MIDS C3 Forum

MIDS Communication, Coordination, and Collaboration (C3) Forum

MIDS CORs & Contractors Monthly Meeting

Tuesday
March 5, 2024
1:00-2:00 PM Eastern

MIDS C3 Forum Agenda

- **Leveraging Artificial Intelligence for Concept Extraction in Measurement Development**
  - Andrew Taylor (Yale)
- **Using Large Language Models (LLMs) to Generate Mechanism Maps**
  - Jeff Geppert (Battelle)
- **Artificial Intelligence and Quality Improvement**
  - Pierre Barker, Gary Kantor, Jeff Rakover, Marina Renton (Institute for Healthcare Improvement)

Leveraging Artificial Intelligence for Concept Extraction in Measurement Development

Andrew Taylor, MD MHS

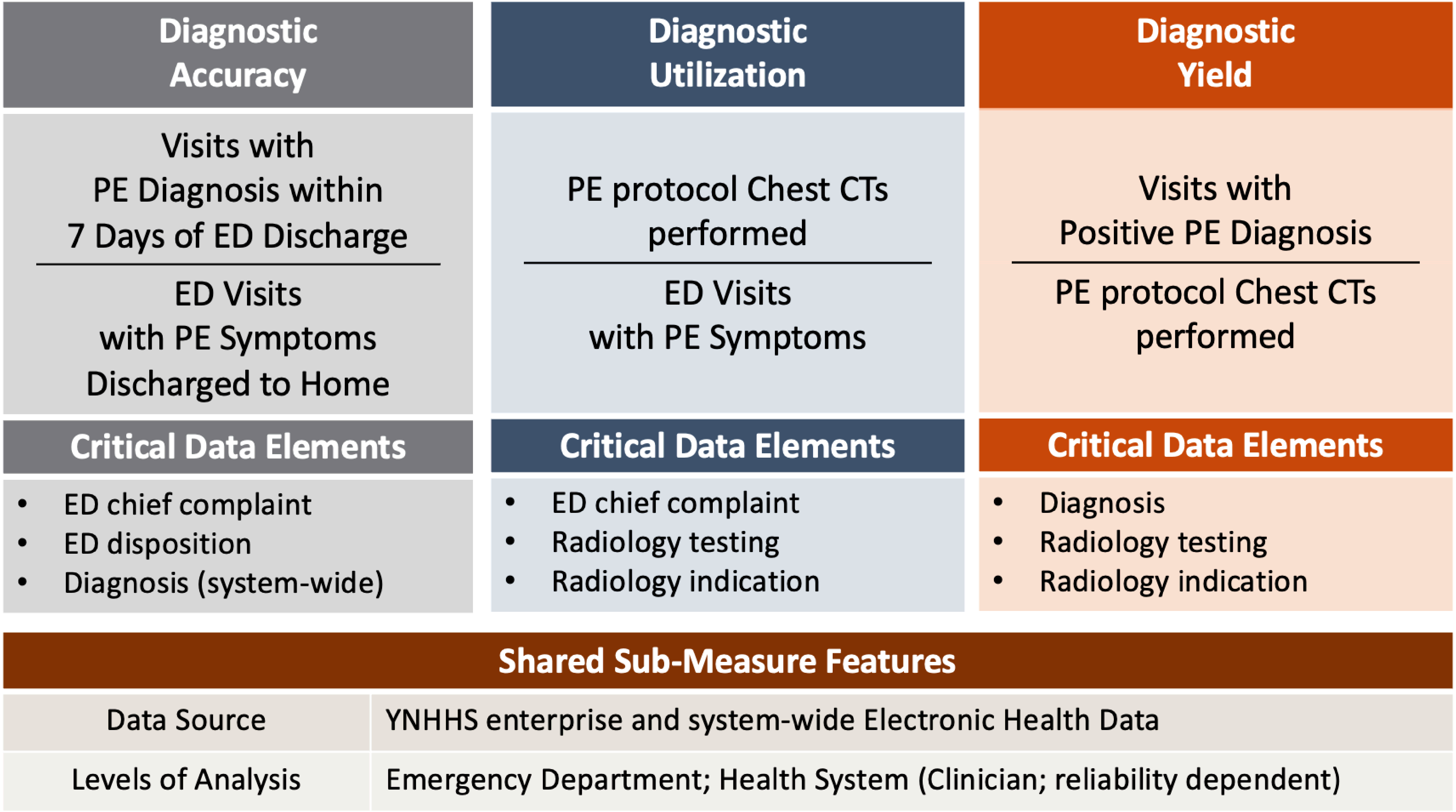
Associate Professor of Emergency Medicine, Biomedical Informatics and Data Science, and Biostatistics, Yale

Outline

Challenges with Measure Concept Extraction

Defining/Extracting Concepts Through Artificial Intelligence

The Challenges with Measure Concept Extraction



Diagnostic Accuracy	Diagnostic Utilization	Diagnostic Yield
$\frac{\text{Visits with PE Diagnosis within 7 Days of ED Discharge}}{\text{ED Visits with PE Symptoms Discharged to Home}}$	$\frac{\text{PE protocol Chest CTs performed}}{\text{ED Visits with PE Symptoms}}$	$\frac{\text{Visits with Positive PE Diagnosis}}{\text{PE protocol Chest CTs performed}}$
Critical Data Elements	**Critical Data Elements**	**Critical Data Elements**
• ED chief complaint • ED disposition • Diagnosis (system-wide)	• ED chief complaint • Radiology testing • Radiology indication	• Diagnosis • Radiology testing • Radiology indication

Shared Sub-Measure Features	
Data Source	YNHHS enterprise and system-wide Electronic Health Data
Levels of Analysis	Emergency Department; Health System (Clinician; reliability dependent)

Concepts are inherently noisy

Concepts are often complex and not captured in a standardized format

Concepts are often only embedded within unstructured data

Defining Concepts Through AI

How Do We Identify/Define Acute Pulmonary Embolism?

Traditional Method

- Utilizes ICD codes for diagnoses in administrative claims data.
- Dependent on the accuracy of the healthcare provider's diagnosis.
- Straightforward for billing and epidemiological tracking.
- Does not account for undiagnosed cases or misdiagnoses.
- Limited to information that is formally documented and coded.

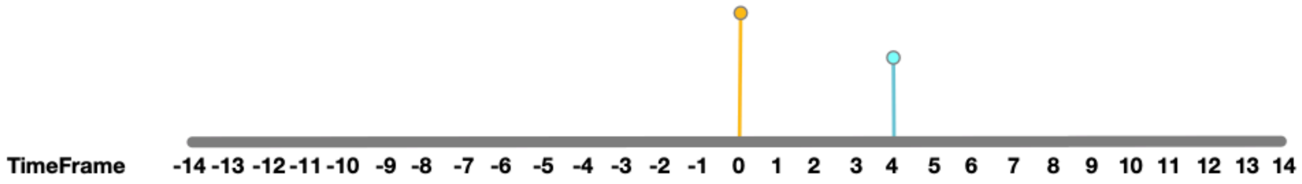
AI Algorithm Using Structured Data

- Analyzes structured data elements like ICD codes, lab results, and vital signs.
- Employs pattern recognition to identify instances of "pulmonary embolism."
- Potentially more accurate by correlating various structured data points.
- Limited by the availability and quality of structured data.
- Does not incorporate clinical nuances present in free-text notes.

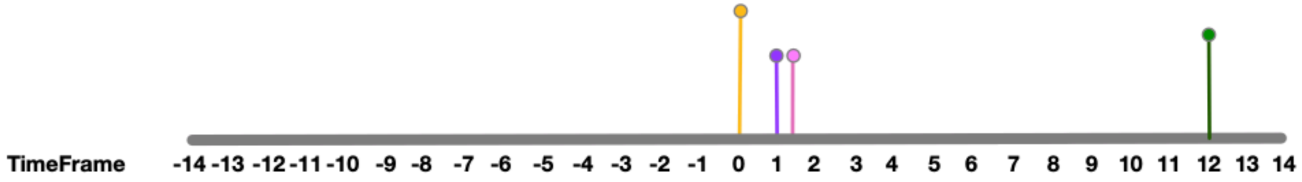
NLP/AI algorithm Using Unstructured Data

- Processes free-text clinical notes, radiology reports, and other narrative documents.
- Extracts contextual information indicative of "pulmonary embolism."
- Can identify risk factors and clinical observations not found in structured data.
- Requires advanced NLP tools to interpret complex medical language.
- Computationally intensive but provides a deeper clinical picture.

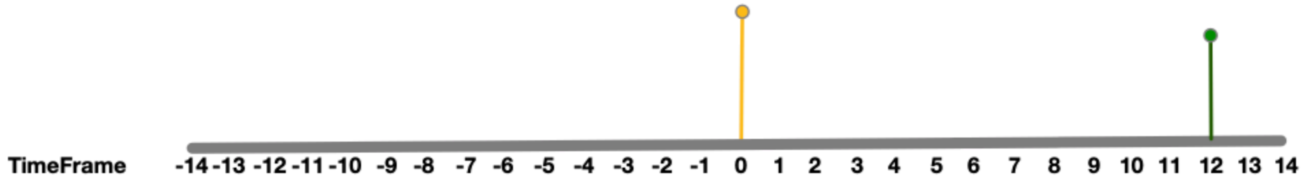
Scenario 1: Prior history of Pulmonary Embolism (PE) 2 months prior, on anticoagulation, presents with chest pain. Computed Tomography (CT) PE negative. Patient documented as having PE on problem list because of prior history.



Scenario 2: No prior history of PE. CT PE ordered as outpatient study by primary doctor. Patients presents to ED after result positive and called in by primary doctor.



Scenario 3: Patient admitted with chest paint. No CT ordered in ED. CT ordered 3 days later in hospital, result positive.



