Clinical Data Wrangling

Session 5: Adding Hypertension

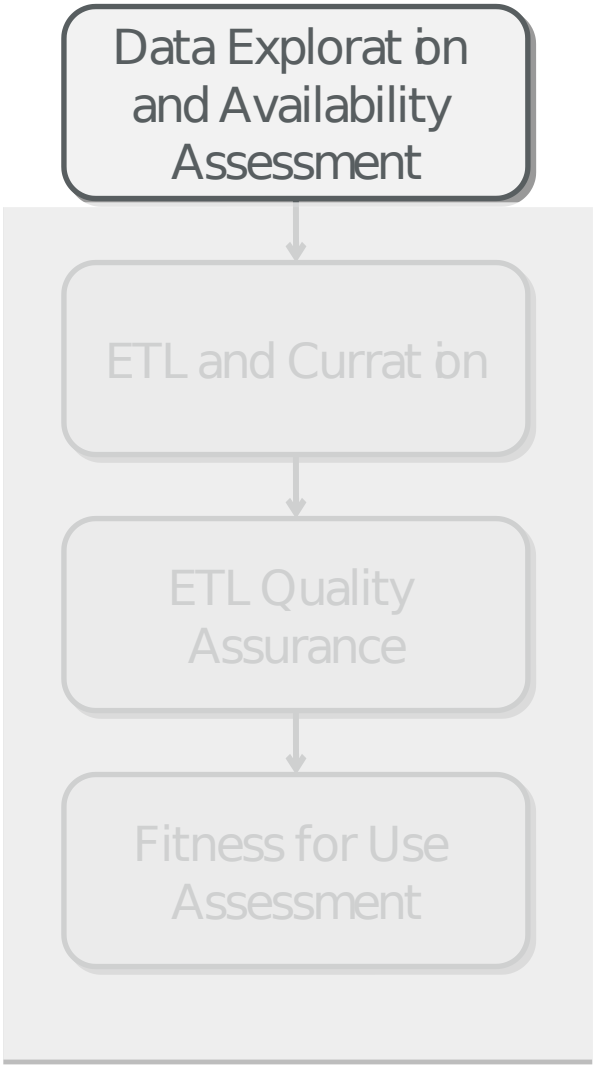
Overview of Wrangling Hypertension

Nicole G Weiskopf, 8/26/18

Wrangling hypertension

Research suggests that hypertension may be an important factor in understanding the impact of sleep apnea on cardiovascular risk.

We’re going to start with a quick overview of what hypertension is and why it might be important in our model from a physiological standpoint. Then we’ll briefly revisit the data wrangling pipeline before you all tackle the process in your groups.


Data Exploration and Availability Assessment
ETL and Curration
ETL Quality Assurance
Fitness for Use Assessment

OHSU

Where would you find a hypertension dx in a patient record?

- **Problem list**
- Admission / discharge diagnoses
- Billing data
- Unstructured data, like notes

OHSU

Decide what information from the EHR you would look for to establish diagnosis of HTN

Underdiagnosis of Hypertension Using Electronic Health Records FREE

Dipanjan Banerjee ✉, Sukyung Chung, Eric C. Wong, Elsie J. Wang, Randall S. Stafford, Latha P. Palaniappan

American Journal of Hypertension, Volume 25, Issue 1, 1 January 2012, Pages 97–102, https://doi.org/10.1038/ajh.2011.179

Published: 01 January 2012 **Article history** ▾

OHSU

Questions based on article:

EXERCISE: Answer the following questions

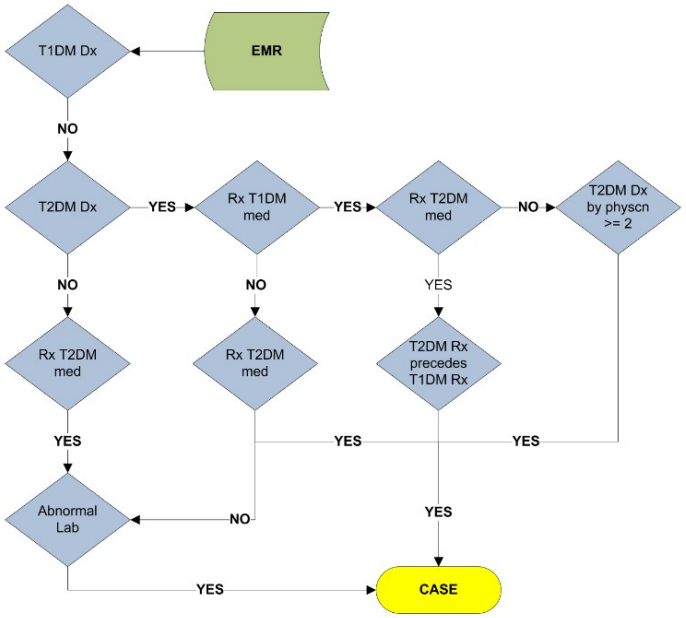
1. Is it sufficient to look just at the coded diagnoses? Why or why not?
2. What other sources of information would you consider? E.g., medications, vitals, labs, etc.
    1. You don’t need to be exhaustive

Generate a VERY simple algorithm for determining if a patient has HTN based on these clinical concepts

Remember the diabetes example we showed you.

