Clinical Data Wrangling

Session 3: Building the basic model

Applying the Data Wrangling Process

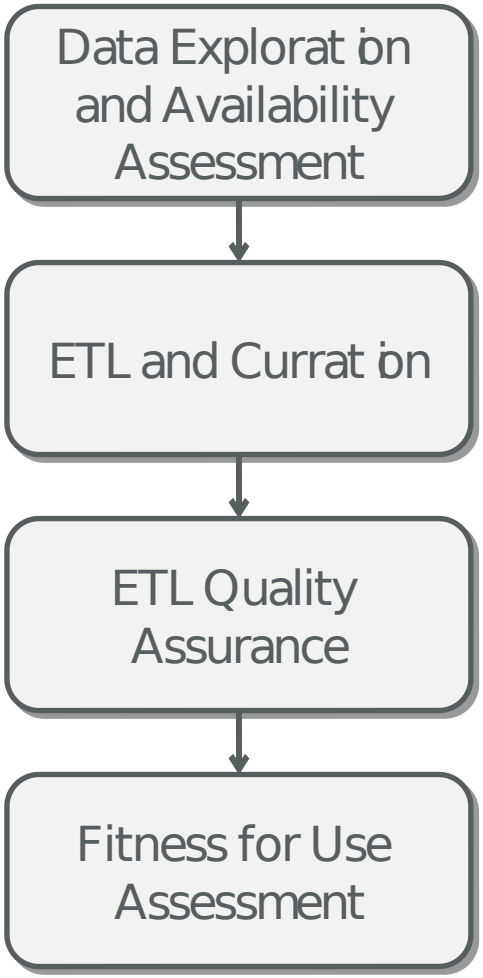
Nicole G Weiskopf, 8/21/18

Wrangling diabetes

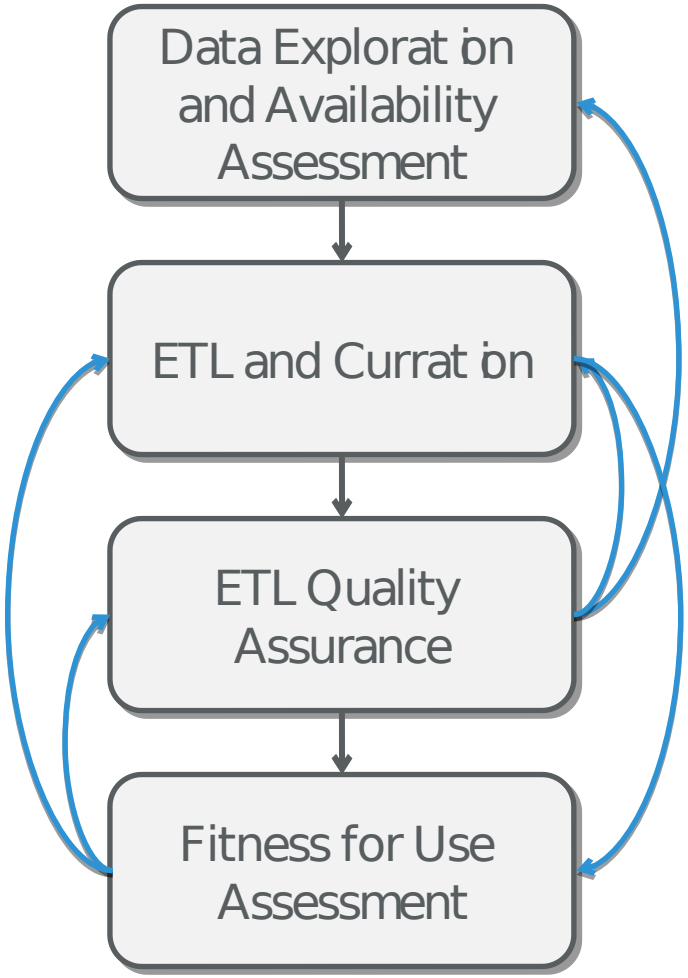
Research suggests that diabetes may be an important factor in understanding the impact of sleep apnea on cardiovascular risk.

Let’s walk through the process of wrangling this concept from a clinical dataset so that we can then determine if it adds predictive value to our model.

