Strategies for Medical Data Extraction and Presentation Part 3: Automated Context- and User-Specific Data Extraction

Bruce Reiner¹

Published online: 2 April 2015
© Society for Imaging Informatics in Medicine 2015

Abstract In current medical practice, data extraction is limited by a number of factors including lack of information system integration, manual workflow, excessive workloads, and lack of standardized databases. The combined limitations result in clinically important data often being overlooked, which can adversely affect clinical outcomes through the introduction of medical error, diminished diagnostic confidence, excessive utilization of medical services, and delays in diagnosis and treatment planning. Current technology development is largely inflexible and static in nature, which adversely affects functionality and usage among the diverse and heterogeneous population of end users. In order to address existing limitations in medical data extraction, alternative technology development strategies need to be considered which incorporate the creation of end user profile groups (to account for occupational differences among end users), customization options (accounting for individual end user needs and preferences), and context specificity of data (taking into account both the task being performed and data subject matter). Creation of the proposed context- and user-specific data extraction and presentation templates offers a number of theoretical benefits including automation and improved workflow, completeness in data search, ability to track and verify data sources, creation of computerized decision support and learning tools, and establishment of data-driven best practice guidelines.

Keywords Data mining · Data extraction · Decision support

✉ Bruce Reiner
breiner1@comcast.net

¹ Department of Radiology, Baltimore VA Medical Center,
10 North Greene Street, Baltimore, MD 21201, USA

Introduction

The relative lack of data integration in conventional medical information technologies results in a great deal of potentially relevant clinical data being overlooked, which can result in medical error, diminution in diagnostic confidence, and recommendation for unnecessary tests (which can result in delayed diagnosis/treatment planning and added expense) [1, 2]. This relative data inaccessibility coupled with manual workflow requirements produces a conundrum for healthcare providers—either spend inordinate amounts of time retrieving data which will adversely affect productivity or work with existing data limitations and run the risk of adverse clinical outcomes. The decision as to how function in this problematic environment is often dictated by the specific task at hand, individual provider's performance standards and preferences, and technology in use (i.e., context, user, and technology specificity).

In a companion article [3], a methodology was described for creating a patient referenceable database, which attempts to retrieve clinical data from multiple disparate information sources into a single all-inclusive database, effectively creating a “single source” of relevant clinical data. This could in theory improve workflow, productivity, completeness of data search, decisiveness, and clinical outcomes. Even if such a single all-inclusive patient referenceable database was successfully created however, the problems of “data overload” and manual data navigation remain as impediments to optimized task performance. The answer therefore lies in synergistic creation of automated data extraction templates which can take into account individual end user requirements and preferences, along with the specific task being performed. If successful, one could envision a scenario in which context- and user-specific data is automatically extracted from the universal patient database allowing for combined data economy,

workflow efficiency, and enhanced clinical outcomes. This data extraction strategy could be further enhanced by customizing data presentation states specific to individual end user's preferences, in an attempt to enhance data comprehension and decision-making in the most time-efficient fashion.

End User Profiling

One of the greatest challenges with conventional medical information system technology and its contained data is the fact that the population of participating end users is extremely diverse. This end user diversity can be defined in accordance with differences in occupation (e.g., physician, technologist, nurse), medical discipline (e.g., radiology, surgery, internal medicine), institutional demographics (e.g., tertiary care hospital, rural hospital, outpatient practice), education/training (e.g., general vs subspecialty), and individual variability (e.g., personality, workload, computer proclivity, clinical experience). The current approach of creating technology (and associated data) in a "one size fits all" approach is counterproductive and forces end users to adapt to the idiosyncrasies of the technology, rather than have the technology adapt to the needs and idiosyncrasies of the individual end user [4]. By creating adaptive and flexible technology, one could in theory improve end user performance through enhanced workflow, productivity, and clinical decision-making.

The first step in end user profiling is to define in general terms differences in data requirements in accordance with occupational status and medical discipline. Table 1 shows representative examples of generalized data allotments for three occupational groups (radiologist, clinician, and technologist), while Table 2 shows the representative breakdown of data categories for these three occupational groups within one specific medical discipline (medical imaging).

