Author's response to reviews

Title: Text data extraction for a prospective, research-focused data mart: implementation and validation

Authors:

Monique E Hinchcliff (m-hinchcliff@northwestern.edu)
Eric Just (emjust@gmail.com)
Sofia Podlusky (s-podlusky@northwestern.edu)
John Varga (j-varga@northwestern.edu)
Rowland W Chang (rwchang@northwestern.edu)
Warren A Kibbe (wakibbe@northwestern.edu)

Version: 2 Date: 28 June 2012

Author's response to reviews: see over
June 28, 2012

Dear Editors for *BMC Medical Informatics and Decision Making*:

We are resubmitting our manuscript entitled: *Text data extraction for a prospective, research-focused data mart: implementation and validation* (Manuscript number: 894171666904102). We have substantially revised the manuscript in response to each of the Reviewers’ comments. We are confident that you will find the manuscript significantly improved. Below are the Reviewers’ comments followed by our responses. Page number references are included to facilitate review.

Responses to critiques by Reviewer Rebecca Crowley:

**Critique 1:**
Although the application area is important and it’s great to see information extraction being used with an EDW, the novelty of the work presented here is limited. The work may be more appropriate for a case study or application of technology paper than a research manuscript, especially if the setting and use of the data as a translational sciences tool is explicitly addressed.

**Response 1:**
The specific application, the transformation of data from a pulmonary function test instrument, building a standard extract transform and load to bring those data into our data warehouse, and the transformation of the data into a research-accessible (and reliable!) data mart is not terribly novel other than the fact that we were able to move this process from a research demonstration project into production. The novelty is in the reusable process and framework we developed, and the ability to couple this pipeline into the existing extract transform and load process in a production-ready form. We have subsequently used this pipeline (with other regular expression strings) to load a variety of data types into the Enterprise Data Warehouse. The regular expression-based approach, while certainly not novel, has had a number of advantages for us over natural language processing methods for instrument data. Conversely, we have found natural language processing methods to be necessary when extracting useful information from clinical notes. We have included a statement to this effect in the discussion section on page 12 first paragraph.

**Critique 2:**
The authors state on p12 that 11 discrete variables were studied, however they enumerate only ten variables on p13. Please clarify.

**Response 2:** Thank you for recognizing this inconsistency. Eleven discrete variables were collected for this study. In the previous submission, we neglected to mention gender. This omission has been corrected. The text on page 9 now reads: Eleven data elements (forced vital capacity and forced vital capacity % predicted from spirometry tests; total lung capacity, total lung capacity % predicted from lung volume tests; and diffusion capacity for carbon monoxide and diffusion capacity for carbon monoxide % predicted from diffusion tests, as well as medical record number, height, weight, gender, and exam date) were selected for each research participant.

**Critique 3:**
The authors do not provide a justification for why these 10 variables were selected from the total 107 parsed fields. Are these the most clinically interesting? Hardest or easiest? Randomly selected?

**Response 3:** Two of the eleven variables included patient and test identifiers (medical record number, and exam date) that were required to conduct the validation study with manual chart review. Clinically relevant variables included three anthropometric variables (height, weight and gender) that directly influence pulmonary function as well as six clinically relevant discrete pulmonary function parameters (forced vital capacity, forced vital capacity % predicted, total lung capacity, total lung capacity %
predicted, diffusion capacity for carbon monoxide and diffusion capacity for carbon monoxide % predicted. Because pulmonary function tests include three main components (spirometry, lung volumes and diffusion capacity for carbon monoxide) we validated two parameters from each of these tests. Physicians treating patients with systemic sclerosis follow forced vital capacity (FVC) and FVC % predicted obtained from the spirometry exam, total lung capacity (TLC) and TLC % predicted from the lung volume exam, and diffusion capacity of carbon monoxide (DLCO) and DLCO % predicted from the diffusion study to screen for lung disease. Please see page 9 for revised text.

Critique 4:
It is also not clear why the validation used such a specific population of scleroderma patients? Why not study a random selection?
Response 4: We apologize for not explaining more clearly how the scleroderma study cohort was selected. One hundred subjects from a cohort >500 scleroderma patients were randomly selected. All participants met the American College of Rheumatology criteria for systemic sclerosis/scleroderma and had consented to partake in medical research. Please see page 9 for revised text.

