

Barriers to the Reuse of Routinely Recorded Clinical Data: A Field Report

Kathrin Dentler^{a,b}, Annette ten Teije^a, Nicolette de Keizer^b, Ronald Cornet^{b,c}

^a Department of Computer Science, VU University Amsterdam, The Netherlands

^b Department of Medical Informatics, Academic Medical Centre, University of Amsterdam, The Netherlands

^c Department of Biomedical Engineering, Linköping University, Sweden

Abstract

Today, clinical data is routinely recorded in vast amounts, but its reuse can be challenging. A secondary use that should ideally be based on previously collected clinical data is the computation of clinical quality indicators. In the present study, we attempted to retrieve all data from our hospital that is required to compute a set of quality indicators in the domain of colorectal cancer surgery. We categorised the barriers that we encountered in the scope of this project according to an existing framework, and provide recommendations on how to prevent or surmount these barriers. Assuming that our case is not unique, these recommendations might be applicable for the design, evaluation and optimisation of Electronic Health Records.

Keywords:

Clinical Data, Reuse of Data, Secondary Use of Data, Data Quality, Pragmatic Interoperability, Electronic Health Record.

Introduction

Today, increasing volumes of clinical data are being routinely recorded and stored in Electronic Health Records (EHRs). The potential benefit from reusing the resulting data sources is enormous, both for individual patients and society in general. In fact, according to a recent report by PricewaterhouseCoopers [1], using data for secondary purposes is one of the most promising ways to improve health outcomes and costs. Such purposes comprise clinical research, the recruitment of eligible patients for clinical trials, the early detection of epidemics, reimbursement, clinical audit, the generation or testing of medical hypotheses and quality monitoring or reporting based on clinical quality indicators. However, reusing clinical data is often challenging in practice.

The Dutch government releases sets of both legally mandatory and voluntary evidence-based quality indicators for various kinds of diseases and interventions. The government requests indicator results for entire reporting years to monitor and compare the quality of care. These indicators typically require data from several sources and are often computed manually, which is error-prone and time-consuming. To enable timely feedback and, where necessary, intervention inside hospitals, the indicators should be computed automatically and in real-time, based on routinely recorded clinical data. For a recent study, we strove to gather all raw source data required to compute a set of indicators for the Gastrointestinal Oncology Centre Amsterdam (GIOCA).

The GIOCA is a specialised outpatient clinic that has been set up to improve the quality of care for patients with (suspected)

cancer of the gastrointestinal tract. Patients who register at the GIOCA are scheduled for an appointment within only seven days at most. During this appointment, examinations to diagnose the patient are carried out, the case is discussed in a multidisciplinary meeting, and a detailed treatment plan is established and communicated to the patient. As this patient-centred rapid diagnosis process reduces the time until treatment starts, the founders of the GIOCA are motivated to measure its performance. We chose the domain of colorectal cancer surgery because it is also the subject of the recently founded Dutch Surgical Colorectal Audit (DSCA) [2]. The DSCA collects all data items necessary to compute the set of indicators. These data items are currently being entered manually by one of our surgeons, but ideally, they would be pre-populated from the underlying information systems, reviewed by the surgeon and then submitted to the DSCA. The GIOCA uses the same information systems as other departments of our hospital, plus additional spreadsheets for internal administration and management.

The goal of this paper is to report on the barriers that we encountered in the attempt to gather all raw source data required to compute the set of indicators. We categorised these barriers and provide recommendations on how they could be prevented or surmounted. A part of these recommendations can support the design of our hospital's new EHR. Supposing that our experiences are similar to data reuse projects in many hospitals, we assume that these recommendations might also help to design, evaluate and optimise systems in other hospitals.

