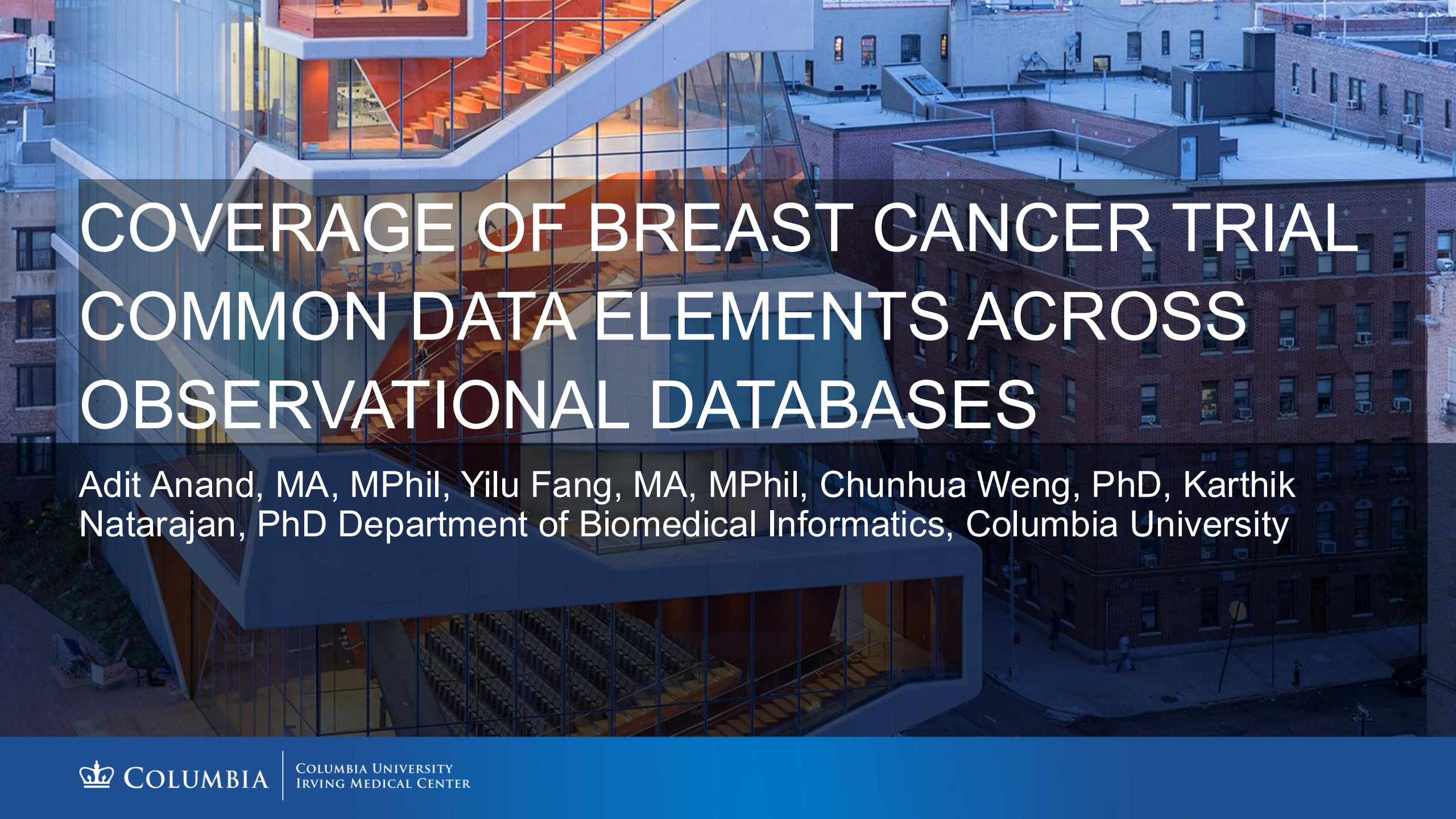




COVERAGE OF BREAST CANCER TRIAL COMMON DATA ELEMENTS ACROSS OBSERVATIONAL DATABASES

Adit Anand, MA, MPhil, Yilu Fang, MA, MPhil, Chunhua Weng, PhD, Karthik Natarajan, PhD Department of Biomedical Informatics, Columbia University

Learning Objectives

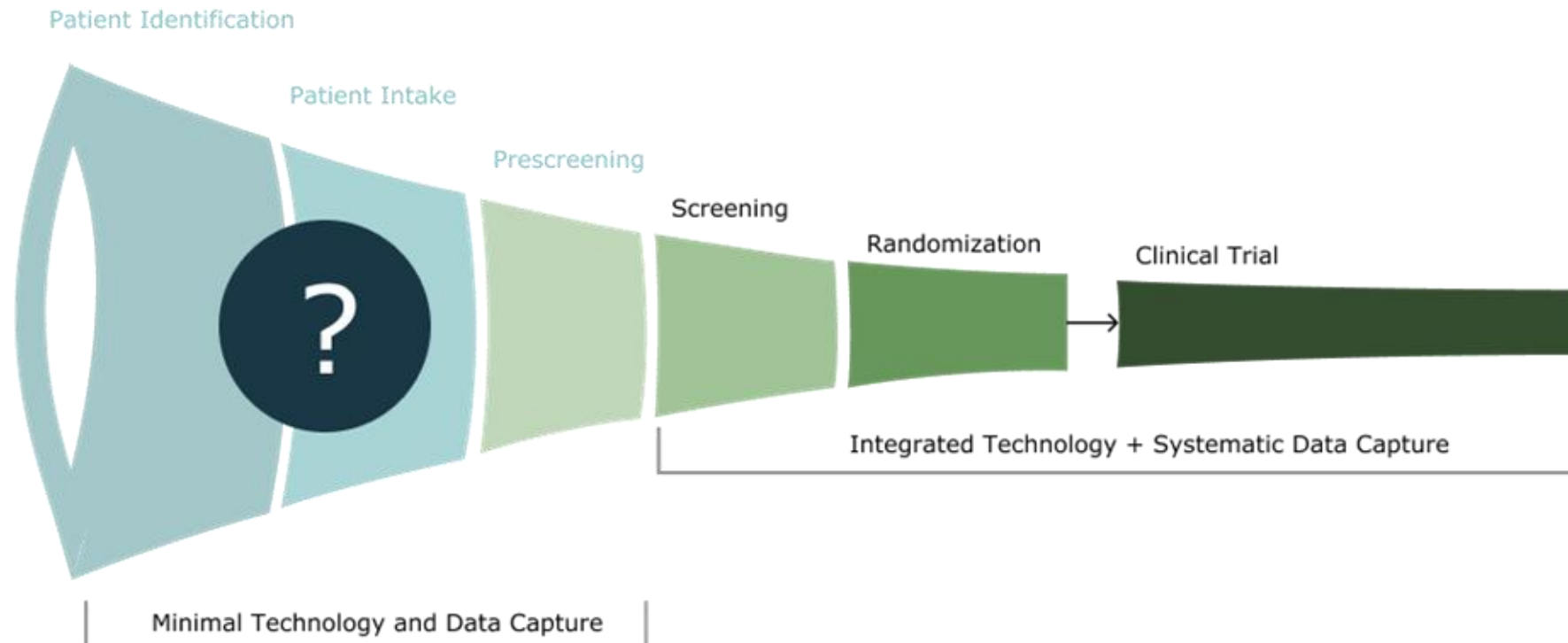
1. Learn the nuances of identifying common data elements from clinical trials' eligibility criteria
2. Understand how to evaluate coverage for common data elements in healthcare databases

Disclosure of Relevant Financial Relationships

- I have no relevant financial relationships to disclose

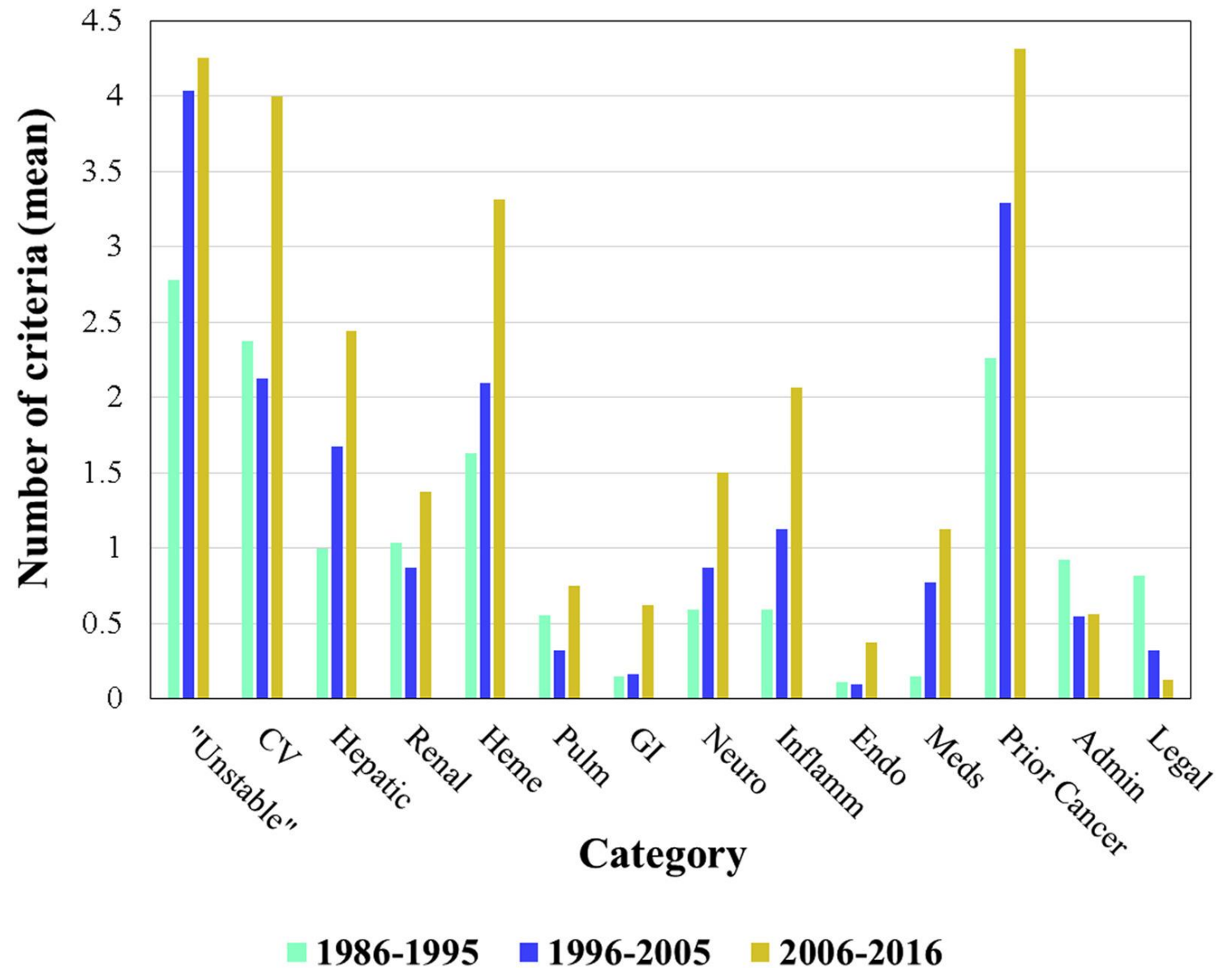
Background

The Bottleneck of Clinical Trial Recruitment



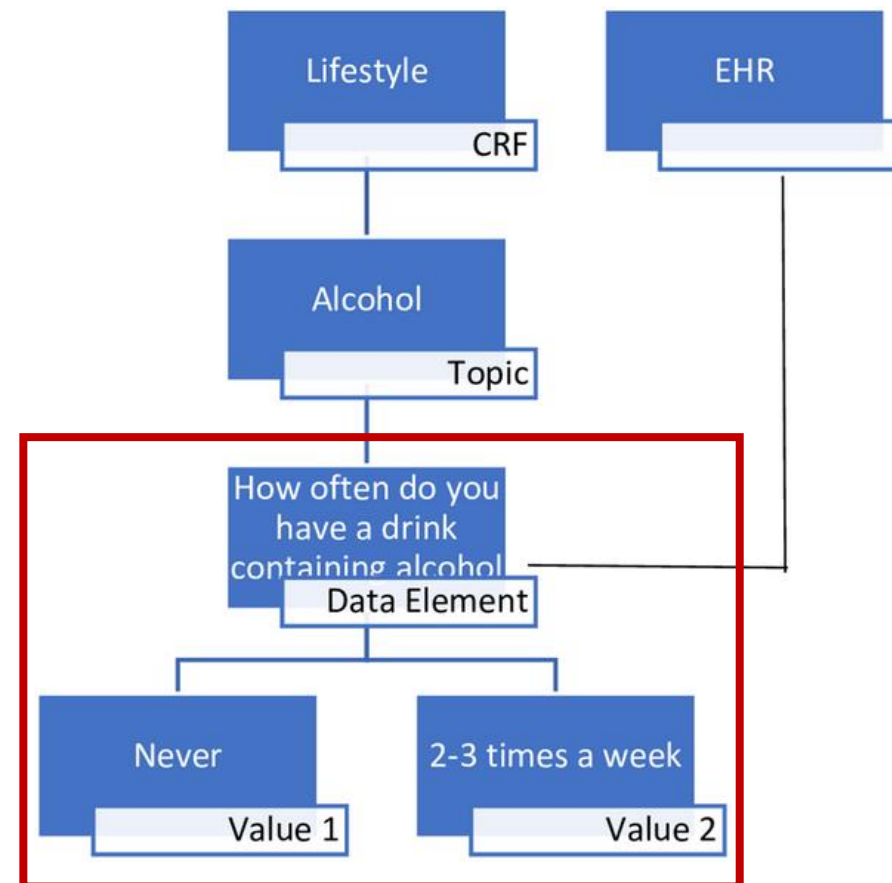
Challenges of Eligibility Criteria

- Free-text by nature
- No standard for computational representation
- Increasing complexity of eligibility criteria



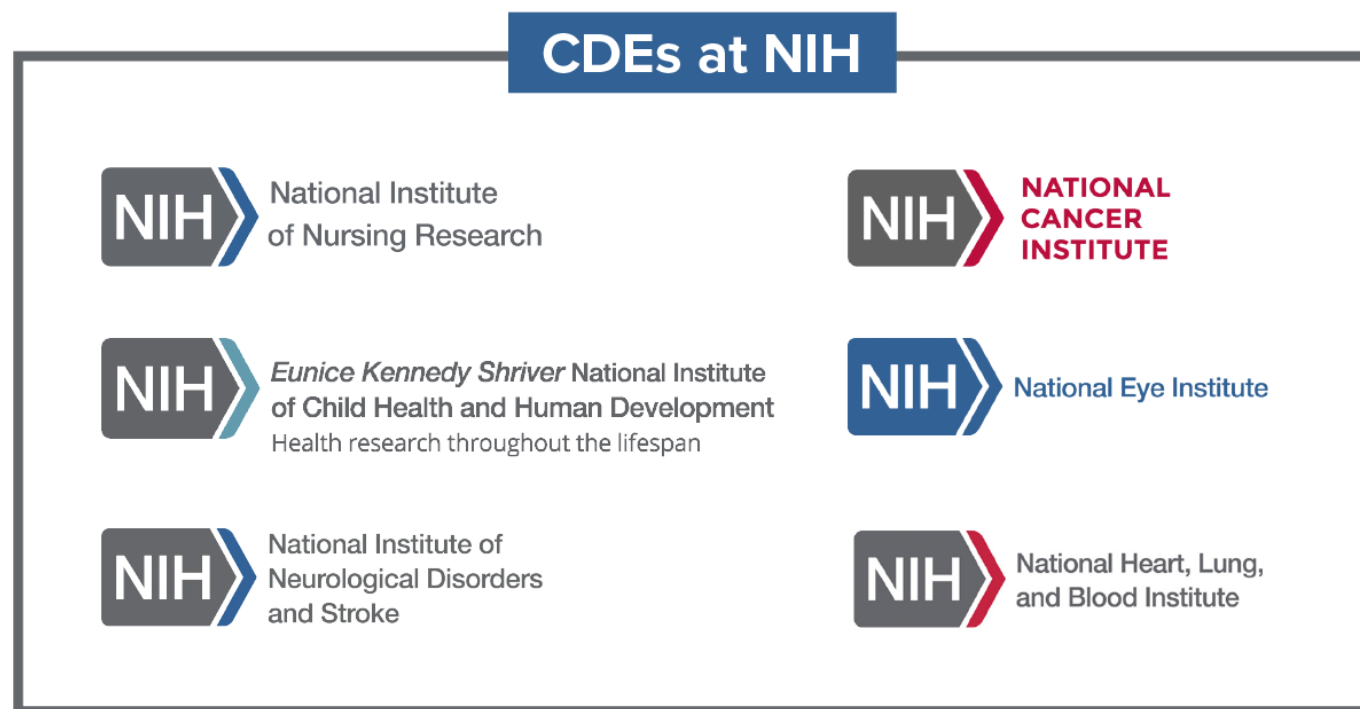
Role of Common Data Elements (CDEs)

- CDEs represent generalizable variables consistently found in datasets
- CDE curation improves data quality and interoperability



CDEs in Healthcare

- CDEs facilitate research in multiple domains
- Curating CDEs in oncology trials' eligibility criteria can inform informatics research and trial teams' efforts



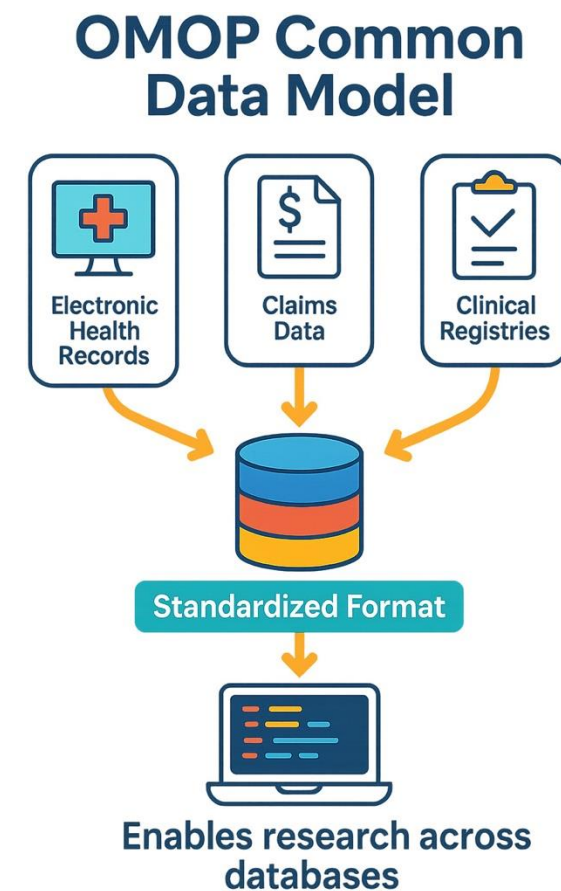
Research Questions Tackled

1. Which CDEs emerge from the eligibility criteria of breast cancer trials processed with traditional named-entity recognition methods?
2. Across structured EHR databases, which identified CDEs do we see captured for breast cancer patients?

Methods

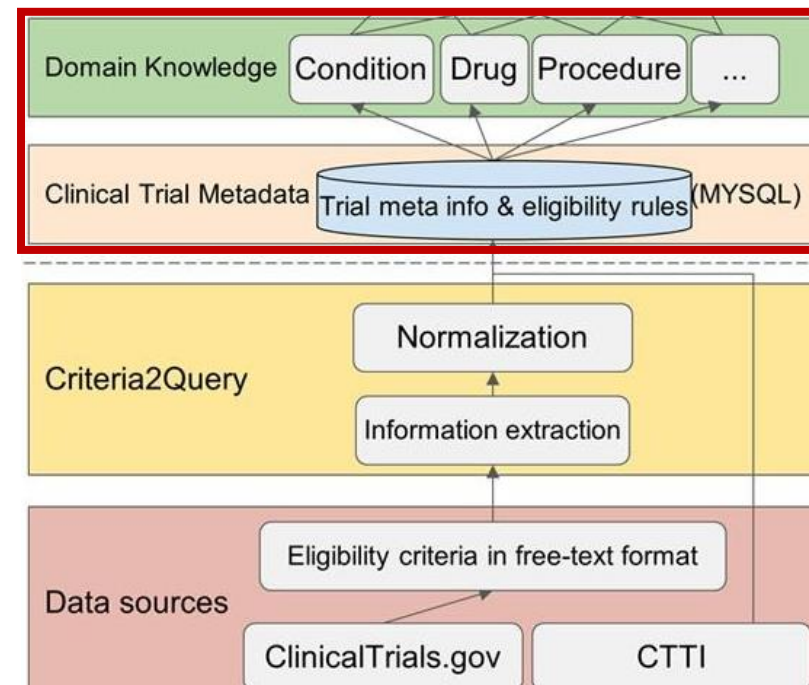
Data Sources – Breast Cancer Patients

- Cohort Definition
 - PheKB's eMERGE breast cancer phenotype
- Databases (Cohort Size)
 - Columbia University Irving Medical Center (CUIMC) EHR (n = 60,837)
 - All of Us (AoU) CDR v8 (n=11,971)



Data Sources – Clinical Trials

- Clinical Trials Knowledge Base (CTKB)
 - Database of OMOP concepts found in eligibility criteria
- 7,984 breast cancer trials from ClinicalTrials.gov



Common Data Element Analysis

- CDE Reporting
 - Stratified across 4 OMOP domains
 - Drug CDEs are rolled up into ATC or HemOnc classifications
- CDE Coverage Calculation
 - Choose 5 most frequent CDEs for each OMOP domain
 - Calculate percentage of patients with ≥ 1 CDE record



Condition



Drug



Measurement



Procedure

Results

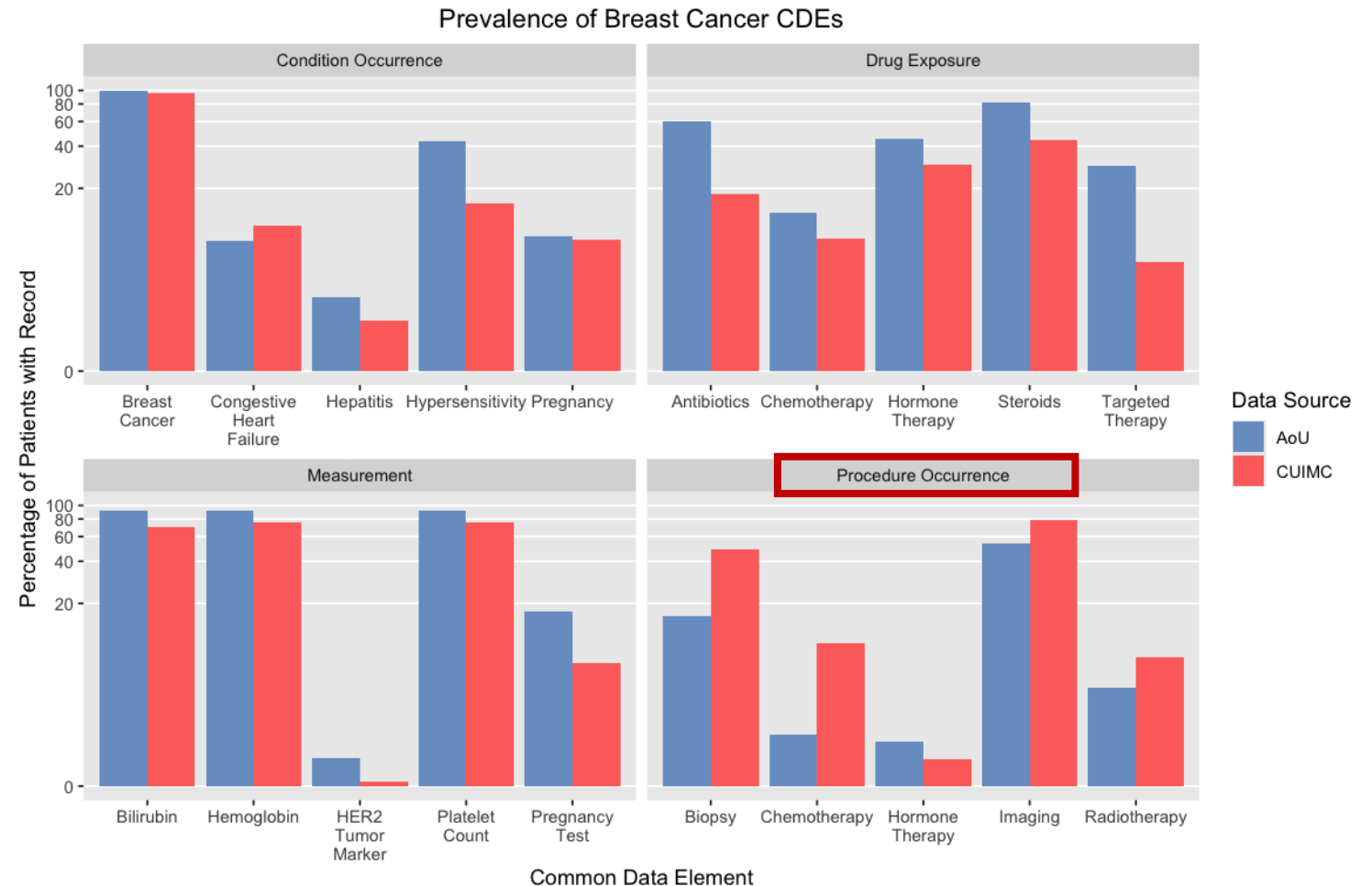
CDEs of Breast Cancer Trials

OMOP Domain	Common Data Element	Trial Count (n = 7,984)
Condition Occurrence (n=4,873)	Breast Cancer	6,285 (78.720%)
	Pregnancy	3,333 (41.746%)
	Hypersensitivity	2,245 (28.119%)
	Congestive Heart Failure	1,543 (19.326%)
	Hepatitis	1,515 (18.975%)
Drug Exposure (n=1,910)	Chemotherapy	2,214 (22.730%)
	Hormone Therapy	1,817 (22.758%)
	Targeted Therapy	1,018 (12.751%)
	Steroids	976 (12.224%)
	Antibiotics	439 (5.498%)

OMOP Domain	Common Data Element	Trial Count (n = 7,984)
Measurement (n=1,247)	Bilirubin	2,179 (27.292%)
	Platelet Count	2,086 (26.127%)
	Pregnancy Test	1,517 (19.000%)
	HER2 Tumor Marker	1,464 (18.337%)
	Hemoglobin	1,271 (15.919%)
Procedure Occurrence (n=1,055)	Chemotherapy	3,645 (45.654%)
	Radiotherapy	2,174 (27.229%)
	Imaging	1,745 (21.856%)
	Biopsy	1,579 (19.777%)
	Hormone Therapy	1,293 (16.195%)

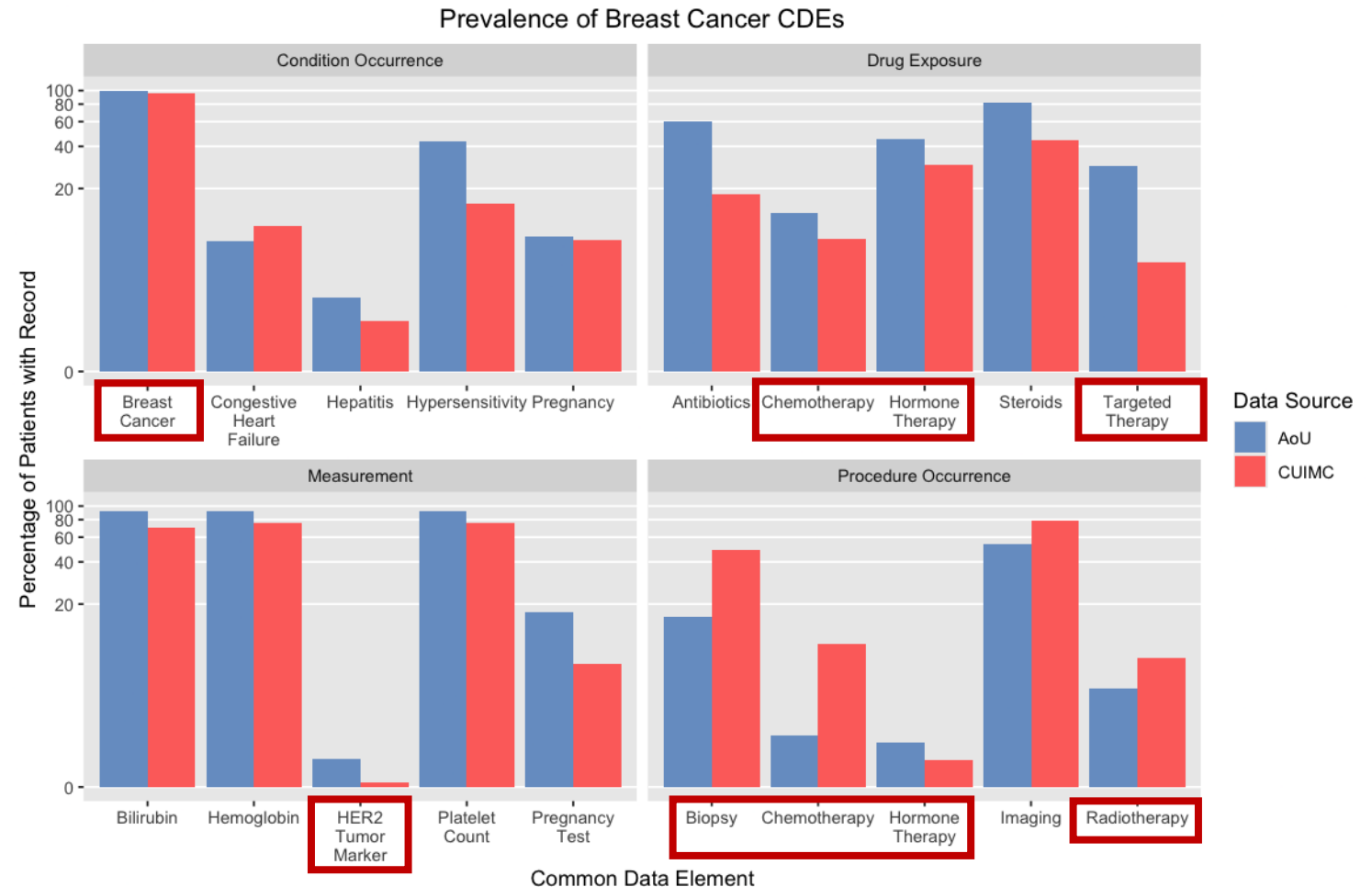
CDE Coverage in Structured Data

- Highest coverage for conditions and measurements
- Lowest coverage across procedures



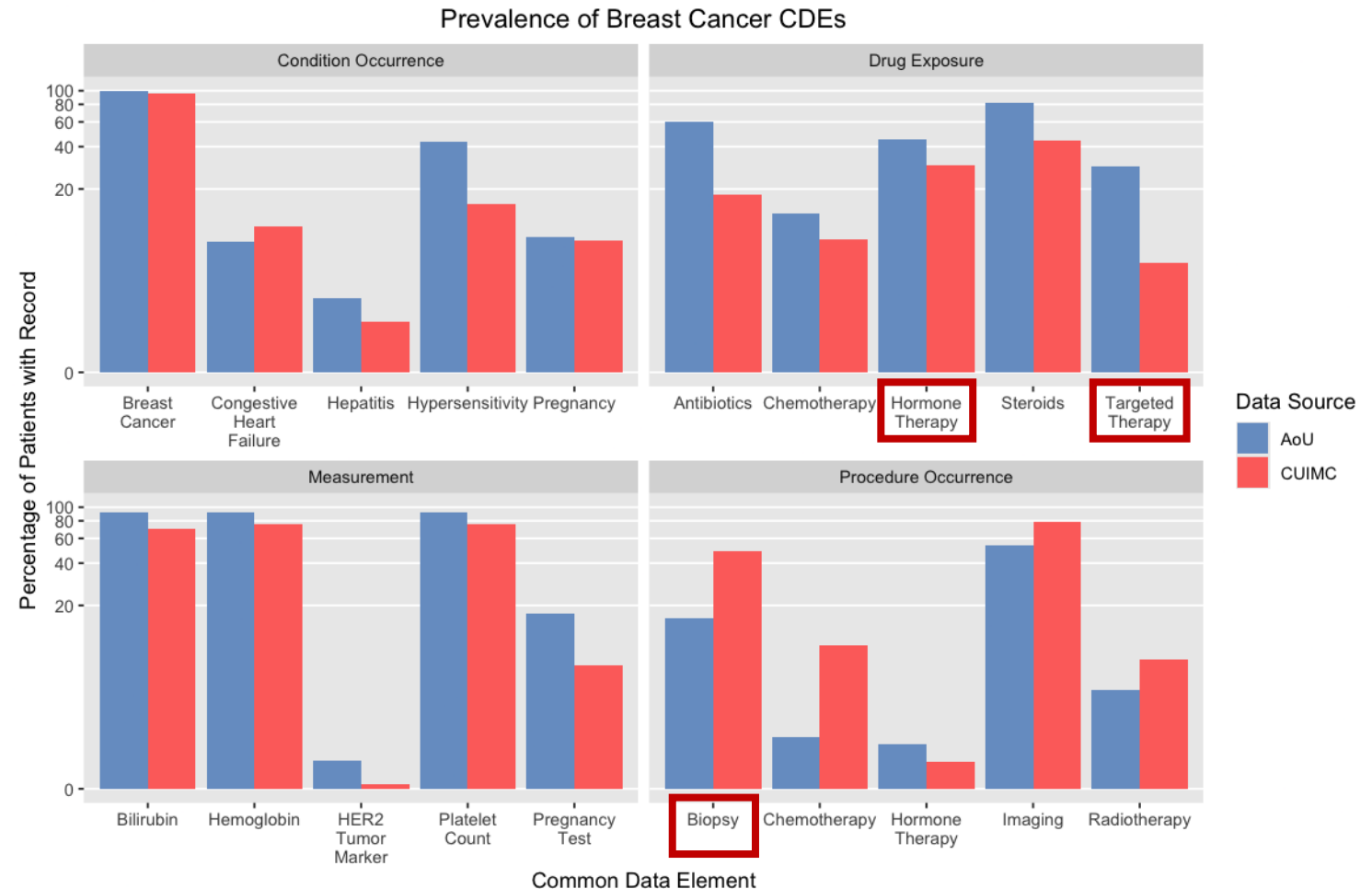
CDE Coverage in Structured Data

- Highest coverage for conditions and measurements
- Lowest coverage across procedures
- Lagging coverage of oncology CDEs



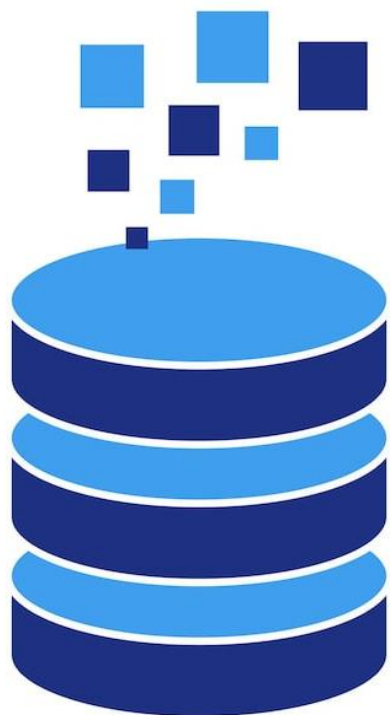
CDE Coverage in Structured Data

- Highest coverage for conditions and measurements
- Lowest coverage across procedures
- Lagging coverage of oncology CDEs
- Variance in CDE coverage across databases



Limitations

Limited eligibility
criteria dataset



Rule-based + deep learning
concept extraction

Entity Extraction

Extract: Location

"**Paris** is a beautiful city
located in **France**."

Location: **Paris**

Location: **France**

Generalizability
across cancer types



Conclusion

- Structured data is insufficient to query cancer CDEs of trial criteria
- Healthcare database characteristics inform their usage for trial recruitment
- The identified CDEs can guide efforts such as
 - Database fitness-for-use evaluations
 - Design of cancer trial recruitment methodologies

Special Thanks To

- Dr. Karthik Natarajan
- Yilu Fang and Dr. Chunhua Weng



COLUMBIA UNIVERSITY
DEPARTMENT OF
BIOMEDICAL INFORMATICS



Thank you for
listening!

Questions?