OHSU

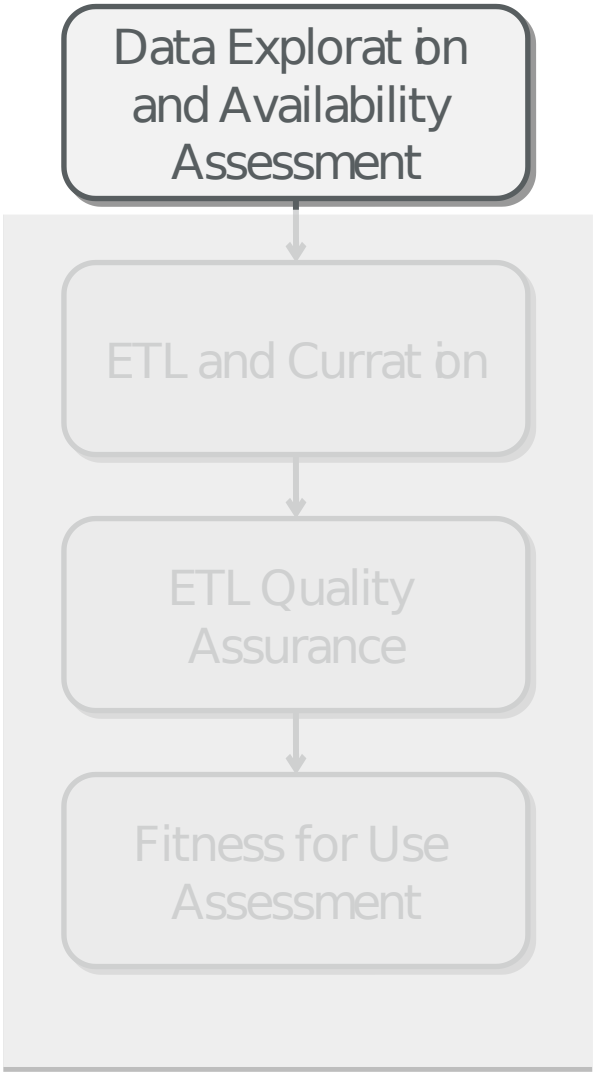

Data Exploration and Availability Assessment
ETL and Curration
ETL Quality Assurance
Fitness for Use Assessment

OHSU



The reality is a bit messier, but the process is *roughly* linear.

OHSU


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

OHSU

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

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

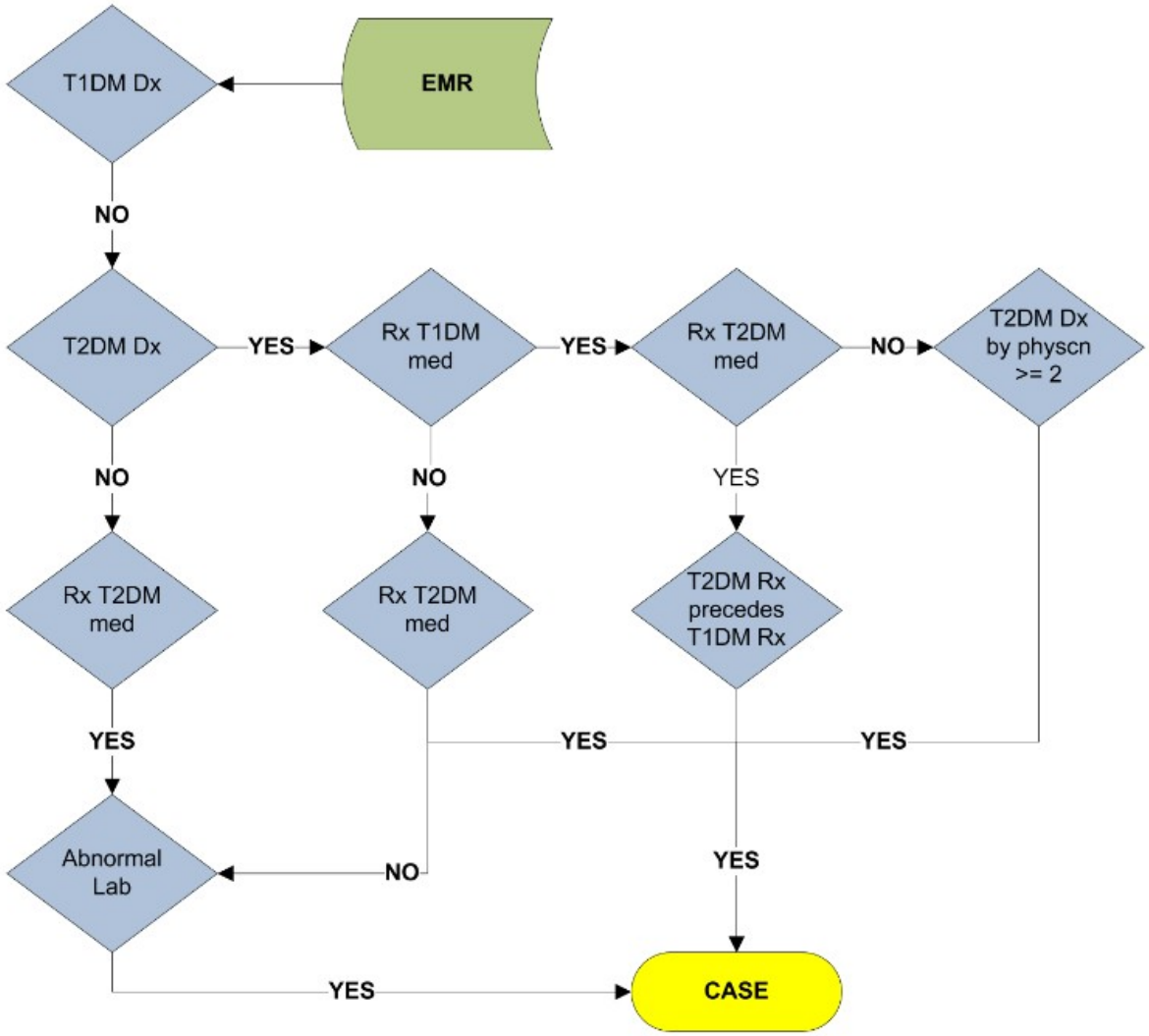
OHSU

Are there other indicators in the record suggesting diabetes?

- Medications:
  - Insulin
- Lab results:
  - HbA1c, blood glucose

OHSU

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



Jennifer Pacheco and Will Thompson. Northwestern University. Type 2 Diabetes Mellitus. PheKB; 2012 Available from: https://phekb.org/phenotype/18

OHSU

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 tracking data entry fields back to their location in the backend database.
- In our case, for the sake of argument, we're going to assume we have the following information:
  - Problem list
  - Most recent HbA1c
  - List of active medications

OHSU

Which concepts are *necessary* to determine if diabetes is present?