🟨 CT imaging event

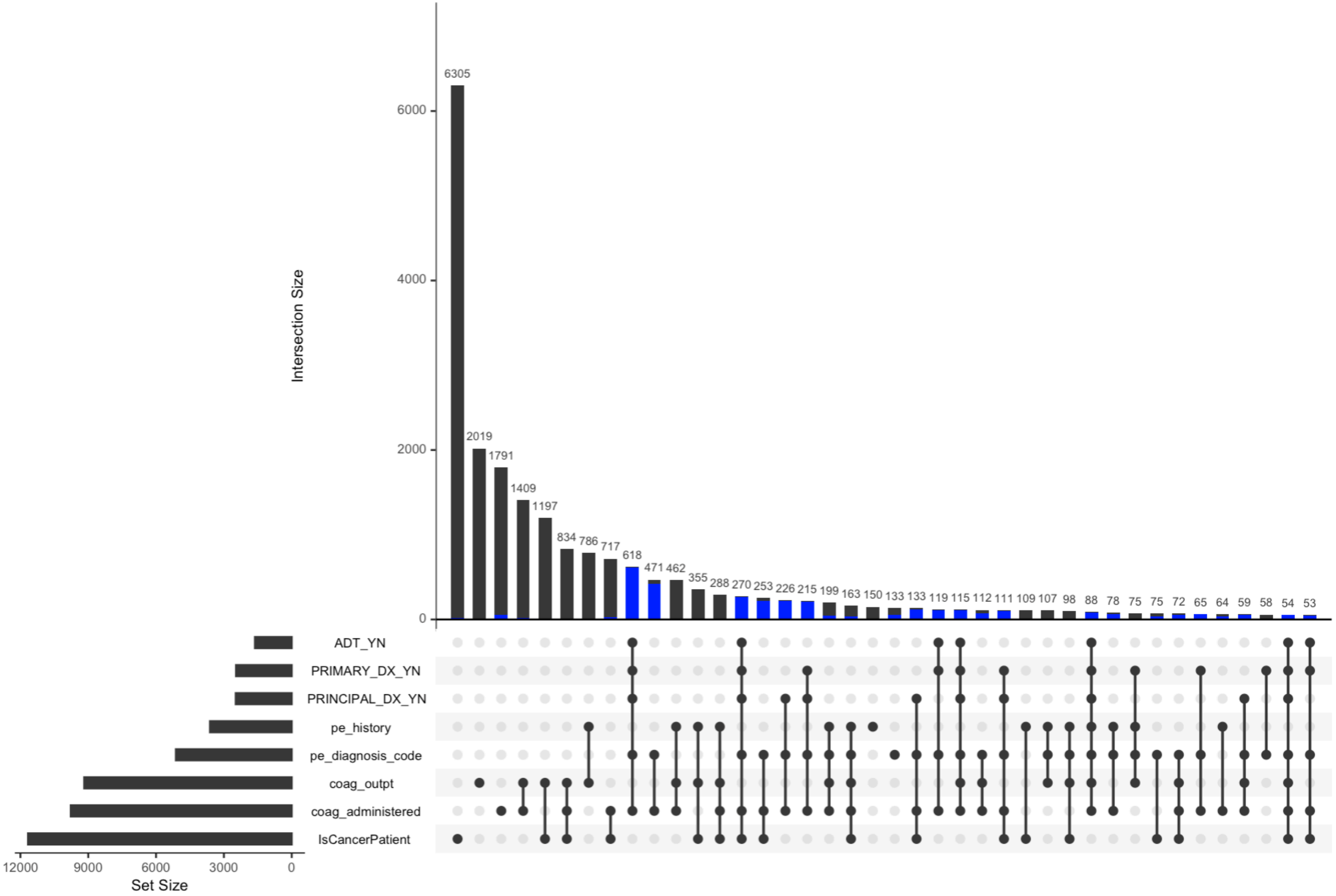
🟩 PRINCIPAL_DX_YN

🟪 PRIMARY_DX_YN

🟥 ADT_YN

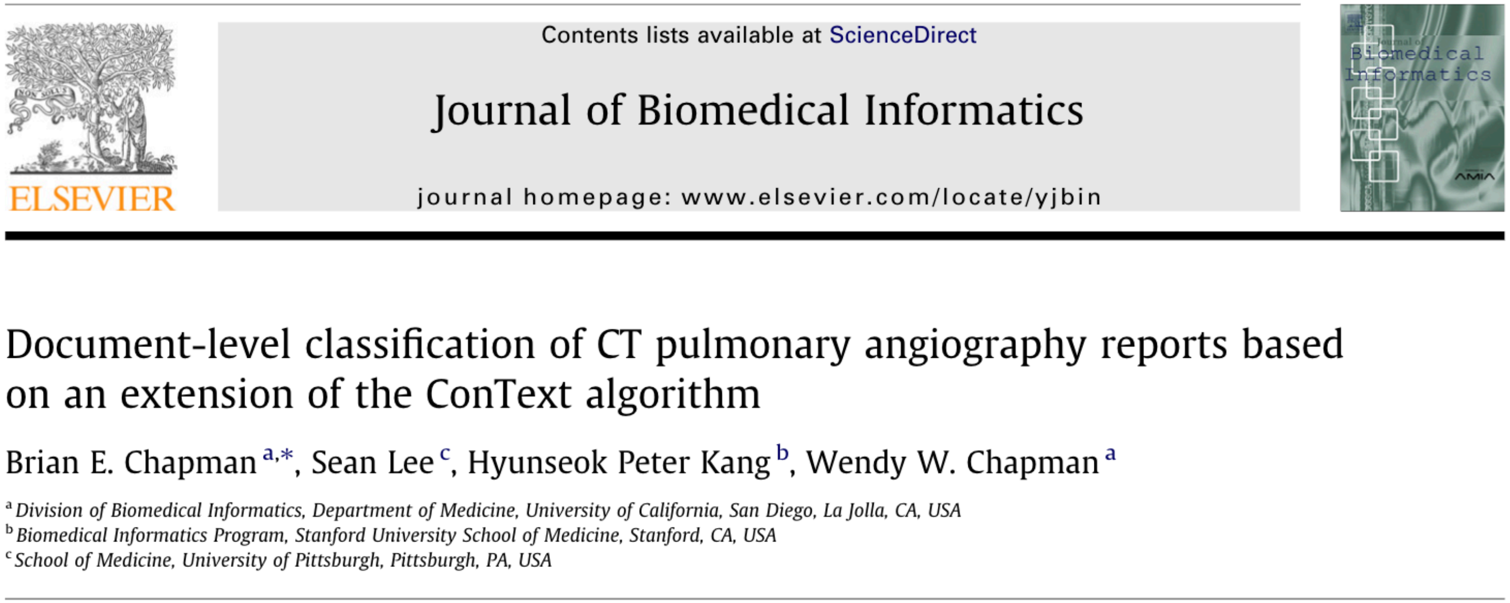
🟦 Other PE code

Set Relationships between Structured Data Elements



AI/NLP Method Unstructured Data

PEFinder: NLP Program and Performance Characteristics



Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm

Brian E. Chapman [a,*], Sean Lee [c], Hyunseok Peter Kang [b], Wendy W. Chapman [a]

[a] *Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, La Jolla, CA, USA*
[b] *Biomedical Informatics Program, Stanford University School of Medicine, Stanford, CA, USA*
[c] *School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA*

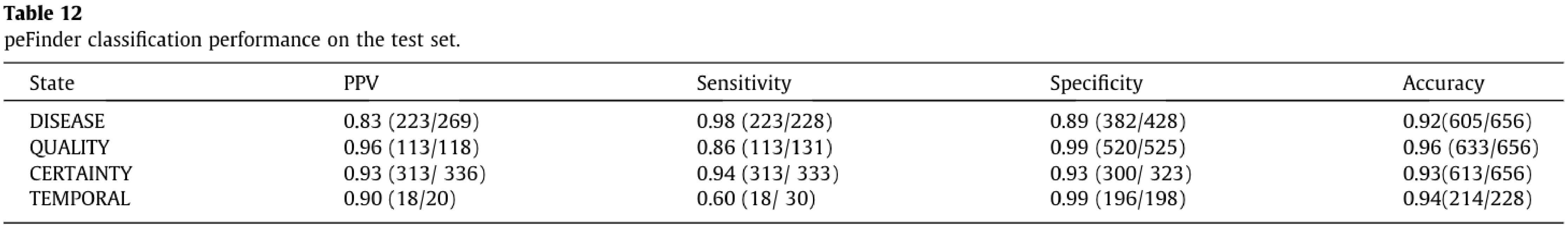
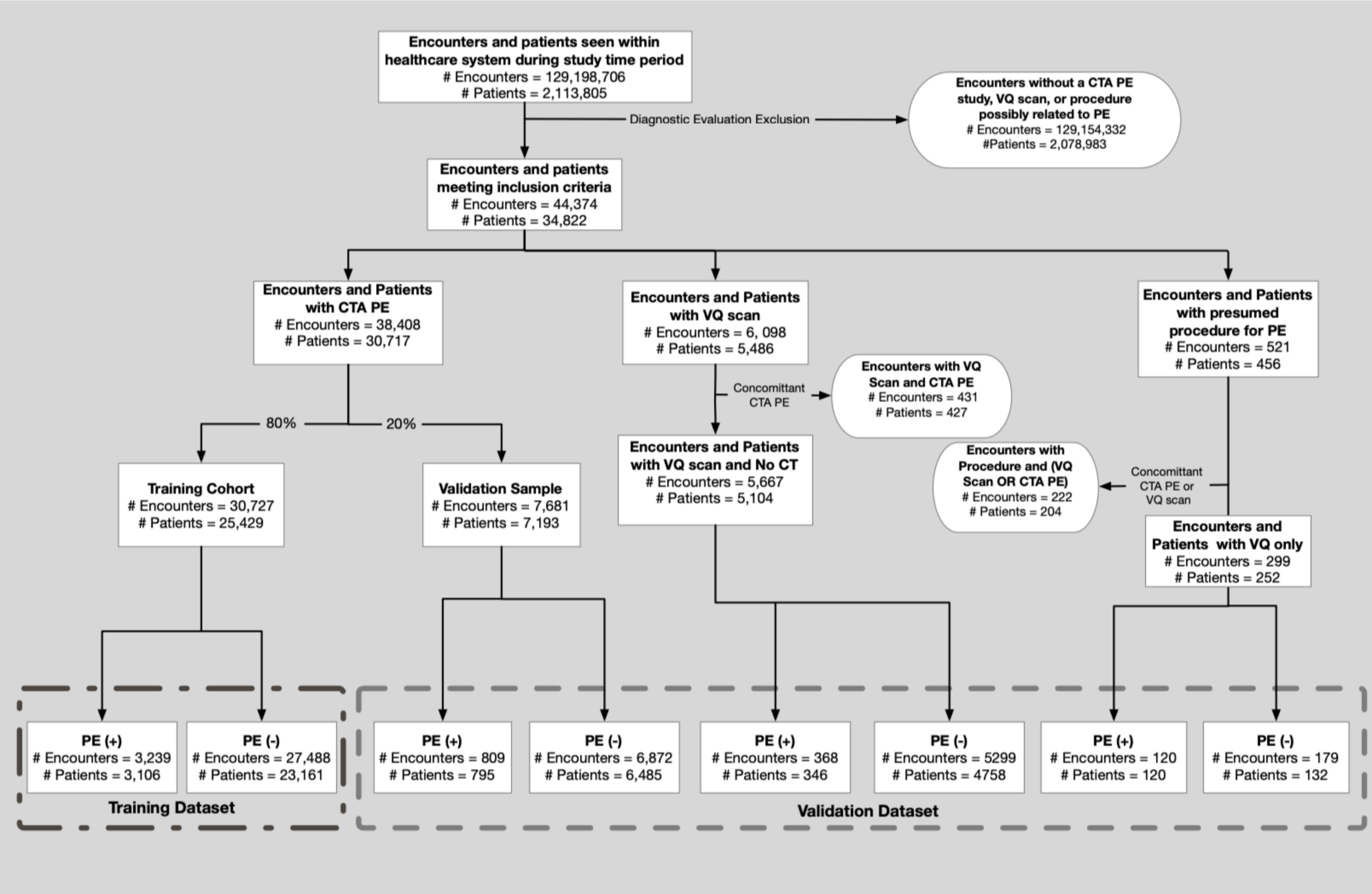
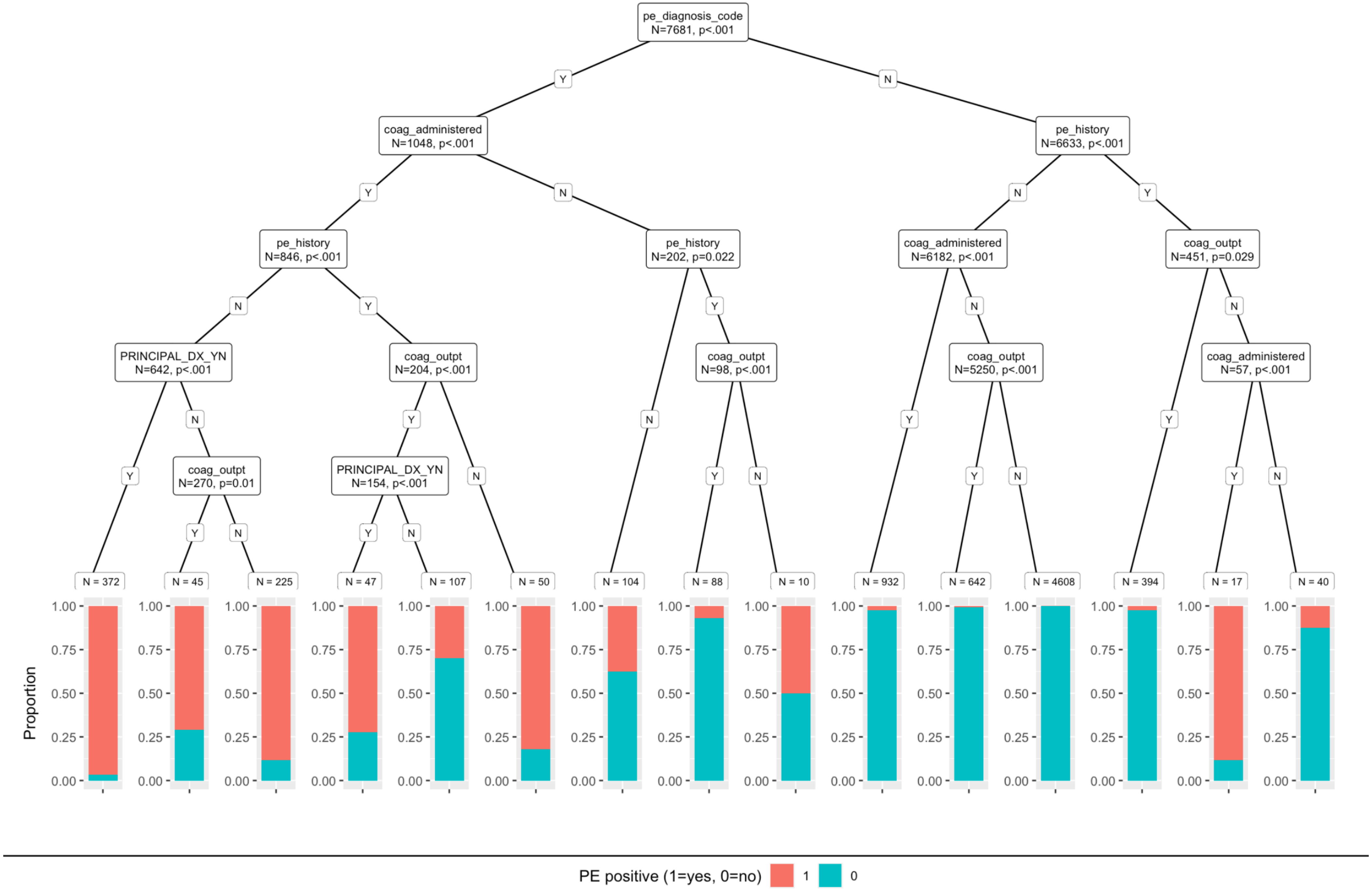


Table 12
peFinder classification performance on the test set.