The generalized data allotments for the three representative occupations provided illustrate fundamental data requirement differences in accordance with occupation. Given the fact that imaging data lies at the core of radiologist and imaging technologist practice, it is logical that the largest data category within their data extraction templates to be dedicated to medical imaging data. A primary care physician (PCP), on the other hand, is responsible for comprehensive medical care of the patient and must therefore access and analyze data from a wider variety of medical data disciplines. As a result, only 10 % of the PCP data extraction template is dedicated to medical imaging data, which in turn requires greater economy of medical imaging data in comparison to the larger medical imaging data footprints for radiologist and technologist data extraction templates. This data "economy" can be accomplished in a variety of ways including a narrower chronology (i.e., shorter time frame for data extraction), prioritization of "actionable" medical imaging data (i.e., selective filtering of

Table 1 Generalized data allotments in accordance with occupational status

1. Radiologist
a. Imaging 55 %
b. Laboratory 5 %
c. Pathology 5 %
d. Surgical and interventional 5 %
e. Medical problem list 10 %
f. Clinical tests 5 %
g. Historical 10 %
h. Other 5 % (pharmacology, physical exam, genetics, dietary, medical devices)
2. Primary care clinician
a. Imaging 10 %
b. Laboratory 10 %
c. Pathology 5 %
d. Surgical and interventional 5 %
e. Medical problem list 20 %
f. Clinical tests 5 %
g. Historical 15 %
h. Physical exam 10 %
i. Pharmacology 10 %
j. Dietary 10 %
3. Imaging technologist
a. Imaging 40 %
b. Laboratory 10 %
c. Pathology 5 %
d. Surgical and interventional 5 %
e. Medical problem list 5 %
f. Clinical tests 5 %
g. Historical 15 %
h. Allergies 5 %
i. Complications 5 %
j. Other 5 % (physical exam, pharmacology, devices)

Table 2 Representative data allotments within an individual data discipline (medical imaging)

1. Radiologist
a. Imaging and report 70 %
b. Technical 5 %
c. Safety 5 %
d. Quality 5 %
e. Clinical 10 %
f. Other 5 % (procedural, administrative, pharmacologic)
2. Primary care clinician
a. Imaging and report 60 %
b. Administrative 15 %
c. Clinical 20 %
d. Other 5 % (procedural, safety, quality, pharmacologic)
3. Imaging technologist
a. Imaging and report data 20 %
b. Technical 40 %
c. Safety 10 %
d. Quality 10 %
e. Clinical 10 %
f. Pharmacologic 5 %
g. Other 5 % (administrative, procedural)

higher priority report data related to radiologic findings, disease, and follow-up recommendations), and preferential weighting to specific types of imaging exams.

The data within each broad data category can be further divided into individual data subcategories. In Table 2, the broader category of “medical imaging data” has been subdivided into a number of more narrowly focused individual data components including imaging and report, technical, safety, quality, clinical, administrative, procedural, and pharmacologic data. A radiologist whose primary task is that of image interpretation will therefore allocate a higher allotment to imaging and report data than a technologist, whose primary task is that of image acquisition, with a correspondingly high allotment to technical data. This illustrates the depth and granularity of data within each broad data category, with the ability to selectively extract data specific to the occupational requirements of different end user groups. When these generalized end user profile data extraction templates are combined with task-specific requirements and individual end user preferences, customizable context- and user-specific data extraction templates can be created.

Context Specificity

The creation of user-specific profiles is intended to create generic data extraction templates which account for the basic requirements of end users in accordance with their occupational status and routine daily task performance. They are intended to sift through and create structure within the myriad of data which is largely uncategorized in conventional practice and information technology databases. Even with the creation of end user profiles, an overabundance of data remains, making data extraction inefficient and largely manual. The goal therefore is to introduce logic into the data extraction process, specific to the nature of each individual task and the specific data requirements associated with successful and complete task performance (i.e., context-specific data extraction).

Context can be subdivided into two broad categories, subject matter and task (Tables 3 and 4). Subject matter refers to

Table 4 Context-specific subject matter

1. Anatomy/organ system
2. Pathologic diagnosis/disease
3. Clinical sign or symptom
4. Pharmaceutical
5. Test result
6. Device
7. Surgical procedure
8. Quality or safety

the primary category of data being evaluated, which most commonly refers to a specific anatomic region, disease entity, or clinical sign/symptom. Task refers to the specific action which is being undertaken in order to facilitate clinical care delivery and is often specific to the occupation of the end user. For medical imaging providers (e.g., technologists, radiologists), common tasks include non-invasive testing and procedures. The subject matter is routinely specific to a given organ system/anatomic region, symptom or sign, or pathologic diagnosis. As an example, a chest CT angiography being performed to evaluate chest pain and possible pulmonary embolus would be categorized as a non-invasive test (i.e., task) specific to the chest (i.e., anatomic region), pulmonary arterial system (i.e., organ system), diagnosis of pulmonary embolus (i.e., pathology), and complaint of chest pain (i.e., symptom). Data extraction would in turn focus on data relevant to the defined anatomy, organ system, diagnosis, and/or symptom, as well as the task being performed (CT angiography). Using artificial intelligence techniques (e.g., neural networks), a numerical score would rank the data retrieved in a hierarchical fashion in order to assign relative priority to individual data elements. Those data elements which are assigned priority rankings exceeding a specific threshold would in turn be incorporated into the context- and user-specific referenceable database specific to the task being performed.