- How do we determine which data we need?
  - Talk to the experts (providers have strong opinions about this kind of thing)
  - Check the literature
  - Direct interrogation of the data

OHSU

What does the literature say?

Google

accuracy of diabetes diagnoses in problem list

All | Images | News | Videos | Shopping | More | Settings | Tools

About 36,600,000 results (0.28 seconds)

Showing results for accuracy of diabetes ***diagnosis*** in problem list
Search instead for accuracy of diabetes diagnoses in problem list

Problem list completeness in electronic health records: a multi-site ...
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4549158/
by A Wright - 2015 - Cited by 20 - Related articles
Jul 17, 2015 - And, of course, an **accurate listing** of **diabetes** on the **problem list** will .... the **problem list** from encounter-based **diagnosis** coding for billing; ...

OHSU

What does the literature say?

Table 2
Diabetes problem list completeness.

Site	Patients with at least 1 HbA1c > 7.0%	Patients with at least 1HbA1c > 7.0% AND diabetes on problem list *N* (%)
1	330	328 (99.4)
2	33,688	32,264 (95.8)
3	11,290	10,346 (91.6)
4	9585	8319 (86.8)
5	3503	2831 (80.8)
6	7337	5880 (80.1)
7	50,022	37,593 (75.2)
8	32,135	20,340 (63.3)
9	2001	1220 (61.0)
10	10,450	6290 (60.2)
Total	160,341	125,411 (78.2)[a]

[a] 78.2% is a weighted average of the completeness across the sites, weighting sites with more patients with high HbA1c's more heavily. The simple average across sites is 79.4%.

Wright A, et al. Problem list completeness in electronic health records: a multi-site study and assessment of success factors. *International journal of medical informatics*.

OHSU

What do our data say?

Select Crosstab Variable (x)

diabetes_dx

Select Crosstab Variable (y)

HbA1c_over_6.5

```
                HbA1c_over_6.5
diabetes_dx          No     Yes
         No        5327      70
        Yes          25     380
```

What do our data say?

Select Crosstab Variable (x)

diabetes_dx

Select Crosstab Variable (y)

HbA1c_over_6.5

	HbA1c_over_6.5	
diabetes_dx	No	Yes
No	5327	70
Yes	25	380

Diagnosis captures 84% of pts with high A1C, misses 16%.

Can we assume everyone with a high A1C has diabetes?

OHSU

What do our data say?

Select Crosstab Variable (x)

diabetes_dx

Select Crosstab Variable (y)

insulin

```
                 insulin
diabetes_dx           No     Yes
         No         5097     300
        Yes           81     324
```

What do our data say?

Select Crosstab Variable (x)

diabetes_dx

Select Crosstab Variable (y)

insulin

	insulin	
diabetes_dx	No	Yes
No	5097	300
Yes	81	324

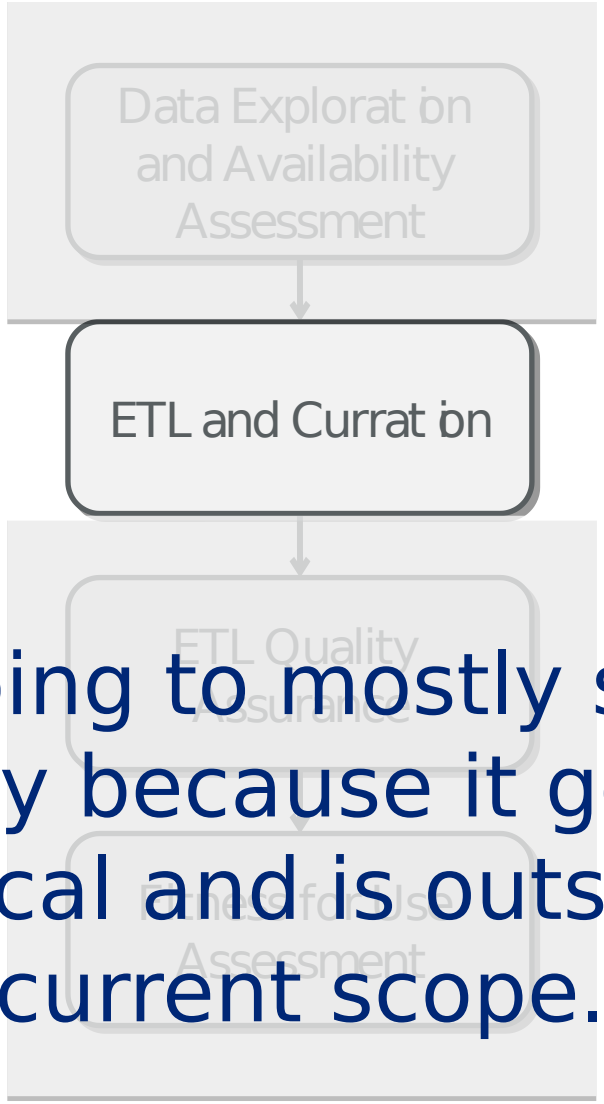
Diagnosis captures 52% of pts with high A1C, misses 48%.

Can we assume everyone on insulin has diabetes?

OHSU

So what’s our final decision about where to find info about diabetes in the EHR?

OHSU



We're going to mostly skip this step today because it gets more technical and is outside of current scope.