State	PPV	Sensitivity	Specificity	Accuracy
DISEASE	0.83 (223/269)	0.98 (223/228)	0.89 (382/428)	0.92(605/656)
QUALITY	0.96 (113/118)	0.86 (113/131)	0.99 (520/525)	0.96 (633/656)
CERTAINTY	0.93 (313/ 336)	0.94 (313/ 333)	0.93 (300/ 323)	0.93(613/656)
TEMPORAL	0.90 (18/20)	0.60 (18/ 30)	0.99 (196/198)	0.94(214/228)

AI Method Structured Data


Encounters and patients seen within healthcare system during study time period
# Encounters = 129,198,706
# Patients = 2,113,805
Diagnostic Evaluation Exclusion
Encounters without a CTA PE study, VQ scan, or procedure possibly related to PE
# Encounters = 129,154,332
#Patients = 2,078,983
Encounters and patients meeting inclusion criteria
# Encounters = 44,374
# Patients = 34,822
Encounters and Patients with CTA PE
# Encounters = 38,408
# Patients = 30,717
Encounters and Patients with VQ scan
# Encounters = 6, 098
# Patients = 5,486
Encounters and Patients with presumed procedure for PE
# Encounters = 521
# Patients = 456
Concomittant CTA PE
Encounters with VQ Scan and CTA PE
# Encounters = 431
# Patients = 427
80%
20%
Training Cohort
# Encounters = 30,727
# Patients = 25,429
Validation Sample
# Encounters = 7,681
# Patients = 7,193
Encounters and Patients with VQ scan and No CT
# Encounters = 5,667
# Patients = 5,104
Encounters with Procedure and (VQ Scan OR CTA PE)
# Encounters = 222
# Patients = 204
Concomittant CTA PE or VQ scan
Encounters and Patients with VQ only
# Encounters = 299
# Patients = 252
PE (+)
# Encounters = 3,239
# Patients = 3,106
PE (-)
# Encouters = 27,488
# Patients = 23,161
PE (+)
# Encounters = 809
# Patients = 795
PE (-)
# Encounters = 6,872
# Patients = 6,485
PE (+)
# Encounters = 368
# Patients = 346
PE (-)
# Encounters = 5299
# Patients = 4758
PE (+)
# Encounters = 120
# Patients = 120
PE (-)
# Encounters = 179
# Patients = 132
Training Dataset
Validation Dataset


pe_diagnosis_code
N=7681, p<.001
Y
N
coag_administered
N=1048, p<.001
pe_history
N=6633, p<.001
pe_history
N=846, p<.001
pe_history
N=202, p=0.022
coag_administered
N=6182, p<.001
coag_outpt
N=451, p=0.029
PRINCIPAL_DX_YN
N=642, p<.001
coag_outpt
N=204, p<.001
coag_outpt
N=98, p<.001
coag_outpt
N=5250, p<.001
coag_administered
N=57, p<.001
coag_outpt
N=270, p=0.01
PRINCIPAL_DX_YN
N=154, p<.001
N = 372
N = 45
N = 225
N = 47
N = 107
N = 50
N = 104
N = 88
N = 10
N = 932
N = 642
N = 4608
N = 394
N = 17
N = 40
Proportion
1.00
0.75
0.50
0.25
0.00
PE positive (1=yes, 0=no)
1
0

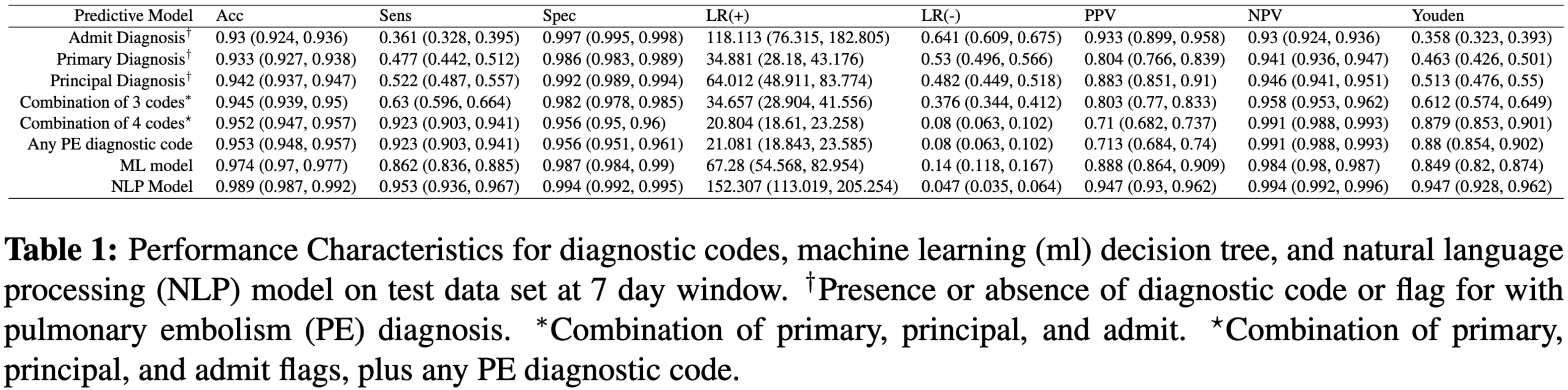
```
IF  Diagnostic Evaluation Trigger Event:
    IF Diagnostic Code for PE*:
        IF Anticoagulation administered*
            IF PE History prior to Event:
        IF receiving anticoagulation as outpatient prior to Event:
                    IF Principal diagnosis PE:
            YES
                    ELSE
            NO
        ELSE
                    YES
            ELSE
                YES
    ELSE
            NO
    ELSE
    IF PE History prior to Event:
            IF receiving anticoagulation as outpatient prior to Event:
                NO
            ELSE:
                IF Anticoagulation administered*:
                    YES
        ELSE:
                    NO
    ELSE:
            NO
ELSE:
    NO
```

Definitions:
Diagnostic Evaluation Trigger Event
    a. CTPE Order
    b. VQ Scan Order
    c. Procedure Code

* (+/- 7 days of Diagnostic Trigger):

Anticoagulation defined by medication list
-PE codes provided by ICD-10 list
-CTPE by CPT and names CTA PE
-VQ Scan by Name
-Procedure by procedure code list

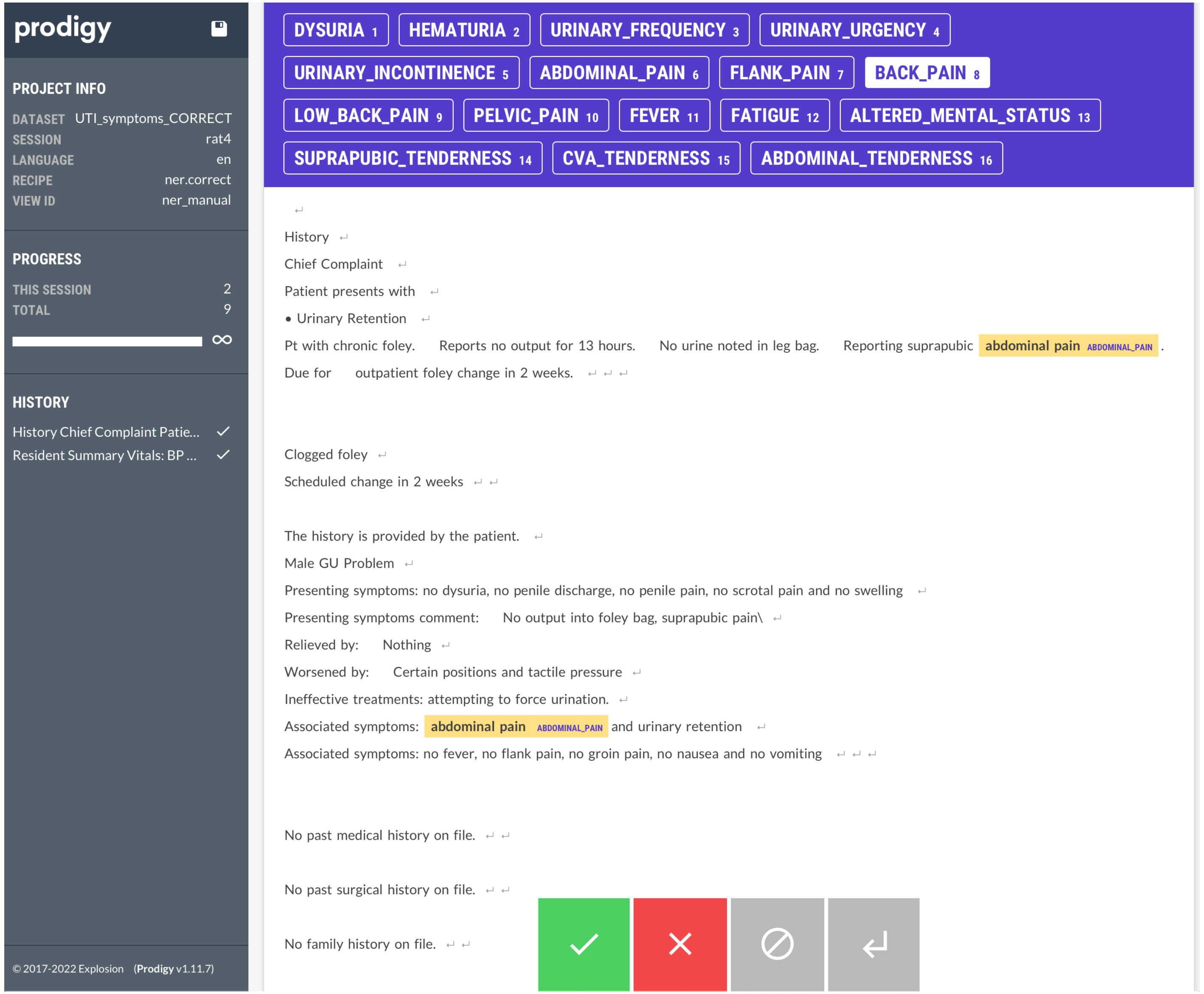
Predictive Performance Across Various Methods



