UNC SCHOOL OF MEDICINE
North Carolina Translational and Clinical Sciences Institute

2023 PPMH Data Literacy Workshop: Session 3

Peter Leese
The Data Science Lab in
NC TraCS Institute

Section I

Identifying patients

Cohorts and convenience

We want to study diabetic patients.
I use E08 when I bill so that's how
we'll identify them.

Grant

...patients with
diagnosis code E08

Good cohorts & translation

- This is hard and complex

Good cohorts & translation

- This is hard and complex
  - My study needs diabetic patients.

Which type → type I, type II, gestational?

Good cohorts & translation

- This is hard and complex

  Which type → type I, type II, gestational?

  – My study needs diabetic patients.

  Defined how → dx, meds, labs, etc?
  And what criteria?

  – My study needs type II diabetic patients.

Good cohorts & translation

- This is hard and complex

Which type → type I, type II, gestational?

  - My study needs diabetic patients.

Defined how → dx, meds, labs, etc?
And what criteria?

  - My study needs type II diabetic patients.

Recent? Once? >Once?

  - My study needs DM II patients from A1C labs.

Cohort choices have effects

- Each definition choice affects cohort
  - DM II based on single diagnosis code
    - Larger but lower confidence (TP and FP high)
  - DM II based multiple, repeated positive labs
    - Smaller and higher confidence (FP very low)
- Important to design a phenotype around needs for sensitivity and specificity

Moving to Phenotyping

- Phenotyping (and cohorting)
  - Process of identifying patients for study
- Computable phenotype
  - Computerized (reusable) queries or algorithms to identify patients, events, or diseases from electronic data
- Phenotyping now the expectation for EHR research (and maybe more)

Ex. Type II Diabetes Phenotype

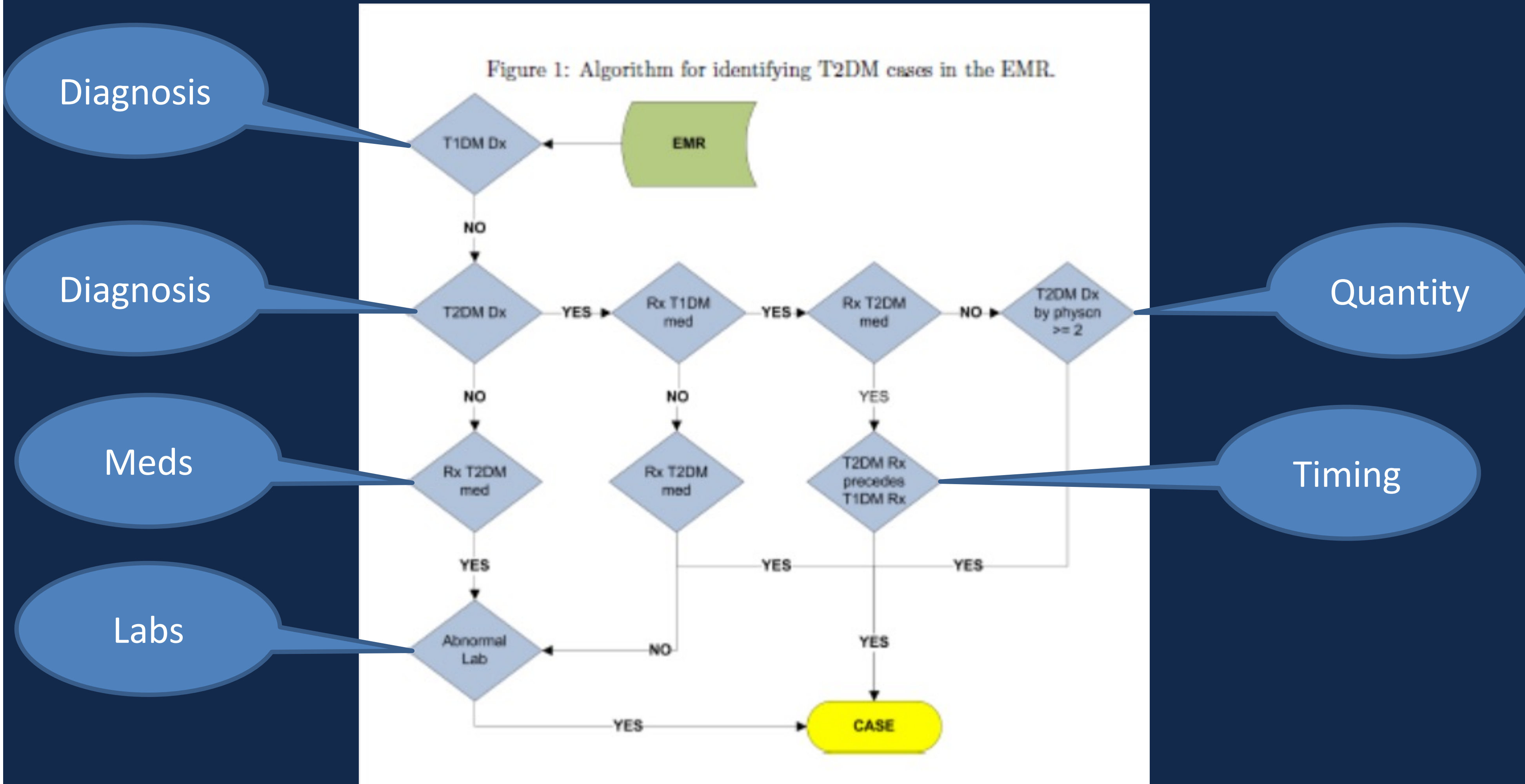


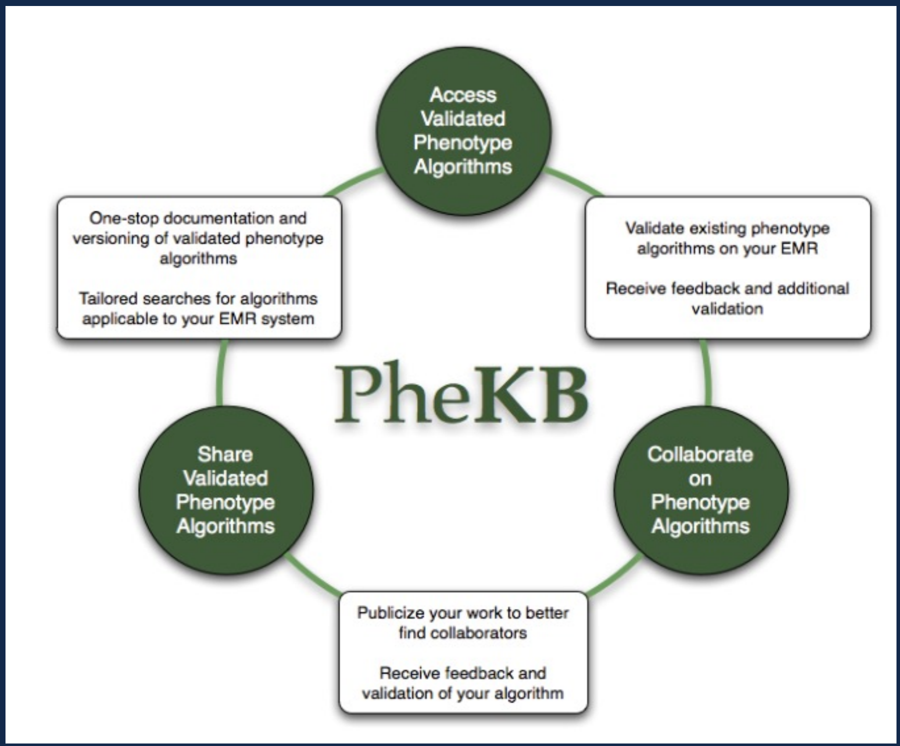
Figure 1: Algorithm for identifying T2DM cases in the EMR.