Methods

With the explicit consent and support of the management of the GIOCA, we cooperated with our hospitals' general ICT service in order to retrieve the data required to compute the set of four clinical quality indicators in the domain of colorectal cancer surgery (the same set as employed in [3]) for the reporting years 2010 and 2011. The indicators are contained in the sets released by the governmental program Zichtbare Zorg and the Dutch Healthcare Inspectorate [4], [5]:

II: Number of examined lymph nodes after resection (process indicator)

Numerator: Number of patients who had 10 or more lymph nodes examined after resection of a primary colon carcinoma.

Denominator: Number of patients who had lymph nodes examined after resection of a primary colon carcinoma.

Exclusion criteria: Previous radiotherapy and recurrent colon carcinomas.

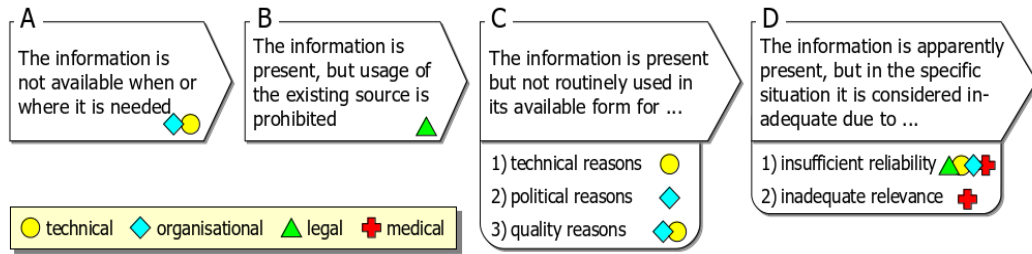


Figure 1-Galster's Framework of reasons why clinical data is not reused.

12: Participation in Dutch Surgical Colorectal Audit (DSCA) (process indicator)

Numerator: Number of surgical resections of a colorectal carcinoma situated in colon or rectum (only count primary carcinomas) for which data has been submitted to the Dutch Surgical Colorectal Audit.

Denominator: Total number of surgical resections of a colorectal carcinoma situated in colon or rectum (only count primary carcinomas).

13: Patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting (process indicator)

Numerator: Number of patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting.

Denominator: Number of patients with rectum carcinoma operated in the reporting year.

Inclusion criterion: Patients who have been operated in the reporting year due to a rectum carcinoma.

Exclusion criteria: Transanal Endoscopic Microsurgery (TEM) resections and recurrent rectum carcinomas. The Dutch Surgical Colorectal Audit states that the presence of a radiologist, a radiotherapist, a surgeon, an oncologist, a colon, stomach and liver physician and a pathologist are required for a preoperative multidisciplinary meeting.

14: Unplanned re-interventions after resection of a primary colorectal carcinoma (outcome indicator)

Numerator: Number of re-interventions during the same admission or during 30 days after the resection (choose longest interval) in the reporting year.

Denominator: Total number of primary resections of a colorectal carcinoma during the reporting year.

Inclusion criteria: Primary colorectal carcinoma = first presentation of a colorectal carcinoma (thus not recurrent); might be the second or next primary presentation.

Exclusion criteria: Transanal Endoscopic Microsurgery (TEM); endoscopic and open polypectomy.

This indicator comes with a list of definitions:

Resection: surgical removal of colon segment where the colorectal carcinoma is situated.

Re-intervention: re-operation in the abdomen or an intervention during which a complication in the abdomen is being treated (inclusive percutaneous incision and drainage, drainage via rectum, embolization of bleedings in the abdomen, et cetera).

Admission: the time the patient spends in a hospital directly after the operation (the same hospital or another one where the patient has been referred), can be longer than 30 days.

Our hospital's general ICT service was currently in the process of investigating the requirements for setting up an operational data store (ODS), which integrates data from several operational databases. As the ICT service was especially interested in investigation of typical data collection requirements from a business intelligence perspective, we started a joint project to gather the required data.

In the absence of a central overview of the data available in our hospital, the goal of the *first phase* of the project was to identify the original sources of the required "raw" data elements, i.e. whether and in which systems these elements are stored, and who are the responsible contact persons. In order to do so, we interviewed the experts who are treating colorectal cancer patients, observed the work- and data flows, and interviewed those responsible for computation of the quality indicators as well as potentially responsible contact persons. In the *second phase* of the project, the team from our hospital's general ICT service worked on the technical design of the ODS and on the actual data retrieval from the various databases. For each of the data elements established in the first phase, they identified its name, type, format and length in the database, and whether it was optional or mandatory. After this phase was completed, we obtained a version of the required data. In the *third phase* of the project, we analysed the data obtained and identified several quality issues that impeded its reuse.

We documented all barriers encountered in the course of this process, and - based on consensus - categorised them according to Galster's framework of causes that impede the reuse of clinical data in clinical settings [6]. Galster's categorisation is based on a literature review and shown in Figure 1. The causes are linked to underlying aspects that have been indicated in the Semantic Health Report [7], i.e. technical, organisational, legal and medical aspects. We also categorised the encountered barriers according to these underlying aspects, as well as the phases of our project.

Our hospital's Institutional Review Board waived the need for informed consent, as individual patients were not directly involved. The use of the data is officially registered according to the Dutch Personal Data Protection Act.

Results

Required data

In the first phase of our project, we identified 12 data elements required to compute the set of 4 quality indicators. After interviewing more than 15 people, including staff members of the GIOCA, those responsible to compute the set of quality indicators, and various database administrators, we identified 9 corresponding source systems as shown in Table 1. All required data elements were stored in a structured digital format, except for relations between diagnoses and procedures, which

are essential to identify procedures that have been carried out for colorectal carcinomas and not for other reasons. These relations are often documented in free-text descriptions such as surgery reports. Most of the identified source systems were stand-alone systems for clinical and administration purposes, with data flows between them. For example, high-detail data from the surgical procedure system flows in less detail into the central procedure register. Two of the sources are external national registers. Please note that several items occurred in several databases, and in principle, one source per item should be sufficient to compute the quality indicators. As we strived to identify the source system with the highest data quality, our initial goal was to retrieve the required data from all identified systems.

The second phase of our project resulted in 5 delivered database tables after 8 months, which are underlined in Table 1. We analysed their quality in the third phase.

Barriers to the reuse of routinely recorded clinical data

In this section, the barriers we encountered are categorised according to Galster's framework as shown in Figure 1, and according to the three phases of our project.

A) Data not available when or where it is needed

Hindered access to data sources (technical and organisational reasons), second phase.

The only way to obtain data from the nationwide histopathology and cytopathology pchive was to request and receive it via email. Furthermore, some of the databases in our hospital are administered by external providers, which do not always guarantee structured and real-time access to our databases.

B) Data present, but usage of the source is prohibited

Patient Numbers (legal reasons), third phase.

The use of data was officially registered according to the Dutch Personal Data Protection Act, under the condition that it was de-identified. Therefore, we received the required data from our hospital's source systems with patient numbers hashed by the ICT service. In a later phase of our project, we could not use these patient numbers to match the patient data with data from our hospital's data warehouse, which uses other hashed patient numbers. Of course, this problem would not have existed if the ICT service and the administrators of the data warehouse had matched the data for us.

C) Data present but not routinely used in its available form

Organisational / Cultural Barriers (organisational reasons), first and second phase.

In our university hospital, we encountered various barriers in the attempt to obtain data for reuse that seemed to be due to insufficient prioritisation and culture of data reuse. First of all, no standard procedure existed to process data requests such as ours. Data can be requested via the ICT service, as in our project, or directly via the database administrators. In the busy environment of the ICT department responsible for critical IT systems, our project did not receive the highest priority, which may have caused some delay. Also, the composition of the team changed several times during the project, hampering smooth communication and progress. Once we identified the relevant source systems and the corresponding responsible persons, there were no clear guidelines and procedures on how to request a database extract; rather, this issue had to be discussed with every database administrator individually.