Critique 5:
The total number of reports (for these 100 patients) is not stated. This is particularly confusing because the results state that there were 1100 extracted values evaluated against manual review, yet in the Methods (p13) it suggests that these patients had multiple PFTs. The 1100 values suggest that there was only 1 report studied per patient (assuming that there were 11 instead of 10 variables extracted, see point 1). Anyways, I am confused. None of these numbers add up.
Response 5: Patients with SSc undergo pulmonary function tests every 6 months on average to screen for and/or monitor the progression of lung disease. We chose to study pulmonary function test results for the pulmonary function test performed closest to the date of the initial visit to our Program for consistency. Only one PFT report per participant was studied. To clarify, we abstracted 107 discrete data points from each PFT. Only eleven were used in the validation study. The eleven most clinically relevant data points were studied for 100 randomly selected patients. Please page 9 for revised text.

Critique 6:
The authors do not state the time period during which these ~38K PFTs were collected, or whether there was any inherent variability in the format of the documents. Since the documents are in fact text generated from the SensorMedics database sent as HL7 to Cerner, it seems likely that every document shared a rigidly identical format (it is machine generated after all). It should not be surprising then that regular expressions could perform nearly perfectly in parsing this data, especially given the ten variables that were selected for study. This is the most significant problem with the study.
Response 6: The PFTs collected for this study were performed between 12/2004 and 9/2010. There was no inherent variability in the format of the documents. Reviewer 1 is correct in her assessment that the results are not surprising. The regular expression parsing is certainly not novel; the novelty is in the incorporation of an easily configurable regular expression pipeline into a standard ETL process that provides production-capable transformation of these data for research purposes. A statement has been added to the discussion on page 12. The dates that the PFT were performed are now included on page 9.

Critique 7:
The authors do not address maintenance. What happens if the SensorMedics model changes with a new version? What if they change PFT machines and software? What if they simply change the format of the SensorMedics report?
Response 7:
The regular expressions would need to be updated. The particular format has not changed in more than 10 years, but this issue is very common in data warehousing extract transform and load processing. Underlying data, mappings, and systems in use in the clinic are in constant change and unfortunately most of these changes require changes to the extract transform and load processes. At the current time, our Enterprise Data Warehouse is importing data from more than 70 clinical systems, and dealing with change synchronization issues is very important for us, please see page 13.

Critique 8:
It’s not clear exactly what is meant by the “Regextractor framework”. Is there something beyond the ETL component wrapping the regular expressions? This doesn’t seem like a framework. Please clarify.
Response 8:
The Regextractor framework is necessary as our clinical enterprise, including the Research-use EDW, is in a .NET environment. The Microsoft SQL Server Integration Services (SSIS) platform (Microsoft’s term for it) is the Microsoft tool for managing ETL into a SQL Server instance. We load between 50 and 750 GB of incremental changes through SSIS each night. Unfortunately, SSIS makes it very difficult to embed regular expressions without recompiling the ETL package. What we have done is create a separate ‘framework’ for embedding regular expressions in an easily updatable text file (the way they were intended to be stored!). While this particular innovation is only relevant (and perhaps necessary) to SSIS, the number of medical data warehouses using SSIS appears to be significant based on surveys made by the Healthcare Data Warehousing Association and the CTSAs. We have added a relevant publication from the CTSAs showing this on page 12[1] .

Critique 9:
Although this is valuable work and important for clinical research, the actual software may not be immediately usable except to institutions using SensorMedics. The generalizability of the approach should be more thoroughly discussed in a limitations section.
Response 9:
Thank you for drawing our attention to our failure to relay the central message of the paper. The take home message of the manuscript is that we have developed a framework for easily embedding regular expressions into SSIS ETLs. Our approach can be adapted to many other electronic health records that contain text data within reports that are not readily amenable to data extraction. The discussion has been revised to reflect this important point.

Critique 10:
Known limitations of regular expressions and other very shallow forms of text processing are not adequately addressed.
Response 10:
Thank you for your comment. We were not trying to replicate NLP or emulate a machine-learning approach with regular expressions. However, there are a large class of problems that are relevant for healthcare data warehousing applications that require simple, reproducible and fast transformation tools that are more sophisticated than simple SQL-based searching but less sophisticated than NLP. Transformation of instrument data is the poster child for regular expression-based transformations. It appears to us that the data warehousing community, or at least the SQL Server-using data warehousing community has forgotten this, or perhaps it is because incorporating regular expressions into ETL workflows is so difficult that it is not done routinely. We added a statement to the Discussion section to indicate that regular expressions only work with well structured data and are not a substitute for NLP: “We recognize that regular expressions are only appropriate for highly structured, machine-produced data and span the gap between structured data entry and natural language processing techniques for semantically interpreting text such as clinical notes”, page 12.
Critique 11:
Figure 2 is not particularly helpful and could be omitted as it is addressed in the text.
Response 11: We have extensively revised Figure 2 to make it more relevant and useful.