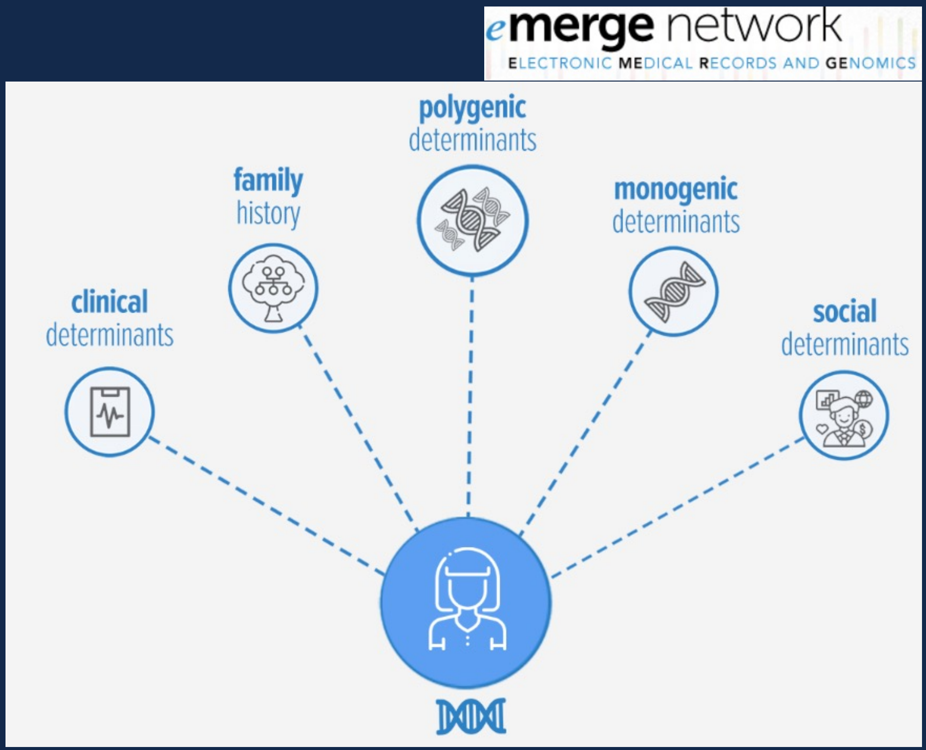
Phenotype characteristics

- Exhaustive criteria
  - Multiple data domains
  - Multiple criteria
  - Often algorithmic
- Scientific over convenient
  - Ideally validated to gold standard
  - Test characteristics ideally measured

Finding high-quality phenotypes

- Literature (pubmed, google scholar)
- Phenotype KnowledgeBase (phekb.org)
- eMERGE network





PheKB

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

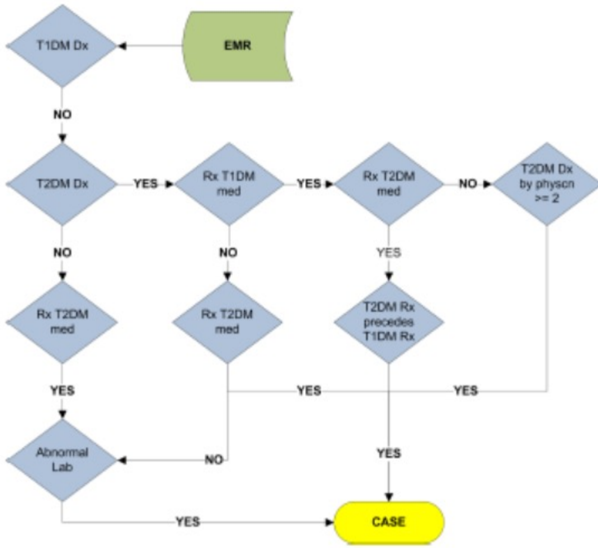
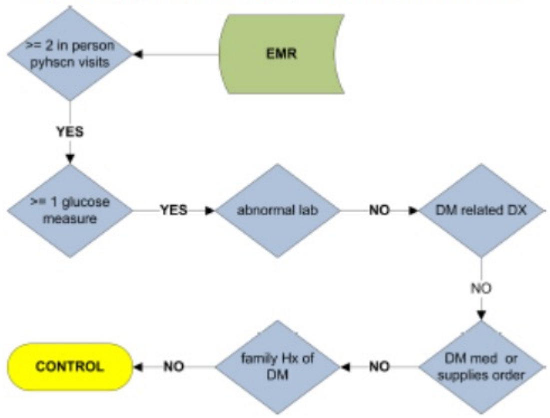


Figure 2: Algorithm for identifying T2DM controls in the EMR.



Phenotype ID: 18

Status:
Final
Do Not List on the Collaboration Phenotypes List

Type of Phenotype:
Disease or Syndrome

Phenotype Attributes:
ICD 9 Codes
Laboratories
Medications

Authors: Jennifer Pacheco and Will Thompson

Contact Author:
Jen Pacheco

Files:
- T2DM Algorithm
- Data Dictionary
- DiabetesChartReview-AbstractionForm7_19_10_Marshfield.doc
- DiabetesChartReview-CodeBook7_22_10_Marshfield.doc
- KNIME workflow with T2DM algorithm logic
- example potential cases file for input into KNIME workflow
- example potential controls file for input into KNIME workflow
- UPDATED list of ICD diagnosis codes inc. ICD-10

Institution:
Northwestern University

Date Created:
Monday, February 6, 2012

URLs:
https://phekb.org/phenotype/emerge-omop-tes
phenotype

Age:
Adult

Network Associations:
eMERGE

Owner Phenotyping Groups:
eMERGE Northwestern Group

View Phenotyping Groups:
eMERGE Phenotype WG

Data Model:
OMOP

PubMed References

1. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus.
Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, Yawn BP, Pacheco JA, Chute CG.
J Am Med Inform Assoc. 2012.
PMID: 22249968

2. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study.
Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, Denny JC, Peissig PL, Miller AW, Wei WQ, Bielins Chute CG, Leibson CL, Jarvik GP, Crosslin DR, Carlson CS, Newton KM, Wolf WA, Chisholm RL, Lowe WL.
J Am Med Inform Assoc. 2012.
PMID: 22101970

Section II

Walk through some examples

Framing

- Goal is to understand the perspective and how this happens – not to become an expert.

Example 1

- I need hypertensive patients.

Example 1

- I need hypertensive patients.
  - Adults, kids, other person level criteria?
    - Does condition look different in different people?

Example 1

- I need hypertensive patients.
  - Adults, kids, other person level criteria?
    - Does condition look different in different people?
  - Diagnoses, labs, meds?
    - You'll need to know (or learn) about these

Example 2

- I need adult acute covid patients.

Example 2

- I need adult acute covid patients.
  - Covid defined as diagnosis or lab?
    - When was dx code available?
    - When were labs available?
      - What happened to labs?
    - Does the strain matter?

Example 2

- I need adult acute covid patients.
  - Covid defined as diagnosis or lab?
    - When was dx code available?
    - When were labs available?
      - What happened to labs?
    - Does the strain matter?
  - Identify by meds?
    - Who gets med and when…and what could that do?

Example 3

- I need adult long covid patients.
  - Long covid defined as…
    - Diagnosis?
    - Lab?
    - Med?

ELSEVIER

The Lancet Digital Health
Volume 4, Issue 7, July 2022, Pages e532-e541

Articles

Identifying who has long COVID in the USA: a machine learning approach using N3C data

Emily R Pfaff PhD [a] *, Andrew T Girvin PhD [b] *, Tellen D Bennett MD [c d], Abhishek Bhatia MS [i], Ian M Brooks PhD [e], Rachel R Deer PhD [j], Jonathan P Dekermanjian MS [f], Sarah Elizabeth Jolley MD [g], Michael G Kahn MD [c], Kristin Kostka MPH [k], Julie A McMurry MPH [h], Richard Moffitt PhD [l], Anita Walden MS [h], Prof Christopher G Chute MD [m], Prof Melissa A Haendel PhD [h]
The N3C Consortium†

Show more

Example 3

- I need adult long covid patients.
  - Long covid defined as…
    - Diagnosis?
    - Lab?
    - Med?

