A Clinical Use Case to Evaluate the i2b2 Hive: Predicting Asthma Exacerbations

Stéphane M. Meystre, MD, PhD, Vikrant G. Deshmukh, MS, Joyce Mitchell, PhD
Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

To evaluate the i2b2 Hive as a tool to query, visualize, and extract clinical data, we selected a use case from the i2b2 airways diseases driving biology project: asthma exacerbations prediction. We analyzed the cohort selection and the extraction of the clinical data used by this asthma exacerbations prediction study. The structured data included the asthma diagnosis, birthdate, age, race, sex, height, weight, and BMI. The smoking status is typically only mentioned in clinical notes, and we evaluated the Natural Language Processing (NLP) application embedded in the i2b2 NLP cell to extract the smoking status from history and physical exam reports.

Querying structured data was possible with the i2b2 workbench for about half the clinical data elements. The remaining had to be queried using a commercial database management system client. The automated extraction of the smoking status reached a mean precision of 0.79 and a mean specificity of 0.90.

Introduction and Background

The growth of Electronic Health Records (EHRs) and of the electronic clinical data they contain offer great opportunities for faster and easier access to clinical data for research purposes. This access is typically provided through data warehouses and requires knowledge of the structure of the data warehouse and of database querying languages. Clinicians and researchers rarely possess this knowledge and have to rely on data warehouse specialists. As a result, their access to clinical data warehouses can be time consuming and expensive. Several attempts to alleviate these issues have been proposed, such as the caGRID Browser[1], REDCap[2], or the i2b2 Hive[3]. Informatics for Integrating Biology and the Bedside (i2b2)[4] is one of the seven National Centers for Biomedical Computing (NCBC) funded by the NIH, and is based at Partners Healthcare (Boston, MA). It focuses on developing a new informatics framework to bridge clinical research data with basic sciences research data, and on driving biology projects that serve as testbeds for this informatics framework. These projects include airways diseases, hypertension, Huntington's disease, diabetes mellitus type II, major depressive disorder, rheumatoid arthritis, and obesity.

The main component of the informatics framework is a modular, user-friendly, data querying and visualization system called the i2b2 "Hive" [3]. This "Hive" is made of interoperable "Cells" in a service-oriented architecture, and of a persistent data storage called the Clinical Research Chart. This datamart (i.e. specialized subset of a data warehouse) aggregates data from a variety of clinical and research data sources. Many cells have a visible client component that users can interact with. These components are grouped in the i2b2 "Workbench", and allow for searching or browsing the terminology provided with the i2b2 Hive, building and storing queries with concepts from this terminology, and visualizing the queried data. Optional cells allow extracting pulmonary function test results from the corresponding reports, and extracting some clinical information (diagnoses, discharge medications, smoking status, and other terms defined by regular expressions[5]) from clinical notes using Natural Language Processing (NLP). The provided terminology includes demographic data, diagnoses, laboratory exams, medications, and procedures.

To evaluate the generalizability of the i2b2 Hive to another academic institution that didn't participate in its development, a partnership between the i2b2 NCBC and the Department of Biomedical Informatics at the University of Utah has been established. The Department of Biomedical Informatics was selected for its significant informatics expertise, and for the availability of a clinical data warehouse and well-developed resources in both the clinical and research domains. The clinical data warehouse captures data from over 200 clinical and administrative data systems and currently has over 2 million patient demographic records.

A local i2b2 Hive was installed and tested at the University of Utah. An evaluation of its technical and functional aspects followed, as described in Deshmukh et al.[6] To evaluate the i2b2 Hive in the context of a concrete clinical research study, we selected one of the i2b2 driving biology projects – airways diseases – as a use case, and focused on an asthma exacerbation prediction study[7] realized in Boston in the context of this driving biology project. This asthma exacerbation prediction study included 12,792 asthma patients with race, sex, height, weight, birthdate, and smoking history data. The smoking history was automatically extracted from clinical notes using a NLP application[8]. Asthma exacerbations were defined as an inpatient admission or emergency department visit with a principal diagnosis of asthma (coded with ICD-9-CM). Patients were divided into a training set and a testing set. The former was used to develop a multivariate logistic regression prediction model, and the latter to evaluate the accuracy of the predictive model, with a resulting area under the ROC curve of 0.67 [7].

