Health data for research: a nationwide privacy-proof system in Belgium

Nicolas Delvaux
Bert Aertgeerts
Johan van Bussel
Geert Goderis
Bert Vaes
Mieke Vermandere

Correspondence: nicolas.delvaux@kuleuven.be
Abstract

Background
Health data collected during routine care has important potential for reuse for other purposes, especially as part of a learning health system to advance quality of care. Many sources of bias have been identified through the lifecycle of health data that could compromise the scientific integrity of these data. New data protection legislation requires research facilities to improve safety measures to ensure privacy.

Objectives
In this paper we address the question on how health data can be transferred from various sources and using multiple systems to a centralized platform, called Healthdata.be, while ensuring accuracy, validity, safety and privacy. We also demonstrate how these processes can be used in various research designs relevant for learning health systems.

Methods
The Healthdata.be platform urges uniformity of data registration at the primary source through the use of detailed clinical models. Data retrieval and transfer is organized through end-to-end encrypted eHealth channels and data is encoded using token keys. Patient identifiers are pseudonymised so that health data from the same patient collected across various sources can still be linked without compromising the deidentification.

Results
The Healthdata.be platform currently collects data for more than 150 clinical registries in Belgium. We demonstrate how the data collection for the Belgian primary care morbidity register INTEGO is organized and how the Healthdata.be platform can be used for a cluster randomized trial.

Conclusions
Collecting health data in various sources and linking this data to a single patient is a promising feature that can potentially address important concerns on validity and quality of health data. Safe methods of data transfer without compromising privacy are capable of transporting these data from the primary data provider or clinician to the research facility. More research is required to demonstrate that these methods improve the quality of data collection allowing researchers to rely on electronic health records as a valid source for scientific data.
Introduction

More than a decade ago, the Institute of Medicine (IOM) introduced the “learning health system” (LHS) in response to the challenges on how to generate and apply the best evidence to guide health care choices [1]. An important aim of an LHS is to enable the use of routinely collected health data for knowledge generation, not only to ensure innovation in health care, but also for quality, safety and value. In the LHS cycle, analysis of routine health data creates new insights, which are then introduced back to health care providers through quality improvement tools such as clinical decision support, feedback or audit reports. Even though the gold standard for measuring effectiveness in health care has always been the randomized clinical trial (RCT), increasing attention is being given to registries and health data to contribute to evidence-based practice [2]. The use of health data recorded in the electronic health record (EHR) for research could help bridge the gap between evidence generated in controlled experiments and its application in daily clinical practice [3].

Even though the type of data recorded for research and the data stored in EHRs is similar, the use of health data poses some important problems. Concerns on data quality and validity, completeness of data capture and lack of interoperability have been identified as important barriers to the use of EHRs for clinical research [4]. Experiences from European and American efforts in the use or reuse of health data from EHRs have identified several important challenges [5,6]. Sources of bias included health care system bias, variations in EHR system functionalities and lay-out, and data extraction tools. These concerns have prompted opinions that the reuse of data for purposes other than that for which they were originally collected may be inappropriate [7]. Moreover, the European General Data Protection Regulation (GDPR) has further restricted reuse of health data for research purposes [8,9]. Recent recommendations and guidance have proposed solutions to these concerns but many challenges remain unanswered [10,11].

Perhaps one of the most important challenges to the use of health data in clinical research is the persistent divide between clinicians (data providers) and researchers (data scientists) [12]. The incapacity to communicate between engineers and researchers on the one hand and clinicians on the other hand has resulted in a disconnect between the world of research in medical informatics and the true problems in health care. In the evaluation and management of the health data for specific research questions, it is important that researchers understand the accuracy, comprehensiveness, retrievability and specificity of health data recorded during routine care. On the other hand, clinicians need to be aware of the potential of the health data they record or manage and the implications that inconsistent or missing recordings may have on reusability. Important facilitators to bridging this divide is enabling semantic interoperability, creating an environment of safe and reliable data transfer, ensuring privacy and security, and incentivizing valid and complete data capture [10]. Additionally, experiences in the use of diverse sources of health data have led to a better understanding of data provenance (understanding of the authoritative source of a given data element of interest). For instance, if the measure of interest is whether a patient took a certain drug, then the best source for this outcome may not be the clinician’s order entry data but instead the nursing medication administration record. In
In this sense, data collection tools that allow the aggregation of health data across sources are important enablers of a LHS.