Example 3

- I need adult long covid patients.
  - Long covid defined as…
    - Diagnosis?
    - Lab?
    - Med?

ELSEVIER

The Lancet Digital Health

Volume 4, Issue 7, July 2022, Pages e532-e541

Articles

Identifying who has long COVID in the USA: a machine learning approach using N3C data

Emily R Pfaff PhD [a] *, Andrew T Girvin PhD [b] *, Tellen D Bennett MD [c d], Abhishek Bhatia MS [i], Ian M Brooks PhD [e], Rachel R Deer PhD [j], Jonathan P Dekermanjian MS [f], Sarah Elizabeth Jolley MD [g], Michael G Kahn MD [c], Kristin Kostka MPH [k], Julie A McMurry MPH [h], Richard Moffitt PhD [l], Anita Walden MS [h], Prof Christopher G Chute MD [m], Prof Melissa A Haendel PhD [h]

The N3C Consortium[†]

Show more

Example 4

- Patients that got ICU care.

Example 4

- Patients that got ICU care.
  - Admitted, discharged? How to identify?

Example 4

- Patients that got ICU care.
  - Admitted, discharged? How to identify?

Example 4

- Patients that got ICU care.
  - Admitted, discharged? How to identify?
  - How to identify transfers?

Example 4

- Patients that got ICU care.
  - Admitted, discharged? How to identify?
  - How to identify transfers?
    - Diagnosis, procedures?
    - What other options?

Example 4

- Patients that got ICU care.
  - Admitted, discharged? How to identify?
  - How to identify transfers?
    - Diagnosis, procedures?
    - What other options?
    - ADT data, granular billing, proxies

Example 5

- Patients that came to ED for avoidable reasons

Example 5

- Patients that came to ED for avoidable reasons
  - What's avoidable?
    - Natural language answer
    - 'Conditions or reasons a clinician would deem not requiring emergency care'

Example 5

- Patients that came to ED for avoidable reasons
  - What's avoidable?
    - Natural language answer
    - 'Conditions or reasons a clinician would deem not requiring emergency care'
    - Data answer → typically requires developing algorithm to replicate clinical knowledge

Example 6

- Homeless patients
  - Diagnosis code?
  - Home address?
  - Documented in note?

CLINICAL DATA LITERACY SERIES: ELECTRONIC HEALTH DATA BASICS

Date	Topic	Instructor(s)
Wed May 10, 2:30-4:00pm	How health care system generates data and how this data is stored in the EHR	Peter Leese
Wed May 17, 2:30-4:00pm	code sets used to record health care data	Emily Pfaff
Wed May 24, 2:30-4:00pm	fundamental units of how health care data is organized in the EHR	Peter Leese & Emily Pfaff
Wed May 31, 2:30-4:00pm	how to design a research question for clinical data	Michael Adams & Anna Jojic

Helpful Resources Handout

Download at bottom of series webpage

https://go.unc.edu/clinical-data-literacy

EHR Data Driven Research:

Progress, not Perfection

Emily Pfaff, PhD, MS
Assistant Professor, UNC Chapel Hill School of Medicine / Co-Director, Informatics & Data Science @ UNC's CTSA

title

"This data is junk!"

Questionable Data = Questionable Science

It is easy to lie with EHR data, whether intentionally or out of ignorance.

Who's to blame? These three scientists are at the heart of the Surgisphere COVID-19 scandal

Author partnership on coronavirus papers is “completely bizarre” and should have been a red flag, former journal editor says

8 JUN 2020 • BY CHARLES PILLER

Surgisphere appears over time to have shifted its efforts into developing a database of hospital records that could be used for research. When the pandemic erupted, Desai declared that his data set could answer key questions about the efficacy and safety of treatments. Speaking about the finding that hydroxychloroquine increases mortality in COVID-19 patients, the main finding from the now retracted *Lancet* paper, he told a Turkish TV reporter, "with data like this, do we even need a randomized controlled trial?" Soon after, the World Health Organization temporarily suspended enrolling patients for its COVID-19 trial of the drug.

https://www.science.org/content/article/whos-blame-these-three-scientists-are-heart-surgisphere-covid-19-scandal

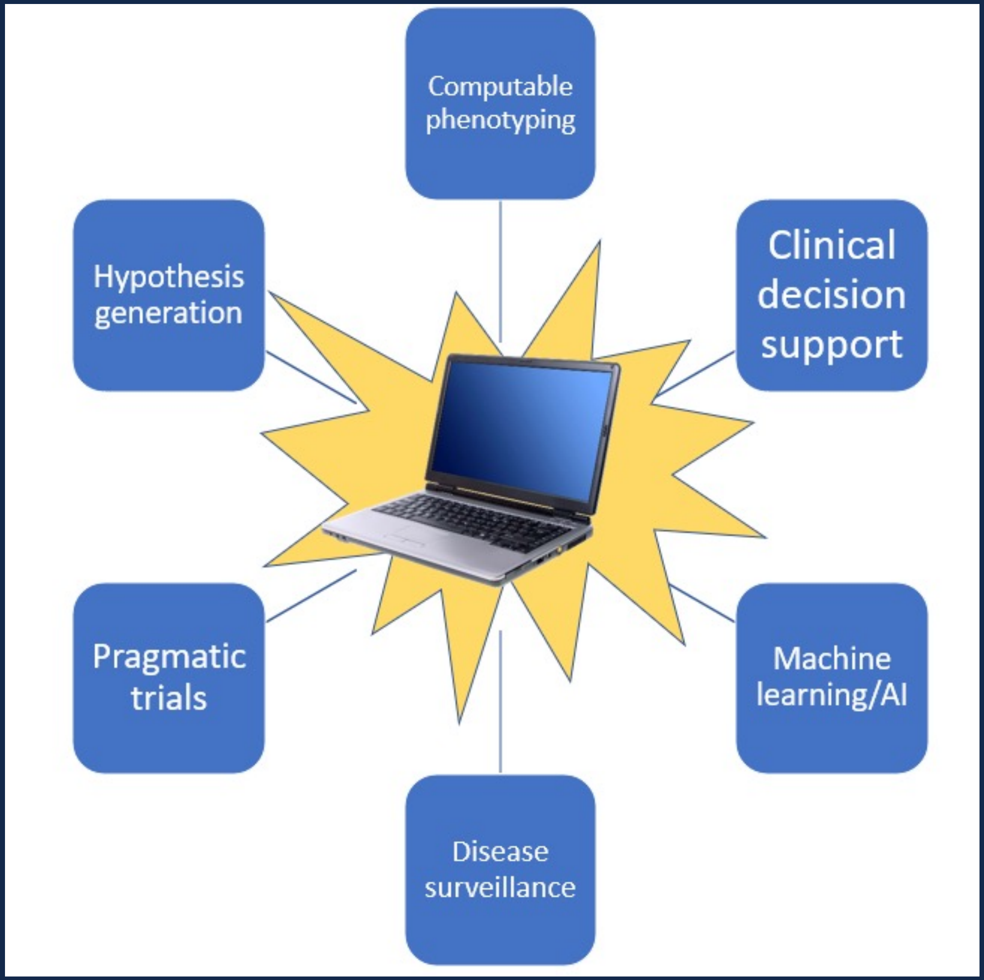
Questionable Data = Questionable Science

Erroneous conclusions can result from:

- Treating the absence of evidence as evidence of absence.
- Unaccounted-for selection bias.
- Lack of understanding of how data are collected.
- Using methods inappropriate for the data.
- Poor quality data.

But EHR research has so much potential!

So, how do we use what's good, and avoid the pitfalls?



UNC SCHOOL OF MEDICINE
North Carolina Translational and Clinical Sciences Institute

METHODOLOGICAL CONCERNS

Missing Data

- The EHR is not a holistic representation of patient health.
- Missing information may be missing for many reasons.
  - Temporal
  - Patient type
  - Technical
- Missing data is unavoidable—your interpretation is what counts.

Did this patient have COVID-19?

HR shows 1
egative PCR test,
/2020

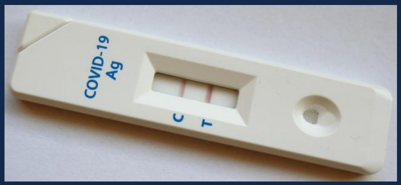
HR shows visit
r fatigue and
yspnea, 7/2022

Missing Data

- The EHR is not a holistic representation of patient health.
- Missing information may be missing for many reasons.
  - Temporal
  - Patient type
  - Technical
- Missing data is unavoidable—your interpretation is what counts.

Did this patient have COVID-19?

HR shows 1
egative PCR test,
/2020



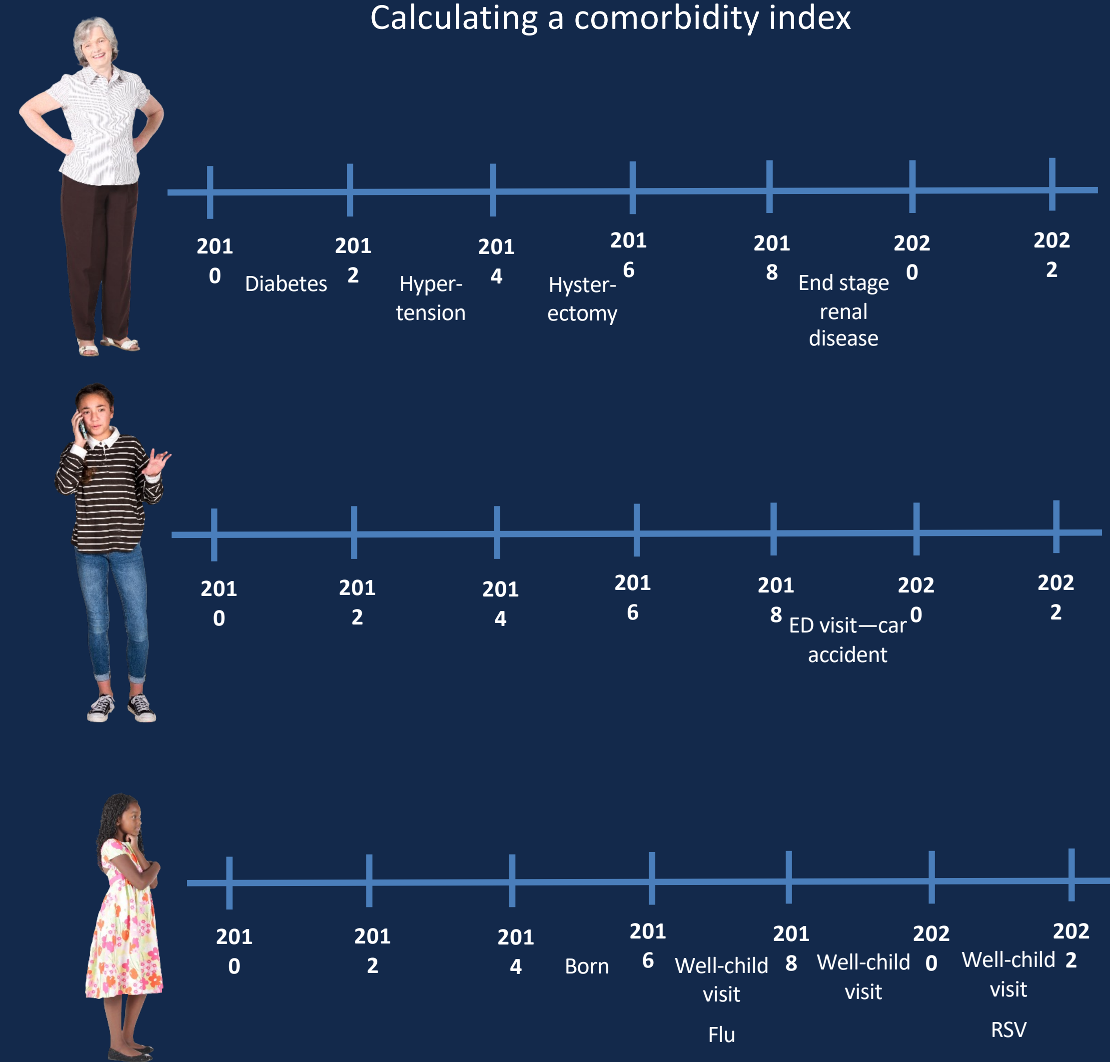
Positive home test, 3/2022

HR shows visit
r fatigue and
yspnea, 7/2022

Missing Data

- The EHR is not a holistic representation of patient health.
- Missing information may be missing for many reasons.
  - Temporal
  - Patient type
  - Technical
- Missing data is unavoidable—your interpretation is what counts.

Calculating a comorbidity index



Selection Bias

- People who seek healthcare are not representative of the population.
- EHR data skews toward sicker patients.
- Essential to remember who is not represented in your data.

ELSEVIER

International Journal of Infectious Diseases
Volume 116, Supplement, March 2022, Page S40

IJID

PS05.04 (947)

RETRACTED: Treatment with Ivermectin Is Associated with Decreased Mortality in COVID-19 Patients: Analysis of a National Federated Database

I. Efimenko [1], S. Nackeeran [2], S. Jabori [3], J.A. Gonzalez Zamora [4], S. Danker [3], D. Singh [1]

Show more

+ Add to Mendeley Share Cite

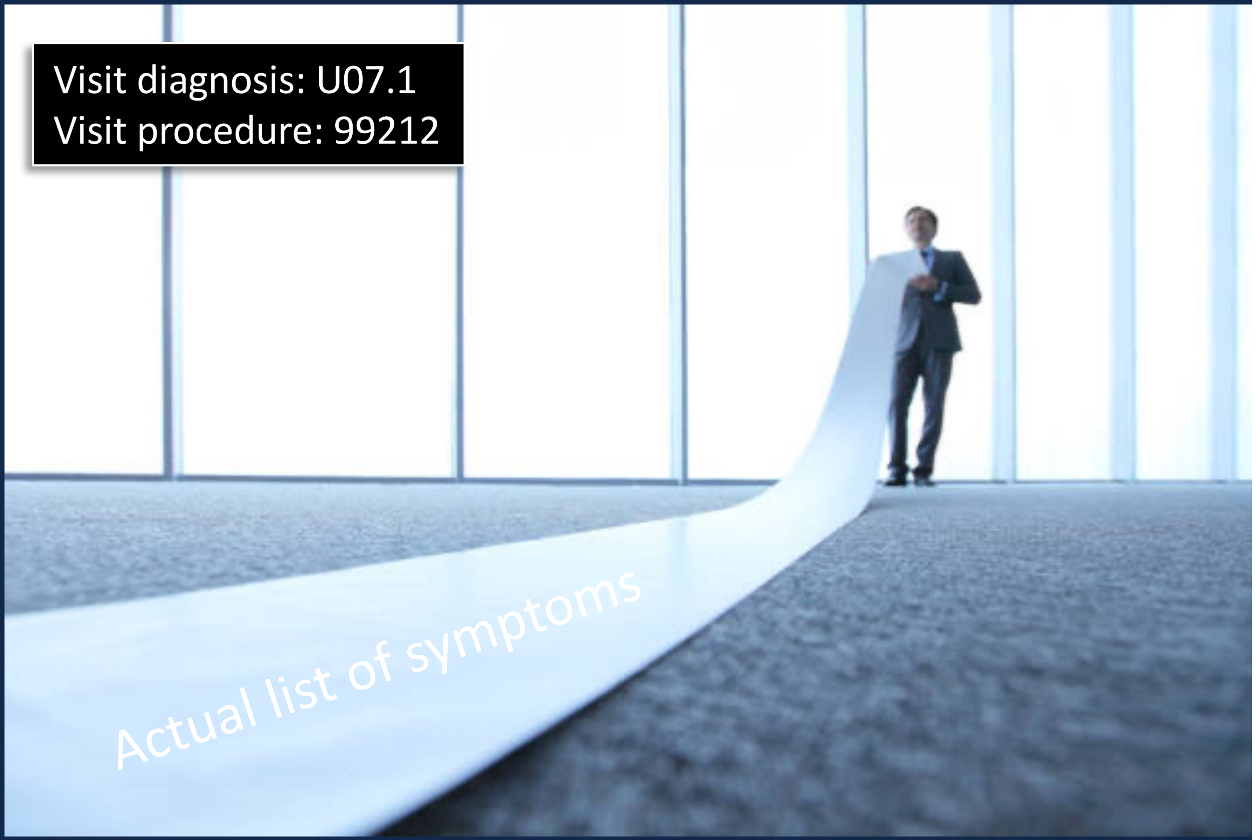
https://doi.org/10.1016/j.ijid.2021.12.096

This article has been retracted: please see Elsevier Policy on Article Withdrawal (https://www.elsevier.com/about/our-business/policies/article-withdrawal).

of studies). As in any retrospective study, we could not control for all the confounding variables, mainly severity of disease in patients treated with ivermectin or remdesivir. Another important caveat is that it was conduct

Data Collection Caveats

- The EHR is for clinical care………. and for billing.
- Some data are entered by coders, not clinicians.
- Some data are entered to justify procedure/lab orders.
- Some data just aren't entered.



Inappropriate Analyses

- Incidence/prevalence
- In many cases, evaluating positive outcomes
- Effects of over the counter drugs
- Outcomes for unvaccinated patients
- Applying ML or scoring algorithms without accounting for bias

The NEW ENGLAND JOURNAL *of* MEDICINE

MEDICINE AND SOCIETY

Debra Malina, Ph.D., *Editor*

Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms

Darshali A. Vyas, M.D., Leo G. Eisenstein, M.D., and David S. Jones, M.D., Ph.D.

Physicians still lack consensus on the meaning of race. When the *Journal* took up the topic in 2003 with a debate about the role of race in medicine, one side argued that racial and ethnic

subtle insertion of race into medicine involves diagnostic algorithms and practice guidelines that adjust or "correct" their outputs on the basis of a patient's race or ethnicity. Physicians use these

UNC SCHOOL OF MEDICINE
North Carolina Translational and Clinical Sciences Institute

A WHIRLWIND TOUR OF EHR DATA QUALITY

DQ Framework: Kahn, et al. (2016)

- **Conformance** – do the values present meet syntactic or structural constraints? (E.g., "Does this table follow the OMOP rules?")
- **Completeness** – what is the level of missingness, when compared with common expectations? (E.g., "Date of death is missing for 55% of deceased patients.")
- **Plausibility** – how believable are the data values? (E.g., adult height should not significantly fluctuate over time.)

So, where do things go wrong?

title

Data transformation can make data more useful, but with each transformation, quality can degrade.

Mapping errors

Simple human error
The concept of "ambulatory" visits in the source system gets mis-mapped to a similar-sounding word during ETL.

VISIT_ID	VISIT_TYPE	VISIT_DATE
34547	AMBULATORY	6/5/2004

Source data

VISIT_ID	VISIT_TYPE	VISIT_DATE
34547	AMBULANCE	6/5/2004

Transformed data

Content knowledge error
Serum and urine creatinine get mapped to the same lab identifier despite being very different tests.

PATIENT_ID	LAB_CD	LAB_NAME
29834723	Y77A89	CREATININE, SER
29834723	B212P0	CREATININE, UR

Source data

PATIENT_ID	LAB_CD	LAB_NAME
29834723	39452	CREATININE
29834723	39452	CREATININE

Transformed data

Granularity Changes

DISCHG_DISP_CD	DISCHG_DISP_NAME
01	HOME
02	EXPIRED
03	TRANSFERED
04	LEFT AGAINST MED ADVICE
05	SKILLED NURS. FAC.
06	HOSPICE
07	REHAB

DISCHG_DISP_CD	DISCHG_DISP_NAME
H	HOME
D	DECEASED
OT	OTHER

- Transformations often “roll up” long lists of codes from a source system into a more manageable list.
- Can be helpful for analysis; aggregated categories should be guided by use case.
- Resulting aggregation may not be granular enough for all use cases.
- Source concepts can be grouped incorrectly—hard to trace back.

Loss of Context

All diagnosis codes are not the same—they have a type. If the type is lost through oversimplification, the data can be used incorrectly in analysis.

PATIENT_ID	DX_CD	DX_TYPE
29834723	E11.3	PATIENT REPORTED
29834723	U07.1	BILLING

Source data

PATIENT_ID	DX_CD
29834723	E11.3
29834723	U07.1

Transformed data

Losing a "status" flag on billing transactions can cause us to mix voided transactions in with non-voided transactions!

BILL_ID	BILL_AMT	STATUS
55476	3255.67	FINAL
55476	546.20	VOID

Source data

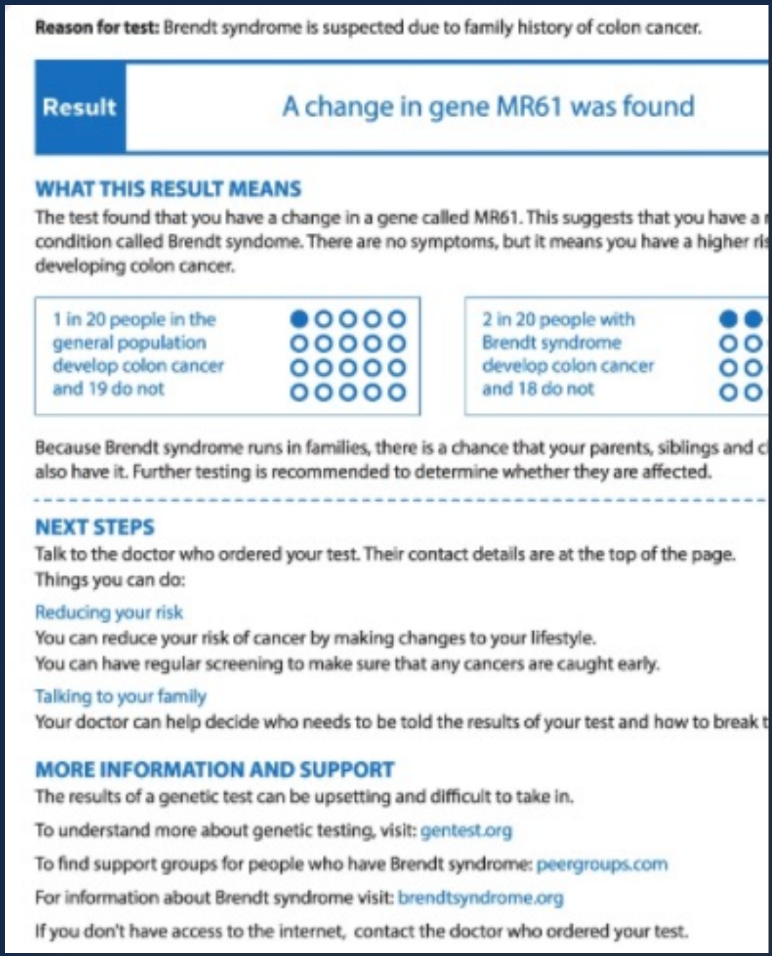
BILL_ID	BILL_AMT
55476	3255.67
55476	546.20

Transformed data

Missing Data

- Not all data are ETL'ed from the EHR in the same way, or at all.
  - e.g., PDFs, death data
- Individual variables may have a high rate of missingness
  - e.g., BMI, race and ethnicity

The transformation is not *wrong,* but the data are confusing/misleading. There may be no "fix," but an explanation is warranted.



Reason for test: Brendt syndrome is suspected due to family history of colon cancer.

Result A change in gene MR61 was found

WHAT THIS RESULT MEANS

The test found that you have a change in a gene called MR61. This suggests that you have a r
condition called Brendt syndome. There are no symptoms, but it means you have a higher ris
developing colon cancer.

1 in 20 people in the general population develop colon cancer and 19 do not

2 in 20 people with Brendt syndrome develop colon cancer and 18 do not

Because Brendt syndrome runs in families, there is a chance that your parents, siblings and c
also have it. Further testing is recommended to determine whether they are affected.

NEXT STEPS

Talk to the doctor who ordered your test. Their contact details are at the top of the page.
Things you can do:

Reducing your risk

You can reduce your risk of cancer by making changes to your lifestyle.
You can have regular screening to make sure that any cancers are caught early.

Talking to your family

Your doctor can help decide who needs to be told the results of your test and how to break t

MORE INFORMATION AND SUPPORT

The results of a genetic test can be upsetting and difficult to take in.

To understand more about genetic testing, visit: gentest.org

To find support groups for people who have Brendt syndrome: peergroups.com

For information about Brendt syndrome visit: brendtsyndrome.org

If you don't have access to the internet, contact the doctor who ordered your test.

Farmer, G.D., Gray, H., Chandratillake, G. *et al.* Recommendations for designing genetic test reports to be understood by patients and non-specialists. *Eur J Hum Genet* **28**, 885–895 (2020). https://doi.org/10.1038/s41431-020-0579-y

Garbage in, garbage out

The data are wrong.
You have mis-mapped your units of measure during transformation.

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	CM

Source data

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	IN

Transformed data

The data reflect the source.
The clinician thought she was entering centimeters, but the EHR was set to inches.

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	IN

Source data

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	IN

Transformed data

UNC SCHOOL OF MEDICINE
North Carolina Translational and Clinical Sciences Institute

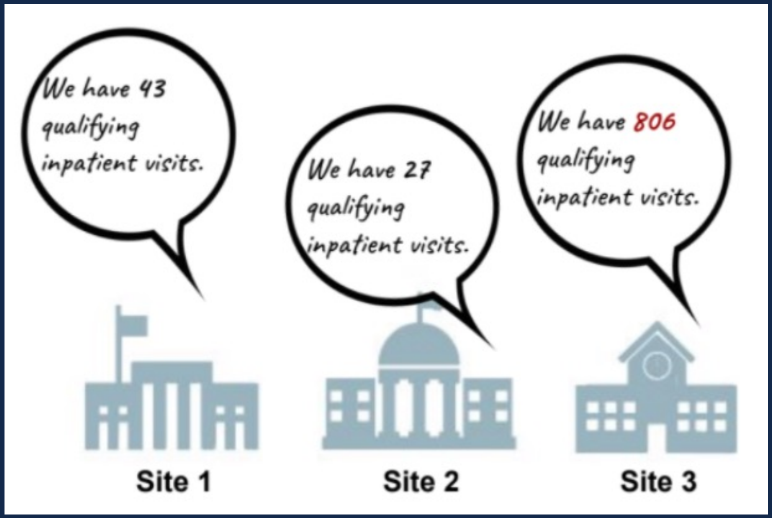
MULTI-SITE EHR DATA QUALITY

About N3C

- N3C is a data resource and collaborative community built for COVID-19 research
- Funded and managed by NCATS, led by the National Center for Data to Health (CD2H)
- The N3C data resource is a national COVID dataset available to researchers across the country
  - EHR data about COVID patients and match controls from 75 health care systems across the country; refreshed weekly
  - Housed at NIH in N3C Enclave, a secure portal for data analysis
  - Variety of analytical tools available for use by researchers
- More information is available at covid.cd2h.org.

Federated Data Quality

- Check conformance to CDM's rules
- Check for anomalies, implausible data, missingness
- Assessment can be shared with the network, but is based on a single site's data.



Site	Patient	Visit Type	Adm. Date	Disc. Date
1	123	IP	7/4/2020	7/8/2020
1	456	IP	5/6/2020	5/20/2020
2	987	IP	8/2/2019	8/7/2019
2	654	IP	9/3/2019	9/14/2019
3	234	IP	**1/26/2021**	1/26/2021
3	234	IP	**1/26/2021**	1/29/2021
3	234	IP	**1/26/2021**	1/30/2021
3	234	IP	**1/26/2021**	1/27/2021

title

Case in point: Harmonizing death data

	OHDSI	pcornet (The National Patient-Centered Clinical Research Network)	TriNetX	ACT (The ACT Network)
Death data supported?	Y	Y	Y	Y
Death date required?	Y	N	Y	Y
Death cause supported?	Y	Y	N	N
Discharge disposition supported?	Y	Y	N	N

N3C Minimum Checks (part 1)

Check Type	Data Checks
Source CDM Conformance	**Must Pass:** All tables required by the native CDM specs are present, with all CDM-required fields populated; fields that use a controlled value set (e.g., “M” for male, “F” for female, etc.) are populated with valid values
Demographics	**Must Pass:** count of patients qualifying for COVID phenotype is reasonable when compared with sites of similar size; sex, race, and ethnicity distributions reasonable for the site’s population; month of birth evenly distributed throughout the calendar year **Heads Up:** > 20% of race or ethnicity is missing or “No Matching Concept”
COVID tests	**Must Pass:** all COVID tests must be coded with an OMOP standard concept (or, for non-OMOP source data, the LOINC equivalent); all COVID test results must be coded with an OMOP standard concept (or, for non-OMOP source data, the equivalent controlled vocabulary term); numbers of negative and positive COVID tests are reasonable when compared with sites of similar size **Heads Up:** High numbers of COVID tests with *null* results
Conditions	**Must Pass**: Clinical encounters are present that are coded with U07.1 (ICD-10 code for COVID), and those encounters are distributed across various visit types (e.g., outpatient, inpatient, emergency)

N3C Minimum Checks (part 2)

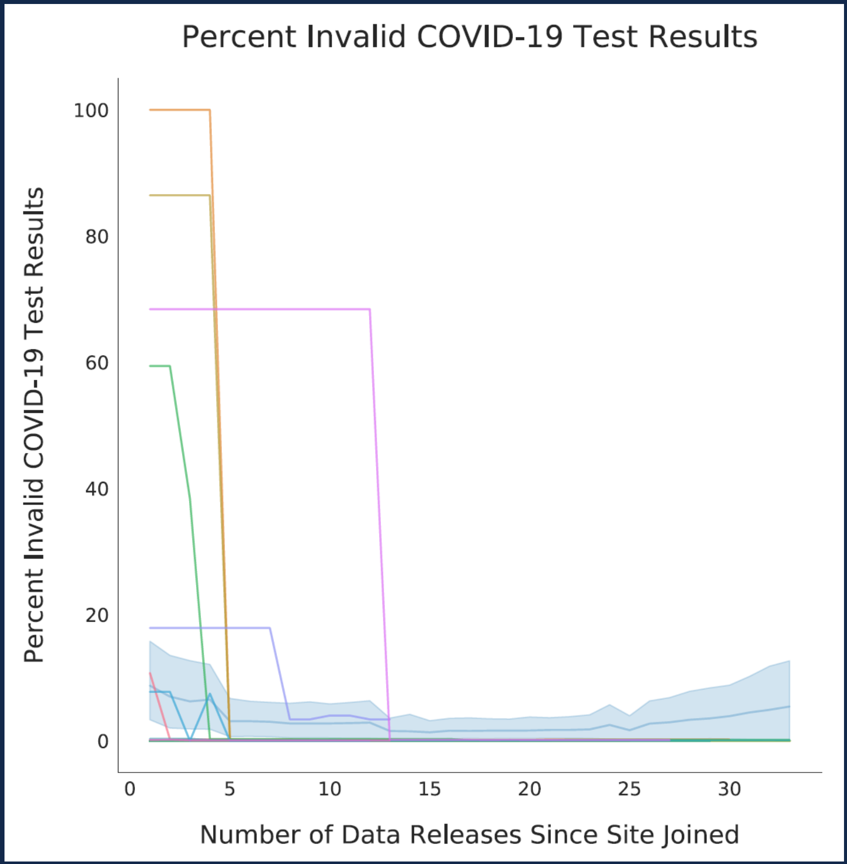
Check Type	Data Checks
Encounters	**Must Pass:** Clinical encounters are distributed across a variety of standard visit types (e.g., outpatient, inpatient, emergency); the distribution of visit types is reasonable when compared with similar sites; the majority of inpatient visits have valid end dates; the mean duration of visits of various types is reasonable for that type of visit; vast majority of visit end dates are later than or equal to the visit start date
Measurements/ Observations	**Heads Up:** The site supports only a small number (e.g., 5-10) of unique measurement or observation types
Coding Completeness	**Must Pass:** No more than 20% of records in any domain are coded with non-standard OMOP concept IDs without further explanation (OMOP sites only); no more than 20% of records in any domain are coded with “0 - No Matching Concept” without further explanation (affects OMOP sites only); the PERSON_ID attached to all records in domain tables must exist in the PERSON table; primary keys are valid (i.e., no duplicate rows in any table); if applied by the site, date shifting is consistent within each patient across all domains
Fitness for Use	Use of the data by researchers often reveals additional DQ issues for one or more sites (e.g., sparsely populated body mass index data, in the context of a study of obesity and COVID). In these cases, we report the findings to sites so that they can take action in their local data if they wish to have their site’s data included in the study.

Data Quality Heuristics

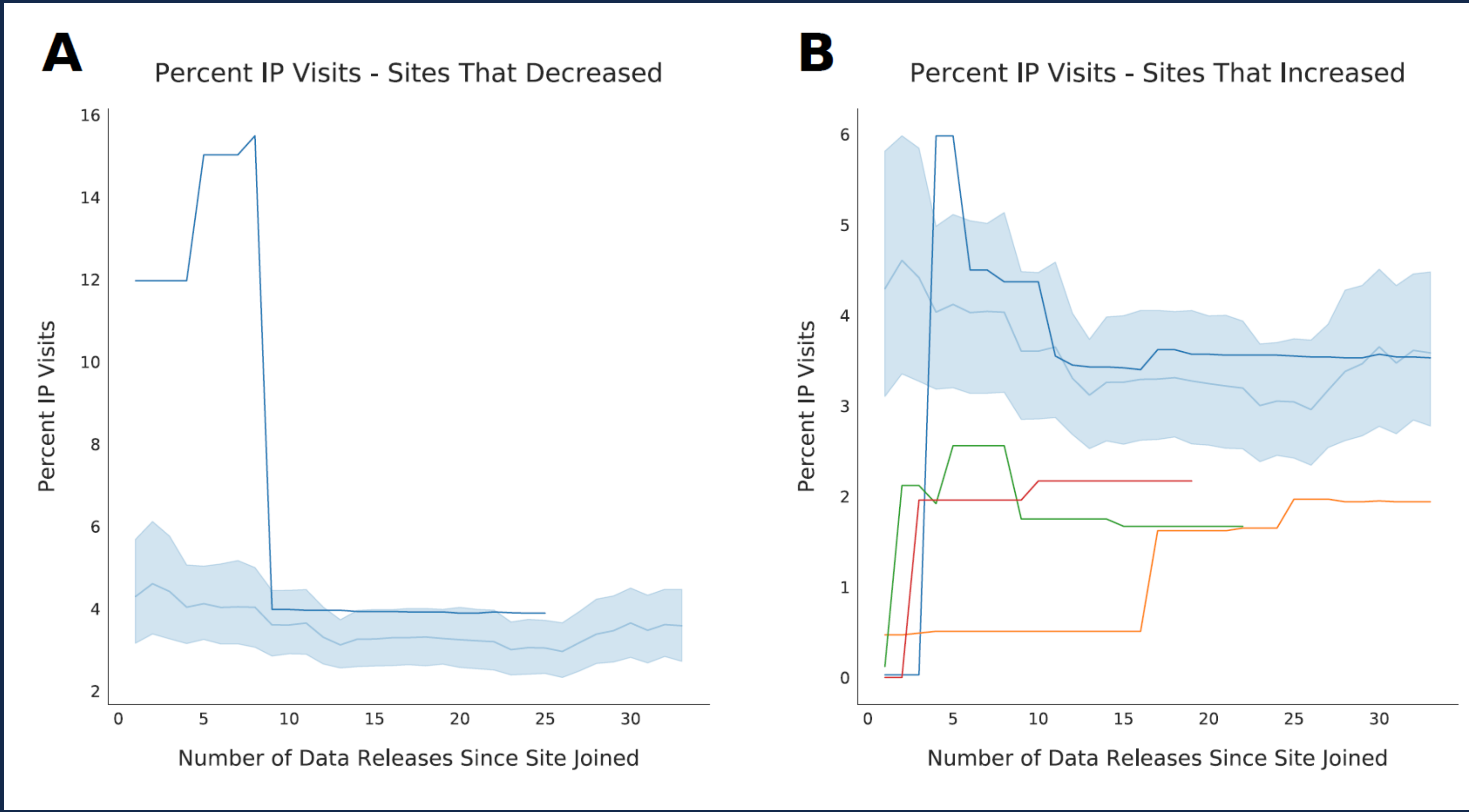
#	Heuristic	Type	# Sites	% Sites*
1	Not using (or improperly using) source CDM's controlled vocabulary in one or more fields	Source CDM Conformance	13	23.2%
2	COVID test result values not standardized or null	COVID tests	11	19.6%
3	Lacking/incorrectly populating field(s) required by source CDM	Source CDM Conformance	9	16.1%
4	Implausible distribution of visit types (e.g., 75% inpatient)	Encounters	7	12.5%
5	Large number of "No Matching Concept" records (OMOP source only)	Coding Completeness	6	10.7%
6	Lacking table(s) required by source CDM	Source CDM Conformance	5	9.0%
7	Many or all inpatient visits lacking valid end dates	Encounters	5	9.0%
8	Few or no clinical encounters coded with U07.1	Conditions	5	9.0%
9	Implausible count of patients qualifying for phenotype	Demographics	3	5.4%
10	Small number of unique measurement/observation types	Measurement/Observation	2	3.6%
11	PERSON_IDs in fact tables that are not in the PERSON table	Coding Completeness	2	3.6%
12	Primary keys are not unique	Coding Completeness	2	3.6%
13	Inconsistent local date shifting causing implausible timelines	Coding Completeness	2	3.6%
14	Implausible demographics (e.g., 100% male patients)	Demographics	2	3.6%
15	Data utility challenges (e.g., missing mortality data)	Fitness for Use	N/A	N/A

*Denominator: 56 sites; 37 unique sites are represented across these categories.

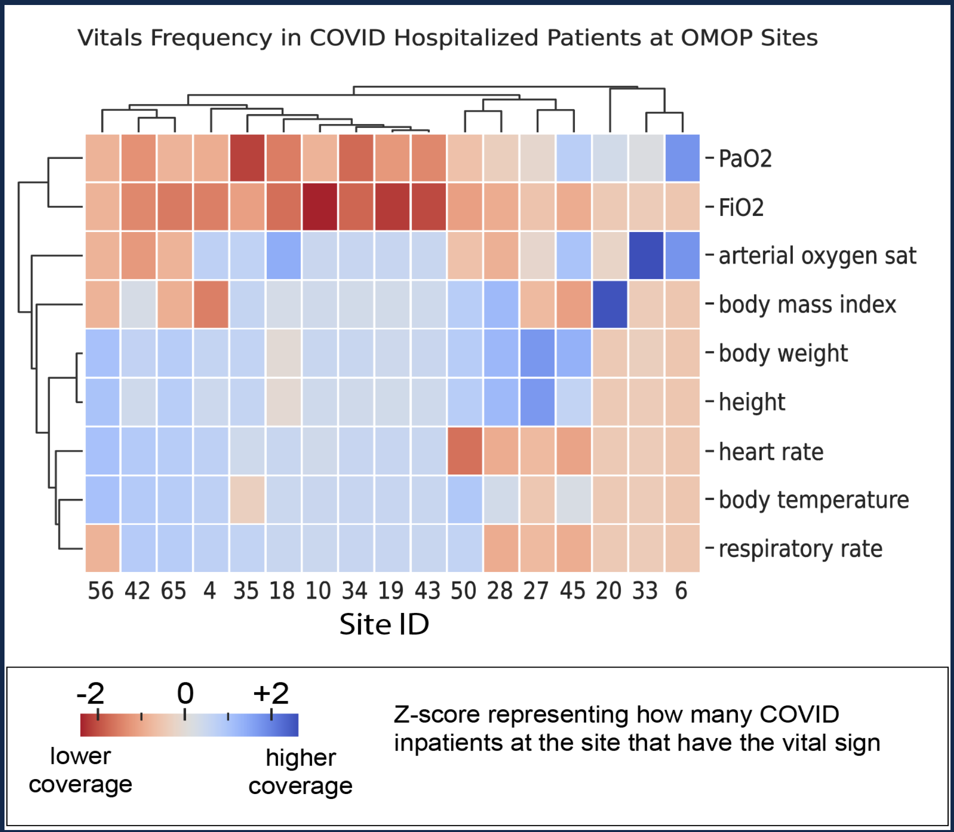
Example: Heuristic #2, COVID test results not standard



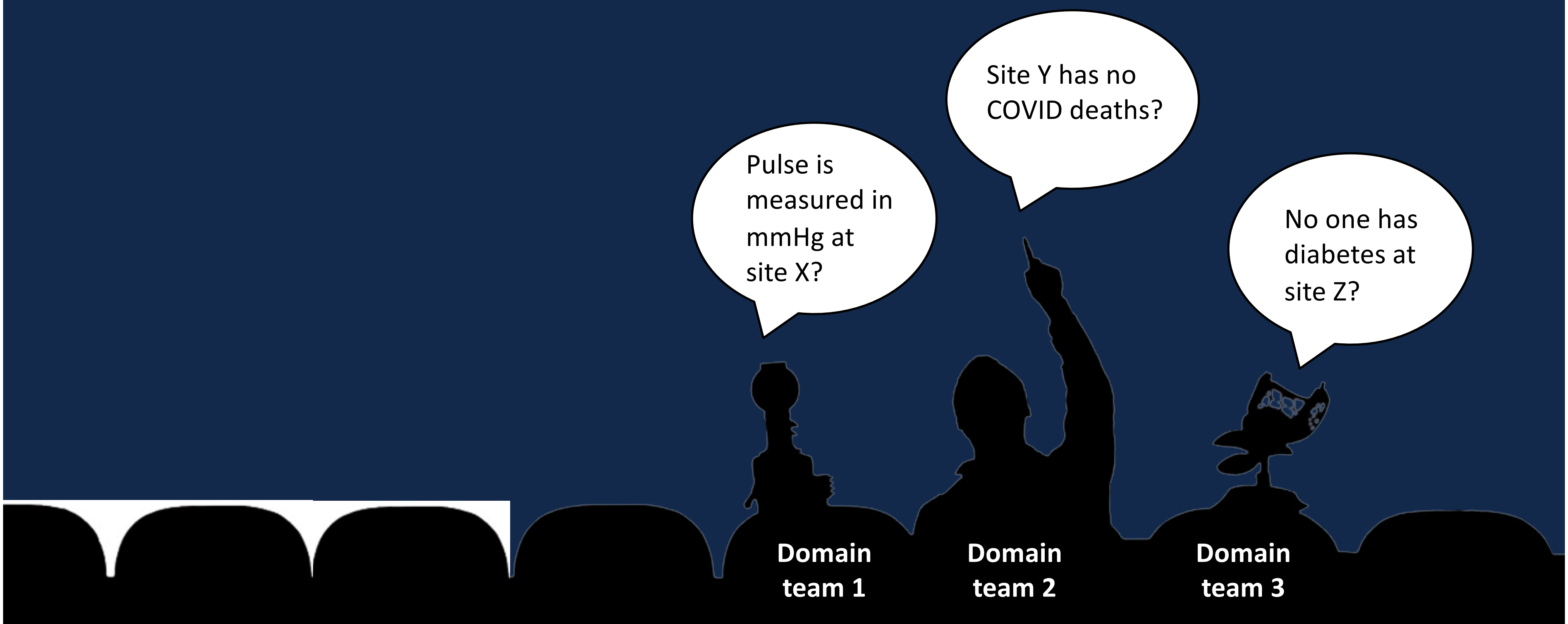
Example: Heuristic #4, Implausible visit type distribution



Site-to-Site Benchmarking



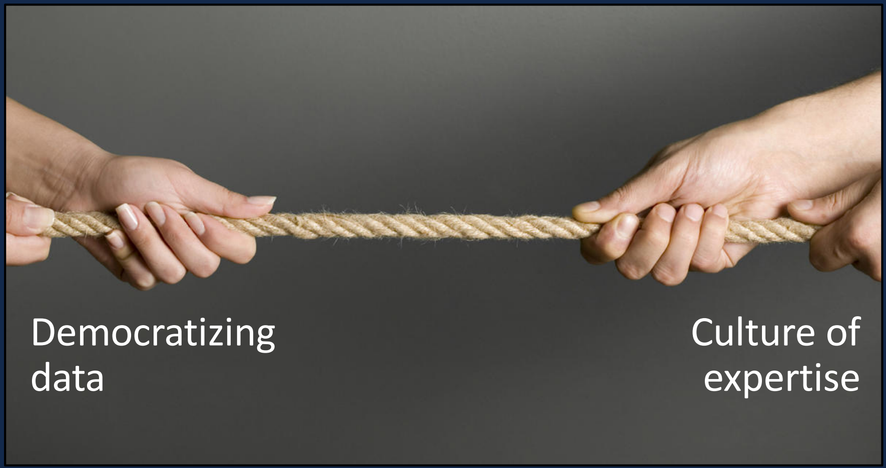
Crowdsourced Quality



title

“So, these data *are* junk!...Right?"

Who wins?



Enter: Team Science

Clinical SME

Research questions

Clinical domain knowledge

Clinical Informaticist

Data engineering/ extraction

Data quality

Data context expertise

Data Scientist

Statistical analysis

Data visualization

Methods expertise

Takeaways

- EHR data can be used for important and novel research.
- It’s also easy to misuse, or misunderstand.
- There is tension between democratizing EHR data and a culture of deep expertise.
- Team science is a promising path forward for clinical informatics using EHR data.

Thank you!

Questions welcome: epfaff@email.unc.edu