, Lee J, Kim K, Lee D. Inferring New Drug Indications using the Complementarity between Clinical Disease Signatures and Drug Effects. J Biomed Inform 2016;59:248-57.
82. Shah BR, Lipscombe LL. Clinical Diabetes Research using Data Mining: a Canadian Perspective. Can J Diabetes 2015;39(3):235-8.
83. Alnazzawi N, Thompson P, Batista-Navarro R, Ananiadou S. Using Text Mining Techniques to Extract Phenotypic Information from the PhenoCHF Corpus. BMC Med Inform Decis Mak 2015;15 Suppl 2:S3.
84. Karystianis G, Sheppard T, Dixon WG, Nenadic G. Modelling and Extraction of Variability in Free-text Medication Prescriptions from an Anonymised Primary Care Electronic Medical Record Research Database. BMC Med Inform Decis Mak 2016;16:18.
85. Taglang G, Jackson DB. Use of "Big Data" in Drug Discovery and Clinical Trials. Gynecol Oncol 2016;141(1):17-23.
86. Tilve Alvarez CM, Ayora Pais A, Ruiz Romero C, Llamas Gomez D, Carrajo Garcia L, Blanco Garcia FJ, et al. Integrating Medical and Research Information: a Big Data Approach. Stud Health Technol Inform 2015;210:707-11.
87. Murphy SN, Herrick C, Wang Y, Wang TD, Sack D, Andriole KP, et al. High Throughput Tools to access Images from Clinical Archives for Research. J Digit Imaging 2015;28(2):194-204.
88. Bakken S, Reame N. The Promise and Potential Perils of Big Data for Advancing Symptom Management Research in Populations at Risk for Health Disparities. Annu Rev Nurs Res. 2016;34:247-60.
89. Owen J, Imel ZE. Introduction to the Special Section "Big'er' Data": Scaling up Psychotherapy Research in Counseling Psychology. J Couns Psychol 2016;63(3):247-8.
90. Mertz L. The Case for Big Data: New York City's Kalvi HUMAN Project Aims to Use Big Data in Resolving Big Health Questions. IEEE Pulse 2016;7(5):45-7.
91. Brennan PF, Bakken S. Nursing Needs Big Data and Big Data Needs Nursing. J Nurs Scholarsh 2015;47(5):477-84.
92. Goldacre B, Gray J. OpenTrials: Towards a Collaborative Open Database of all Available Information on all Clinical Trials. Trials 2016;17:164.
93. Hecht A, Busch-Heidger B, Gertzen H, Pfister H, Ruhfus B, Sanden PH, et al. Quality Expectations and Tolerance Limits of Trial Master Files (TMF) - Developing a Risk-based Approach for Quality Assessments of TMFs. Ger Med Sci 2015;13:Doc23.
94. Nicholls SG, Quach P, von Elm E, Guttmann A, Moher D, Petersen I, et al. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines. PLoS One 2015;10(5):e0125620.
95. O'Reilly EK, Hassell NJ, Snyder DC, Natoli S, Liu I, Rimmler J, et al. ClinicalTrials.gov Reporting: Strategies for Success at an Academic Health Center. Clin Transl Sci 2015;8(1):48-51.
96. Dissemination Beyond Publication. Nurse Res. 2015;22(6):5.
97. Wu DT, Hanauer DA, Mei Q, Clark PM, An LC, Proulx J, et al. Assessing the Readability of ClinicalTrials.gov. J Am Med Inform Assoc 2016;23(2):269-75.
98. Dufka FL, Munch T, Dworkin RH, Rowbotham

MC. Results Availability for Analgesic Device, Complex Regional Pain Syndrome, and Post-stroke Pain Trials: Comparing the RReADS, RReACT, and RReMiT Databases. Pain 2015;156(1):72-80.
99. Williams RS, Lotia S, Holloway AK, Pico AR. From Scientific Discovery to Cures: Bright Stars within a Galaxy. Cell 2015;163(1):21-3.
100. Kapp JM, Hensel B, Schnoring KT. Is Twitter a Forum for Disseminating Research to Health Policy Makers? Ann Epidemiol 2015;25(12):883-7.
101. Cihoric N, Tsikkinis A, Miguelez CG, Strnad V, Soldatovic I, Ghadjar P, et al. Portfolio of Prospective Clinical Trials Including Brachytherapy: an Analysis of the ClinicalTrials.gov Database. Radiat Oncol 2016;11:48.
102. Jeong S, Han N, Choi B, Sohn M, Song YK, Chung MW, et al. Construction of a Database for Published Phase II/III Drug Intervention Clinical Trials for the Period 2009-2014 Comprising 2,326 Records, 90 Disease Categories, and 939 Drug Entities. Int J Clin Pharmacol Ther 2016;54(6):416-25.
103. Waffenschmidt S, Guddat C. Searches for Randomized Controlled Trials of Drugs in MEDLINE and EMBASE Using Only Generic Drug Names Compared with Searches Applied in Current Practice in Systematic Reviews. Res Synth Methods 2015;6(2):188-94.
104. Cohen JF, Korevaar DA, Wang J, Spijker R, Bossuyt PM. Should we Search Chinese Biomedical Databases when Performing Systematic Reviews? Syst Rev 2015;4:23.
105. Grande AJ, Hoffmann T, Glasziou P. Searching for Randomized Controlled Trials and Systematic Reviews on Exercise. A Descriptive Study. Sao Paulo Med J 2015;133(2):109-14.
