Clinical Research Informatics: Supporting the Research Study Lifecycle

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

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.

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

### 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
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 111 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 bibliometrics [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 classified “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

9. De Moor G, Sundgren M, Kaipa D, Schmidt A,


74. Goldec B, Gray J. OpenTrials: Towards a Collaborative Open Database of all Available Information on all Clinical Trials. Trials 2016;17:164.


80. Dufka FL, Munch T, Dworkin RH, Rowbotham


110. Perneger TV. Online Accesses to Medical Research Articles on Publication Predicted Confidence Ranked Classification of Randomized Controlled Trial Articles: an Aid to Evidence-Based Medicine. J Am Med Inform Assoc 2015;22(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: johnso@med.cornell.edu