Mining the Electronic Health Record

Steve Simon, Department of Biomedical and Health Informatics, UMKC

How many researchers…?

image of a light bulb

Let's start with a bad joke. How many researchers does it take to screw in a lightbulb? Fifteen. One to screw in the lightbulb and fourteen to serve as co-authors.

I have to admit I'm one of those fourteen. I written some of my own articles, but some of my best work is when I'm third, fifth, or eight co-author.

The best thing about being a statistician…

John Tukey

I'm in strong agreement with the famous statistician, John Tukey, who said "The best thing about being a statistician is that you get to play in everyone's backyard."

Backyard #1

Children's Mercy Hospital building

Here are some of the local backyards I've gotten to play in. Children's Mercy, …

Backyard #2

Cleveland Chiropractic College building

…, Cleveland Chiropractic, …

Backyard #3

MRI Global building

..., MRI Global, ...

Backyard #4

North Kansas City Hospital building

…, North Kansas City Hospital, …

Backyard #5

Saint Luke's Hospital building

..., Saint Luke's Hospital, ...

Backyard #6

Truman Medical Center building

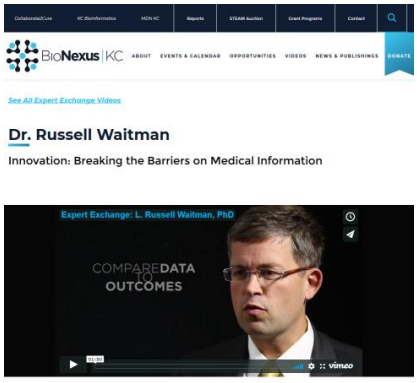
..., and Truman Medical Center, ...

My favorite two backyards

UMKC and KUMC mascots

… but my two favorite backyards are UMKC and KUMC. UMKC pays most of my salary, but I have worked on the committee that provided oversight to all the Data Safety and Monitoring Boards at KUMC. I co-authored a several publications on Bayesian models with KUMC faculty. I helped out with the National Database of Nursing Quality Indicators. I provided informal support to one of the first PhD graduates from the Biostatistics Department. So I consider myself to be half Kangaroo and half Jayhawk. Would that be a Kangjay or a Hawkaroo?

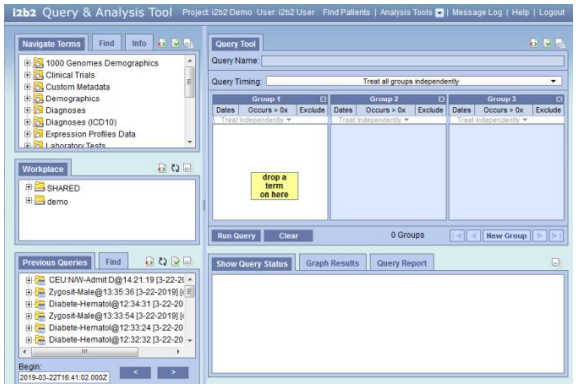
New backyard: Russ Waitman



Russ Waitman

So when this guy [pause] offered me a chance to work at KU Med Center in Medical Informatics, I said “Heck yes!” Now this wasn’t easy. I had child care commitments and so to find the time to work with Russ I needed to drop some really good things at UMKC. But I had to do this, because Medical Informatics is just like Statistics. They also get to play in everyone’s backyard.

i2b2 software



Screenshot of i2b2 software

The common standard for research using the electronic health record is i2b2.

It's an open standard which should be attractive to anyone in this audience who is a student. You learn i2b2 here, and everything you learn will transfer directly to the job you get after you graduate.

They are using i2b2 at some of the local playgrounds that I showed earlier: Children's Mercy, Saint Luke's, and Truman Medical Center. But the i2b2 system at KU Med Center is really great. They've included records from both the billing side and the clinical side. How cool is that! They've integrated the electronic health record with a cancer registry and a trauma registry. They have a slick system called data builder, which dumps the i2b2 records into a SQLite database that you can import into R, and they even let mere mortals like me query directly from the actual database.

The i2b2 software is great for targetted research involving a well structured research hypothesis. But for data mining, which is a more unstructured approach, you have to dig a bit deeper.

The database structure behind i2b2

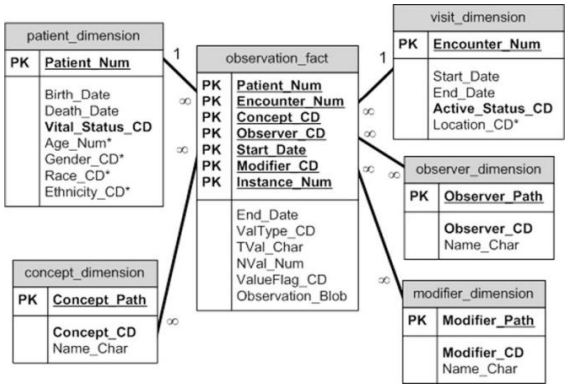


Diagram of i2b2 schema

Here's the database schema for i2b2. For something as complex as the Electronic Health Record, it's a very spartan design. It has to be, as we'll see in a minute.

How many surgeries?

```
select_surgeries <-
  "SELECT name_char FROM blueherondata.concept_dimension
     WHERE name_char LIKE '%ectomy%'"


dbGetQuery(c_connect, select_surgeries) %>%   # Extract records
  use_series(NAME_CHAR) %>%                   # Convert to vector
  strsplit(" ") %>%                           # Split into words
  unlist %>%                                  # Re-convert to vector
  tolower %>%                                 # Force to lower case
  grep("ectomy", ., value=TRUE) %>%           # Toss extraneous words
  gsub("[[:punct:]]", "", .) %>%              # Remove punctuation
  gsub("ectomy.*", "-", .) %>%                # Remove ectomy suffix
  unique %>%                                  # Remove duplicates
  sample(100, replace=FALSE) %>%              # Select 100 random
  sort %>%                                    # Arrange
  paste(collapse=", ")                        # Delimit with commas
```

Screenshot of SQL and R code

Here's something simple you can only do easily if you can directly query the database. This is a simple program for getting a list of certain types of operations. You're looking for the suffix "ectomy" which is Greek for "cut it out." Here's a bit of SQL code and some post-processing in R. You should be able to do this in SAS as well, but I haven't tested it yet.

How many surgeries?

"acromion-, adenoid-, alveol-, apic-, apico-, arthr-, arytenoid-, astragal-, ather-, burs-, capsul-, carp-, clitor-, coccyg-, crani-, dacryoaden-, dacryocyst-, diaphys-, disarticulationhemipelv-, disk-, diverticul-, endarter-, epididym-, epiglottid-, epiplo-, ethmoid-, fasci-, fistul-, frenul-, ganglion-, gastr-, gingiv-, gloss-, hemigastr-, hemigloss-, hemilamin-, hemilaryng-, hemiphalang-, hemorrhoid-, hepat-, hymen-, hyster-, infundibul-, irid-, labyrinth-, lamin-, lip-, lump-, mucos-, my-, myom-, nephr-, nephroureter-, oophor-, osteophyt-, pannicul-, patell-, phalang-, pharyngolaryng-, phleb-, pleur-, plex-, pneumon-, postadenoid-, postcholecyst-, postgastr-, postlymphaden-, postmastoid-, postpolyp-, postprostat-, postsplen-, prostat-, rectosigmoid-, salping-, salpingoophor-, scler-, segment-, sequestr-, sialoaden-, sigmoid-, sphenoid-, sympath-, synov-, tenon-, tenosynov-, trabecul-, trachel-, trisection-, trisegment-, turbin-, tyl-, tympanomastoid-, umbil-, urethr-, uvul-, vagin-, valv-, vas-, vesicul-, vulv-"

The full list of "ectomies" would be about twice as big. And that illustrates a key challenge with this type of data. It is very sparse. There are hundreds of things that a surgeon can cut out of you, but you should be very grateful that the doctor only chops out one or two things at the most. Looking at the drugs that you could get, even the worst polypharmacy cases would be a small fraction of thousands of drugs that are available. The same thing for diagnosis codes. So the design matrix for any regression model in this area becomes really huge and most of the entries are zeros.