The specific task being performed will have a profound effect on the complexity and granularity of the data search and extraction process. In the prior example of a chest CT angiography, all imaging and clinical data in the patient’s referenceable database would be included in the search process, including prior chest imaging exams, interventional procedures and clinical tests specific to the chest and/or vascular system, laboratory data related to the pulmonary and/or cardiac system, and historical and physical exam data related to the pulmonary, cardiac, and vascular systems. This could obviously entail a substantial amount of data in the processes of data search, extraction, and review. If, on the other hand, we were to take the example of a chest X-ray also performed for evaluation of chest pain and possible pulmonary embolus, the complexity and granularity of the data extraction process

Table 3 Context-specific tasks

1. Consultation
2. History and physical exam
3. Non-invasive testing
4. Routine evaluation
5. Follow-up assessment
6. Medical treatment planning
7. Emergency evaluation
8. Surgical/procedural planning
9. Quality assurance

would be far more superficial, in large part due to the fact that the data requirements for these two non-invasive imaging tests are far different from one another. In the example of a chest X-ray, data extraction will in large part be limited to a defined number of recent chest imaging exams and reports, along with cursory clinical data. In order to accommodate these marked differences in data granularity, rules must be established for each different task (which can in turn be customized to the specific needs and preferences of individual end users) to define the depth and breadth in data extraction requirements.

Customization, Learning, and Decision Support Features

In addition to providing the ability to create data profiles in accordance with occupation, task, and subject matter, data extraction templates should also provide the capability of customizing data commensurate with the unique needs to individual end users. Some customization features may be fixed or static, such as preferences related to data display (i.e., display presentation states) or requisite number of comparative historical data points. Other customization features may be dynamic in nature, providing the ability to modify data extraction requirements in real time.

One example of a real-time customization feature is an application or tool which can increase or decrease the extent of context-specific data granularity through a simple input command. This could be accomplished in a number of ways ranging from a three-part option for data granularity (limited, moderate, expansive) to a tool which allows the end user to manually dial up or dial down data granularity beyond the default setting. This option may be of particular value at higher levels of workload or fatigue, when end users may feel vulnerable to data overload. This illustrates the concept of adaptive and flexible technology, by not only taking into account context and user specificity but also providing an easy-to-use tool for modifying the technology (and associated data) in real time, based upon everyday fluctuation changes in workflow and stress.

Customization features need not be restricted by active input of the end user but can also be derived from computer learning and workflow analysis. Technologies such as electronic auditing tools and eye tracking software could in theory be used to monitor data usage and visualization in the course of everyday workflow and create a data workflow profile for individual end users. As specific trends of individual users are identified, the computer can provide feedback along with recommendations as to how the data extraction templates could be modified to best accommodate that individual user's data requirements and workflow. If, for example, a specific component of extracted data is routinely being overlooked or underutilized, the user may elect to eliminate or modify this data component from their profile.

Alternatively, if this component of overlooked or underutilized data has been shown to adversely affect clinical outcomes (e.g., errors in diagnosis, excessive follow-up recommendations), then the data extraction/presentation profile may be modified to facilitate improved visualization and analysis of the data in question. One way to accomplish this may be to visually highlight (e.g., alteration in font, color) the specific component of data routinely being overlooked in an effort to increase attention. An alternative application could provide an automated prompt prior to task completion to the end user alerting them the fact that specific portions of the extracted dataset were overlooked.

The ability of the database to serve as a medical reference is also a unique feature of the data extraction tool. In contemporary practice, when a physician creates a medical document (e.g., radiology report, consultation note, discharge summary), he/she will not typically reference the data source. In the event that the report data is found to be erroneous, it is difficult to determine whether the conclusions reached were the result of faulty decision-making or due to erroneous supporting data. Suppose, for example, a radiologist issuing an abdominal CT report performed for acute abdominal pain erroneously stated that the patient is status post appendectomy and the observed inflammation is the right lower quadrant and is the result of inflammatory bowel disease. In reality, however, the patient had no prior abdominal surgery and the inflammatory process of concern is in actuality due to a ruptured appendix, which requires emergency surgery. Along with the life-threatening medical error, the question arises as to how the radiologist came to the conclusion that the patient had previous appendectomy. Was it the result of erroneous information that he had been supplied with, and if so, what was the data source? The ability to audit all data within the referenceable database provides an important quality assurance tool to ensure that data sources are reliable, inaccurate data is identified and corrected, and medical decision-making is predicated on verifiable and accurate data. One can even go one step further and create references and electronic hypertext links between report data and the database from which it was derived. These links would not be restricted to textual data alone but also include graphical (e.g., flow charts), pictorial (e.g., pathology and radiology images), and numerical (e.g., laboratory) data. This could in theory create greater data accountability and assist in outcomes analysis, while also providing medical educational and training resources.

The referenceable database and context-specific data extraction capabilities described create a unique ability to confirm or refute a specific diagnosis. Computerized decision support tools have previously been described to assist in diagnosis at the point of care by defining clinical and/or imaging criteria [5, 6]. An expanded approach could be adapted to the context-specific database, in which established clinical and imaging criteria are presented to the end user at the point of care in an effort to improve diagnostic confidence and accuracy. As an example, in the evaluation of acute appendicitis, a

number of established clinical (e.g., white blood cell count, pain characterization [i.e., localization, duration, intensity], fever) and imaging (e.g. appendicular size, localized infiltration of fat, wall thickening, free fluid) criteria have been established for diagnosis. While these predefined criteria can be presented at the time of clinical and/or imaging evaluation as a form of clinical decision support, in conventional practice, it is largely left up to the individual healthcare provider to manually search available data sources, which is often time consuming, problematic, and incomplete. The ability to create a centralized referenceable database coupled with context-specific data extraction provides a data infrastructure which could improve workflow, satisfaction, and completion of data search; enhance diagnostic confidence and accuracy; and provide a tool for outcomes analysis. The predefined data elements specific to the diagnosis of concern could theoretically prompt an automated search of the database to determine which data elements are available, highlight and automatically extract these “high-priority” data elements in the database, and calculate the statistical likelihood of the diagnosis in question. In addition, the source of the highlighted data elements could be automatically recorded to validate data accuracy, and these combined data elements and sources could be integrated with the corresponding report for longitudinal analysis. If a specific data element is missing from the database which is deemed essential for accurate diagnosis (e.g., serial white blood cell counts), the absence can be documented in the report and provide justification for additional testing. The goal is to create an automated “data trail” which can be used for decision support, education, utilization review, provider performance analysis, and clinical outcomes analysis.

Conclusion

While not a trivial undertaking, the creation of a centralized patient referenceable database could provide healthcare providers with a single integrated data source. Automated extraction of this database is critical to ensure that essential data is not overlooked or bypassed and workflow efficiency is maintained. While current technology development strategies tend to utilize a “one size fits all” approach, which creates relatively inflexible and static software applications and tools, a preferred approach is to create data extraction templates which are customizable and adaptive to the specific needs and preferences of individual end users. On a macroscopic level, these data extraction templates can be designed in accordance with occupational requirements (i.e., end user profile groups),

which can be subsequently modified in accordance with the specific task being performed and clinical context. On a microscopic level, individual end users within different user profile groups can customize the data extraction and presentation templates in accordance with their own unique preferences.

On a more granular level, one could in theory study how individual end users actually interact with the data through eye tracking analysis, electronic auditing tools, and processing of “linked” data (which has been effectively “cut and pasted” from the data extraction templates into individual medical reports). These types of analyses would provide much needed insight and understanding as to how the most successful end users utilize data in the performance of their activities, which can eventually be used to create “best practice” context-specific data extraction and presentation templates. At the same time, data extraction templates can be shared between different end users specific to the individual task being performed, in a manner analogous to sharing speech recognition templates. The goal of this analysis is not to limit individual end users in the manner in which they create data extraction templates, but instead to empower them so that they can better understand the relative pros and cons of different data extraction options. In the end, the object is to adapt the technology to the needs and preferences of the individual end user and task being performed, while also providing data analytics and educational tools for iterative refinement, workflow optimization, and improved clinical outcomes.

References

1. Bates DW, Cohen M, Leape LL, et al: Reducing the frequency of errors in medicine using information technology. *J Am Med Inform Assoc* 8:299–308, 2001
2. Reiner B: Strategies for medical data extraction and presentation. Part 1: current limitations and deficiencies. *J Digit Imaging* 2014 (in press)
3. Reiner B: Strategies for medical data extraction and presentation. Part 2: creating a customizable context and user-specific patient referenceable database. *J Digit Imaging* 2014 (in press)
4. Reiner B: One size (doesn't) fit all. *J Am Coll Radiol* 4:567–570, 2008
5. Garg AX, Adhikari NKJ, McDonald H, et al: Effect of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 293:1223–1238, 2005
6. Kyle RA, Rajkumar SV: Criteria for diagnosis, staging, risk stratification and response assessment of multiple myeloma. *Leukemia* 23: 3–9, 2009