Predictive Model	Acc	Sens	Spec	LR(+)	LR(-)	PPV	NPV	Youden
Admit Diagnosis†	0.93 (0.924, 0.936)	0.361 (0.328, 0.395)	0.997 (0.995, 0.998)	118.113 (76.315, 182.805)	0.641 (0.609, 0.675)	0.933 (0.899, 0.958)	0.93 (0.924, 0.936)	0.358 (0.323, 0.393)
Primary Diagnosis†	0.933 (0.927, 0.938)	0.477 (0.442, 0.512)	0.986 (0.983, 0.989)	34.881 (28.18, 43.176)	0.53 (0.496, 0.566)	0.804 (0.766, 0.839)	0.941 (0.936, 0.947)	0.463 (0.426, 0.501)
Principal Diagnosis†	0.942 (0.937, 0.947)	0.522 (0.487, 0.557)	0.992 (0.989, 0.994)	64.012 (48.911, 83.774)	0.482 (0.449, 0.518)	0.883 (0.851, 0.91)	0.946 (0.941, 0.951)	0.513 (0.476, 0.55)
Combination of 3 codes*	0.945 (0.939, 0.95)	0.63 (0.596, 0.664)	0.982 (0.978, 0.985)	34.657 (28.904, 41.556)	0.376 (0.344, 0.412)	0.803 (0.77, 0.833)	0.958 (0.953, 0.962)	0.612 (0.574, 0.649)
Combination of 4 codes⋆	0.952 (0.947, 0.957)	0.923 (0.903, 0.941)	0.956 (0.95, 0.96)	20.804 (18.61, 23.258)	0.08 (0.063, 0.102)	0.71 (0.682, 0.737)	0.991 (0.988, 0.993)	0.879 (0.853, 0.901)
Any PE diagnostic code	0.953 (0.948, 0.957)	0.923 (0.903, 0.941)	0.956 (0.951, 0.961)	21.081 (18.843, 23.585)	0.08 (0.063, 0.102)	0.713 (0.684, 0.74)	0.991 (0.988, 0.993)	0.88 (0.854, 0.902)
ML model	0.974 (0.97, 0.977)	0.862 (0.836, 0.885)	0.987 (0.984, 0.99)	67.28 (54.568, 82.954)	0.14 (0.118, 0.167)	0.888 (0.864, 0.909)	0.984 (0.98, 0.987)	0.849 (0.82, 0.874)
NLP Model	0.989 (0.987, 0.992)	0.953 (0.936, 0.967)	0.994 (0.992, 0.995)	152.307 (113.019, 205.254)	0.047 (0.035, 0.064)	0.947 (0.93, 0.962)	0.994 (0.992, 0.996)	0.947 (0.928, 0.962)

Table 1: Performance Characteristics for diagnostic codes, machine learning (ml) decision tree, and natural language processing (NLP) model on test data set at 7 day window. †Presence or absence of diagnostic code or flag for with pulmonary embolism (PE) diagnosis. *Combination of primary, principal, and admit. ⋆Combination of primary, principal, and admit flags, plus any PE diagnostic code.

How Do Identify PE Signs and Symptoms?



PROJECT INFO

DATASET UTI_symptoms_CORRECT
SESSION rat4
LANGUAGE en
RECIPE ner.correct
VIEW ID ner_manual

PROGRESS

THIS SESSION 2
TOTAL 9
∞

HISTORY

History Chief Complaint Patie... ✓
Resident Summary Vitals: BP ... ✓

© 2017-2022 Explosion (**Prodigy** v1.11.7)

History
Chief Complaint
Patient presents with
- Urinary Retention

Pt with chronic foley. Reports no output for 13 hours. No urine noted in leg bag. Reporting suprapubic abdominal pain ABDOMINAL_PAIN .
Due for outpatient foley change in 2 weeks.

Clogged foley
Scheduled change in 2 weeks

The history is provided by the patient.
Male GU Problem
Presenting symptoms: no dysuria, no penile discharge, no penile pain, no scrotal pain and no swelling
Presenting symptoms comment: No output into foley bag, suprapubic pain\
Relieved by: Nothing
Worsened by: Certain positions and tactile pressure
Ineffective treatments: attempting to force urination.
Associated symptoms: abdominal pain ABDOMINAL_PAIN and urinary retention
Associated symptoms: no fever, no flank pain, no groin pain, no nausea and no vomiting

No past medical history on file.

No past surgical history on file.

No family history on file.

Takeaways

Thank you!

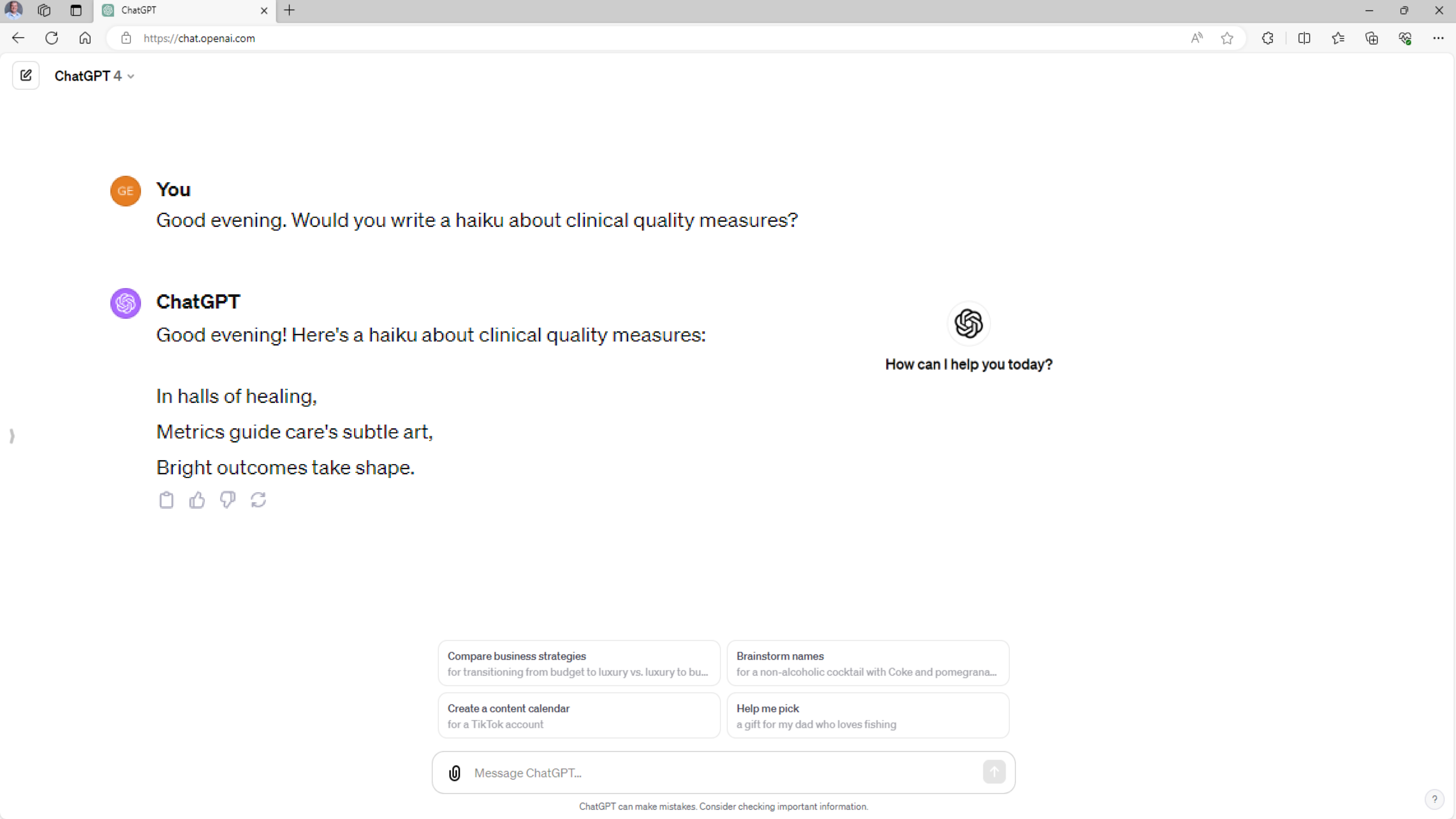
richard.taylor@yale.edu

Using Large Language Models (LLMs) to Generate Mechanism Maps

Jeff Geppert, geppertj@battelle.org
Battelle
March 5, 2024

What is Generative AI?

- Definition: AI models that emulate the structure and characteristics of input data to generate derived synthetic content, including images, videos, audio, text, and other media
  - Uses Large Language Models (LLM)
  - E.g., ChatGPT (https://chat.openai.com)
- Emerging Gen AI Use Cases
  - Create content – enhance transparency, equity, up-skill, "first-draft"
  - Elaborate, annotate, synthesize – enhance reasoning
  - Automate – enhance productivity
- Gen AI Best Practices
  - Ontologies and verifying claims – maximize benefits, mitigate errors


ChatGPT
https://chat.openai.com
ChatGPT 4
GE
How can I help you today?
Compare business strategies
for transitioning from budget to luxury vs. luxury to bu...
Brainstorm names
for a non-alcoholic cocktail with Coke and pomegrana...
Create a content calendar
for a TikTok account
Help me pick
a gift for my dad who loves fishing
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.
?

You

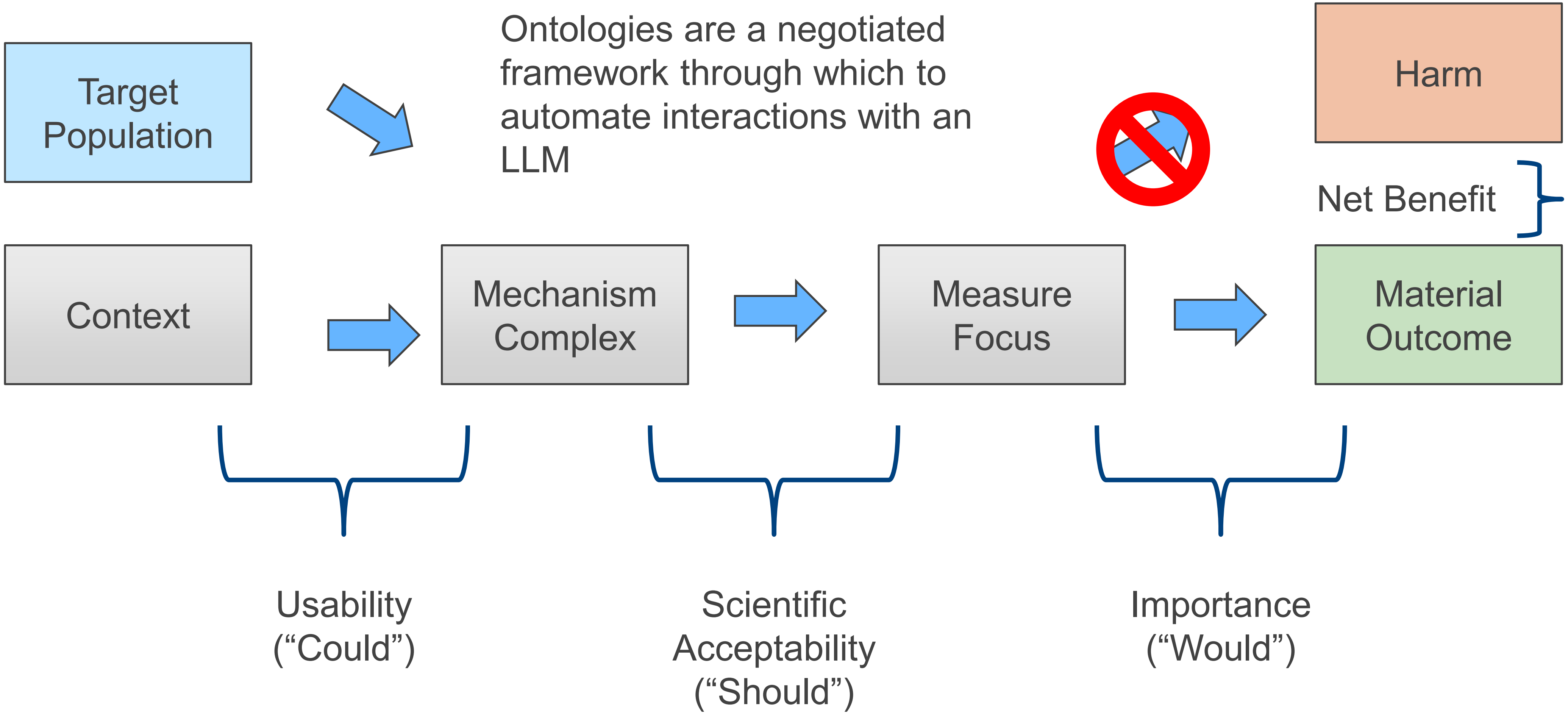
Good evening. Would you write a haiku about clinical quality measures?

ChatGPT

Good evening! Here's a haiku about clinical quality measures:

In halls of healing,
Metrics guide care's subtle art,
Bright outcomes take shape.

