Decision Support for Data Segmentation (DS2): Application to Pull Architectures for HIE

Carl A. Gunter (Univ. of Illinois), Mike Berry (HLN), and Martin French (Concordia Univ.)

Architecture and protocol changes in the sharing of health records lead to a need for decision support functions that enforce privacy preferences of patients. These needs can be addressed with a collection of techniques called Decision Support for Data Segmentation (DS2).

Segmentation of patient medical records was once a side effect of paper media stored in separate locations. For instance, a patient might have a paper record at their primary provider and another record at a substance abuse treatment facility. There were two general types of sharing. In first case the paper record was in a place where it could be accessed by physicians at a provider facility, such as a hospital or clinic. This enabled sharing among the physicians at the provider. In a second case, when there was a need to share the records between two distinct providers (or other parties, such as the patient's insurer), a representative of the originating provider would select relevant parts of the patient paper record and FAX them to the consuming provider. This act was usually done with patient consent and a knowledge of the goal for sharing the record, such as referral to a specialist or new primary care provider. For instance, a patient could decide if the consuming provider should receive substance abuse records from an originating provider.

In the last decade the medical community has begun to shift to the use of computerized rather than paper records. Systems designed primarily to digitize the paper record are known as Electronic Medical Records (EMRs) and typically offer the same kinds of sharing models as with paper records – via printed paper and FAX. There has been steady progress on developing ways to convey medical records using standardized XML formats and codes for medical diagnoses, treatments, drugs, etc. These advancements are embodied in the concept of an Electronic Health Record (EHR). The EHR distinguishes itself from the EMR by incorporating features beyond mere digitization of the paper record, principal among them interoperability between providers even with different EHR vendors. In particular, this enables a digital analog of the FAX in which an email-like message can be sent from one provider to another. This can be described as a *push* communication technique.

Digital push techniques share many of the privacy advantages of the old FAX techniques. However, along with the FAX, they suffer from significant limitations. In particular, they assume that there is some agent that is contacted to make the push. This process adds bureaucracy and makes it too slow for some applications like treating an emergency. Also, while it is ideally the case that the originating provider knows what the consuming provider wants, this is not always so. Finally, an originating provider may know, or be well-equipped to support, the privacy policy of a patient.

Many of these and other such issues can be addressed by the use of a *pull* communication technique: that is, the consuming provider directs a specific query to the originating provider and receives relevant portions of EHR records in response to this request. This allows the consuming provider to request the necessary records as the patient's relationship with the consuming provider develops. Since there is no need for a patient request to the originating provider, the records can be obtained quickly (basically instantly at or before the point of care). With proper support the request can be made to some or all of the providers that hold records for a given patient, thereby providing access to a real-time on-demand longitudinal record.

Both push and pull approaches have infrastructure requirements. The push case is easier since it can be met with a variant on secure electronic mail that addresses the public key infrastructure problem for providers so messages can be signed and encrypted. The pull case is harder because it offers more capabilities and this requires more infrastructure. Fortunately there is an emerging set of standards to support the pull approach. A case study in their application can be seen in Health Information Exchange (HIE) systems being implemented in many of the fifty states of the United States. A common architecture for these HIEs establishes an index of patient records and a broker that uses this index to pull records by collecting them from a central repository or obtaining them on-demand from a distributed collection of originating providers, and consolidating them for the consuming provider. Another impetus for this direction was a report from the President's Council of Advisor on Science and Technology (PCAST), which recommended the development of a nationwide pull system based on a broker architecture of this kind.

The advantage of the pull approach is the many benefits it will offer for the treatment of patients by increasing the knowledge that will be readily available to their healthcare

providers. However, there is a threat that this ready flow of information from one context to another will be open to abuse and break down some of the protections that were a side effect of the inconvenience of sharing paper records. In particular, without some regulation, records from a substance abuse treatment facility might be included with other records and collected by a provider that does not need or want them. Patients view this as a threat to their privacy and may feel the need to opt out of the benefits of the system.

This threat has been met with the development of additional standards for representing patient consents and for adding meta-data tags to records to support patient consents. In particular, it is possible to label certain types of structured information as being extra-sensitive so that a patient can authorize sharing it only with limited parties. This works in conjunction with standardized formats for the patient consents themselves, so, for instance, an HIE can consult a patient consent document to determine whether information of a certain type should be shared with a given consuming provider. For instance, if an originating provider has some substance abuse treatment data and some other data then the former can be tagged as sensitive and the patient can consent to share the non-sensitive part of the record.

This strategy provides traction on the problem only if there is a way to tell how to place the meta-data tags. Ideally this can be done as close to the origin of the record as possible, such as by one of the patient's physicians. This process has many limitations, not the least of which is the time and expertise it will require from physicians. The process can be beneficially aided by automation in which certain codes are identified by standards bodies as likely to be viewed as sensitive. For instance, medical ontology systems can be used to identify substance abuse codes so they can be automatically marked as sensitive.

The Decision Support for Data Segmentation (DS2) project focuses on a strategy for addressing a significant limitation of this sensitive code list approach. The problem is that the ability to infer sensitive conditions is influenced by context. For instance, an HIV diagnosis may be marked as sensitive, but if there are certain other associated types of diagnoses and medications present then HIV can be inferred from these. The problem cannot be readily solved by just making a large enough list of sensitive codes since it is the *combination* of codes that is the problem and it is infeasible to make a list of all possible sensitive combinations.

The concept of DS2 is to generalize from the idea of a sensitive code list to the idea that the determination of sensitivity is a form of medical decision support. That is, the approach is to formulate algorithms for determining sensitivity or other features of interest and implement these algorithms as modules within an extensible system for decision support. A discussion of two fundamental concepts underpinning DS2 illustrates its potential. A *predicate* is a function that looks at a patient record and gives a (binary) decision about whether it implies a specified inference. For instance, an HIV predicate looks at a patient record and infers whether the patient is likely to be living with HIV. By contrast, a *reducer* is an operator that redacts (removes, segments, sequesters, etc.) parts of a record so that a given condition, viewed as sensitive to report in a given context, cannot be readily inferred from the remaining portion of the record. A *simple* predicate is one that looks for the presence of sensitive codes taken from a list. The corresponding reducer removes those codes from the record. However, predicates and reducers can be more complex. For instance, a probabilistic predicate uses machine learning techniques to develop a classifier that predicts whether a sensitive condition is implied. A corresponding reducer removes codes that are likely to imply the presence of a sensitive condition. The DS2 project has explored a variety of ways to produce predicates and reducers and compared these techniques to simple predicates. Details can be found on the web page[1] for the project. The site offers reports, software downloads, and a demo of an *inference analyzer*, which is tool that indicates the probability of inferring a condition based on other conditions as determined by a large data set of discharge records.

Complex predicates cannot generally be implemented manually by the parties that create the patient records because the originator of any single record cannot tell what will be implied by the consolidated record that is given to the consuming provider. Predicates and reducers can be associated with the broker system so that they can be used to analyze the longitudinal record. Such tools may also be useful in intra-organizational contexts as well. For instance, as hospital organizations have grown due to acquisitions and relationships like Accountable Care Organizations (ACOs), there is an increasing need to control the flow of sensitive information within the organization. This is a situation often encountered with mental health records, where the psychiatric division of a hospital aims to limit general access to records often with techniques like EMR support for access control and audit functions. Tools and architectures to detect and regulate information flows will be valuable additions to many contexts where pull technologies dominate.

[1] https://sharps-ds2.atlassian.net