ETL and Curration Basics

- **E**xtract: pull desired data from source(s)
- **T**ransform: process extracted data into appropriate format
- **L**oad: insert transformed data into target resource

OHSU

Some Example SQL

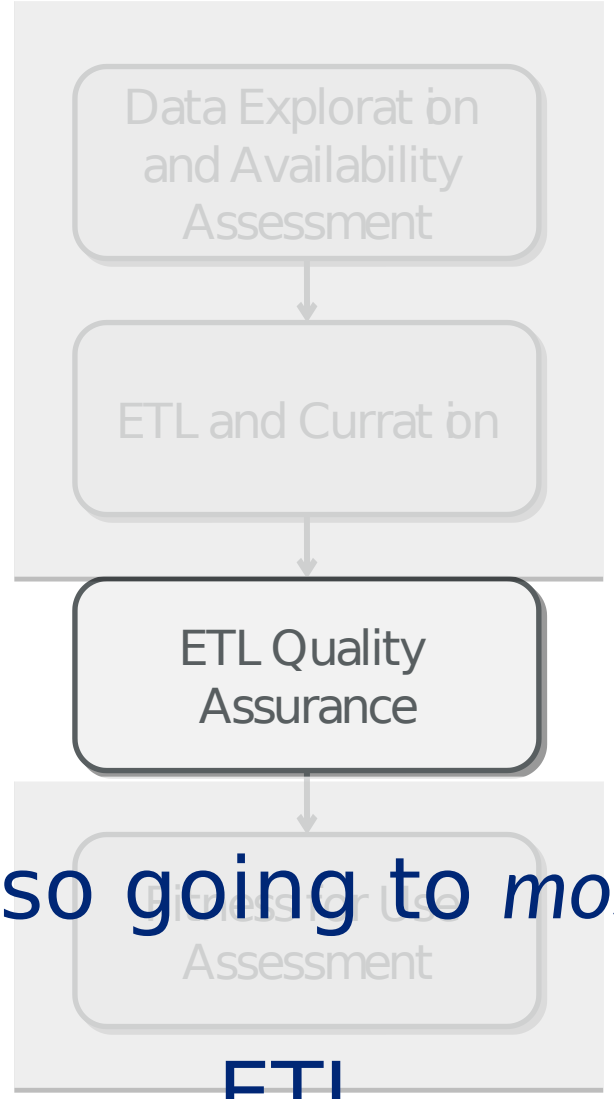
```
SELECT DISTINCT pid
FROM problemList
WHERE dxName IS LIKE "%Diabetes%"
```

This is bad. Don't do this.

```
UNION

SELECT DISTINCT pid
FROM labs
WHERE labName = "HbA1c"
    AND labValue > 6.5
```


OHSU



We're also going to *mostly* skip ETL.

t

Assessing ETL quality

Goal is to ensure you didn't lose or corrupt information during the ETL process. There is always the chance that you will identify preexisting data quality problems at this stage.

Here are some simple steps in order of increasing resource (time, effort) intensiveness.

1. Check that simple descriptive statistics (e.g., counts) match between final resource and source database
2. Check counts over time if you have temporal data
3. Look at the actual values! Do some simple distributions, bin the values, etc.
4. Spot check against the source data (e.g., manual chart review)

OHSU

Example of an ETL quality problem

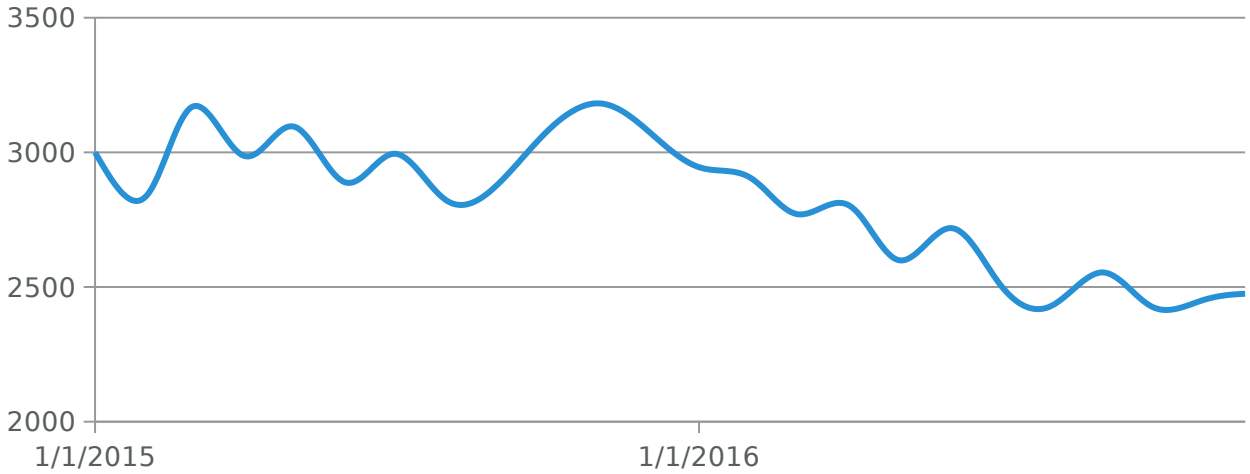
```
SELECT pid, labDate, labValue
FROM labs
WHERE labName = “HbA1c”
   AND labValue > 6.5
```

- Simple stats: counts of records match, but overall seem higher than we might expect
- Temporal trend: higher counts in earlier data
- Actual values: …

Example of an ETL quality problem

```
SELECT pid, labDate, labValue
FROM labs
WHERE labName = “HbA1c”
   AND labValue > 6.5
```