Model Input – CMO Ontology



Target Population

Ontologies are a negotiated framework through which to automate interactions with an LLM

Harm

Net Benefit

Context

Mechanism Complex

Measure Focus

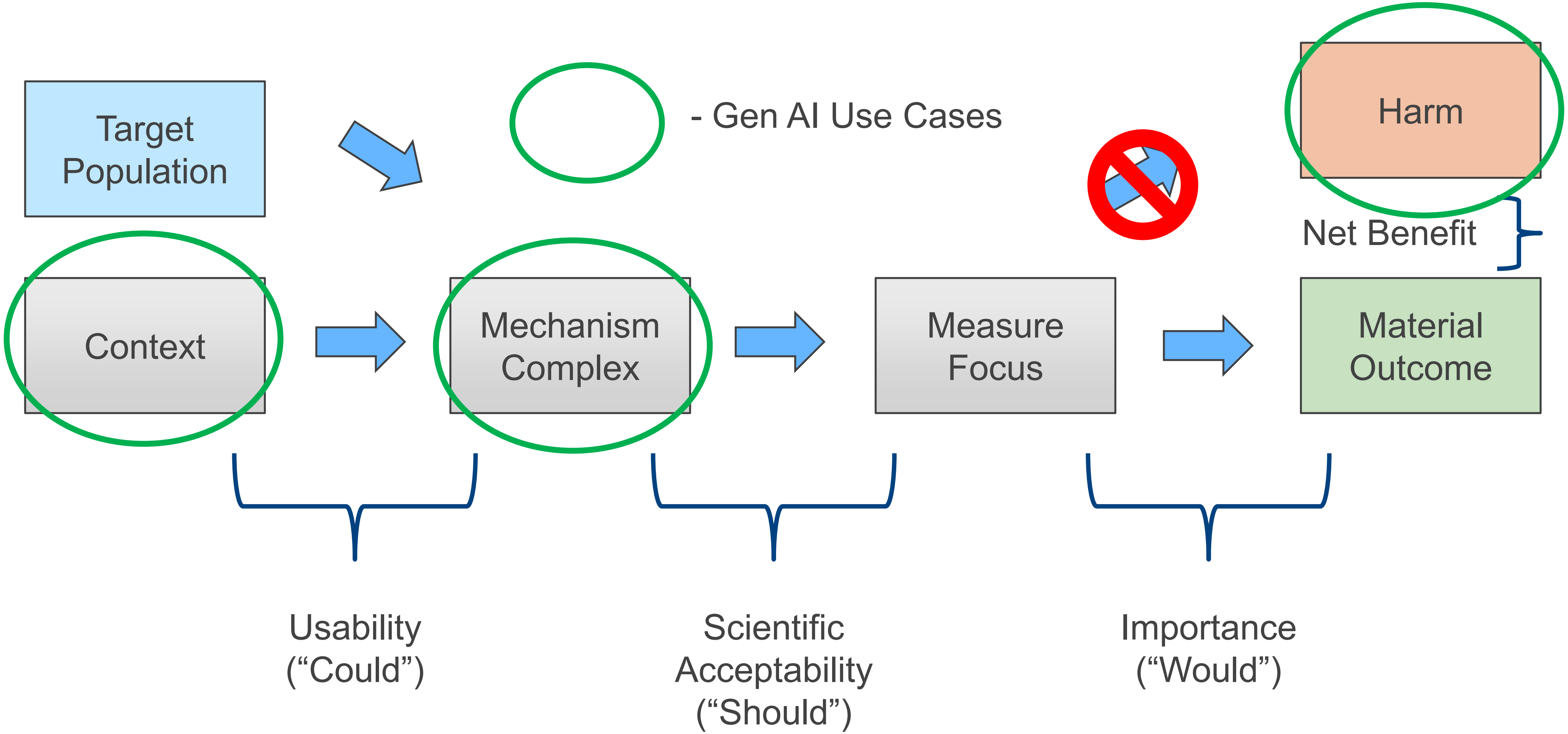
Material Outcome

Usability (“Could”)

Scientific Acceptability (“Should”)

Importance (“Would”)

Model Input – CMO Ontology Cont.



Target Population

- Gen AI Use Cases

Harm

Net Benefit

Context

Mechanism Complex

Measure Focus

Material Outcome

Usability
("Could")

Scientific
Acceptability
("Should")

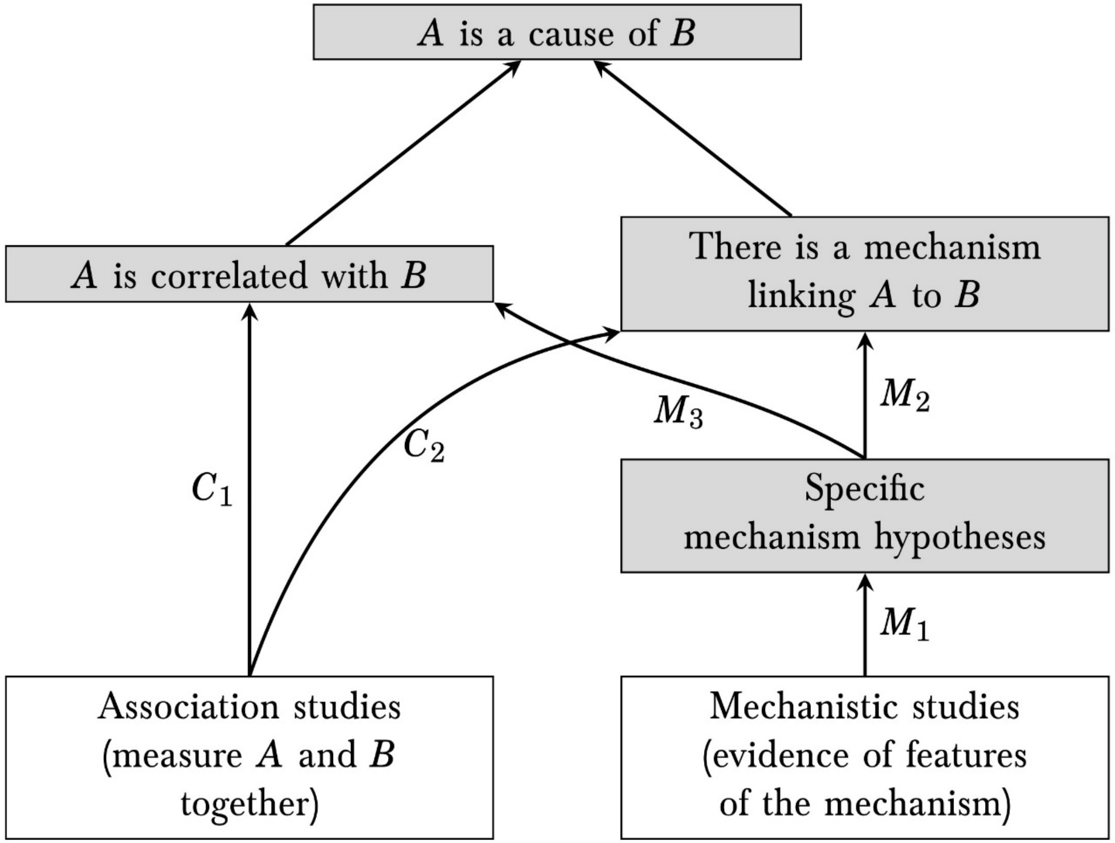
Importance
("Would")

Model Output - Verifying Claims

- Measure developers and/or measure stewards make certain explicit or *implicit* assertions or claims about the potential benefits and risks/harms associated with measure use
- In general, there are three top-level claims related to measure properties necessary for a measure to yield positive net benefit to persons and entities:
  - *Would* claim: Person or entity *would* make decisions based on the measure because the measure focus is associated with a material outcome (end/importance)
  - *Should* claim: There are known and effective ways of selection or improvement that the person or entity *should* use (ways/scientific acceptability)
  - *Could* claim: Any barriers or facilitators to whether the person or entity *could* use those ways are known and addressed (means/usability)

What is made explicit? Causal Claims

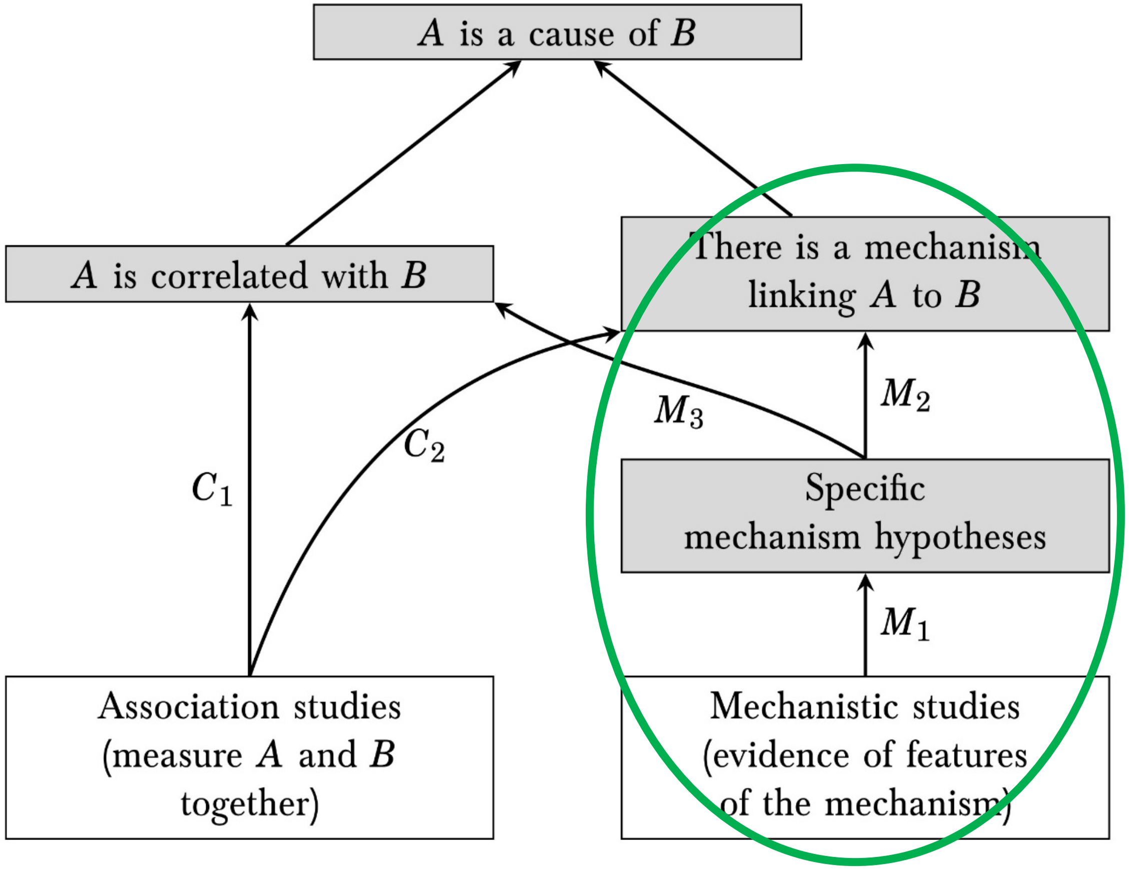
- A is a cause of B
  - A: an intervention (drug, device, procedure, quality program, service delivery model, payment model)
  - B: an outcome (mortality, morbidity, harm, functional status, patient experience, workforce burden, measure focus)
- Association claims
  - A is correlated with B
- General mechanism claims
  - A is responsible for B
  - Accounts for the association


A is a cause of B
A is correlated with B
There is a mechanism linking A to B
M_3
M_2
C_2
C_1
Specific mechanism hypotheses
M_1
Association studies (measure A and B together)
Mechanistic studies (evidence of features of the mechanism)

Source: Shan, Y., Williamson, J. (2023). Evidential Pluralism in the Social Sciences. United States: Taylor & Francis.

What is made explicit? Causal Claims Cont.

- A is a cause of B
  - A: an intervention (drug, device, procedure, quality program, service delivery model, payment model)
  - B: an outcome (mortality, morbidity, harm, functional status, patient experience, workforce burden, measure focus)
- Association claims
  - A is correlated with B
- General mechanism claims
  - A is responsible for B
  - Accounts for the association



A is a cause of B
A is correlated with B
There is a mechanism linking A to B
M_3
M_2
C_2
C_1
Specific mechanism hypotheses
M_1
Association studies (measure A and B together)
Mechanistic studies (evidence of features of the mechanism)

Source: Shan, Y., Williamson, J. (2023). Evidential Pluralism in the Social Sciences. United States: Taylor & Francis.