Minor Essential Revisions

Critique 12:
The Methods section is improperly placed in this document. The current pages 12-14 should be placed before the results section on page 7.
Response 12: Thank you for drawing our attention to this error. The Methods section is now in the proper place.

Responses to critiques by Reviewer Joanne Luciano:
Critique 1: I do not generally work with in Enterprise Data Warehouse (EDW) environment, but I do have extensive experience and knowledge in the area of computer science and translational research. As a reader of this sort, I would be better by a decreased use of Three Letter Acronyms (TLA). Spelling out, defining terms such as data warehouse and data mart, and how they are different from the more familiar database would improve understanding of the paper and make it easier and more enjoyable to read. For example, that a data warehouse is optimized for querying the data while a database is optimized to record data and will be confused by this. The same is true for EHRs, chart review etc. Adding a few clarification sentences will greatly improve the readability and accessibility of the manuscript because they will give the reader a more solid grounding in the concepts introduced, which will support the acronyms are used throughout. For example, the title of the paper is “Text data extraction for a prospective, research-focused data mart: implementation and validation” but the term data mart is never defined, nor is any prose given to provide the reader with an appreciation of what it is and why its useful, and how it relates to a data warehouse. Furthermore, it would be helpful to include with the definition of terms, examples of what they mean and why they are useful and important in this context.
Response 1: This is always a difficult issue in writing a manuscript that is both accessible by a broad audience and yet targeted to the aficionado. Your specific examples are very helpful and we clarified our terminology including providing a definition of data mart on page 5 in the Background section.

Critique 2:
It would be beneficial if there were a way to try tool outside of the SSIS (SQL Server Integration Services) or SSRS (SQL Server Reporting Services) environment (perhaps with a demo) to reproduce the results, or otherwise explore the software in order to determine its utility. I can imagine a web-based service that performs the regex parsing could be useful to a broader community.
Response 2: You are correct, however most other Extract, Transform, Load environments make the incorporation of regular expressions much easier than SSIS. The necessity of having to develop a SSIS package (Regextractor) to load a text file of regular expressions versus having to recompile the SSIS package every time the regular expression changes is uniquely a SSIS issue. While I am a fan of webservice-based implementations of many types of transformations, these are not easily integrated into an ETL environment by any vendor currently.

Critique 3: The example of the Pulmonary Function Test needs to be fleshed out more. For example, “Pulmonary function tests (PFT) [a] are administered using a SensorMedics Vmax Encore PFT Autobox Pro machine [b] and data is captured using SensorMedics software version IVS-0101- 21-1A (both CareFusion Corporation, San Diego, CA). PFT data is stored in a proprietary [5] database used by the SensorMedics software.
[a] It would improve the paper to provide a description of what the PFT test specifically measures (how is “function assessed”), why the test is performed, and what the data measurements obtained tell me, would make this sentence interesting and informative.

[b] SensorMedics Vmax Encore PFT Autobox Pro machine?? a what? I am sure the reader will understand if you would give them a chance – I know that it doesn’t matter, and the only thing that is important for your software is that the instrument generates “data”, but I would like to learn something about the example too, since you are illustrating it. Therefore, a scenario would be useful to help the reader understand the software, by example, and in context, the problem that software solves and how it solves it. The section on clinical utility could provide a bit more – something along the lines of what’s in Wikipedia (“is a systemic autoimmune disease or systemic connective tissue disease that is a subtype of scleroderma (a chronic systemic autoimmune disease (primarily of the skin) characterized by fibrosis (or hardening), vascular alterations, and autoantibodies). [2] It is characterized by deposition of collagen in the skin and, less commonly, in the kidneys, heart, lungs & stomach. Female to male ratio is 4:1. The peak age of onset is between 30-50 years.” And maybe a little about the prevalence So I learn something and to warrant the paper being in a BMC journal rather than an ACM. That being said, it is commendable that the scleroderma investigators have been able to build additional queries and use for other research projects.

Response 3:

a) Thank you for your comment. To make the paper more readable and clinically relevant we have revised the manuscript extensively. On page 7 under Pulmonary Function Test Data Flow, we now provide an explanation of pulmonary function tests (PFT). Figure 1 (a PFT report from the electronic health record) and Table 1 (description of the PFT measures) are now referenced so the reader can see an actual report and the discrete variables that are extracted from the report.

b) On pages 8 and 9, we define scleroderma and provide more background information about the disease and the utility of PFT to screen for and monitor lung disease progression-the leading cause of death in this patient population. In Table 1 we provide definitions of the various component tests of the PFT. Additionally, we provide a photo of a patient within an autobox so that the reader can see how PFTs are performed (Figure 2). We describe manual chart review (former method used to collect PFT data) and list the limitations of the approach as well as introduce our new system for PFT data retrieval (Enterprise Data Warehouse approach):

“Patients with systemic sclerosis/scleroderma, a rare connective tissue disease that causes skin and internal organ fibrosis especially in the lungs, vascular disease, and autoantibody production, undergo PFTs to screen for the development and progression of lung disease. Scleroderma predominately affects middle-aged women, and lung disease is the leading cause of death. As a result, aggregate PFT data are required for many scleroderma translational research projects and provide a measure of disease progression in patients with scleroderma. In the past, scleroderma research assistants printed PFT reports and manually entered data into data capture tools such as spreadsheets. Because this process was time consuming, ecologically unfriendly and error prone, a new system was developed in conjunction with Northwestern Medical Enterprise Data Warehouse data architects to generate a PFT data mart within the Warehouse that can readily be
queried and maintained through automated processes.”

**Critique 4:**
Figure 1 is good. An excerpt of it should be used to walk the reader through an example. Show how Table 1 and Figure one relate.

**Response 4:** Thank you for your suggestion of how to help the reader make the connection between Table 1 and Figure 1. Figure 1 now references Table 1 data. We hope that this is more clear.

**Critique 5:**
Figure 3 is unreadable - the font is too small. Also, there is no annotation. In the text show the text and the regular expression that matches the pattern.

**Response 5:**
The regular expressions will be made available as a more readable appendix and the entire package including the regular expressions are available at http://regextractor.codeplex.com.

**Critique 6:**
Table 1 could use more detail in the description. As it stands, I do not get much out of it.

**Response 6:** Thank you very much for asking for this clarification. A more detailed description of Table 1 data has been provided. Please see the revised Table 1.

**Critique 7:**
Would Figure 3 be better represented as a graphic with indicating a cycle since it occurs every day (or am I reading it wrong?) It would also be more informative to indicate where on this diagram the software tool is operating. I am guessing that it is in step going into to the EWD Discrete PFT data mart (and is the data mart within the EWD?) A graphic would be helpful here.

**Response 7:** In response to your helpful comment and that of the other reviewer as well, we have revised Figure 2 to make it more meaningful and relevant.

**Minor Essential Revisions**

**Critique 8:** Cut down on the use of three letter acronyms. It’s not that big a paper or that cumbersome, especially for the terms that are used less frequently, such as ODS. OK to put it in parenthesis after the initial use, but better to not use the TLA.

**Response 8:** We have lessened the use of TLAs to make the paper more enjoyable to read. Thank you for your comment.

**Critique 9:** In the Technical Generalizability of Text Processing Workflow section, please show or describe how “The EDW makes it easy to construct regular expression parsing pipelines for text.”

**Response 9:**
We changed this to state that:
“The ‘Regextractor’ SQL Server Integration Services package makes it easy to modify the regular expressions without recompiling and incorporate the parsing pipeline into existing extract, transform and load processes”, please see page 13.

**Critique 10:**
RE: Downloaded 315 times since 6/15/2011 – that date is nearly a year old. I’m wondering why? ....

**Response 10:**
As of June 1, 2012, the open source software has been downloaded 552 times. A more up to date tally has been added to the manuscript on page 2 and page 9.

Discretionary Revisions

Critique 11:
I’m also wondering if this technology is “legacy” and how it compares to or fits in with other translational research pipelines such as i2b2 or if the output can be adjusted to generate linked open data?

Response 11:
i2b2 is a datamart strategy, and ETL is a way to populate data warehouses that tools like i2b2 can use. Linked open data is typically used after the data warehouse is populated through a tool like i2b2.

Critique 12:
I’m interested (as a reader) in knowing what the details of 6 inconsistencies were – what were the mistakes in the manual abstraction (and why, especially since the chart abstractor was familiar with PFT)?

Response 12: The 6 inconsistencies were mistakes made during data entry. The manual data extractor made typographical errors that accounted for five/six inconsistencies. These errors resulted in two inconsistent medical record numbers, two incorrect test dates, and one inconsistency in height. The final error was due to incorrect coding for gender during manual data entry. Female was entered instead of male. Our finding of 6/1100 mistakes in manual data entry is well below the reported 10% error rate for manual data extraction using a paper intermediate, and yet underscores the importance of designing systems to electronically collect and integrate research data. Please see page 14.

If you have any additional questions or concerns, please do not hesitate to contact me. I can be reached at 312-503-0495.

Sincerely,

Monique Hinchcliff, MD MS
Northwestern University
Feinberg School of Medicine
Division of Rheumatology
Northwestern Scleroderma Program

Reference: