Clinical Research Informatics: Recent Advances and Future Directions

M. Dugas[1,2]
[1] Institute of Medical Informatics, University of Münster, Germany
[2] European Research Center for information systems (ERCIS)*

Summary

Objectives: To summarize significant developments in Clinical Research Informatics (CRI) over the past two years and discuss future directions.

Methods: Survey of advances, open problems and opportunities in this field based on exploration of current literature.

Results: Recent advances are structured according to three use cases of clinical research: Protocol feasibility, patient identification/recruitment and clinical trial execution.

Discussion: CRI is an evolving, dynamic field of research. Global collaboration, open metadata, content standards with semantics and computable eligibility criteria are key success factors for future developments in CRI.

Keywords

Clinical research informatics, protocol feasibility, patient recruitment, clinical trial execution, open metadata, semantic annotation

Yearb Med Inform 2015;10:174-7
http://dx.doi.org/10.15265/IY-2015-010
Published online August 13, 2015

Introduction

In the IMIA yearbook 2013, Peter Embi [1] reviewed advances in Clinical Research Informatics (CRI). He identified 6 categories of CRI: Data and Knowledge Management, Discovery and Standards; Clinical Data Re-Use for Research; Researcher Support and Resources; Participant Recruitment; Patients/Consumers and CRI; Policy, Regulatory and Fiscal Matters. He concluded that "the field of CRI is broad and rapidly advancing". This survey focuses on Data Management of CRI towards interoperability. It is based on experiences from a large-scale European project in this topic area and addresses the following questions: What are significant developments in CRI over the past two years? What are open problems and opportunities?

Methods

This is a survey article, i.e. not a formal, systematic review. It is rather a subjective selection of important publications from the past two years based on practical experience in this field, in particular from the European project "Electronic Health Records for Clinical Research (EHR4CR)" [2, 3]. EHR4CR is one of the largest public-private partnerships with 33 partners (academic and industrial), aiming at providing adaptable, reusable and scalable solutions for reusing data from EHR systems for Clinical Research. The description of recent advances in CRI will be structured according to three use cases of clinical research: Protocol feasibility, patient identification and recruitment and clinical trial execution. Basically, these three use cases cover the full range of clinical research. Future directions are derived from those papers, again in a non-formalized, simplified and most probably incomplete manner.

* http://www.ercis.org

Recent Advances

Protocol Feasibility

A key process in clinical research is protocol feasibility. The task is to estimate how many patients are available according to a set of feasibility criteria (e.g. diabetes type II patients, aged 18-60, HbA1c >8%) in a defined setting (e.g. hospitals A, B and C) and time frame (e.g. within past 12 months). A clinical study can only be successful, if a patient cohort of adequate size is existing. Patient counts are usually sufficient to answer this question, i.e. aggregated, irreversibly de-identified data.

Various successful projects regarding protocol feasibility were reported in the literature, for example Doods et al. report about a protocol feasibility platform with real EHR data in five countries [4]. In the context of EHR4CR, a generic query language (ECLECTIC: Eligibility Criteria Language for Clinical Trial Investigation and Construction) was developed and implemented [5]. Key challenges for multi-site systems are extraction, transformation and loading (ETL) of data into data warehouses and mappings of local codes to a central terminology, as described by Hussain in a European context [6] and McMurry in a US context [7]. First versions of key data elements for protocol feasibility have been defined in Europe [8].

Patient Identification and Recruitment

Once a clinical study is initiated, eligible patients need to be identified. It is well-known that a large proportion of clinical trials are delayed or not successful due to issues with

patient recruitment. In contrast to protocol feasibility, aggregated patient counts are not sufficient to support patient identification and recruitment. Candidate patient lists need to be generated and communicated to treating physicians [9]. In a second step, local study teams get involved. Recently, generic architectures and system functionalities for patient recruitment systems were defined, for example in a German setting [10, 11]. Data completeness of electronic health records for patient recruitment was described [12]. There are several recent reports that CRI tools can support the recruitment process, both for specific diseases [13, 14] and on a general level [15, 16]. In analogy to protocol feasibility, ETL of clinical data and mapping of local codes to a central terminology are key steps for such systems.

Clinical Trial Execution

Data management in clinical trials is costly due to the high documentation workload – on average 180 pages per patient in a trial [17] – and the need for high data quality. To support clinical trial execution, data can be transferred from EHR systems into electronic data capture (EDC) systems. In the past, the feasibility of this approach has been demonstrated, for example by El Fadly in a French setting [18]. Recently, first reports about cost-benefit [19] and efficiency of this method in specific clinical studies (e.g. regarding non-cardiac surgery [20]) were published. CRI systems – like any other IT system – are associated with a significant setup and maintenance cost, therefore more evaluations regarding economic aspects would be useful. In analogy with protocol feasibility and patient recruitment, ETL of local data and mapping of codes are critical steps. Content standards are being developed to foster EHR-EDC data exchange, for example by the American Heart Association and the American College of Cardiology [21].

Future Directions and Discussion

The sheer amount of concurrent publications in CRI – for example JAMIA and Biomedical Informatics [22] dedicated special issues to this topic – is already indicating the activity and scientific relevance of this field. In the past years, a large-scale deployment of EHR systems took place all over the world in economically developed countries. This has major implications on clinical research, because nowadays many important clinical findings are available exclusively in EHR systems and not on paper any more. However, many current CRI systems have a prototype character and are limited to small-scale settings. Scalability of CRI approaches will therefore be a key topic for the next years.

In the following some – maybe provocative – theses regarding future directions of CRI are presented and discussed.

The Landscape of Medical Documentation: Global Collaboration Is Needed for CRI

Medical documentation is very granular and complex. On a global basis, patients report their symptoms in 200+ languages, there are 20.000+ hospitals and 1.000+ EHR systems. ICD-10 lists 13.000+ diagnoses and for each diagnosis a suitable documentation approach with an appropriate data model should be implemented.

In principle, each data item on a case report form (CRF) could be derived from one medical concept (e.g. patient age). In SNOMED there a 300.000+ non-synonymous concepts available. A typical CRF consists of approximately 40 data items. This corresponds to 1,5E171 possible CRFs, i.e. **there are much more possible CRFs than atoms in the universe (~ 1E80)** (Fig. 1). This explains why **interoperability of clinical information systems will never happen by chance**. Instead, global collaboration is needed to design CRI systems in the future. As medical informatics community we should learn from our colleagues in high-energy physics, who set up a global collaborative effort with thousands of researchers to explore such an abstract topic like the Higgs-Boson. In our domain we have exciting challenges with a global impact such as informatics for personalized medicine.

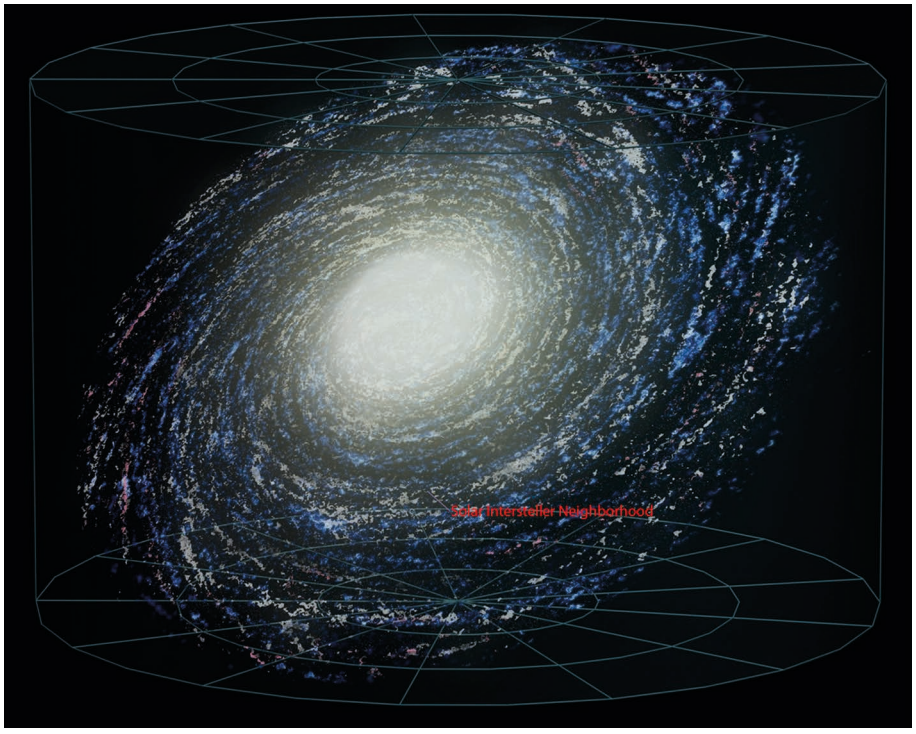

Fig. 1 Design of medical documentation is a large scale problem and requires global collaboration: Based on 300.000 concepts per data item and 40 items per form there are more possible CRFs than atoms in the universe. [picture source: Wikipedia, Andrew Z. Colvin]

Transparency of CRFs Is Required: Open Metadata for CRI

Clinical trials put patients at risk to foster medical research. From an ethical point of view, this can only be justified if trials are designed and conducted in the best possible way. Currently, empty case report forms are mostly business secrets, i.e. they are not available to the scientific community (Fig.2). Even eligibility criteria are not fully transparent [23]. IMIA has demanded transparency for trustworthy reuse of health data: "The cornerstone of data sharing and reuse is trust ... one of the important components of trust is transparency" [24]. From an informatics perspective, the secrecy around CRFs leads to re-inventing the wheel for trial design and documentation in thousands of studies worldwide. Currently, there is a public debate and ongoing legislation about transparency in clinical trials as addressed in the European clinical trials regulation (536/2014) [25]. The European Medicines Agency (EMA) plans to publish clinical study reports for new drug applications, which would be a great step forward towards transparency. As informaticians we should point out that sharing of metadata, in particular medical forms, is a key step for transparency and a prerequisite for a broad discussion about best practice in trial design and documentation. Actually, metadata sharing is mandatory to enable data sharing. **Open metadata** can contribute to interoperability between EHR and EDC systems and therefore **should be the norm** [26, 27]. Public portals for open metadata are already available, for instance https://medical-data-models.org with 4.000+ CRFs (as of August 2015) [28].

Fig. 2 Currently most CRFs are business secrets, which is blocking efficient data exchange between information systems. Open metadata is a key step to foster interoperability.

Shorter and Smarter CRFs Are Needed: Content Standards with Semantics

As of August 2015, Clinicaltrials.gov lists 195.000+ trials. From a medical perspective, very similar information is collected in EHR and EDC systems: Basically, the goal is to provide complete patient documentation. Interestingly, most data elements are collected to demonstrate absence of adverse events. However, excellent documentation is capturing all medically relevant facts and at the same time needs little documentation effort, i.e. much less than 180 pages per patient. Therefore design of efficient documentation is really an art and much more sharing of best documentation practice is needed in the future. There are several initiatives aiming to bridge the interoperability gap between clinical care and research highlighting the need for semantic mappings, such as eMERGE [29], SHRINE [30] and SHARPn [31]. From my perspective, missing semantic annotation in databases is the root cause for data integration problems [32]. Semantically annotated data items and forms facilitate to compare documentation approaches [33, 34], help to avoid redundant data entry by integration of information systems [35] and foster data analysis [36]. Content standards with semantics are evolving, both from regulatory bodies (e.g. SDTM from FDA) and medical scientific societies.

Eligibility Criteria Need to Be Computable in the Future

Computable eligibility criteria (EC), i.e. EC expressed in a "computable query language ... for a clinical trial that can be evaluated from patient data without human intervention" [5] have many benefits. However, currently these EC are presented in free text, in a non-computable manner. Many excellent literature is available regarding natural language processing (NLP) of these criteria [37, 38] and frequent medical concepts in EC were described [39]. Several important shortcomings of existing EC were identified. From my perspective, we don't need better NLP, instead we need computable criteria from the very beginning. Clearly, informatics expertise is missing in many institutional review boards, accepting underspecified criteria like "patient has no major disease" or "patient is eligible according to clinical judgement". These vague criteria hamper analysis both by computers as well as physicians. As informaticians, we have to point out that unclear, non-computable eligibility criteria lead to unethical trials, because it is not known, for what patient cohort the results of the trial are actually valid. Therefore methodological input from the CRI community can help to improve the design of future clinical trials.

Conclusion

CRI is an evolving, dynamic field of research. This survey addressed Data Management of CRI in the context of interoperability. There

is a strong need for global collaboration to address the huge challenges of efficient and effective data capture in clinical research. Open metadata, content standards with semantic annotation and computable eligibility criteria are key success factors for the future of CRI.

References

1. Embi PJ. Clinical research informatics: survey of recent advances and trends in a maturing field. Yearb Med Inform 2013;8(1):178-84.
2. 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 EHR-4CR project. J Biomed Inform 2015;53:162-73.
3. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunityes for clinical research. J Intern Med 2013;274(6):547-60.
4. Doods J, Bache R, McGilchrist M, Daniel C, Dugas M, Fritz F; Work Package 7. Piloting the EHR4CR feasibility platform across Europe. Methods Inf Med 2014;53(4):264-8.
5. Bache R, Taweel A, Miles S, Delaney BC. An Eligibility Criteria Query Language for Heterogeneous Data Warehouses. Methods Inf Med 2015;54(1):41-4.
6. Hussain S, Sun H, Sinaci A, Erturkmen GB, Mead C, Gray AJ, et al. A framework for evaluating and utilizing medical terminology mappings. Stud Health Technol Inform 2014;205:594-8.
7. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One 2013;8(3):e55811.
8. Doods J, Dugas M, Fritz F on behalf of EHR4CR WP7. A European inventory of common EHR data elements for clinical trial feasibility. Trials 2014;15:18.
9. Dugas M, Lange M, Berdel WE, Müller-Tidow C. Workflow to improve patient recruitment for clinical trials within hospital information systems - a case-study. Trials 2008;9:2.
10. Trinczek B, Köpcke F, Leusch T, Majeed RW, Schreiweis B, Wenk J, et al. Design and multicentric Implementation of a generic Software Architecture for Patient Recruitment Systems reusing existing HIS tools and Routine Patient Data. Appl Clinl Inform 2014;5(1):264-83.
11. Schreiweis B, Trinczek B, Köpcke F, Leusch T, Majeed RW, Wenk J, et al. Comparison of Electronic Health Record System Functionalities to support the patient recruitment process in clinical trials. Int J Med Inform 2014; 83:860-8.
12. Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak 2013;13:37.
13. Rahimi A, Liaw ST, Taggart J, Ray P, Yu H. Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in Electronic Health Records. Int J Med Inform 2014;83:768-78.
14. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014;21(5):801-7.
15. Afrin LB, Oates JC, Kamen DL. Improving clinical trial accrual by streamlining the referral process. Int J Med Inform 2015;84(1):15-23.
16. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014; 21(2):221-30.