Our aim was to use the same criteria to select a group of patients and retrieve the same clinical data as if we would evaluate the generalizability of part of this asthma exacerbation prediction study (prediction model excluded). Here, we describe and discuss this process, including the automated extraction of the smoking status from clinical notes using the Natural Language Processing application embedded in the NLP cell of the i2b2 Hive.

A substantial part of the medical record is made of narrative clinical text documents that represent patient history and reports of therapeutic interventions or clinical progress[9]. Clinical and basic sciences research create a need for structured and coded data instead. As a possible answer to this issue, NLP can convert free text into coded data. The automated extraction of information from clinical text is a recent application of information extraction methodologies, as described in Meystre et al.[10] Techniques for automatically encoding textual documents from the medical record have been evaluated by several groups. Examples are SymText[11], and MedLEE (Medical Language Extraction and Encoding system)[12]. The i2b2 Hive integrates another NLP application called HITEx, evaluated when extracting diagnoses and the smoking status of patients[8]. To extract the smoking status, HITEx implements a "pipeline" that splits the document in sections (sectionizer), in sentences (sentence splitter), and in tokens (text tokenizer), and then selects the sentences with mentions of "smok", "tobacco", or "cigar" (e.g. "Quit *smok*ing 15 years ago after an 80- to 100-pack-year history"). Finally, a Support Vector Machine [13] classifier, already trained on about 8500 smoking-related sentences [8], determines the smoking status associated with each selected sentence. The automated extraction of the smoking status was also the focus of a competition in 2007: the i2b2 "smoking challenge"[14].

Methods

The local installation and the technical and functional evaluation of the i2b2 Hive are described in details in another publication[6]. Our installation included version 1.2 of the i2b2 Hive, on separate application and database servers. For this study, we selected a cohort of about 5,000 patients with a diagnosis of asthma (i.e. 493.xx ICD-9-CM codes) from our local clinical data warehouse, and extracted demographic and clinical data from these patients. These data were then loaded in a transition datamart, and then in part in our local i2b2 datamart (i.e. Clinical Research Chart (CRC)). This process required significant terminology mapping and extension of the terminology provided with the i2b2 Hive, as described in Deshmukh et al.[6]. The i2b2 datamart included demographic data, diagnoses, procedures, laboratory tests, and medications. Its terminology was enriched with new medications and laboratory tests to accommodate our local clinical data, but biometric data like height, weight, and BMI were absent from the terminology provided with the i2b2 Hive and could therefore not be stored in the i2b2 datamart; they were only stored in the transition datamart.

The asthma exacerbation prediction study[7] that we chose as a use case for clinical data querying and extraction was based on the clinical data elements listed in Table 1. Structured data were queried with the i2b2 workbench or with a commercial database management system (DBMS) client (SQL*Plus; from Oracle Corp., Redwood City, CA) if not possible with the i2b2 workbench, as depicted in Figure 1. Unstructured data (i.e. smoking status) were extracted from history and physical (H&P) exam reports using HITEx.

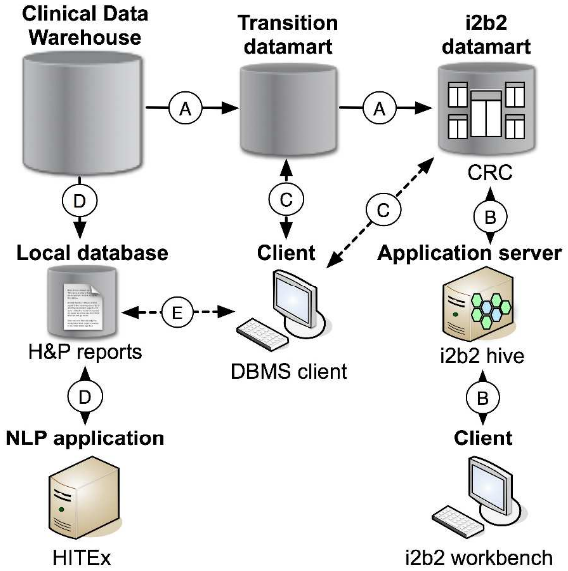

Figure 1: Data loading, querying and extraction processes. (A: structured data loading; B: querying with the i2b2 workbench; C: querying with the commercial DBMS client; D: H&P reports extraction and analysis; E: querying smoking status).

Patient cohort selection: The inclusion criteria were: adults (at least 18 years old) with race, sex, height, weight, and smoking history data. If available, the

Clinical data	Data format	Values
Asthma diagnosis	Structured	ICD-9-CM codes (493.xx)
Visit type	Structured	Inpatient admission or outpatient emergency department visit
Birthdate	Structured	Date
Age	Structured	Categorized: 18-44, 45-64, 65-74, 75 and more.
Race	Structured	White, Black, Hispanic, Asian.
Sex	Structured	Male, Female, Unknown.
Height	Structured	Numeric [m]
Weight	Structured	Numeric [kg]
Body Mass Index	Structured	Numeric [kg/m^2] and then categorized: underweight (BMI<18), normal (18≤BMI<25), overweight (25≤BMI<30), obese (30≤BMI<40), morbidly obese (BMI ≥ 40).
Smoking history	Unstructured (cited in text)	Categorized as: Positive history, Negative history

Table 1: Clinical data used for asthma exacerbations prediction.

BMI (Body Mass Index) was also used, and height and weight were used to calculate it otherwise.

If patients had at least one instance of asthma exacerbation (an inpatient admission diagnosis of asthma or an emergency department visit with a primary diagnosis of asthma), they were classified as case, and as control otherwise. For the cohort selection task, we first tried using the i2b2 workbench, and used the commercial DBMS client if the workbench was not sufficient (read below for more details).

Structured data query and extraction: We also used the i2b2 workbench as much as possible to query and extract the structured clinical data listed in Table 1. The i2b2 workbench (version 1.2) didn't allow downloading data and couldn't therefore be used for data extraction. The commercial DBMS client was used for the data extraction and to query the data that the i2b2 workbench wasn't able to query.

We measured the prevalence of each clinical data element in our cohort (i.e. the proportion of patients in our cohort with the corresponding clinical data stored in the database), when queried with the i2b2 workbench and when queried with the commercial DBMS client.

Unstructured data extraction: The smoking status was extracted from clinical notes using the Natural Language Processing application embedded in the i2b2 NLP cell and called HITEx. The smoking status is mentioned in several different clinical notes such as the discharge summary or the history and physical exam report. The latter cites the smoking status the most frequently, and generally in more details. We therefore chose to extract the smoking status from history and physical exam reports. These reports could be stored in the i2b2 datamart, but we did not import them in this datamart; we extracted them from our clinical data warehouse and stored them in a local MySQL (from MySQL AB, Uppsala, Sweden) database used by HITEx. The latter was then used to analyze these reports, and we used it in batch mode with the default configuration for smoking status extraction. The output of HITEx was a list of sentences that mentioned the smoking status of the patient, with the corresponding extracted smoking status (Current_Smoker, Past_Smoker, or Non-Smoker) and details about the location of the sentence, the reports, etc. In the asthma exacerbation prediction study[7], values for the smoking status were: positive history of smoking, or negative history of smoking. We therefore simply categorized the Current_Smoker and Past_Smoker statuses as a positive history of smoking, and the Non_Smoker status as a negative history of smoking.

To measure the accuracy of the smoking statuses extracted by HITEx, sentences and the corresponding smoking statuses were selected from the output using a stratified random sampling method (to ensure equal representation of each smoking status). To measure the accuracy with a confidence level of 95% and a confidence interval of about 3%, we estimated that 300 sentences for each smoking status were required (total of 900 sentences). Two reviewers independently examined each of these sentences, and selected the smoking status they deemed correct. If they disagreed, the disputed case was discussed to choose the smoking status considered the most correct.