- Simple stats: counts of records match
- Temporal trend: number of results decreases over time…

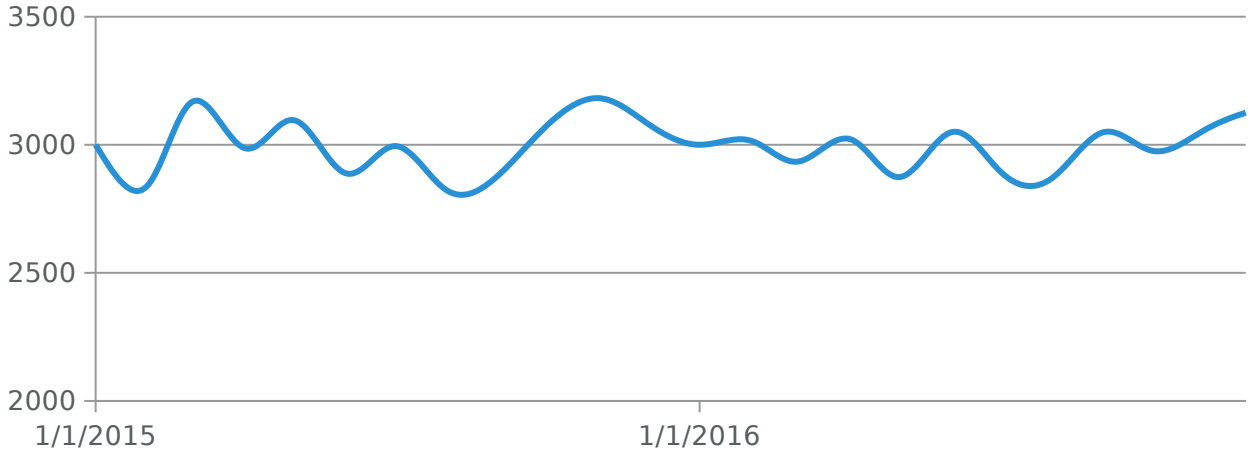


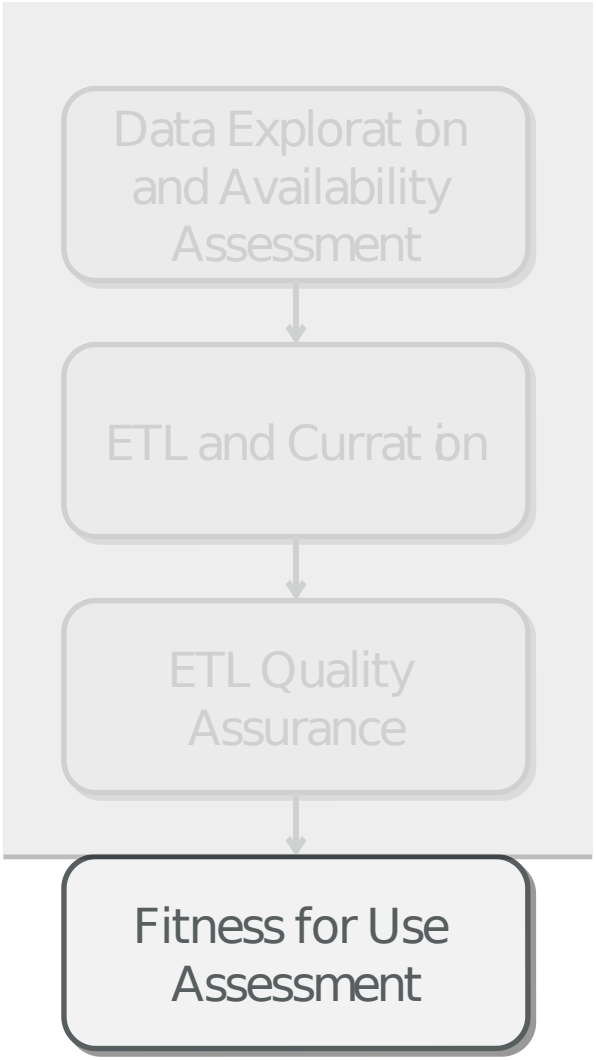
Example of an ETL quality problem

```
SELECT pid, labDate, labValue
FROM labs
WHERE (labName = "HbA1c" OR labCode = "4548-4")
   AND labValue > 6.5
```

Possible explanation: lab began relying more on LOINC codes.

Solution: run your queries again including LOINC code




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

OHSU

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?

Provides definitions and approaches to assess quality of data internally (“verify”) and externally (“validate”), against other sources of data or knowledge.

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

OHSU

Checking Conformance

- Check if all data are same **type**
  - If categorical, check that all values are permitted
- If you're using a data **standard**, check that all values are actually recorded in that standard

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

OHSU

Example of a conformance problem

```
SELECT pid, labName,
MAX(labValue)
FROM labs
WHERE labName = "HbA1c"
GROUP BY pid, labName
```

This query gives us the highest HbA1c value for each patient that has at least one HbA1c result

pid	labName	MAX(labValue)
123445	HbA1c	Done
124234	HbA1c	Done
123256	HbA1c	Done
765784	HbA1c	Done
453463	HbA1c	Done
458474	HbA1c	Done
456723	HbA1c	Done
999555	HbA1c	Done
839843	HbA1c	Done

What happened? How do we fix it?

```
SELECT pid, labName,
MAX(labValue)
FROM labs
WHERE labName = "HbA1c"
   AND valueType = "numeric"
GROUP BY pid, labName
```

pid	labName	MAX(labValue)
123445	HbA1c	Done
124234	HbA1c	Done
123256	HbA1c	Done
765784	HbA1c	Done
453463	HbA1c	Done
458474	HbA1c	Done
456723	HbA1c	Done
999555	HbA1c	Done
839843	HbA1c	Done

OHSU

Checking Plausibility

- There is **concordance** between different variables (e.g. diagnoses and lab results)
- **Distributions** of values match expected distributions
  - Can be based on general knowledge, other clinical data sources, registry data, etc.