In Belgium, more than 150 clinical registries actively collect health data from multiple sources, such as primary care facilities, laboratories, hospitals, and radiology centers. Moreover there are multiple information systems or EHRs available for each of these sources. For example, for primary care practices alone there are at least eight different EHRs available. In 2012 the Scientific Institute of Public Health was charged with centralizing and improving these clinical registries as part of the national eHealth action plan in a new platform named Healthdata.be. The challenge for this task was to develop a system that allows the integration of data from diverse sources and collects them through multiple systems by clinicians during routine care, while ensuring accuracy, validity, safety, and privacy of the data. In this paper we address the following questions:

1. How can health data be transferred from various original sources of entry to a centralized platform for reuse and what efforts can be done to limit sources of bias?
2. How can health data within an LHS be used for various research designs?

In this paper we will describe the elements of the Healthdata.be project designed for data extraction, data transfer and data processing. Subsequently we will demonstrate how Healthdata.be was used in the INTEGO primary care morbidity registry [13] and in a cluster randomized trial in primary care [14].

Materials and methods

Data structure and semantic interoperability

Health data are at the core of both EHRs and clinical research registries. However, to collect these data in a meaningful manner, these data must have the same structure, use interoperable terminologies and be documented using a detailed clinical model (DCM) [15,16]. DCMs provide detailed specifications of medical concepts in a given context and specify precisely the terminology to be used in terms of technical standards, reference models and platforms [17,18]. They define all structured elements and attributes of a concept, including their relationships to the root concept, their data types and the code lists which can be used. Where possible, code lists include internationally accepted coding such as LOINC, SNOMED CT, ICD-10, etc. Figure 1 illustrates the DCM for the concept blood pressure, including all the associated data elements and their code lists as designed by the Netherlands Federation of University Medical Centres. Content standardization of scientific data collections, using a DCM, contributes to enhanced data quality and correct interpretation of data for research [19]. In addition, a substantial part of the information collected in the context of clinical registries, are continuity of care concepts (e.g. diagnoses, medication, laboratory test results...). This means that there is a certain overlap in content between different registries. Aligning all registries with DCMs therefore enables harmonization across projects, allowing for a maximal reuse of existing data. As they are independent from technical aspects such as message format, system and network, the DCMs can be considered as technically neutral. Therefore, they can be used as a semantic layer between communication standards on the side of the primary source, and the parameters
defined within different registries on the other side. Given the technological complexity and the variety of legacy systems, a stepwise approach for registry standardization is proposed. In a first phase, existing registries are simply mapped to the DCMs. In order to maximize the degree of semantic interoperability, the registry structure can then be adjusted to comply with the logic defined in the DCMs. Registry variables that do not align with the DCM logic can still be completed using the manual interface. In a last phase, registries will be fully DCM-based, allowing the data provider to complete all registrations at the point of care.

Systems used by data providers are being urged to comply with these DCMs and these elements are being included in local certification standards. When data providers or researchers are confronted with a concept for which no DCM exists, they can apply for the development of one to enable automated provisioning of registries.

**Data extraction**

When shaping the principles of LHSs, the IOM reiterated the need to reflect on the burden data collection can be on health care professionals and the importance of limiting this burden to the issues most important to patient care and knowledge generation [20]. Healthdata.be has developed an electronic data capture (EDC) system called HD4DP (Healthdata for data providers), which enables data extraction at the primary source of the clinician. The logic defined in the DCMs is used in an application programming interface (API) that serves as an interface between data providers’ primary source systems and the EDC. Supported input messages, based on health exchange standards, are mapped to the DCMs, upon which structured and coded information from primary source systems can be automatically prefilled in the EDC. This reflects the principle that health data should be recorded only once, and is expected to significantly reduce administrative burden for health professionals. When data are entered through this data collection form (which can be web-based but also installed locally), a comma separated value (csv) file is generated which can then be transferred on a patient-by-patient basis. Data can also be transferred in batch if the data provider’s system can generate this data set with the variables automatically prefilled. This method for data capture is illustrated in figure 2 in the box titled HD4DP.

**Data transfer, safety and privacy**

Transfer of sensitive health data is challenging with regards to technicity, safety and privacy. In Belgium, data transfer between healthcare professionals is organized through existing eHealth channels by an end-to-end (E2E) encryption [21]. For this data transfer, patient and provider identification is ensured through their unique social security identification number (SSIN) and data are encrypted with an algorithm that uses the SSIN of the sender and the receiver. This implies that only the sender or receiver can decrypt the message with the use of their electronic identification card (eID) and pin code. The message is sent through an application called the eHealthBox [22], which can only be accessed using the same eID and pin code. Hence, this eHealthBox functions as a mailbox for encrypted health data. Institutions such as Healthdata.be
also have an eHealthBox, which can be only accessed by individuals with the necessary security clearances. Transfer of health data from primary data providers to the Healthdata.be platform uses these channels of encrypted data transfer.

An important feature in enabling the linkage of health data extracted from different systems or settings is the ability to identify data from the same patient. When health data is being sent from one healthcare provider to another in the context of clinical care, the content of the message is encrypted but the identity of the patient remains known. However, for research purposes, the content of the message is encrypted but the identity of the patient must be blinded. This poses an important challenge when health data for a single patient is collected across sources. To enable this linkage without unblinding the coded data, an extra step is introduced in the data transfer. Where masking of identifiers is required, the eHealth platform acts as a trusted third party (TTP) and uses an algorithm to pseudonymize this data element [23]. The algorithm used will always code the same patient identifier in the same way, ensuring that data from multiple sources from a single person will always be linkable but still coded. A critical moment in this chain of data transfer is the point where the data are received by the eHealth service and decrypted, before the identifier is coded. To ensure that a breach in this chain does not result in the loss of sensitive data, the whole message, with the exception of the patient identifier, is encrypted a second time using a token key encryption (ETK) known only by Healthdata.be [21]. This double encryption (ETK encryption and E2E encryption) ensures safe and blinded data whilst allowing for linkage of health data from multiple sources from a single person. All these processes, including the encryption and coding happen automated and require no human input. Figure 3 illustrates each of the encryption, coding, transfer and decryption steps.

Healthdata.be can process the data collection for very diverse specialties or research facilities in healthcare. To ensure that the requested data is in accordance with the research question or aim of the project, a thorough screening of the project is organized. Each project submits a research protocol including a list of specific data variables being collected. An internal steering committee, an ethics committee associated with a research center and the National Privacy Commission’s Sector Committee for eHealth review this submission. Only when all authorities have approved the project, can the data collection commence.

Data analysis
The Healthdata.be platform not only enables safe data transfer but also provides a secure environment for data handling and data analysis for research purposes. Coded data is received by the HD4RES (Healthdata for research) service, which shows the data as it was sent by the data provider. The interface of HD4RES is almost identical to that of HD4DP, with the exception that identification details are coded. Upon arrival in HD4RES, the data is not yet stored in the datawarehouse (DWH) of Healthdata.be. The DWH has three separate entities: the validation environment, the analysis environment and the reporting environment, and uses SAS Enterprise Guide ® (SAS Institute Inc., North Caroline, USA) to visualize and process the data. It is first stored in a validation table where data quality controlled. Healthdata.be allows for semi-automated processes so that the validation of a continuous data capture can be operationalized.
Once validated, data is then promoted to the analysis environment of Healthdata.be. Access to HD4RES and the separate environments of the DWH is secured through a two factor authentication and can be restricted depending on the needs of the researcher. Data processing and reporting can also be operationalized in order to accommodate a continuous data flow in ongoing registers.

Feedback
A pitfall to accepting data from various sources is the possibility of missing or erroneous data. Erroneous data can be prevented by introducing restricted possibilities, ranges or syntaxes for the data that is transferred through HD4DP. For example, validation rules that detect out-of-range data, missing data or alphanumeric results for a numeric value can already prevent the transfer of these errors at the site of the data provider. However, it may still be possible that an aberrant value is transferred to HD4RES that needs correction. To allow for this correction by the data provider, a feedback loop has been designed. This feedback loop uses the same channels and encryption methods for data transfer and includes a decoding of the SSIN by the TTP of eHealth so that the primary data provider can identify the person for whom a corrected data variable is requested. This feedback loop is illustrated in figure 3.

Results
Case for Intego
INTEGO is a primary care morbidity registry, which was founded over 20 years ago. [13] The INTEGO registry has collected data from more than 100 general practitioners (GPs) and 400,000 patients since its start. Participating GPs all used Medidoc® (Corilus, Aalter, Belgium) for documenting their clinical practice, and were skilled in structured registration. Eligibility for participation in the INTEGO network required GPs to record on average at least one new diagnosis per patient per year, less than 5% non-coded diagnoses, and these two previous requirements should remain stable for at least three years. In 2017, the IT support for Medidoc® was terminated and all its users were urged by the vendor to migrate to a new EHR software, called CareConnect® (Corilus, Aalter, Belgium), a cloud based EHR. This transition marked the moment to redesign the data collection for this registry, which had not changed since its start. To comply with the new GDPR legislation, the Healthdata.be platform was identified as partner for this task.
INTEGO does not require any data collection besides that being collected for daily clinical practice. However, the quality of the collected data is expected to be of high quality. The data transferred to INTEGO not only includes basic concepts such as diagnoses or problems, procedures, prescriptions, laboratory tests, parameters or vital signs and personal information. It also includes intricate attributes such as longitudinal care for the same problem (problem-oriented medical registration), causal relationships between diagnoses and prescriptions or the evolution of a health issue from a symptom into a diagnosis over time. Although many aspects of this registry were already described in existing DCMs, many of these attributes required additional coding and mapping to maintain their meaningfulness.
The INTEGO procedures were approved by the KU Leuven ethics committee (nr. ML1723) and by the National Privacy Commission’s Sector Committee for eHealth (decision nr. 13.026 of March
19th, 2013). The procedures to collect data by Healthdata.be were approved by the Belgian Privacy Commission on April 17th, 2018. To date almost all GPs have migrated to CareConnect® and the first data export is being prepared and tested. On the one hand a ‘core INTEGO’ will be constructed, based on the original eligibility criteria to participate in the INTEGO network, to perform epidemiological research. On the other hand an ‘extended INTEGO’ will be constructed, without eligibility criteria to participate, to perform research on the quality of registration, quality of care and impact of audit and feedback.

Case for ELMO trial
The ELMO trial is a practical cluster randomized trial investigating the effects of decision support on the quantity and quality of laboratory test ordering behavior by GPs [14]. Data is collected from three separate sources on in total around 11 500 patients. Data on laboratory tests is provided by three private laboratories, each with a separate internal coding system for laboratory tests. The Belgian Terminology Center has introduced a national subset of LOINC codes which are increasingly in use [24]. For the tests being investigated, mappings to LOINC codes were realized before the start of the trial to ensure interoperability. Data collection includes laboratory tests (type of test, value for test, units), indications for laboratory test ordering, total cost of the test, identification of the physician and identification of the patient. We also collected patient specific data directly from the GPs. GP investigators used several different EHR software for registration of clinical practice. To ensure uniformity in data collection, we designed a clinical report form (CRF) detailing the exact information we wished to extract from the EHR and which data would need to be added manually. To facilitate data extraction, we designed the CRF so that more than 60% of data would be automatically extracted from the EHR, meaning that these data already complied with one of more existing CBBs as defined by healthdata.be and in use by most EHRs. The CRF was programmed and distributed to all EHRs through an application named eForms. eForms are a tool that extract the requested data directly from the EHR and populate the CRF with this data. Only data requests that could not be mapped to a DCM needed to be added manually. Examples of the requested patient specific data were diagnoses or problems (including ICPC codes, date of diagnosis), procedures performed or ordered, referrals to specialist care, pre- and posttest probabilities of disease, diagnostic error. These data were then sent to the Healthdata.be platform through the described eHealth channels.
Finally, for a subset of patients, data was obtained directly from the patients. A similar CRF was designed which surveyed patients on data similar to the data requested from the investigating GPs. Additional information on socioeconomic status was requested. To ensure uniformity and to avoid technical issues, the CRF was not sent directly to participating patients but a telephone interview was conducted by a research assistant who completed the CRF based on the patient’s responses. All this data was collected in separate SAS data sets on the Healthdata.be platform which was accessible through a secured server. Access to various parts of the data sets was dependent on the role of the investigator, where the data managers had access to the staging data sets, and the statisticians had access to the analytics data sets. The chief investigator had access to all data sets and managed the authorizations of the entire team.
Discussion

Healthdata.be has successfully connected a myriad of data providers on a centralized platform through a secure and private method of interoperable data transfer across settings and systems. This was done by enabling interoperable data collection, encrypted data transfer and coded data collection whilst still allowing to connect data from the same patient collected from multiple sources through a system of pseudonymisation. Healthdata.be has largely been able to bridge the disconnect between clinicians and researchers. Alongside the centralization of more than 100 clinical registries, the platform can also be used in clinical trials or studies using routinely collected data at the point of care. Additional data, which is not defined through a DCM and specific to the trial or study, can be added manually to the data collection tool. These features make Healthdata.be an important facilitator of an LHS and help drive quality improvement in health care.