Do something simpler than this.

Figure 1: Algorithm for identifying T2DM cases in the EMR.



EXERCISE: Create a simple algorithm to identify HTN cases in the EHR. Could be graphical or plain text, whatever is easiest for you.

7 Pacheco & Thompson. Northwestern University. Type 2 Diabetes Mellitus. PheKB; 2012.

Which of these clinical concepts are *available*?

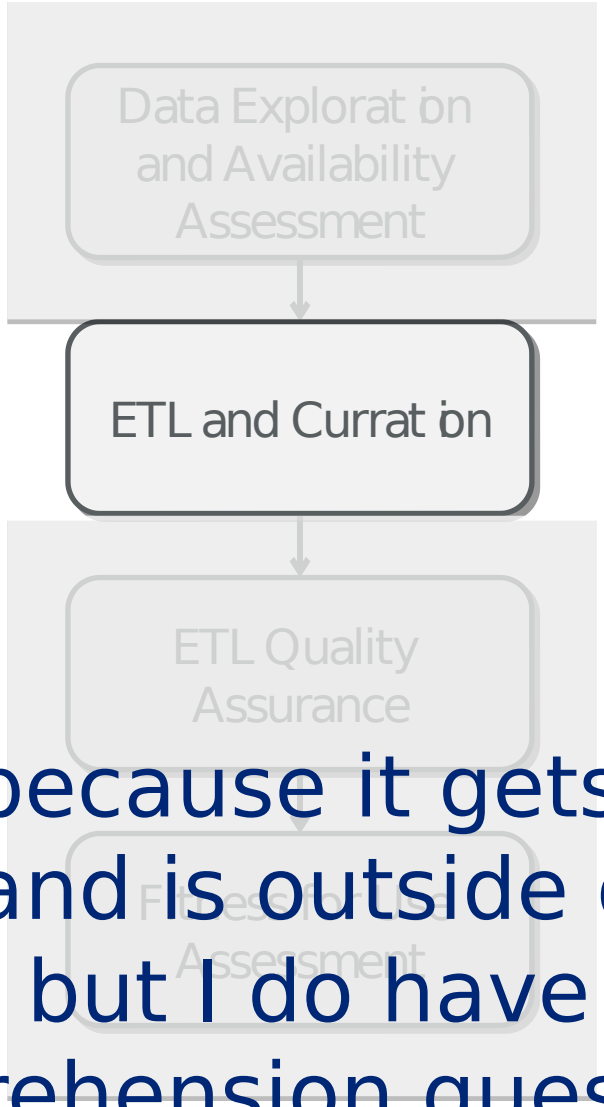
- In real life, this is a complex question to answer and can require a lot of digging through the EHR and talking to clinicians
- In our case, for the sake of argument, we're relying on the SHHS dataset.
- EXERCISE: Which covariates that you identified above are available in the dataset you've been working with?

What do our data say?

Exercise: Answer the following questions using the data explorer.

- How many patients have “official” hypertension in the SHHS dataset?
- Based on the other concepts you identified above, how many patients *should* have a diagnosis of hypertension? Hint: use the crosstab tool
- Spoiler alert: look at the definition of the HTN variable in the SHHS dataset

Exercise: based on what you found in the article and in your data, what's your final "algorithm" for determining who has hypertension?