Another major problem was that no central overview of the various data sources, their governance and content, including employed code systems and data dictionaries, existed, and that the management of our hospital did not envisage such an overview.

Insufficient quality (organisational reasons), third phase.

Analysing the database tables that we received from the ICT service, we encountered the following quality issues.

- *Incompleteness on database level.*

We did not receive any data on radiotherapy. For double-recorded elements (e.g. two sources for multidisciplinary meetings, which are recorded in the EHR, but due to the setup of the GIOCA can also be inferred from the patient's first visit), we received only one of the sources. Additionally, we encountered a problem that can probably not be generalised to other hospitals. The dataset did not include data for a complete reporting year, which would have been essential as quality indicators are computed per reporting year. We received data for 2010 and 2011, but for 2010 information on lymph node examinations and multidisciplinary meetings was missing, while for 2011, information on admission and discharge dates was missing.

- *Incompleteness on data element level.*

Table 1-Required data items and their source systems

Data Element	Source System	Indicator
Surgical procedure: date, type, anatomic location	<u>Surgical procedure system</u> , procedure register	11, 12, 13, 14
Diagnosis: anatomic location, type (primary, recurrent)	<u>Diagnosis register</u>	11, 12, 13, 14
Radiotherapy: date	Radiotherapy system, appointment register, Procedure register	11
Lymph node examination: date, number of examined lymph nodes (in pathology report)	<u>Nationwide histopathology and cytopathology data archive</u>	11
Surgical procedure submitted to Dutch Surgical Colorectal Audit (DSCA)	National register	12
Preoperative multidisciplinary meetings: date	<u>EHR</u> , appointment register	13
Admission: admission and discharge date	<u>Admission register</u>	14

While the surgical procedure data was delivered for both reporting years, it was probably incomplete, with the last surgical procedure of the year 2011 being on the 16th of December. Furthermore, around half of the multidisciplinary meetings had no date recorded, and were therefore unusable.

- *Incorrectness.*
Some data elements were obviously incorrect, such as dates in the far future (e.g. year 2101).
- *Lack of interlinking of data in various sources.*
To reuse data for indicator computation, it is essential to know which procedures have been carried out for which diagnoses, but these relations are not recorded in our hospital in a structured format. The only relation between the different data sources is the (hashed) patient number.
- *Missing provenance of data.*
To reuse data, its provenance might be of interest, especially when data can stem from several sources. However, the data that we received did not contain any provenance information, so that we had to schedule further meetings to obtain a clear overview.
- *Lack of inside-knowledge of "meaning of data".*
Our hospital employs national and local code systems instead of international standard terminologies, and a central metadata registry is absent, making it hard to identify the meaning of the respective data elements. For example, we encountered diagnosis-treatment codes such as 3314, 554 or 11, diagnosis codes such as 13862, 29798, 7155, and specialisations such as KGA, AUD or KEC. Likewise, procedure codes such as 335127, 989899 or 338533Y were not interpretable without knowledge of the coding system.

Problem of selecting patients in one system and querying their data in another system (technical reasons), third phase.

Our hospital's data infrastructure is based on several small source systems instead of one large system, which makes the execution of queries - which are automatically optimised for integrated systems - harder. When querying several systems, one has to identify a suitable starting point to obtain a basic set of relevant patients, and then query other systems based on the identified patient numbers. However, it might not always be clear which one is the most suitable system to start with, and querying separate datasets can lead to a large number of irrelevant results. With regard to colorectal surgery indicators, for example, we would search for all patients who had a colectomy or a resection of rectum due to a colorectal carcinoma. In order to do this, we must query the surgical procedure database for all patients with relevant procedures and the diagnosis database for all patients with a colorectal carcinoma, and then construct the intersection of both query result sets.

D) Data apparently present, but in the specific situation it is considered inadequate

Because the data did not cover a complete reporting year, we did not attempt to compute the set of indicators, and therefore did not analyse the relevance and reliability of the data. We assume that all data was relevant, but that we might have encountered reliability issues, including obviously wrong procedure years such as 2101. Reliability issues are especially visible when data is recorded twice instead of being reused, such as in our hospital and in the DSCA, and double recorded items are inconsistent. We are currently investigating such issues in a subsequent study.

Discussion

Main findings

In our study, we identified a number of barriers that hinder the (timely) reuse of routinely collected clinical data. Even though all data that we required was in principle available in a digital format, and most of it within our hospital, it took a long time until we received a version of the requested data, and the data itself was of insufficient quality. The barriers that we identified cover all four of Galster's categories of why clinical information is not reused. However, category C, "Data present but not routinely used in its available form," contained the most problems, mainly due to underlying organisational/cultural and data quality reasons.

Due to the identified data quality issues, we proceeded in gathering the required data ourselves, with the explicit consent of both our hospital's ICT service and the management of the GIOCA. We started using our freshly launched hospital-internal data warehouse for research, which turned out to satisfy our requirements, with two exceptions: the data items radiotherapy and multidisciplinary meeting. We made contact with another data gathering and analysis initiative in the scope of the GIOCA, which already gathered radiotherapy and multidisciplinary meeting data and willingly shared it with us, so that we finally had a solid basis to compute first quality indicators.

Related Work

Holzer and Gall [8] compiled a catalogue of eight core requirements for secondary use of EHRs. Because the authors follow a document-oriented approach to data reuse (as opposed to our structured database-oriented approach), most of their core requirements, such as "possibility to formulate queries within the retrieved documents" cannot be related directly to our findings. However, other requirements, such as "use of standards and terminologies" fit ours.

Prokosch and Ganslandt [9] identify three challenges in the context of reusing EHRs for clinical research: to establish comprehensive clinical data warehouses that can be harvested with data mining methods, to establish an IT infrastructure that supports clinical research and to integrate and link medical record systems and clinical trial databases. We argue that data warehouses are advantageous but not imperative if source systems can be accessed directly, and also support the claim that a hospital's IT infrastructure should support clinical research. Their last challenge - to link medical record systems and clinical trial databases - falls outside the scope of our study.

Ancker et al. [10] observed that secondary use of data might require a higher degree of data integrity than the original primary use. The Semantic Health Report [7] claims that to fully realise the potential of EHR systems, the data they contain should be of high quality, and that timely and secure access to those entitled has to be ensured. To reuse clinical data, systems must be able to exchange data, preserving its meaning. These requirements are also reflected in our recommendations that we compiled based on the barriers that we encountered.

Strengths and limitations

It might be regarded as a limitation of this study that it included only one hospital. However, we assume that our project can be seen as analogous to data reuse projects in many hospitals, which are likely to encounter similar problems and barriers, and therefore might profit from our recommendations.

It is also questionable to what extent the computation of quality indicators is typical for general data reuse. We argue that

the general challenge that characterised our project was to retrieve high-quality data from several sources within and outside our hospital, and we assume that this challenge underlies most data reuse projects. Real-time access to clinical data and integration of feedback with EHRs is a further challenge, which would not only be desirable for the computation of quality indicators, but also indispensable for secondary uses such as clinical decision support.

Recommendations

Ensure availability of data and accessibility of data sources

When choosing external providers for EHR systems, data accessibility and reuse should be considered in order to avoid “data silos”, in which it is easy to insert data, but hard to extract it. Likewise, only copies of high-quality local data should be submitted to external registers, ensuring the hospital’s ownership of the data.

Ensure patients’ interests, privacy and security while allowing for re-use

Even though this is not a direct finding of our study, it should be noted that the patients’ rights, privacy and security must be protected. Patient data should always be de-identified, unless the patients’ identity is absolutely necessary and of high value, such as in the recruitment of eligible patients to clinical trials, which might require informed consent.

Set-up a reuse-friendly organisation and culture

Especially in university hospitals, data reuse should be prioritised, and this prioritisation should be part of the hospital culture. Standard procedures should be set up to request data for reuse and financial resources should be made available to extract data for research that might benefit both the hospital and its patients. In order to facilitate reuse, a central overview of available data sources should be administered, including their governance and content as well as employed code systems and data dictionaries. Such a metadata registry could for example be based on ISO/IEC 11179, an international standard for representing metadata.

Increase data quality

Data quality comprises completeness and correctness, but also the recording of relations between diagnoses and procedures, and the use of standard terminologies and information models that enable meaning-based retrieval and facilitate the “Collect once, use many” paradigm [11]. In order to increase the quality of elements that are required for reuse, those responsible for recording the respective elements should be made aware of foreseeable secondary uses. Data quality also comprises metadata and provenance. In the scope of our project, it would have been helpful to know which systems the data stemmed from, as well as who recorded the data, and when and why it was recorded.

Allow for cross-database querying

While one monolithic overarching hospital-internal EHR might be desirable, in practice, the IT infrastructures of many hospitals consist of several dedicated source systems. In principle, this should not be a problem as long as data in all systems can be accessed and seamlessly integrated. However, querying several systems is harder and in our case required manual work. The ability to execute hospital-wide federated queries would alleviate this barrier.

Conclusion

In this paper, we categorised barriers encountered in the attempt to reuse data from our hospital for clinical indicator computation and provide recommendations that might support the design, evaluation and optimisation of EHRs. Patient data can be considered one of the most valuable resources that a hospital has at its disposal, and therefore its reuse should be facilitated while preserving the patient’s privacy, security and interest.

References

- [1] PricewaterhouseCoopers. Transforming healthcare through secondary use of health data. 2009.
- [2] Dutch Surgical Colorectal Audit: <http://www.clinicalaudit.nl/>
- [3] Kathrin Dentler, Annette ten Teije, Ronald Cornet, and Nicolette de Keizer. Towards the Automated Calculation of Clinical Quality Indicators. Knowledge Representation for Health-Care, LNCS 6924:51–64, 2012.
- [4] Colorectal Indicator set 2011 of Zichtbare Zorg: <http://www.zichtbarezorg.nl/mailings/FILES/htmlcontent/Ziekenhuizen/2011/Indicatoren/Verplicht/Indicatorenengids%20Colorectaal%20Carcinoom%202011%20def%20nieuw.pdf>
- [5] Indicator set 2011 of the Dutch Healthcare Inspectorate: http://www.igz.nl/Images/2010-07%20Basisset%20kwaliteitsindicatoren%20ziekenhuizen%202011_tcm294-283436.pdf
- [6] Galster G. Why is clinical information not reused? In Stud Health Technol Inform. 2012;180:624-8.
- [7] V.N. Stroetmann, D. Kalra, P. Lewalle, A. Rector, J.M. Rodrigues, K.A. Stroetmann, G. Surjan, B. Ustun, M. Virtanen, and P.E. Zanstra. Semantic Interoperability for Better Health and Safer Health Care. European Commission, (January), 2009.
- [8] K. Holzer and W. Gall. Utilizing IHE-based Electronic Health Record systems for secondary use. Method Inform Med, 50(4):319–25, January 2011.
- [9] Hans-Ulrich Prokosch and T. Ganslandt. Perspectives for Medical Informatics. Method Inform Med, pages 38–44, 2009.
- [10] J.S. Ancker, Sarah Shih, M.P. Singh, and A. Snyder. Root Causes Underlying Challenges to Secondary Use of Data. AMIA Annual, pages 57–62, 2011.
- [11] James J. Cimino. Collect Once, Use Many: Enabling the Reuse of Clinical Data through Controlled Terminologies. Journal of AHIMA, 78(2):24–29, 2007.

Address for correspondence

Kathrin Dentler, AI Department, Division of Mathematics & Computer Science, Faculty of Sciences, VU University Amsterdam, Room U 3.03, de Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. k.dentler@vu.nl