OHSU

Checking Plausibility

Plausibility (aka correctness) is difficult to check. We have no gold standard with clinical data reuse. The underlying biomedical state of a patient cannot be observed, but only approximated via the data we have available.

- There is **concordance** between different variables (e.g. diagnoses and lab results)
- **Distributions** of values match expected distributions
  - Can be based on general knowledge, other clinical data sources, registry data, etc.

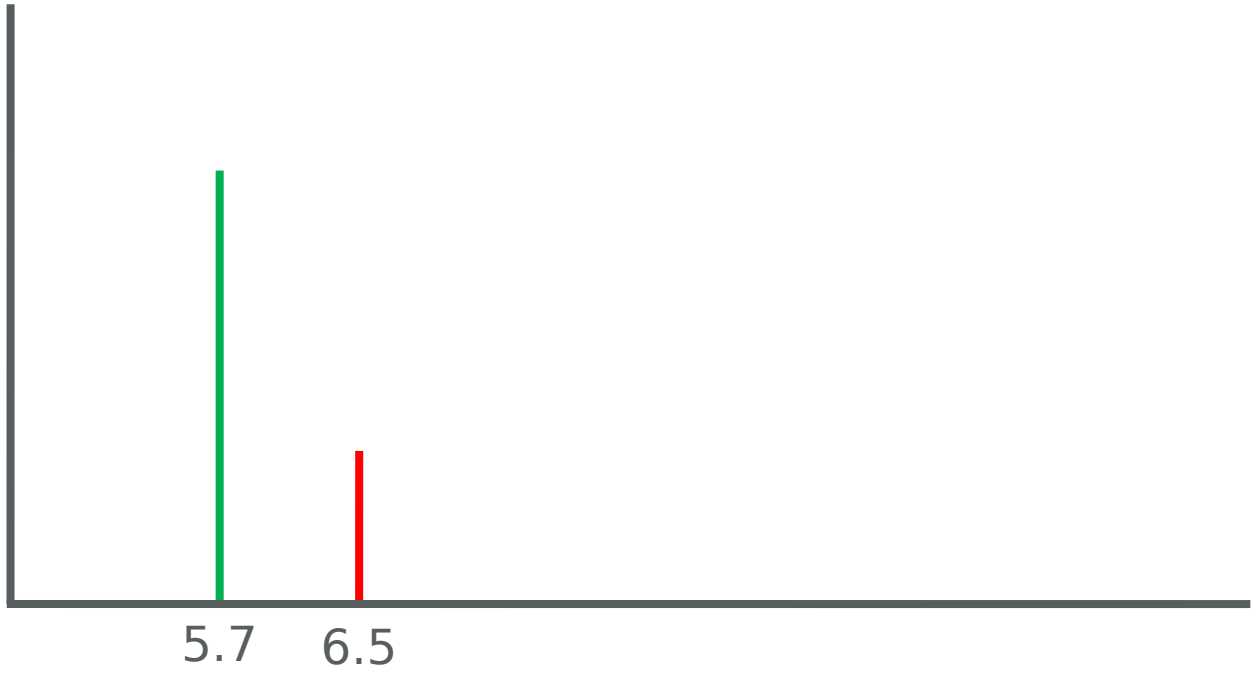
OHSU

Checking Plausibility

- How could we check the plausibility of HbA1c values? Can we compare to expected distributions?
- I looked but couldn't find a good reference distribution for a1c values.
  - Reported ranges are mostly either of healthy cohorts or people with diabetes.
- But we *do* know the percent of the population with diabetes, so we can bin the a1c values and see if they reflect that.

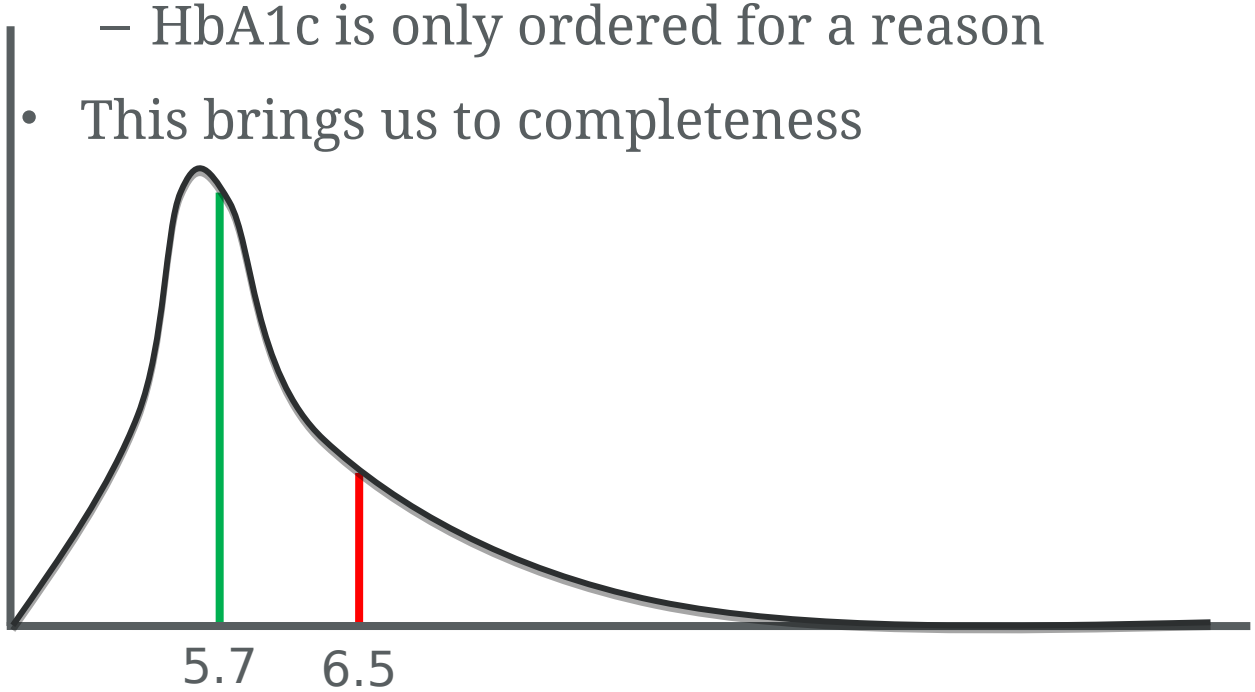
Checking Plausibility

- From CDC, about 9.4% of the population has diabetes.
- What do our data show?



Checking Plausibility

- What does this mean? Are the A1c values “bad”?
- Possible explanations:
  - Our patients are sicker than average population
  - HbA1c is only ordered for a reason
- This brings us to completeness



Checking Completeness

- Well under half our patient population has a numeric HbA1c lab result
  - By some definition, it is *missing* for most patients
- What form of missingness do you think this is?
  - Some combination of MAR and MNAR

OHSU

Checking Completeness

Options for managing missing A1c:

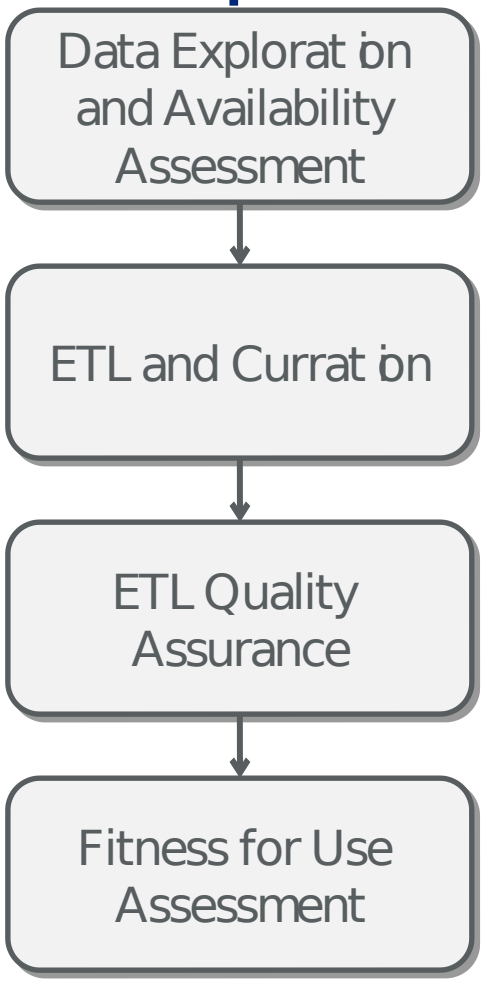
- Don’t use the variable
- Drop all patients without it
- Assume that the *absence* of A1c has inherent meaning
  - This is essentially what we’re doing when we combine A1c and diabetes diagnosis as a single dichotomous variable
  - Be careful not to conflate omission and negation

OHSU

Make a final decision about fitness for use of diabetes concept

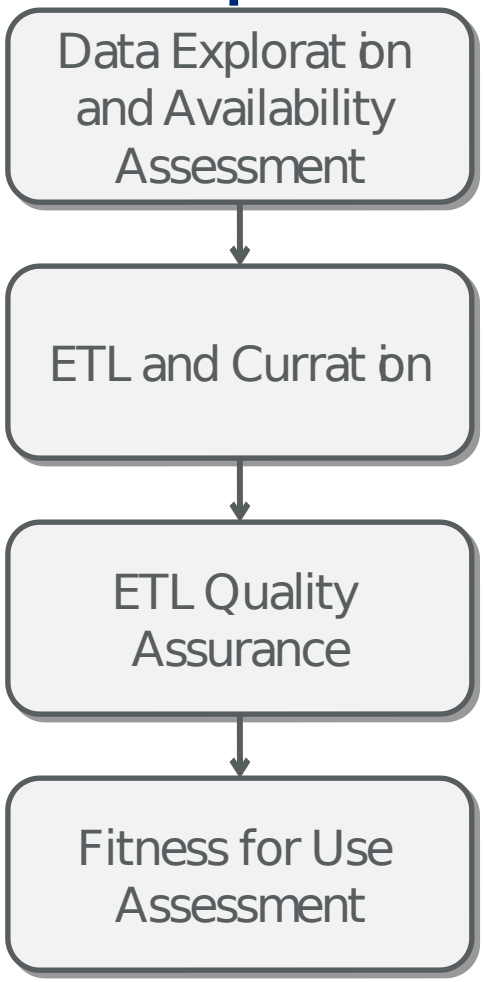
Note: we are not deciding if diabetes should be included in the model, only if the data are good enough if we want to include it.