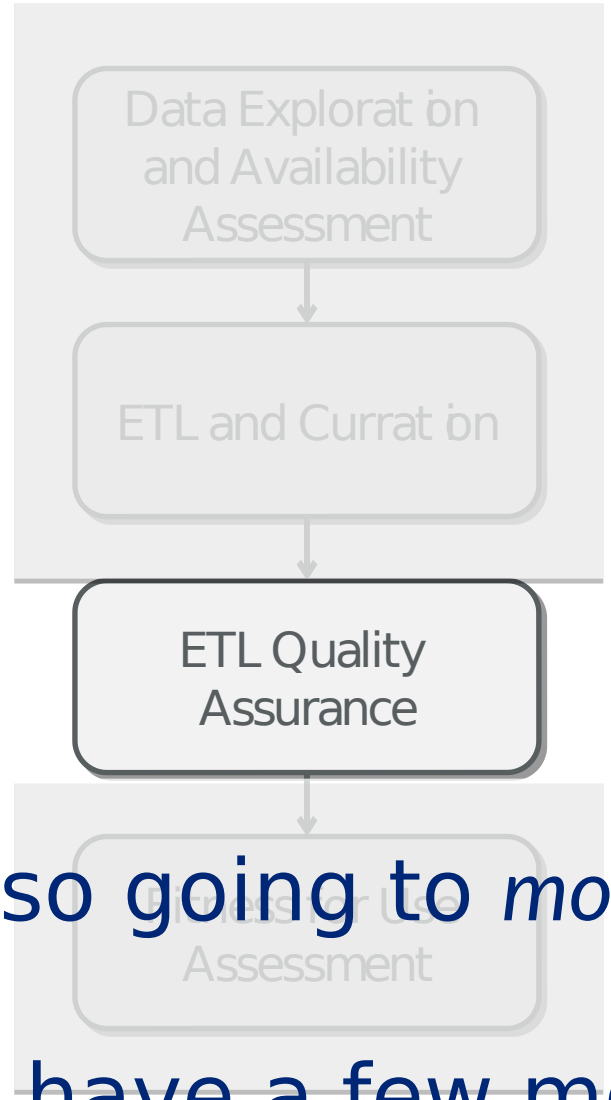
w

today because it gets more technical and is outside of current scope, but I do have a few comprehension questions.

OHSU

Exercise: ETL and Curration Questions

1. If you were trying to identify patients with hypertension in the EHR, would you do a text string search or search for specific diagnostic codes?
2. What does ETL stand for?
3. Most patients have more than one blood pressure recorded. How would you determine hypertension in such cases? Mean value, highest value, most recent value, etc.? And why?



We're also going to *mostly* skip

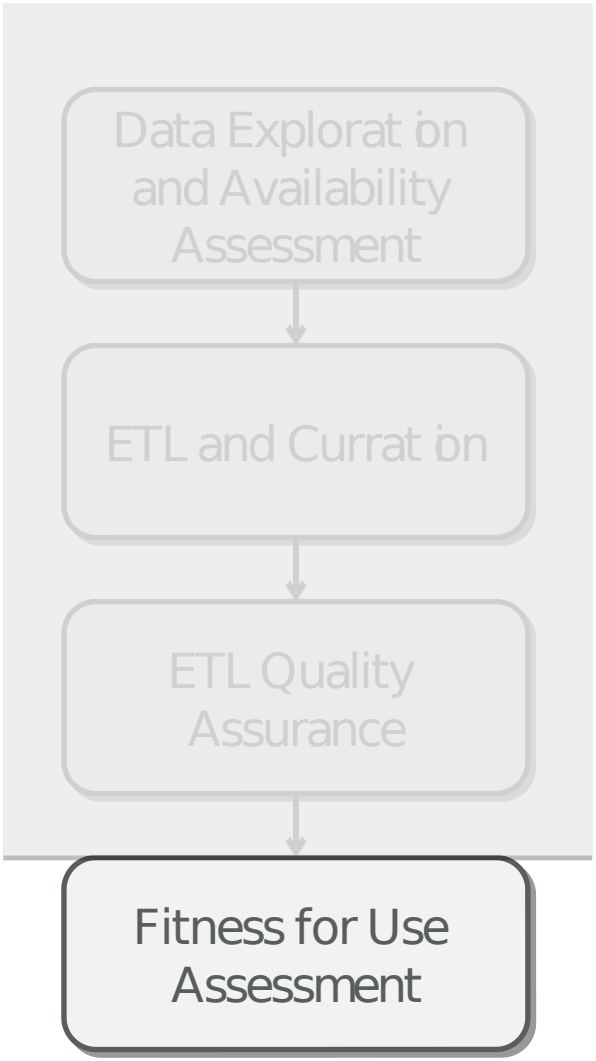
t

ETL, but I have a few more basic questions.

Exercise: Assessing ETL quality

1. Say you identify 1,000 patients in your EHR with a problem diagnosis of hypertension, but when you pull all systolic blood pressure values over 140, you have over 10,000. What possible reasons could there be for this? (Hint: think about question 3 from the ETL and Curration questions)
2. You plot your counts of coded HTN diagnoses from the EHR over time and notice a significant jump in 2017. What might have happened? (Hint: think about Eilis's intro to HTN)
3. You've double checked your ETL process and trust it. Your counts of "derived" HTN cases are quite a bit higher than your coded diagnosis, and you want to double check this. Assuming you have access to the EHR, what can you do?