17. Getz K. Protocol Design Trends and their Effect on Clinical Trial Performance. RAJ Pharma 2008;5: 315-6.
18. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, et al. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform 2011 Dec;44 Suppl 1:S94-102.
19. Bruland P, Forster C, Breil B, Ständer S, Dugas M, Fritz F. Does single-source create an added value? Evaluating the impact of introducing x4T into the clinical routine on workflow modifications, data quality and cost-benefit. Int J Med Inform 2014 Dec;83(12):915-28.
20. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch HU, et al. Secondary use of routinely collected patient data in a clinical trial: An evaluation of the effects on patient recruitment and data acquisition. Int J Med Inform 2013;82(3):185-92.
21. Cannon CP, Brindis RG, Chaitman BR, Cohen DJ, Cross JT Jr, Drozda JP Jr, et al. 2013 ACCF/AHA key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes and coronary artery disease: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Clinical Data Standards. Circulation 2013;127(9):1052-89.
22. Embi PJ, Payne PR. Advancing methodologies in Clinical Research Informatics (CRI): Foundational work for a maturing field. J Biomed Inform 2014;52:1-3.
23. Bhattacharya S, Cantor MN. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. J Biomed Inform 2013;46(5):805-13.
24. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inform 2013;82(1):1-9.
25. REGULATION (EU) No 536/2014 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 April 2014 on clinical trials on medicinal products for human use. Available from: http://ec.europa.eu/health/human-use/clinicaltrials/ regulation/ [accessed 2015 August 03]
26. Dugas M. Why we need a large-scale open metadata initiative in health informatics - a vision paper on open data models for clinical phenotypes. Stud Health Technol Inform 2013;192:899-902.
27. Dugas M, Jöckel KH, Friede T, Gefeller O, Kieser M, Marschollek M, et al. Memorandum "Open Metadata". Open Access to Documentation Forms and Item Catalogs in Healthcare. Methods Inf Med 2015 Jun 25;54(4).
28. Breil B, Kenneweg J, Fritz F, Bruland P, Doods J, Trinczek B, et al. Multilingual Medical Data Models in ODM Format. Appl Clin Inform 2012;3: 276-89.
29. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;20(e1):e147-54.
30. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One 2013;8(3):e55811.
31. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 2013;20(e2):e341-8.
32. Dugas M. Missing semantic annotation in databases - the root cause for data integration and migration problems in information systems. Methods Inf Med 2014; 53(6):516-7.
33. Dugas M, Fritz F, Krumm R, Breil B. Automated UMLS-based Comparison of Medical Forms. PLoS One 2013;8(7): e67883.
34. Krumm R, Semjonow A, Tio J, Duhme H, Bürkle T, Haier J, et al. The Need for Harmonized Structured Documentation and Chances of Secondary Use - Results of a Systematic Analysis with Automated Form Comparison for Prostate and Breast Cancer. J Biomed Inform 2014; 51:86-99.
35. Cimino JJ, Ayres EJ, Remennik L, Rath S, Freedman R, Beri A, et al. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): Design, contents, functionality and experience to date. J Biomed Inform 2014;52:11-27.
36. Zhu Q, Freimuth RR, Lian Z, Bauer S, Pathak J, Tao C, et al. Harmonization and semantic annotation of data dictionaries from the Pharmacogenomics Research Network: a case study. J Biomed Inform 2013;46(2):286-93.
37. Miotto R, Weng C. Unsupervised mining of frequent tags for clinical eligibility text indexing. J Biomed Inform 2013;46(6):1145-51.
38. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform 2014;52:112-20.
39. Varghese J, Dugas M. Most Frequent Medical Concepts in Clinical Trial Eligibility Criteria and their Coverage in MeSH and SNOMED-CT. Methods Inf Med 2015;54(1):83-92.

Correspondence to:
Prof. Dr. Martin Dugas
Institute of Medical Informatics
University of Münster
Albert-Schweitzer-Campus 1 | A11
D-48149 Münster, Germany
Tel: +49 251 83 55262
E-mail: dugas@uni-muenster.de