Use LLMs to Generate Mechanistic Claims

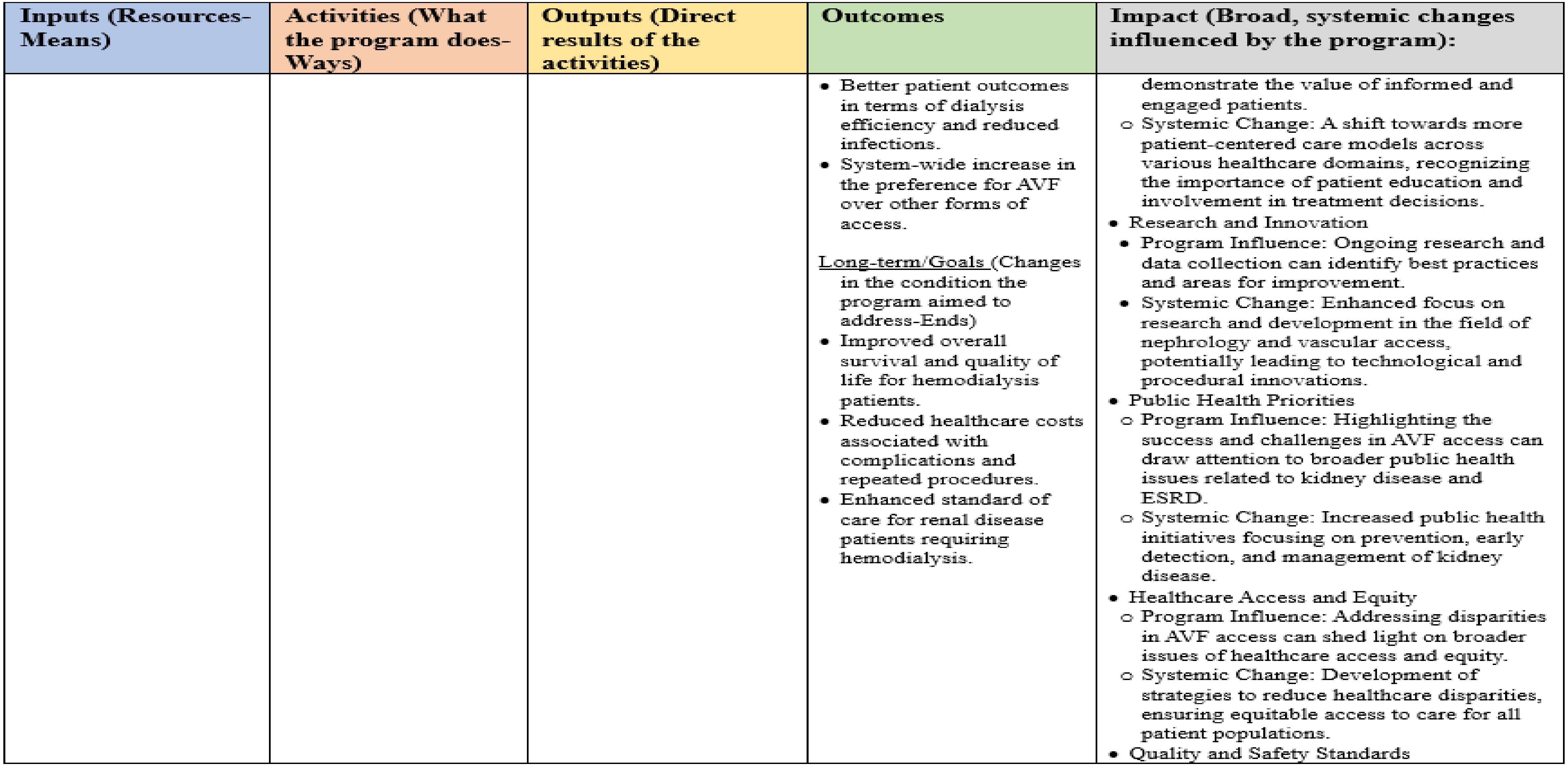
- Would you describe a logic model for [increasing, decreasing] the likelihood of [measure focus] in the [target population]?
- For example,
  - Measure focus: persons receiving hemodialysis
  - Target population: receiving Arteriovenous fistula (AVF) vascular access
  - Would you describe a logic model for increasing the likelihood of receiving Arteriovenous fistula (AVF) vascular access in this persons receiving hemodialysis ?

Certainly! Hemodialysis is a life-saving treatment for patients with end-stage renal disease (ESRD). To perform hemodialysis, patients require a reliable and efficient means of vascular access, which allows blood to flow from the patient to the dialysis machine and back. There are three main types of vascular access for hemodialysis:

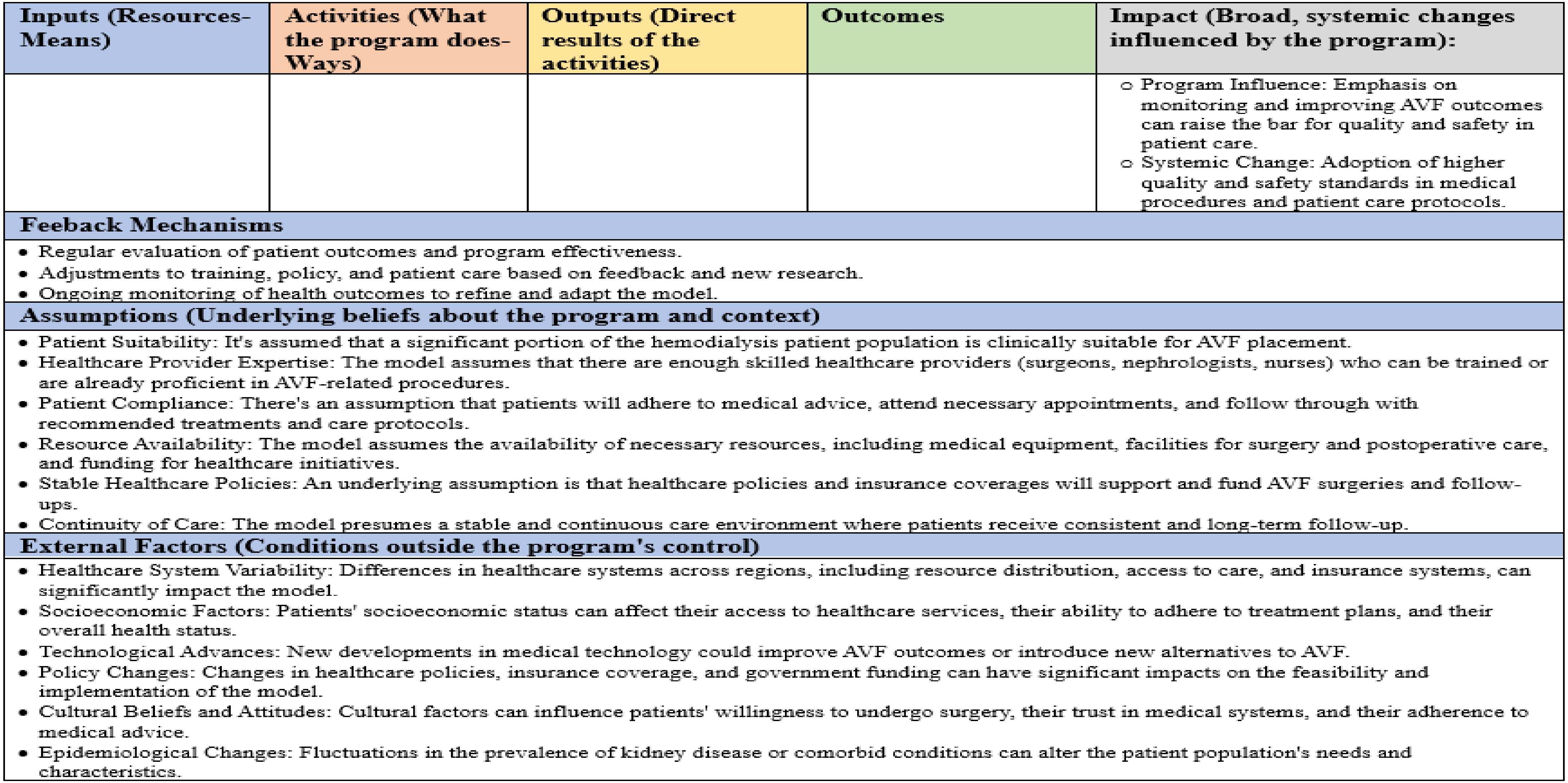
1. Arteriovenous fistula (AVF)
2. Arteriovenous graft (AVG)
3. Central venous catheter (CVC)

The arteriovenous fistula (AVF) is created by directly connecting a patient's artery to a vein. Over time, the vein enlarges and thickens, facilitating the drawing and return of blood during dialysis.

Inputs (Resources-Means)	**Activities (What the program does-Ways)**	**Outputs (Direct results of the activities)**	**Outcomes**	**Impact (Broad, systemic changes influenced by the program):**
• Skilled healthcare professionals (nephrologists, surgeons, nurses). • Training programs for AVF placement and maintenance. • Medical equipment and facilities for surgery and follow-up care. • Patient education materials. • Funding for healthcare initiatives. • Support from healthcare policy and administration. • Access to patient data and healthcare records for monitoring.	• Early screening and identification of patients for AVF. • Preoperative vascular mapping to assess suitability for AVF. • Surgical creation of AVF. • Postoperative monitoring and care for AVF maturation. • Ongoing training and education for healthcare providers. • Patient education and counseling about the benefits and care of AVF. • Policy advocacy for supporting AVF use.	• Number of patients screened for AVF suitability. • Number of AVFs surgically created. • Number of healthcare providers trained in AVF-related procedures. • Educational sessions conducted for patients. • Policy changes or implementations supporting AVF use.	Short-term (Changes resulting from the outputs): • Increased awareness among patients and healthcare providers about the benefits of AVF. • Improved patient selection for AVF placement. • Enhanced skills among healthcare providers for creating and maintaining AVFs. • Improved patient readiness and compliance for AVF surgery. • Policy and systemic changes facilitating increased AVF use. Intermediate term (effects observed as the program matures) • Increased rate of successful AVF placements. • Reduced complications and failures in AVF post-surgery.	• Healthcare Policy and Funding ○ Program Influence: Advocacy and demonstrated success of the program can lead to changes in healthcare policies, prioritizing funding for AVF procedures and postoperative care. ○ Systemic Change: Shift in national or regional healthcare funding and policies to support early and efficient access to AVF for eligible patients. • Standardization of Care Practices ○ Program Influence: Implementation of best practices for AVF creation and maintenance could set a benchmark for care quality. ○ Systemic Change: Adoption of these standards across healthcare systems, leading to a more uniform approach to hemodialysis vascular access. • Training and Workforce Development ○ Program Influence: The focus on training and continuous education can highlight the need for specialized skills in nephrology and vascular surgery. ○ Systemic Change: Changes in medical education and professional development requirements, ensuring a well-trained workforce proficient in AVF management. • Patient Education and Engagement ○ Program Influence: Comprehensive patient education initiatives can



Inputs (Resources-Means)	Activities (What the program does-Ways)	Outputs (Direct results of the activities)	Outcomes	Impact (Broad, systemic changes influenced by the program):
			• Better patient outcomes in terms of dialysis efficiency and reduced infections. • System-wide increase in the preference for AVF over other forms of access. Long-term/Goals (Changes in the condition the program aimed to address-Ends) • Improved overall survival and quality of life for hemodialysis patients. • Reduced healthcare costs associated with complications and repeated procedures. • Enhanced standard of care for renal disease patients requiring hemodialysis.	demonstrate the value of informed and engaged patients. ○ Systemic Change: A shift towards more patient-centered care models across various healthcare domains, recognizing the importance of patient education and involvement in treatment decisions. • Research and Innovation • Program Influence: Ongoing research and data collection can identify best practices and areas for improvement. • Systemic Change: Enhanced focus on research and development in the field of nephrology and vascular access, potentially leading to technological and procedural innovations. • Public Health Priorities ○ Program Influence: Highlighting the success and challenges in AVF access can draw attention to broader public health issues related to kidney disease and ESRD. ○ Systemic Change: Increased public health initiatives focusing on prevention, early detection, and management of kidney disease. • Healthcare Access and Equity ○ Program Influence: Addressing disparities in AVF access can shed light on broader issues of healthcare access and equity. ○ Systemic Change: Development of strategies to reduce healthcare disparities, ensuring equitable access to care for all patient populations. • Quality and Safety Standards



Inputs (Resources-Means)	Activities (What the program does-Ways)	Outputs (Direct results of the activities)	Outcomes	Impact (Broad, systemic changes influenced by the program):
				○ Program Influence: Emphasis on monitoring and improving AVF outcomes can raise the bar for quality and safety in patient care. ○ Systemic Change: Adoption of higher quality and safety standards in medical procedures and patient care protocols.

Feeback Mechanisms

- Regular evaluation of patient outcomes and program effectiveness.
- Adjustments to training, policy, and patient care based on feedback and new research.
- Ongoing monitoring of health outcomes to refine and adapt the model.

Assumptions (Underlying beliefs about the program and context)

- Patient Suitability: It's assumed that a significant portion of the hemodialysis patient population is clinically suitable for AVF placement.
- Healthcare Provider Expertise: The model assumes that there are enough skilled healthcare providers (surgeons, nephrologists, nurses) who can be trained or are already proficient in AVF-related procedures.
- Patient Compliance: There's an assumption that patients will adhere to medical advice, attend necessary appointments, and follow through with recommended treatments and care protocols.
- Resource Availability: The model assumes the availability of necessary resources, including medical equipment, facilities for surgery and postoperative care, and funding for healthcare initiatives.
- Stable Healthcare Policies: An underlying assumption is that healthcare policies and insurance coverages will support and fund AVF surgeries and follow-ups.
- Continuity of Care: The model presumes a stable and continuous care environment where patients receive consistent and long-term follow-up.

External Factors (Conditions outside the program's control)

- Healthcare System Variability: Differences in healthcare systems across regions, including resource distribution, access to care, and insurance systems, can significantly impact the model.
- Socioeconomic Factors: Patients' socioeconomic status can affect their access to healthcare services, their ability to adhere to treatment plans, and their overall health status.
- Technological Advances: New developments in medical technology could improve AVF outcomes or introduce new alternatives to AVF.
- Policy Changes: Changes in healthcare policies, insurance coverage, and government funding can have significant impacts on the feasibility and implementation of the model.
- Cultural Beliefs and Attitudes: Cultural factors can influence patients' willingness to undergo surgery, their trust in medical systems, and their adherence to medical advice.
- Epidemiological Changes: Fluctuations in the prevalence of kidney disease or comorbid conditions can alter the patient population's needs and characteristics.

Next Steps

- LLMs are useful for generating claims, particularly in areas that are not as fully addressed in the published literature: harms, mechanisms, and context
- A more explicit articulation of mechanisms is a complementary (substitute?) approach for establishing validity in addition to correlation (association) studies
  - A claim that a correlation between Measure A and Measure B establishes the validity of either Measure A or Measure B should be supported by a mechanism map that justifies that claim
- Claims generated by LLMs are “first drafts” and should be supported with evidence, which *includes* expert review
  - A mechanism map/logic model should be an output of the Technical Expert Panel (TEP)
  - A more robust form of “face validity”
  - The aim is “plausibility” and “usefulness” for making decisions

AI & QI

Pierre Barker (Senior Sponsor), Gary Kantor (Faculty), Jeff Rakover (Research Lead), Marina Renton (RA)

Institute for Healthcare Improvement

Before We Begin

The tools we discuss today are fun and easy to use, but be sure to check your internal IT policies about which tools you may use for professional purposes

Thanks to our thought partners (interviewees, colleagues, and those joining us today

Terminology: "Artificial intelligence" (AI) can mean many things—from robotics to machine learning. We are focused here on the AI tools that have gained prominence in the past ~one year after the release of OpenAI's ChatGPT—these are "large language models" (LLMs) that use statistical processes to generate (mainly) textual and numeric responses to user queries.

Project Aim and Background

(1) Understand potential benefits of AI for QI work

(2) Build use cases

(3) Understand risks and brainstorm mitigation strategies

(4) Build guidance for the field

Note on Scope: Protecting ourselves from the risks of AI (across sectors) will require a robust public policy response and a proactive response from organizational leaders. We are almost certainly far behind where we need to be. Our research focus was not public policy, but we acknowledge its importance to the current and future AI outlook.

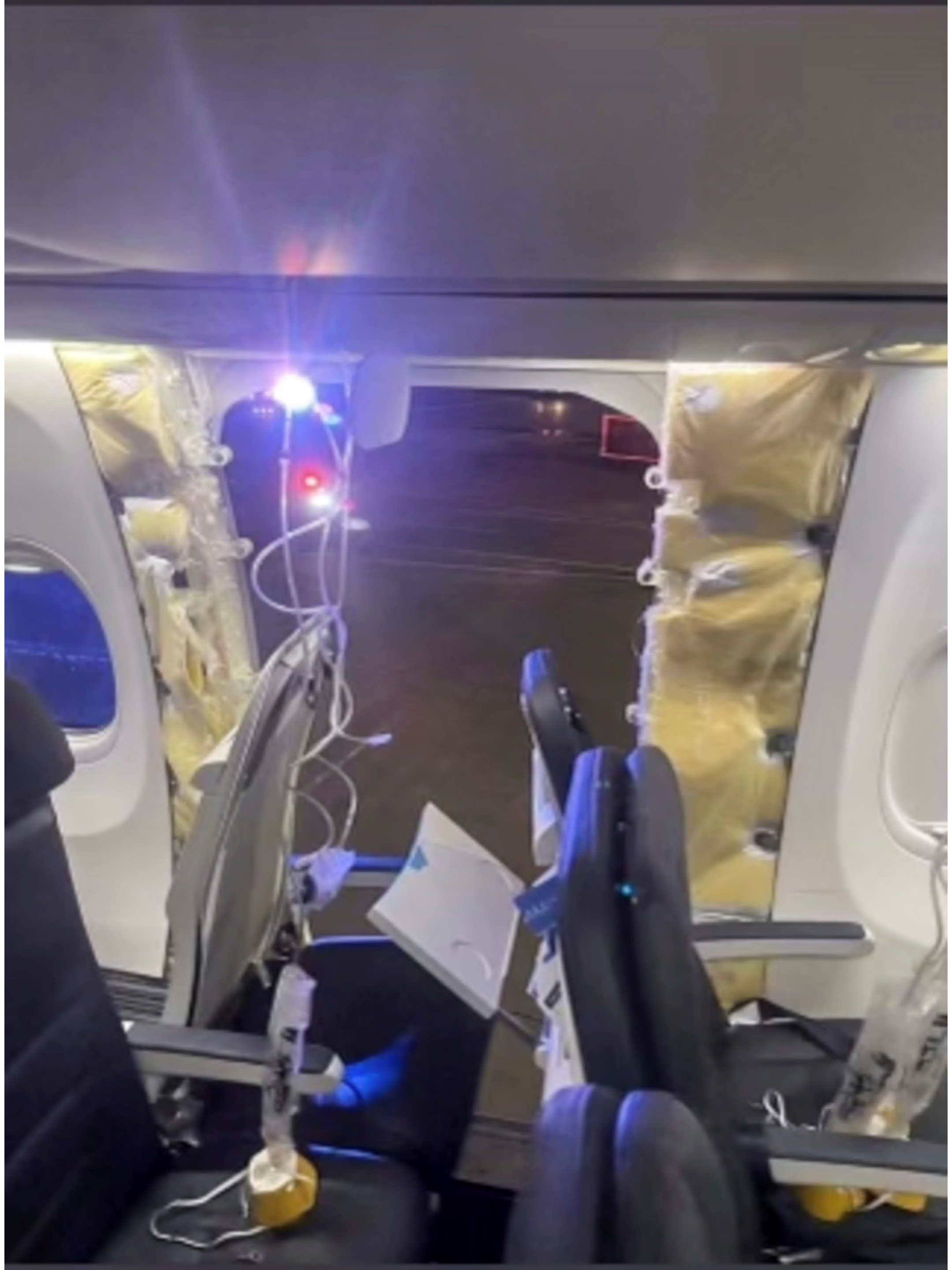
INTERVIEWS

LIT SCAN

REAL TIME TESTING

What does quality mean today?

Quality is everywhere in the news…



The AI connection

We believe that generative AI is going to radically transform how healthcare systems (and other organizations) approach quality.

Why?

Because AI helps with activities like:

- Collecting data (in real time) on the behavior of service producers and service users (e.g., ambient listening, remote monitoring, documentation)
- Conducting rapid analyses, identifying problems, making suggestions
- Ultimately making quality control, improvement, and planning more accessible to everyone

How pervasive is Gen AI now?

AMA Survey: 38% of physicians were using AI as of August 2023 (most common uses: creating discharge instructions, care plans, progress notes; documenting billing codes, medical charts, or notes; translation; assistive diagnosis)

KLAS Research: “In general, the number of organizations adopting generative AI solutions is relatively small, and most who have adopted say it is still too early to report outcomes….Although only 25% of interviewed respondents have implemented generative AI solutions, 58% say their organization is likely to implement or purchase a solution within the next year.”

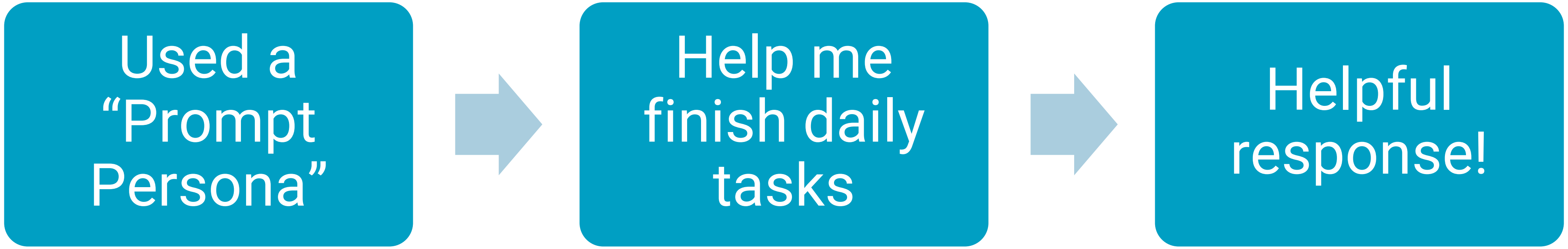
AI and Quality outside Healthcare

- GenAI already coming into play as quality tool outside healthcare
- Quality assurance for those who work in customer service (after the fact review of calls for quality of interactions)
- Co-pilot for programmers: making suggestions in real time to prevent problems/errors

Likely direction for healthcare…

Quality Control & Assurance	Quality Improvement	Quality Planning
• Monitoring clinical interactions against standard operating procedures and clinical best practice to identify gaps as they happen and prompt intervention • Customized interventions that facilitate care transitions (e.g., discharge preparation/education) • Potentially, use of video monitoring to identify risks before they turn into problems (e.g., central line insertion)	• Application of Gen AI to make QI work faster, more accessible • Automatic data review (e.g., use of AI agents that run data on an intentional, periodic basis) to identify problems, opportunity for quality improvement (e.g., automatically updating run charts or SPC charts and identifying unwanted variation) • Making QI teaching more accessible (via AI-assisted personalized coaching)	• Chatbots enable broader touchpoints for patient and family feedback (e.g., Juno tool for AI-user led interviews); more iterative/interactive than survey questions • AI enables faster synthesis of large amounts of qualitative feedback (e.g., focus groups) • AI can support brainstorming to translate feedback into actionable opportunities (e.g., measures to include, strategic aims or goals)

QI Chatbot as a Coach



Used a "Prompt Persona"

Help me finish daily tasks

Helpful response!

Bot: It sounds like there might be an opportunity to adjust your workload and pacing expectations. Quality improvement often involves making changes to the system, rather than focusing solely on individual performance. If you're finding your own deadlines challenging to meet, it's possible that your expectations may need to be adjusted. Start by reviewing your own task list and deadlines. Consider if these are realistic and achievable. If necessary, adjust them to a more manageable pace. This not only sets a good example for your team member, but also ensures that your expectations for both yourself and your team member are attainable. Does this seem like a good next step for you?

Me: Maybe it would make sense for me to start my own quality improvement project on my tasks! But I'm not sure how to set this up.

Bot: Absolutely! Starting a quality improvement project on your own tasks is a great idea. Let's begin with first few steps:

Step 1: Define the problem in measurable terms. For instance, you might define the problem as "the number of my tasks that are not completed by their initial deadline.“ Can you confirm if this is the problem you would like to tackle, or is there a different measurable issue you're encountering with your tasks?

Me: Yes, I think this makes sense.

Bot: Great! Step 2: Next, start collecting some data. Track the tasks that are not completed by their original deadline. This could be as simple as keeping a tally on a piece of paper or in a spreadsheet. For each task that isn't completed on time, note down any reasons or barriers that contributed to the delay. This information will be helpful later when you're trying to figure out what changes to make. Does this sound doable for you? Would you like any more guidance on this step?

Me: I can probably do this for this one project yes. How many days should I do this?

Bot: Step 3: The length of data collection can vary based on the frequency and number of tasks. A good starting point could be two weeks to a month. This should give you a reasonable amount of data to start identifying patterns and potential areas for improvement. Remember, the goal is not to collect data indefinitely, but to gather enough information to understand the problem and start making improvements. Does a two weeks to a month timescale sound reasonable to you? Would you like guidance on what to do once you've collected this data?

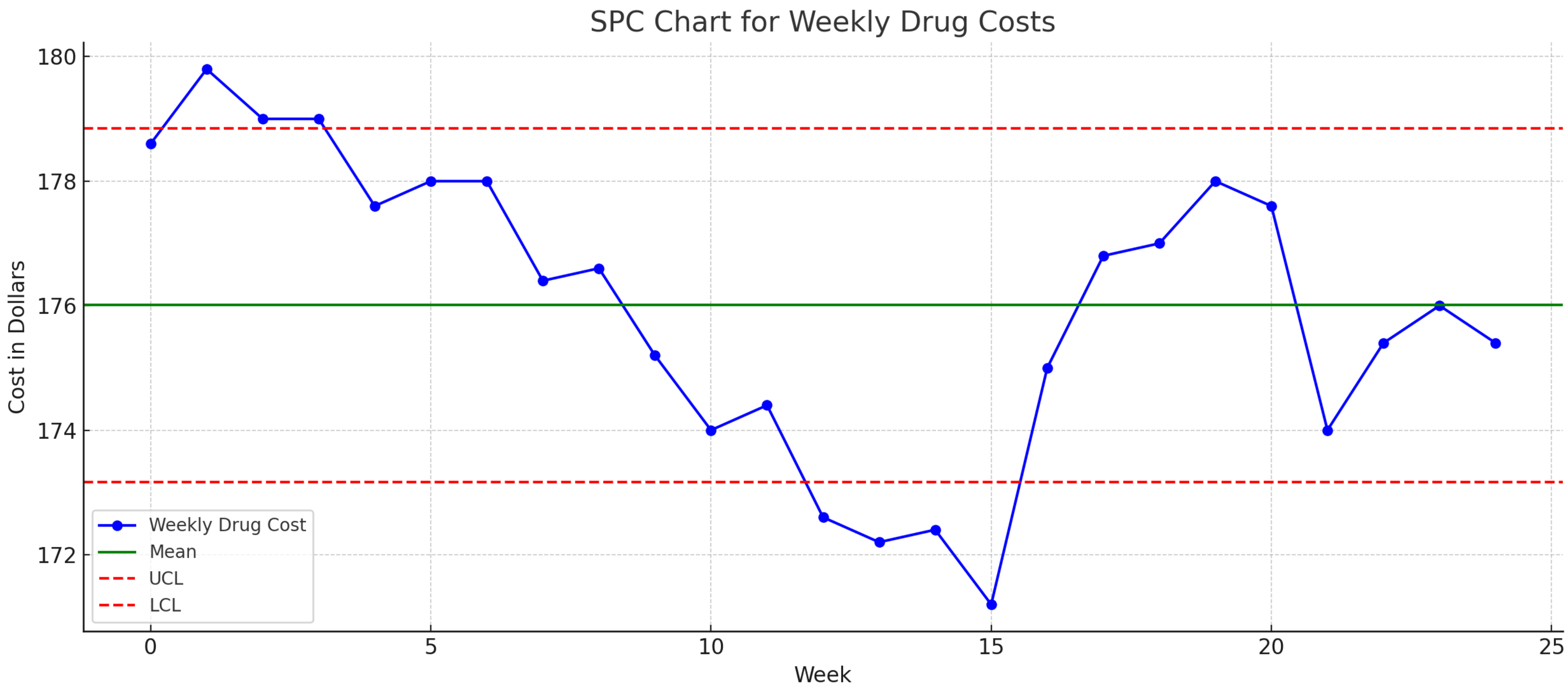
Me: well the project is only 3 months long. Isn't two weeks a bit long? I can do this though if that's what you think would be best.

Bot: Given the duration of your project, a two-week data collection period might indeed be too long. In this case, a shorter period, such as one week, might be more appropriate. This still should give you a reasonable amount of data to help understand the problem. Once you've collected the data, the next step will be to analyze it and identify patterns or common issues that could be contributing to the delays. Does this adjusted timeframe sound more manageable for your project?

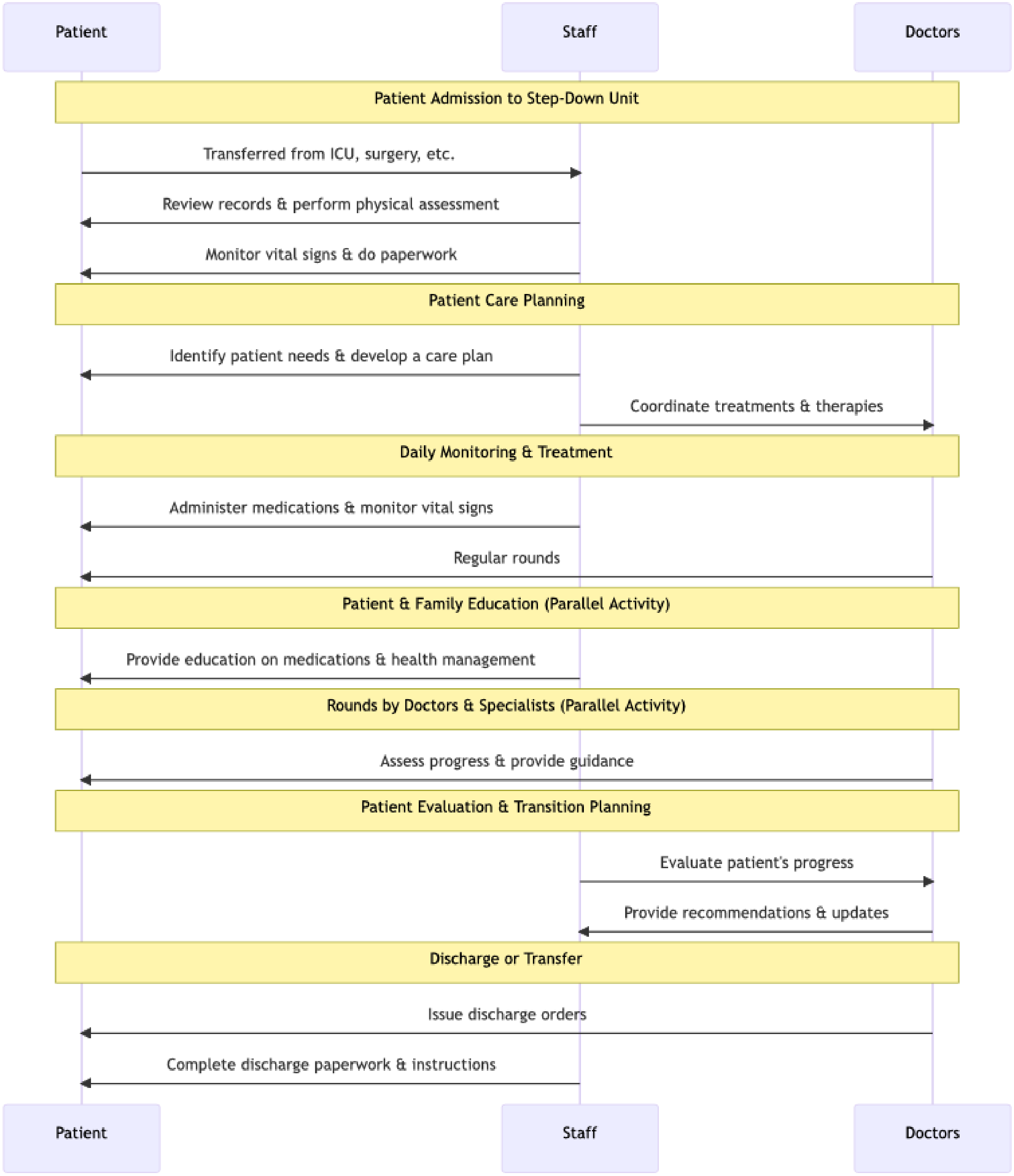
Me: Sure I can do that for one week.

Bot: Excellent!

SPC Chart


SPC Chart for Weekly Drug Costs
Cost in Dollars
180
178
176
174
172
Week
0
5
10
15
20
25
Weekly Drug Cost
Mean
UCL
LCL

Prompt: *"Build a process map for a hospital stepdown unit. Note that some of the processes (e.g., process care planning, daily monitoring and treatment) happen in parallel."*



Patient

Staff

Doctors

Patient Admission to Step-Down Unit

Transferred from ICU, surgery, etc.

Review records & perform physical assessment

Monitor vital signs & do paperwork

Patient Care Planning

Identify patient needs & develop a care plan

Coordinate treatments & therapies

Daily Monitoring & Treatment

Administer medications & monitor vital signs

Regular rounds

Patient & Family Education (Parallel Activity)

Provide education on medications & health management

Rounds by Doctors & Specialists (Parallel Activity)

Assess progress & provide guidance

Patient Evaluation & Transition Planning

Evaluate patient's progress

Provide recommendations & updates

Discharge or Transfer

Issue discharge orders

Complete discharge paperwork & instructions

Patient

Staff

Doctors

Other Examples

1. Pareto analysis
2. Run chart
3. Cause & effect diagram
4. Driver diagram

Six key risks and problems

1. Inaccuracy: AI is an associational predictive model; sometimes its predictions are wrong. E.g., 1+1 = 3
2. Hallucination: AI makes things up sometimes (e.g., a citation to a paper that doesn’t actually exist) that look credible
3. Privacy violation: Many AI tools take ownership of any data you input (e.g., they use the data to generate more predictions). This means the information you put in comes into the public domain.
4. Status quo bias: AI tools make use of what already exists in information systems like the internet. That means they often produce information that reflects the status quo rather than coming up with creative or new thinking.
5. Inequity: Related to #4, AI can exacerbate inequities by continuing to reenforce biases existent in information systems like the Internet (what is out there is already biased).
6. Lack of evidentiary discernment: Different from #1, AI may not weight the relative quality of information it provides. If you ask it for change ideas, it may not provide you with the most evidence-based ones, or give you a sense of which have the best data behind them.
7. Surveillance: Especially for QA/QC, these technologies may turn healthcare settings into ones where providers feel constantly monitored, which could be disastrous for staff experience

AI & QI Guardrails – A Starting Point

1) Use AI as an “extender” – it should complement what current staff are doing. Do not use it as a reason to downsize or eliminate already resource-strapped quality departments.
2) Assume that AI is centering dominant perspectives (e.g., white, male. Heteronormative, cis). Be ready to counter this by utilizing suggestions from our work on anti-racism and QI. Apply an anti-racist lens to its products, and deploy context experts and people with lived experience to analyze and scrutinize what it develops in order to counter bias and support innovative thinking.
3) Assume that AI poses an inherent risk to patient safety as a fallible system that hallucinates and produces inaccurate information. Whenever applying AI in a context that is patient-care-sensitive (e.g., producing change ideas that have possible safety implications), ALWAYS scrutinize the output together with subject matter experts.
4) Assume AI wants your data. Never share proprietary information with AI, and absolutely never share PHI.
5) Ensure humans are making decisions. Use AI to help when tasks are relatively clearcut and you can define an “A to B.”

BATTELLE

It can be done