OHSU


Data Exploration and Availability Assessment
ETL and Curration
ETL Quality Assurance
Fitness for Use Assessment

Fitness for Use

"Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features."

Redman, T (2001) Data quality: the field guide. Based on Juran's work.

OHSU

Fitness for Use

A combination of **data quality** assessment and assessment of **sufficiency** ("Do I have the data I need to answer the questions I want to answer?"). Our goal is to decide if the data of interest are "fit" for inclusion in our model.

For the *intrinsic* data quality component, Kahn et al (2016) is a good resource, though more complicated than you need at this stage.

OHSU

Basics of the Kahn et al. (2016) Harmonized DQ Model

Conformance:	Do data adhere to specified standards and formats?
Completeness:	Are data values present?
Plausibility:	Are data values believable?

Kahn MG et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of EHR

Exercise: Checking Conformance

Imagining that you are using EHR data, go through each of the clinical concepts you identified for inclusion in the HTN algorithm you developed above. For each variable:

1. What type (e.g. string, numeric, etc.) would you expect each variable to be.
   - Use the data explorer to check this for **one** of the variables
2. Identify which of these standards might be appropriate: ICD10, RxNorm, LOINC (Hint: you should be able to figure this out with a quick internet search)

Exercise: Checking Plausibility

1. What is the expected rate of hypertension in the overall US population?
2. What is the expected rate in EHR data (you can use the paper from above)?
3. How does the HTN rate in the SHHS dataset compare to these expected rates?
4. Based on these comparisons, do you trust the HTN data in SHHS? Why or why not?

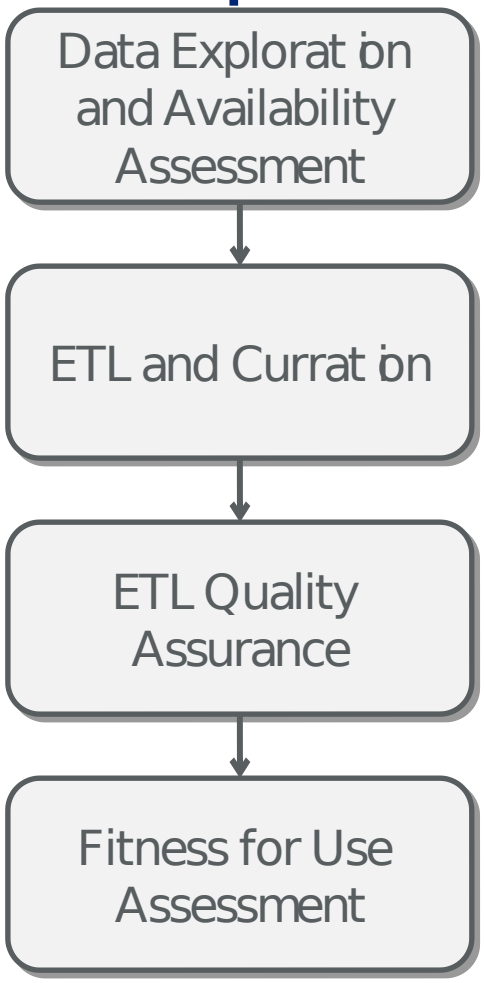
Exercise: Checking Completeness

1. For each of the variables you identified above to derive the presence of HTN, what percentage are missing or NA in the SHHS dataset?
2. Focusing just on the HTN variable in the SHHS dataset, explore missingness
   - Is there a relationship between the outcome variable and missingness of HTN?
   - What about the other important covariates. Do any of them drive missingness of HTN? Especially consider demographic covariates.

Make a final decision about fitness for use of diabetes concept

Reminder: we are not deciding hypertension diabetes should be included in the model, only if the data are good enough if we want to include it.

Make a final decision about fitness for use of diabetes concept



Did we find the appropriate sources for the concept of diabetes?

Do we believe that our ETL process was reliable and valid?

Do our data *conform* to required formats and standards?

Are the values of our data *plausible*?

Are our data sufficiently *complete*?

OHSU

Final Exercise: Determine fitness for use

1. Focusing specifically on the data in SHHS, would you consider the HTN variable fit for use?
2. Imagine that we were working with EHR data, like those described in the Banerjee et al. paper. Would you consider these "derived" HTN data fit for use?
3. For both of the above data sources, what caveats or assumptions would you keep in mind and include in a paper based on these data?