106. Durao S, Kredo T, Volmink J. Validation of a Search Strategy to Identify Nutrition Trials in PubMed Using the Relative Recall Method. J Clin Epidemiol 2015;68(6):610-6.
107. Scherer RW, Huynh L, Ervin AM, Dickersin K. Using ClinicalTrials.gov to Supplement Information in Ophthalmology Conference Abstracts about Trial Outcomes: a Comparison Study. PLoS One 2015;10(6):e0130619.
108. Nankervis H, Devine A, Williams HC, Ingram JR, Doney E, Delamere F, et al. Validation of the Global Resource of Eczema Trials (GREAT Database). BMC Dermatol 2015;15:4.
109. Fredericks S. Questioning the Efficacy of 'Gold' Open Access to Published Articles. Nurse Res 2015;22(6):8-10.
110. Perneger TV. Online Accesses to Medical Research Articles on Publication Predicted Citations up to 15 Years Later. J Clin Epidemiol 2015;68(12):1440-5.
111. Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with Results Reporting at ClinicalTrials.gov. N Engl J Med 2015;372(11):1031-9.
112. Williams RJ, Tse T, DiPiazza K, Zarin DA. Terminated Trials in the ClinicalTrials.gov Results Database: Evaluation of Availability of Primary Outcome Data and Reasons for Termination. PLoS One 2015;10(5):e0127242.
113. Viergever RF, Li K. Trends in Global Clinical Trial Registration: an Analysis of Numbers of Registered Clinical Trials in Different Parts of the World from 2004 to 2013. BMJ Open 2015;5(9):e008932.
114. Schoenthaler M, Miernik A, Wilhelm K, Schlager D, Schoeb DS, Adams F, et al. Level of Evidence, Sponsorship, Conflict of Interest Policy and Commercial Impact of PubMed-listed Clinical Urolithiasis-related Trials in 2014. BJU Int 2016;117(5):787-92.
115. Vodicka E, Kim K, Devine EB, Gnanasakthy A, Scoggins JF, Patrick DL. Inclusion of Patient-Reported Outcome Measures in Registered Clinical Trials: Evidence from ClinicalTrials.gov (2007-2013). Contemp Clin Trials 2015;43:1-9.
116. DiMasi JA, Hermann JC, Twyman K, Kondru RK, Stergiopoulos S, Getz KA, et al. A Tool for Predicting Regulatory Approval After Phase II Testing of New Oncology Compounds. Clin Pharmacol Ther 2015;98(5):506-13.
117. Vucic K, Jelicic Kadic A, Puljak L. Survey of Cochrane Protocols Found Methods for Data Extraction from Figures not Mentioned or Unclear. J Clin Epidemiol 2015;68(10):1161-4.
118. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: Evaluation of a System for Automatically Assessing Bias in Clinical Trials. J Am Med Inform Assoc 2016;23(1):193-201.
119. Hao T, Weng C. Adaptive Semantic Tag Mining from Heterogeneous Clinical Research Texts. Methods Inf Med 2015;54(2):164-70.
120. Mo Y, Kontonatsios G, Ananiadou S. Supporting Systematic rReviews Using LDA-based Document Representations. Syst Rev 2015;4:172.
121. Polepalli Ramesh B, Sethi RJ, Yu H. Figure-associated Text Summarization and Evaluation. PLoS One 2015;10(2):e0115671.
122. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, et al. Automated Confidence Ranked Classification of Randomized Controlled Trial Articles: an Aid to Evidence-Based Medicine. J Am Med Inform Assoc 2015;22(3):707-17.
123. Liu RL. Passage-Based Bibliographic Coupling: an Inter-Article Similarity Measure for Biomedical Articles. PLoS One 2015;10(10):e0139245.
124. Demner-Fushman D, Mork JG. Extracting Characteristics of the Study Subjects from Full-Text Articles. AMIA Annu Symp Proc 2015;2015:484-91.
125. Shao W, Adams CE, Cohen AM, Davis JM, McDonagh MS, Thakurta S, et al. Aggregator: a Machine Learning Approach to Identifying MEDLINE Articles that Derive from the Same Underlying Clinical Trial. Methods 2015;74:65-70.
126. Neville J, Kopko S, Broadbent S, Aviles E, Stafford R, Solinsky CM, et al. Development of a Unified Clinical Trial Database for Alzheimer's Disease. Alzheimers Dement 2015;11(10):1212-21.
127. Sydes MR, Johnson AL, Meredith SK, Rauchenberger M, South A, Parmar MK. Sharing Data from Clinical Trials: the Rationale for a Controlled Access Approach. Trials 2015;16:104.
128. Vickers AJ. Sharing Raw Data from Clinical Trials: What Progress Since we First Asked "Whose Data Set is it Anyway?". Trials 2016;17(1):227.
129. Blankers M, Smit ES, van der Pol P, de Vries H, Hoving C, van Laar M. The Missing=Smoking Assumption: A Fallacy in Internet-Based Smoking Cessation Trials? Nicotine Tob Res 2016;18(1):25-33.
130. Hugh-Yeun K, Cheung WY. Leveraging the Power of Pooled Data for Cancer Outcomes Research. Chin J Cancer 2016;35(1):74.
131. Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB. Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases. Clin Pharmacol Ther 2016;99(3):325-32.

Correspondence to:
Stephen B. Johnson, PhD
Healthcare Policy and Research
Weill Cornell Medicine
425 East 61st Street, DV 305
New York, NY 10065
USA
Tel: +1 646 962 9403
E-mail: johnsos@med.cornell.edu