

i2b2 AIRWAYS DISEASE DBP: PATIENTS, NLP PHENOTYPES, MODELS, AND DATA VIEWER



Ross Lazarus¹, Shawn Murphy², Qing Zeng³, Margarita Sordo³, Lee-Jen Wei⁴, Anne Fuhlbrigge¹, James Signorovitch⁴, Susanne Churchill⁵, Scott Weiss^{1*}, Isaac Kohane⁶

1. Channing Laboratory, Brigham and Womens Hospital and Harvard Medical School; 2. Department of Neurology, Massachusetts General Hospital; 3. Decisions Systems Group, Harvard Medical School; 4. Department of Biostatistics, Harvard School of Public Health; 5. i2b2 NCBC; 6. Informatics Program, Children's Hospital Boston

ABSTRACT

Translating basic research into clinical practice in asthma, a common respiratory illness, is the focus of the airways disease DBP for the i2b2 NCBC. Our overall goal is to elucidate genetic variants associated with individual patient responses to common asthma medications, so we are assembling detailed patient characteristics and DNA samples for genotyping.

In collaboration with the Tools core, we have developed a data warehouse (DW), populated from the existing Partners patient data repository. This DW is a functioning prototype for the patient clinical data component of the i2b2 clinical research chart and contains extensive clinical data on more than 97,000 asthmatics. Some important patient characteristics were not available in coded form from the DW. Collaborating NLP experts from the Scientific Core have provided smoking history, medication use and additional comorbidity measures derived from progress notes and discharge summary texts. They have developed a novel tool which allows generalization of their work, HITEX, which will soon be made freely available as part of the i2b2 toolkit.

We are beginning to directly measure relevant patient characteristics, and store DNA samples for future genotyping from a substantial sample of Partners asthma patients. The cohort is anticipated to be between 1500 and 2000 subjects, recruited through collaboration with physicians located at Partners asthma outpatient specialist services. We are measuring lung function and bronchodilator responsiveness (BDR), and obtaining a past medical history, smoking history, and medication use history as well as a sample of blood for extraction and storage of genetic material. In addition to serving as phenotypes in our genetic analyses, direct measurement of patient characteristics will be used to validate the DW data. Most of these patients will stay in the system and clinical information about them will continue to accrue, providing a rich source of patient phenotype data, which will allow future genetic and pharmacogenetic studies based on the stored genetic material. For example, genetic variants contributing to variation in BDR may be useful for predicting responsiveness, allowing more appropriately targeted medication.

Some asthmatics experience serious and sudden worsening of their illness. These exacerbations can become life threatening and may be preventable. In collaboration with statisticians from the Scientific core, we have developed models predicting risk of acute exacerbation among childhood asthmatics, using detailed and high-quality measurements obtained as part of a clinical trial (CAMP). Much of the information about exacerbation risk can be obtained from demography and past history. BD responsiveness gave additional useful information and specialised measurements gave only small additional incremental information gain. We will test these models in the Partners clinical data repository, and among our cohort to evaluate their generalisability.

Finally, in developing statistical models and evaluating the NLP phenotypes, it quickly became clear that a novel visualization tool was required. A prototype data viewer for the i2b2 clinical research chart developed by the Tools core, is now in production use for data validation and for exploring temporal relationships between events for groups of asthmatic patients.

Supported by grant U54LM008748.

Overview and Goals



The Figure shows the core components of our DBP. Centrally to our mission is a data warehouse (DW), populated with extensive clinical data on more than 97,000 asthmatics extracted from the existing Partners patient data repository (RPDR) shown on the right of the Figure). Many useful patient characteristics are already available in coded form, but some important confounding phenotypes needed for genetic and pharmacogenetic studies such as tobacco smoking history, were not available in coded form from the RPDR.

Our Natural Language Processing (NLP) expert colleagues from the Scientific Core 1 have extracted three of these, described in more detail below. In order to perform pharmacogenetic and genetic studies, we need a blood sample from a large number of asthmatics. These are currently being collected from the Partners Asthma Cohort described in more detail below. Acute asthma exacerbations are important and potentially preventable through more appropriately targeted treatment, so we have modelled these in the CAMP study described below and we will test those models using data from the DW and from the Partners Asthma Cohort (PAC). Additionally, we will validate the DW data using data obtained directly from the PAC. Finally, in order to better understand the DW data, we asked our Tool Core colleagues to develop visualization tools which allow us to see the relatively sparse clinical data and to quickly drill down to explore potential hypotheses.

Asthma DataMart – “Cough’n Shop”

The data warehouse (DW), populated from the existing Partners patient data repository is serving as a functioning prototype for the patient clinical data component of the i2b2 clinical research chart. It contains extensive clinical data on more than 97,000 asthmatics. Some important patient characteristics were not available in coded form from the DW, but these are being extracted using NLP methods (see below). Quality of the DW data will be evaluated by comparing information obtained from a small subset (about 2%) of the subjects who are currently being recruited as part of the PAC described in more detail below. The DW contains pulmonary function data where this has been recorded, as well as longitudinal collections of clinical progress notes in the Partners Longitudinal Medical Record (LMR) collected at each outpatient visit, and in discharge summaries which are prepared after every inpatient stay. The DW will allow us to continue to gather emergency room (ER) visits, outpatient and inpatient episodes and ongoing clinical measurements and text notes, making the PAC ever more valuable as data accrues over time with virtually no additional investment or cost.

In collaboration with physicians from Partners Asthma clinics, we are recruiting patients for the Partners Asthma Cohort (PAC) study. We are measuring lung function and bronchodilator responsiveness, obtaining questionnaire responses to establish life-long tobacco use history, medication history and medical emergency history, and we are obtaining a blood sample from which we will extract DNA and archive it in the Channing Laboratory sample storage facility. The PAC is an important part of our work because it will allow us to validate the quality of data in the DM by comparing information recorded in the DM with values obtained directly from the patient, and because it will allow us to search for association between genetic variants and a range of outcomes including pharmacological (eg BD responsiveness) and epidemiological (eg acute exacerbation risk).

NLP Phenotypes

In addition to conventional coded patient information for our analyses, the DM contains an extensive collection of LMR and discharge summary text which is an important source of patient information. Extracting this information requires Natural Language Processing statistical approaches. Using a variety of methods, and building on the GATE open-source NLP platform, a web service has been exposed which can be used to automate sophisticated NLP processing, creating coded phenotypes and posterior probabilities from text notes. To date, we have parsed 310,998 BWH notes and 552,245 MGH LMR notes as well as discharge summaries. The relatively unstructured notes offer a substantial challenge to NLP techniques, many of which have been designed to operate in relatively structured texts with limited variability. Negation is an important challenge in free-text semantics and we have evaluated four separate methods - NegEx, Neg-Expander, NegSVM, NegNaiveBayes. Each of the four algorithms' output was compared to the negation status assessed by two independent human reviewers. The regular expression and synthetic processing-based algorithms appeared to have better agreement (Kappa = 0.77 to 0.79) with the human reviewers than the classification-based algorithms (Kappa = 0.57 to 0.75), and so have been adopted for further use. We learned that the UMLS, which we are using as our source vocabulary, has many instances of terms which arise in highly specialized contexts and can give completely misleading matches in the LMR notes. For example, the common abbreviation Mr. is mis-interpreted as mental retardation. Other examples are shown in the figures below. We used HITEX to extract diagnoses and medications to a set of 300 outpatient notes, with the suppressible mappings suppressed and not suppressed. The unsuppressed parsing returned 9746 diagnoses and medications, while the suppressed parsing returned only 5126. A preliminary analysis of 100 of the findings omitted by the suppressed approach showed that 98 % of them were indeed wrong findings. The problem of assigning attributes to appropriate theme roles is also providing a substantial NLP challenge. As shown in the middle and right figures below, it can be difficult to set up rules which appropriately identify which party specific statements in a complex sentence should be assigned to. We are making substantial progress in this assignment problem, with code soon to be freely available in the forthcoming HITEX release.

Acute Exacerbations

Acute exacerbations of asthma are an important and potentially life threatening challenge. It is probable that many of these distressing and resource intensive episodes are potentially preventable through better management and possibly through more appropriate personalized targeting of medication to patients with genetic variations which make them more likely to gain benefit from specific drugs. For this reason, one of our goals is to study the effect of genetic variation on acute exacerbation risk as well as on measured response to specific medications. We have developed predictive models using demographic and other patient characteristics using the Childhood Asthma Management Program (CAMP) clinical trial data, in which detailed and very high quality phenotype data were collected over more than 4 years of the trial and to date, for more than 8 additional years of follow up. Taking the CAMP data as our Gold Standard, we will test the fit of these models in the DW data and in data collected by direct observation in the PAC. In brief, we found that most of the predictive information was contained in some relatively simple and inexpensive measures, while the BD response added useful additional information. Response to methacholine and other more expensive tests yielded small additional incremental information only.

Visualizing Data

The DM contains clinical and administrative data for more than 97,000 Partners patients who ever had an asthma ICD9 code assigned. This includes all hospital admission dates and ICD9 codes and discharge summaries, all outpatient visit LMR notes, all laboratory tests such as serum IgE and pulmonary function. We have added the NLP phenotypes derived from text notes to the material already available from the Partners RPDR and thus have an extremely rich source of patient characteristics for our work.

Although rich, this data is not readily amenable to descriptive analysis using the usual epidemiological tools. For example, one patient may have had lung function measured in 1993 and a serum IgE measured in 2002. If we try to look at all of the data for a large group of patients using a traditional cross-tabulation, taking the very important temporal element into account, we might show each measurement of interest for each year of observation – unfortunately, the vast majority of the cells of the table would be empty because of the different trajectories each patient has over time, the different way each physician behaves in terms of ordering tests and regular visits, and individual variation in the natural history of the illness. Thus our DM data are sparse when laid out using the usual descriptive methods.

This challenge has been addressed by a special purpose visualization tool developed by the Tool Core in response to our need to “see” the patient data in a readily comprehensible fashion. Using a timeline metaphor and building on the ontology tools already in existence for the Partners RPDR, our collaborators are developing a very generalized tool for exploring the clinical data we have in our airways disease DM as a prototype for other diseases and for incorporation into the i2b2 Clinical Research Chart.

The tool has already proven to be very valuable to our NLP expert colleagues, allowing us to identify NLP phenotypes which require closer investigation. For example, by looking for patients with inconsistent smoking status NLP phenotypes over time, such as a record of being a current smoker followed at some later date by a record indicating the patient was a lifelong non-smoker, we are able to go back to the original text notes to learn why the inconsistent results were created. This has helped our NLP colleagues identify specific challenges (described in the NLP Phenotypes section) such as negation and attribution which they are using to improve the accuracy of their tools. In response to this specific need, the Tool Core added a facility to the viewer which allows the user to click on any data point and see the original text note from which the value was derived. The texts are encrypted so the viewer displays unreadable text if the user does not have the appropriate encryption token, thus allowing the viewer to be used by a wide range of researchers while protecting patient privacy and IRB requirements – only authorized researchers will have access to the decryption keys required for specific patients. This will be an enormously important and valuable resource for clinicians and researchers alike.

Below are some ‘screenshots’ from the browser. The first shows a typical timeline display showing ICD9 asthma diagnoses on one line for each patient and NLP LMR and discharge summary text derived asthma diagnoses on another. Clearly, the NLP phenotype is seen far less often than the ICD9 code, but sometimes the NLP diagnosis is seen on an occasion when the ICD9 code was absent, suggesting greater sensitivity could be obtained by combining the two data sources. The second shows how multiple patient characteristics can be collected for display. The third shows the resulting display of a line indicating the first date and the last date of any record about each patient, together with two asthma diagnosis variables, allows us to examine the relationship while seeing where left and right censoring might be taking place – we cannot expect to see any diagnoses before the patient was first seen or after they stopped being seen.