Make a final decision about fitness for use of diabetes concept



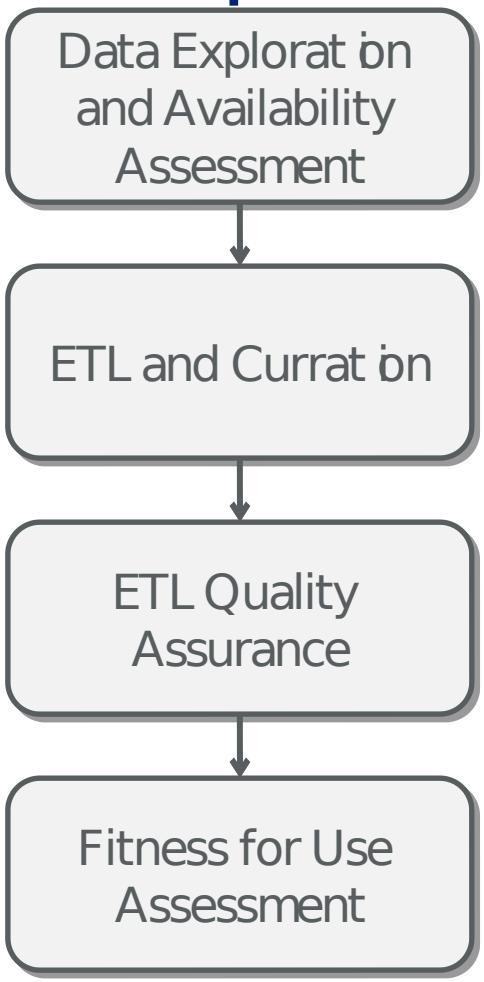
OHSU

Make a final decision about fitness for use of diabetes concept



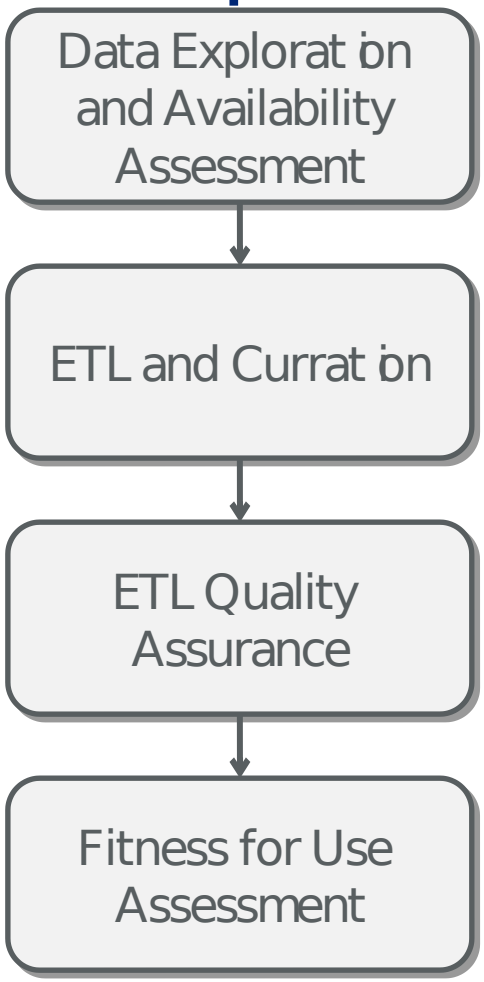
Did we find the appropriate sources for the concept of diabetes?

Make a final decision about fitness for use of diabetes concept



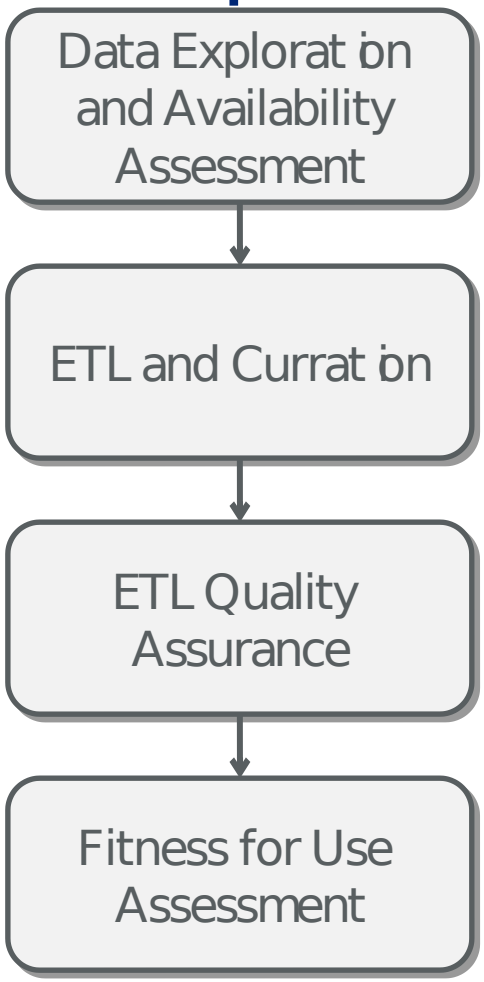
Do we believe that our ETL process was reliable and valid?

Make a final decision about fitness for use of diabetes concept



Do our data *conform* to required formats and standards?
Are the values of our data *plausible*?
Are our data sufficiently *complete*?

Make a final decision about fitness for use of diabetes concept



What would you do?