Each extracted smoking status was categorized as true positive (TP; status mentioned in the sentence and extracted by HITEx), false positive (FP; status not mentioned in the sentence but extracted by HITEx), and true negative (TN; status not mentioned in the sentence and not extracted by HITEx). We then calculated the specificity (TN/(TN+FP)) and the precision (i.e. positive predictive value; TP/(TP+FP)) for each smoking status. Since we only analyzed sentences with smoking statuses detected by HITEx, we could not evaluate the sensitivity or the recall of this detection.

Results

Patient cohort selection: Among the data used as selection criteria to define the cohort of patients, only the asthma diagnosis, the age, the race, and the sex could be queried with the i2b2 workbench, as shown in Table 2. We had to use the commercial DBMS client to query the data corresponding to the other selection criteria, as explained for the structured data query below. The asthma exacerbations (i.e. inpatient admissions or emergency department visits with a principal diagnosis of asthma) also had to be queried with the DBMS client.

Structured data query and extraction: For each structured data elements needed to replicate the asthma exacerbation prediction study, we first tried using the i2b2 workbench to query them. As listed in Table 2, this tool allowed querying for the asthma diagnosis, the age, the race, and the sex. The asthma diagnosis and the sex could be easily selected from the terminology and added as query criteria. For the race, the terminology contains more values than used in the asthma exacerbation prediction study, and we had to manually add only the four values we were interested in (i.e. White, Black, Asian, and Hispanic). The terminology doesn't include the birthdate, the visit type, the height, the weight, or the BMI, and these data therefore had to be queried with the commercial DBMS client. The birthdate and the visit type were stored in the i2b2 datamart (in the patient and visit dimensions), but couldn't be queried with the i2b2 workbench. The height, weight, and BMI could not be stored in the i2b2 datamart.

The prevalence of the data that could be queried with the i2b2 workbench was the same when queried with the commercial DBMS client, which shows the quality of the queries executed in the i2b2 Hive.

As mentioned above, the i2b2 workbench (version 1.2) provided no means to download the queried data. The extraction of clinical data had to be performed with the commercial DBMS client.

Clinical data	**Stored in the transitional datamart**	**Stored in the i2b2 datamart**	**Queried with the i2b2 workbench**
Asthma diag.	●	●	●
Visit type	●	●	—
Birthdate	●	●	—
Age	●	●	●
Race	●	●	●
Sex	●	●	●
Height	●	—	—
Weight	●	—	—
BMI	●	—	—

Table 2: Structured data storage and query.

Unstructured data extraction: Two different reviewers, one physician/informaticist and one informaticist/data architect, reviewed each sentences to create the reference standard. Their agreement was almost perfect, with a Cohen's kappa of 0.91.

The automated extraction of the smoking status mentioned in history and physical exam reports reached a mean specificity of 0.899 and a precision of 0.788, as described in Table 3.

Smoking status	**Specificity**	**Precision**
Current_Smoker	0.855	0.701
Past_Smoker	0.865	0.698
Non_Smoker	0.978	0.963
Positive smoking history	0.860	0.700
Negative smoking history	0.978	0.963
Mean	0.899 (0.73-1)	0.788 (0.41-1)

Table 3: Smoking status extraction specificity and precision (mean with 95% confidence intervals).

Discussion

This evaluation of the i2b2 Hive to query and extract clinical data for a specific use case - predicting asthma exacerbations – suggests that such a user-friendly data querying and visualization tool can address part of the needs of researchers, and improve their access to clinical data. With the terminology provided with the i2b2 Hive and locally enriched with laboratory exam and medication concepts, about half the clinical data required for this use case could be queried with the i2b2 workbench. With further modifications of the terminology, all data elements could probably have been queried with the workbench alone. The extraction of the data had to be realized with the commercial DBMS client, since the version of the i2b2 Hive that we evaluated (version 1.2) didn't offer this functionality. The latest version of the i2b2 Hive (version 1.3) allows downloading some clinical data.

When using the i2b2 workbench, we found the terminology searching and browsing functionalities very intuitive and easy to use. The query building by drag-and-dropping terminology concepts into criteria-grouping zones was also very intuitive. The timeline visualization of the data has limited interest, such as possibly seeing temporal relationships. However, other ways of presenting the data, such as spreadsheets, could be more suitable for researchers.

The smoking status extraction was based on the NLP cell, but the cell interface only allowed for the manual analysis of one document at a time. We therefore chose to use the NLP application embedded in the NLP cell (i.e. HITEx) directly.

The reference standard was of good quality, thanks to the excellent agreement between reviewers. The precision and specificity that we measured – about 79% and 90% - were comparable with another published evaluation of HITEx that reported a mean precision of about 81% and a mean specificity of about 97% [8]. During the i2b2 smoking challenge on 2007, the best systems reached a precision of 90% [14].
Among errors HITEx made, a common one was to attribute a Current_Smoker status when the reference was Non_Smoker (about 39% of the errors), and this was most often due to a negation detection error: *denies* was not considered a negation term, as in "He denies any smoking". Another common error was to attribute a Past_Smoker status when the reference was Current_Smoker (about 36% of the errors), and this was often related to a mention of smoking in the past, like in "He smoked about a pack per day but recently decreased to a couple of cigarettes per day."
The smoking status extraction, if examined as the agreement of HITEx with the reference standard, reached a Cohen's kappa of 0.68, and was therefore a substantial agreement. A limitation of this study is the fact that the smoking status classifier provided with HITEx was trained on discharge summary sentences from Partners Healthcare, when we used it with history and physical exam reports from the UUHSC. The sentence splitting is probably very similar, but phrases expressing a smoking status might be different.

Acknowledgments
We would like to thank Isaac Kohane and Susanne Churchill at the i2b2 NCBC for the subcontract award; Shawn Murphy, Vivian Gainer, Michael Mendis, and Lori Phillips for their help with the i2b2 Hive; Qing Zeng-Treitler and Sergey Goryachev for their help with HITEx; Scott Weiss and Blanca Himes for their collaboration with the asthma study; and Cheri Hunter and Jim Livingston at the UUHSC for the infrastructure and implementation resources.
This investigation was approved by the University of Utah Institutional Review Board, and supported by the i2b2 subcontract (NCBC U54 LM008748), the National Center for Research Resources Public Health Service Research Grant (UL1-RR025764, C06-RR11234), the Department of Biomedical Informatics at the University of Utah, and by the Health Sciences Information Technology Services at the UUHSC.

References

1. Saltz J, Oster S, Hastings S, et al. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics. 2006 Aug 1;22(15):1910-6.
2. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2008 Sep 30.
3. Murphy SN, Mendis M, Hackett K, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. Proc AMIA Symp. 2007:548-52.
4. i2b2. Informatics for Integrating Biology and the Bedside. 2008. Available from: https://www.i2b2.org/
5. The Open Group. Regular Expressions. The Single UNIX Specification, Version 2 1997
6. Deshmukh V, Meystre SM, Mitchell J. Evaluating the Informatics for Integrating Biology and the Bedside Patient Cohort Selection Tool at the University of Utah. (submitted for publication). 2009.
7. Himes BE, Kohane IS, Ramoni M, Weiss ST. Characterization of Patients who Suffer Asthma Exacerbations using Data Extracted from Electronic Medical Records. Proc AMIA Symp. 2008:308-12.
8. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak; 2006. p. 30.
9. Pratt AW. Medicine, Computers, and Linguistics. Advanced Biomedical Engineering. 1973;3:97-140.
10. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics. 2008:128-44.
11. Haug P, Koehler S, Lau LM, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. Proc Annu Symp Comput Appl Med Care. 1994:247-51.
12. Friedman C, Johnson SB, Forman B, Starren J. Architectural requirements for a multipurpose natural language processor in the clinical environment. Proc Annu Symp Comput Appl Med Care. 1995:347-51.
13. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods: Cambridge University Press; 2000.
14. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):14-24.