Laboratory Measures, Data Cleaning, and CKD

08 May 2017

LABORATORY MEASURES

Problems with lab data

- What are laboratory measures?
- How do they change over time?
- When are they measured?

What are laboratory measures?

- A doctor orders a test and gets a value back.
- Generally speaking, this is the data available.
- For example, a clinician requests to know serum creatinine and receives the result of 1.0 mg/dL
- What's missing here?

Labs: Method

- Does the EHR report what method was used?
- In the serum creatinine example, there are two common methods:
  - Jaffe
  - Enzymatic
- These approaches can have different biases in general and in certain circumstances.
- Can yield differences in downstream analyses

"**Measurement of Serum Creatinine – Current Status and Future Goals"**
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1784008/

Labs: Reference Range

- What is an appropriate value for this individual?
- There are generic ranges that are expected
  - For example, 0.7-1.1 mg/dL serum creatinine would be acceptable
- Often, this can vary on factors like age and sex
  - 0.7-1.3 mg/dL for men
  - 0.6-1.1 mg/dL for women
- Was it calculated at first?
- Was that calculation correct?
- Are there other factors in play?

Labs: Limit of detection

- There are frequently limits to how small or large a value can be detected by a method or reported by a laboratory system
- Was there really 0 creatinine?
- How was reaching that limit represented?
  - Simply listed at 0?
  - <LOD?

Labs: Error bounds

- The accuracy of a value may vary on a number of factors
- The actual method of the test:
  - Who performed the test?
  - Where was it performed?
- The variation in the individual:
  - Rhythms of lab values through the day
  - Were they really fasting?
  - Unmeasured covariates

Labs: Units

- The same lab result can be returned with a variety of format changes.
- Units can change over time:
  - “Silent” changes: g/L to mg/mL
- The result type can change as well:
  - Positive and negative for some antibody tests as compared to titer values.

Changes over time

- There are many reasons any of these lab factors can change
  - Can be broad changes over time
  - Can change within the day
- Dealing with these changes can be complicated
  - Normalization
  - Adjustment

Normalization

	Normalization	Adjustment
Method	Yes	Yes
Limit of detection	Lossy	Value + Limit
Error bounds	-	Weighting
Units	Yes	-
Result type	Yes	-

CLEANING DATA FOR REUSE

Overview

- All data types share some factors
- Electronic health records are designed for treating patients, not research
  - Standardization is not always the most important
  - Humans can understand some natural variation better than computers
- Patients visit when the want or need to
  - May be healthy for long stretches
  - May be in the hospital for long stretches

Billing code data

- Billing systems are used to collect money
  - Rates can be negotiated
  - Revise previous reported values
- Can track diseases over time
- Typically fairly clean, though there are revisions over time
- One notable error is trailing and leading zeroes

Laboratory measures

- Laboratory measures were just covered in detail
- Combining across names is a big concern
- Outliers in results can be problems
  - Sometimes they are real
  - Often they are typos or other errors
- Lab systems can report “junk” values
  - “Canceled”
  - “Ordered in Error”
  - “Specimen Insufficient”

Medications

- Medication data has two primary cleaning issues:
  - Missing information
  - Likelihood of treatment
- To completely identify a medication, one needs to consider:
  - Dose
  - Route
  - Strength
  - Frequency
  - Duration

Medication sources

- Prescription orders
  - Free text
  - Electronic ordering systems
- Administration records
- Clinical note text
  - Asking patients
  - Doctors mention potential medications
- Pharmacy fill data

Likelihood of treatment

- How likely is the patient to have truly taken a medication?
- Was it mentioned as option or prescribed?
- Did the individual actually acquire the drug?
- Did they take it as directed?
- Did they take it at all?
- What if it was PRN?

Problem Lists

- How reliable is what's found here?
- Lag in updates
- Missing information

Clinical Notes

- Having the raw note text can be straight forward
- Cleaning may include:
  - HTML tags used in formatting the note
  - Splitting sections
- Normalizing note types
  - How hard is it to identify all CT scans?
  - How hard is it to identify all scans of the abdomen?

Other systems

- Are there other “silos” of data?
- Local registries
- National registries
- Clinical research repositories

STUDY ON CHRONIC KIDNEY DISEASE

CKDGen

- We were looking to participate in a consortium studying the genetics of Chronic Kidney Disease
- The study focused primarily on laboratory measures of kidney function:
  - Serum creatinine
  - Urine creatinine
  - Urine albumin
  - BUN
  - Uric Acid
- Several covariates
  - Hypertension
  - Gout
  - Diabetes

Clean lab values

- First off, I normalized laboratory measures
- Find and map lab tests to the study measures needed
- Clean out spurious results, eg “error values”
- Normalize the units for each test

Lab inclusion: outpatient only

- Due to the complications in inpatient data, we wanted to only select outpatient data
- Any lab value within an inpatient admission was removed
- Additionally, we removed those within 7 days of an inpatient stay
  - Could be leading to admission
  - The inpatient encounter may not be recorded completely

Lab inclusion: Altered status

- We also need to remove lab measures when an individual has had a kidney transplant or is on dialysis
- Remove values from two years prior or later from the first evidence of either
- Kidney Transplant Criteria
  - CPT of ('50360','50365')
  - ICD VCode 'V42.0'
- Dialysis Criteria
  - ICD VCode V56 or V45.1 or any of their child codes
  - ICD10 code Z99.2
  - CPT of ('90935','90937','90945','90947','90993')

Identifying confounding conditions

- For hypertension, selected the first date for either a code or medication.
- Individuals needed a hypertension ICD code (401.* and I10.*) AND any anti-hypertensive medication.
- For gout, the first date for one of the following codes: ICD9: ('274.0', '274.1', '274.8', '274.9') and ICD10: ('M10.0', 'M10.3', 'M10.4', 'M10.9')
- For diabetes, it was the first date of any 2 of the following:
  - Billing codes for any diabetes
  - Any medication for treating diabetes
  - Hemoglobin A1c meeting criteria

Inclusion

- Any individual with:
  - At least one non-excluded lab
  - Genotype data
- The first measure was selected as the baseline
- If there was a measure at least 2 years after baseline, that was used as a followup
- Previously had used median values

Using the measures

- eGFR was calculated using an adjusted method
- Some measures were analyzed using a residual approach
  - A regression model was fit predicting the measure using age, sex, and race
  - The difference between the predicted value and observed value (residual) was used
  - This value is essentially “How high or low is this person based on what we would expect given their age, sex, and race”
- Stratified based on diabetes status (and a merged test)