Facilitating access to reliable health data may be crucial to LHSs, but several situations have illustrated that there may be boundaries to this easier access. The Danish General Practice Database [25] was long considered an outstanding example of a clinical registry, but was suspended due to concerns on privacy and security. Similarly, a large data sharing database linking GP records with hospital data by the British National Health Service (NHS) was terminated due to lack of public confidence [26]. It does not appear to be a coincidence that public concerns on privacy coincide with patients’ increasing access to their own medical data [27]. In response to these concerns, Healthdata.be requires that all clinical registries or trials that wish to use its platform, receive approval from the National Privacy Commission's Sector Committee for eHealth. This authority scrutinizes each application on security and privacy and determines whether the data collection is appropriate for the purposes of the project. Projects that have not obtained approval are not operationalized within Healthdata.be.

Limitations

An important limitation to the Healthdata.be platform is rooted in the decentralized data provision. Despite the efforts to standardize data collection using DCMs, variability in data collection at the point of care is inevitable. Even though DCMs may clearly define a clinical concept, there may still be variations in its use in documenting daily clinical practice. To be fully interoperable, DCMs must also be integrated in a conceptual model such as problem oriented medical registration (PROM). These conceptual models not only include interoperable standards for individual concepts, but also the relations between concepts. Moreover, even when a concept is well defined and documented within the same conceptual model, interrater differences still remain. This is a feature that is common to the way the narrative of a patient is translated into an EHR documentation. Recording guidelines are required and training on how to put these into practice is imperative, but harmonizing system designs and user interfaces may prove to be crucial. Many of the robust registries, such as the British General Practice Research Database [28] or the Dutch Sentinel General Practice Network [29] import their data from a minimal number of different systems to ensure validity. Many of the registries using the Healthdata.be service are also based on data from a single system, but it is also possible to use authentic sources for several data elements to ensure that the most accurate data is available.
For instance, administrative data can be fetched from social security services in real time ensuring that data such as date of birth are correct. Similarly, like in the ELMO study, it is possible to access multiple systems and collect data where results are most likely to be accurate. In this sense, it seems logical to collect data on laboratory tests from the laboratory information system rather than from the GP EHR. However, no comparisons of the validity between registries collecting their data from a single source versus multiple sources have been done.

Conclusion

The reuse of health data collected as part of routine clinical care can further research and improve health care. By ensuring semantic interoperability, safe data transfer and trustworthy data handling, important sources of bias can be avoided. Concerns on data quality and validity can be addressed by collecting data from those sources where the data capture is bound to be most complete and linking these data from multiple sources through pseudonymisation. Further research is required to assess whether these methods truly address concerns on data quality. To date, patients have only limited access and cannot add or change health data in their own patient record. When these features become more widespread, it would be interesting to evaluate how this may influence data quality and validity.

References


22. eHealthBox v.2 Consultation Web Service Cookbook Version 2.4 [Internet]. Brussels: eHealth platform; 2018 Apr. Available from: https://www.ehealth.fgov.be/ehealthplatform/nl/data/file/view/6491265b50d20ba0edea26a2ece4912534f43ec4?name=eHBox%20v2%20Consultation%20WS%20-%20Cookbook%20v2%20v2.4%20dd%2023042018%20.pdf


26. Limb M. Controversial database of medical records is scrapped over security concerns. BMJ 2016 Jul 7;354:i3804. PMID:27387682


Figures


Figure 2: Data capture, data transfer, data encryption, and data reception through the Healthdata.be platform. HD4DP: healthdata for data providers; HD4RES: healthdata for research. Available from https://healthdata.wiv-isp.be/en/services.

Figure 3: Illustration of the data encryption, coding and decryption steps. Step 1: data encryption (with exception of the identifiers); Step 2: encryption of whole message for data transfer through eHealth; Step 3: data transfer to TTP; Step 4: pseudonymization by TTP; Step 5: data encryption for transfer through eHealth; Step 6: data transfer to Healthdata.be; Step 7: decryption of health data: SSIN: social security identification number; HD4DP: Healthdata for data providers; HD4RES: Healthdata for researchers; CSV: comma separated value; ETK: eHealth token key; E2E: end-to-end. Available from https://healthdata.wiv-isp.be/en/services.

Figure 4: Illustration of the feedback loop in case of missing or erroneous data using the eHealth services. HD4DP: Healthdata for data providers; CSV: comma seperated value; HD4RES: Healthdata for researchers. Available from https://healthdata.wiv-isp.be/en/services.