Clinical Research Informatics: Recent Advances and Future Directions

M. Dugas¹,²
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

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.

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 are 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].

**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

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