The other thing about the electronic health record is that each operation, each drug, each diagnosis code sits on a separate record. It has to, or the record would string out so long that it would be unmanageable. So you can't say "Give me the record for a patient with sleep apnea getting Propofol anesthetic for a septoplasty." Instead, you have to match the apnea record with the Propofol record with the septoplasty records. This means using a self-join. Self-joins are tricky. They usually require using nested queries and they can be very inefficient.

I don't want to scare you away though. Mining the electronic health records has been one of the most fun backyards I've had a chance to play in.

Mei Liu, Acute Kidney Injury

Chen et al. BMC Medical Informatics and Decision Making 2018, 18(Suppl 1):13
https://doi.org/10.1186/s12911-018-0597-7

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Causal risk factor discovery for severe acute kidney injury using electronic health records

Weiqi Chen[1,2*], Yong Hu[1,2*†], Xiangzhou Zhang[1,2†], Lijuan Wu[1,2], Kang Liu[1,2], Jianqin He[1,2], Zilin Tang[1,2], Xing Song[3], Lemuel R. Waitman[3] and Mei Liu[3*]

From The 3rd China Health Information Processing Conference
Shenzhen, China. 24-25 November 2017

Abstract

Background: Acute kidney injury (AKI), characterized by abrupt deterioration of renal function, is a common clinical event among hospitalized patients and it is associated with high morbidity and mortality. AKI is defined in three stages with stage-3 being the most severe phase which is irreversible. It is important to effectively discover the true risk factors in order to identify high-risk AKI patients and allow better targeting of tailored interventions. However, Stage-3 AKI patients are very rare (only 0.2% of AKI patients) with a large scale of features available in EHR (1917 potential risk features), yielding a scenario unfeasible for any correlation-based feature selection or modeling method. This study aims to discover the key factors and improve the detection of Stage-3 AKI.

Methods: A causal discovery method (MCDSL) is adopted for causal discovery to infer true causal relationship between information buried in EHR (such as medication, diagnosis, laboratory tests, comorbidities and etc.) and Stage-3 AKI risk.

Mei Liu, next to one of her research publications

Now, what I've shown you so far isn't too fancy, but there's someone else in Medical Informatics, Mei Liu. She has several research publications on Acute Kidney Injury, a big NSF grant, and a PostDoc working with her. She's a lot smarter than I am, but I know all the good light bulb jokes.

Requirements needed for mining the electronic health record

Technical requirements

- Working familiarity with SQL
- Data wrangling skills

Non-technical requirements

- Lust for data
- An interesting backyard

So if you want to work in this area that I call "Mining the electronic health record" you need four things. You need to have a working knowledge of SQL. Not at the level of a database administrator, but you do have to know how to use the WHERE clause, the various types of joins, and a few other basic things. You also have to be pretty good at data wrangling. Data wrangling is a term that some guy (I'm sure it was a guy) developed to make the term data management sound more macho.

More important than these technical skills is that you have to get excited about data. If you weren't salivating at the code that pulled every whatever-ectomy out of the database, then maybe this isn't for you. The final requirement, though, and one that I am sure you already have, is an interesting backyard. I'm not a doctor, and all the examples I come up with on my own for mining the electronic health record are pretty trivial from a medical perspective. If you can bring me an interesting medical question, I can help you with the SQL and the data wrangling.

Informatics meetup

Flier for May 23 Informatics meetup

If you want to work in this area, we have scheduled an informatics meetup on May 23. We'll have speakers from the beginner's intermediate, and advanced spectrum of data mining. Mei Liu and I will be running the show and we'll bring along with a bunch of others in Russ's department. If you've got an interesting backyard, we want you to be there, too.

Where you can find a copy of this talk.

This presentation was developed using R Markdown. You can find all the important stuff at

- https://github.com/kumc-bmi/heron-i2b2-analytics

In particular, look for

- doc/mining-v2-image-credits.txt
- doc/mining-v2-slides.pptx
- doc/mining-v2-speaker-notes.pdf