Needs Assessment / Data Analysis Investigation (Slack Chatlog)

Pauline [4:36 PM 10/30/2018]
@Johnny Bender, Here's our notes about our understanding of the data we need to present.

I think we should focus on the "table 1 of most epidemiological studies" and we should be able to derive everything else like snapshot prevalence and incidence over a period of time from table 1. It's tricky to define table 1. I think one way is to start by something like this (also see the attached sketch)?

1. For each Condition, find the Patient reference
   1.1 Extract gender, age, etc from Patient
2. For the same Condition as 1., find the Encounter reference,
   2.1 From Encounter, find the Observations references. We'll have Observations if, for example, a body check is performed. We may want to provide option to hide any of these.
3. By now we should have a database with as many rows as the number of items in Condition.ndjson
4. When we display on the dashboard, collapse all conditions with the same name and combine all observations values (e.g. take the mean?) It would be nice if we could sort by the SNOMED hierarchy tree, so that multiple similar/related conditions show up under one filter.
5. Snapshot prevalence and incidence over a period of time can then be derived from this database.

For example, we can look at Condition: 9c5b6e51-cd85-4854-b355-6c3e1af9e711 which is "Coronary Heart Disease". We can see that it's linked to "Encounter/9e1cd431-e771-4e9c-83cd-c56100980666". And then if we search for that encounter in Observation.ndjson we'll see a list of params like body weight, height, BMI, blood pressure, glucose, urea nitrogen, creatinine, calcium. The types of observations could be really long so we may want to think about how to display or not display them.

And then for the needs assessment tool, with large amount of data we could perhaps identify abnormal observations that may be a warning sign. Though honestly I'm still not really sure what's the exact exception of this needs assessment tool.

Thoughts?

Johnny Bender [12:16 PM 11/1/2018]
@Pauline Sorry it took me a bit to respond; I've been thinking about what you sent over. I'm not sure there's any benefit to stratifying by observations _within_ the condition identified. Rather, I think it would be beneficial to report out any population findings from these observations. For example, using BMI, we can calculate the prevalence (num/den/%) of *morbid obesity*, *obesity*, and *normal weight* for the population (although BMI is not

consistently measured, so it would be nice to supplement by using height / weight to calculate BMI on the fly).

If investigators were looking to understand associations between conditions and observations, then the comprehensive table you presented in the sketch would make sense. But I think it makes it easier to report on conditions stratified by *whole population*, *male/female*, and *age groups*

Are you all looking for a list of conditions to list in the table?

I could also see being able to sort the table on the fly based on the most prevalent conditions but that would require you to calculate prevalence for all conditions

I was thinking something like this for the table:
Prevalence (Stratify by age groups, sex):
- Numerator / Denominator (%) with health insurance
- Numerator / Denominator (%) without health insurance
- Numerator / Denominator (%) accessing healthcare
    o Should reflect sample represented by data collected from clinics / hospitals
    o Numerator = number of people with records at hospital / clinics in region
    o Denominator = number of people in region
    o Assume you have comprehensive healthcare data coverage for the region
- Numerator / Denominator (%) morbid obesity
- Numerator / Denominator (%) adult obesity
- Numerator / Denominator (%) adult overweight
- Numerator / Denominator (%) normal weight
- Numerator / Denominator (%) adolescent morbid obesity
- Numerator / Denominator (%) adolescent obesity
- Numerator / Denominator (%) adolescent overweight
- Numerator / Denominator (%) adolescent normal weight
- Numerator / Denominator (%) type-2 diabetes
- Numerator / Denominator (%) hypertension
- Numerator / Denominator (%) cardiovascular disease
- Numerator / Denominator (%) elevated cholesterol
- Numerator / Denominator (%) chronic liver disease
- Numerator / Denominator (%) chronic kidney disease
- Numerator / Denominator (%) end-stage renal disease
- Numerator / Denominator (%) kidney failure
- Numerator / Denominator (%) mental health diagnosis

Incidence (Stratify by age groups, sex):
- Number of cancer cases per 100,000 diagnosed in the past month
    o Would be good to rank / stratify by cancer types
- Number of tuberculosis cases per 100,000 diagnosed in the past month
- Adolescent births per 1,000 women aged 15-19 in the past month

These were the conditions mentioned most often in the needs assessments from Cameron County hospitals and Texas Region 5 (which includes Cameron County)
This is an example for incidence, but we could use a smaller timeframe than years, like months. Also, we would only include data from healthcare facilities at this time; data from outside sources could be added later to contextualize the local data.

Johnny Bender [12:49 PM 11/1/2018]
It would also be nice to report confidence intervals for each of the descriptive statistics
Resolving duplicates will be out of scope for this semester, but it's functionality we can add in the future

Johnny Bender [5:04 PM 11/1/2018]
Also @Pauline mapping out these statistics (like you had with the map of Texas) might be useful, but I'd like to be able to drill down to the hyper-local level. Based on the role of the user, they may even have access to the household-level data underlying these statistics. Household- or neighborhood- level data may be useful for targeting interventions at the local level. The prevalence of disease will probably track closely with the distribution of poverty.
I don't think a map is necessary for this semester, but if you really want to map (I agree, it's lots of fun), I would focus in on mapping Cameron County, Texas and representing granular census divisions (zip, city, neighborhood, block, tract, etc.) within the county level.

In the future, it would be cool to be able to map these stats on open-source maps that we could call on the fly, so public health agencies across the country could connect EHRs using Bulk FHIR and choose their relevant population from a picklist or something, or the analysis tool could geomap the patients and figure out the best map to display for the data sample
Also, final thing, if you all decide to use the mockup I provided, a team from a previous semester already created the UI code for the cards (allows you to drag and drop, etc.). You can find the code here: https://github.com/johnnybender/NutriFHIR (edited)

Also, on the UI above I got lazy and stopped using new numbers for the prevalence stratified by gender, etc. These will obviously be replaced with the real data :grin: