Trustworthy Machine Learning for Healthcare

Stanford CS329T Fall 2023

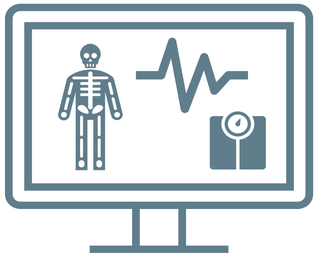
LAYER HEALTH

Monica Agrawal
Divya Gopinath

Electronic Health Records (EHRs)

EHRs contain a wealth of patient data.

And they have seen rapid adoption in the US:



	Hospitals with EHRs	Office Physicians with EHRs
2011	28%	34%
2021	96%	78%

Via https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records

Potential of EHRs

Real-world evidence in EHRs can facilitate personalized medicine.

Clinical trials can't answer every question:

- What drug would lead to the **best outcome** for **this patient**?
- What is the patient's expected **disease trajectory**?
- What **adverse events** might come from this drug combination?

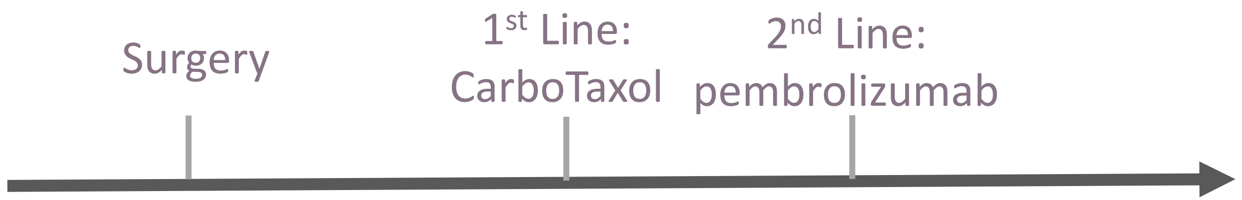
Variables of Interest

- **Disease**

Stage IV endometrial cancer

- **Disease Status**

- **Interventions**



- **Symptoms/ side effects**

- **Confounders**

Cirrhosis, CHF

The challenge

Many of these variables are not in structured data, but trapped in **messy, free-text** clinical notes:

Efficiency of documentation

Splintered care

Deviation from original care plan

How messy can notes be?

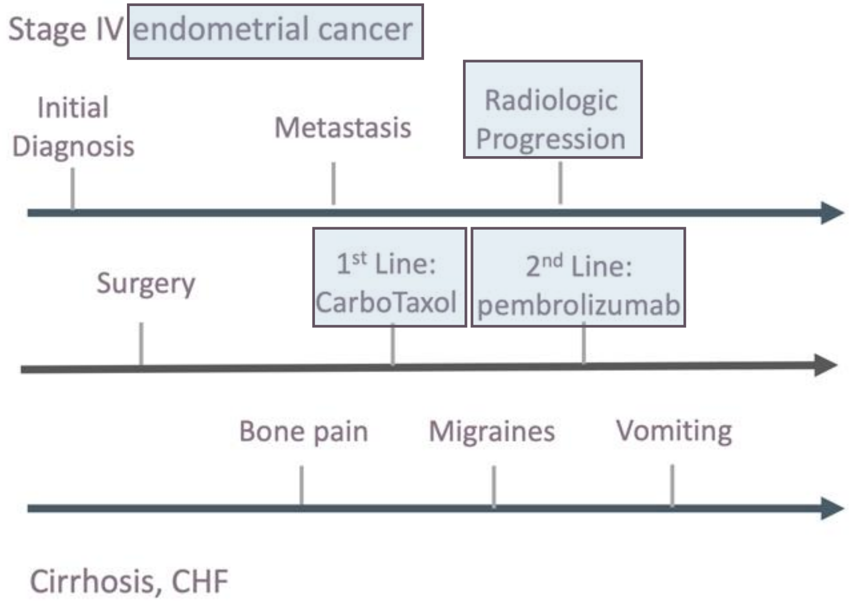
"...pt progressed after 5 mos of CarboTaxo for EC. Will dc and discuss pembro..."

Deciphering clinical text

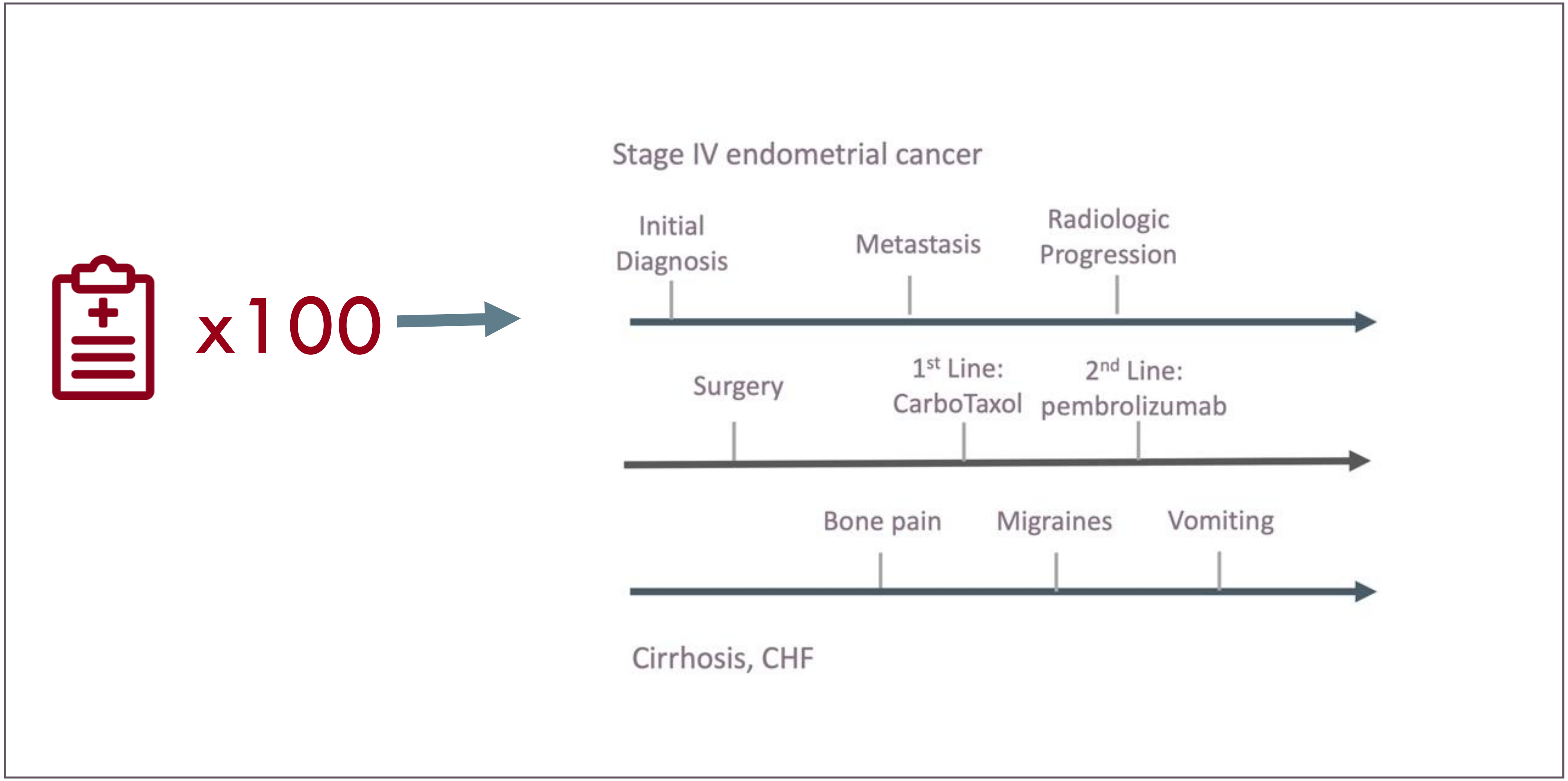
"...pt progressed after 5 mos of CarboTaxo for EC. Will dc and discuss pembro..."

"Patient progressed after 5 months of carboplatin/paclitaxel for endometrial cancer. Will discontinue for pembrolizumab"

Medication	**Carboplatin + paclitaxel**	**pembrolizumab**
Reason	**Endometrial cancer**	**Endometrial cancer**
Status	**discontinued**	**starting (implicit)**
Reason for Stop	**progression**	
Duration	**Past 5 months**	

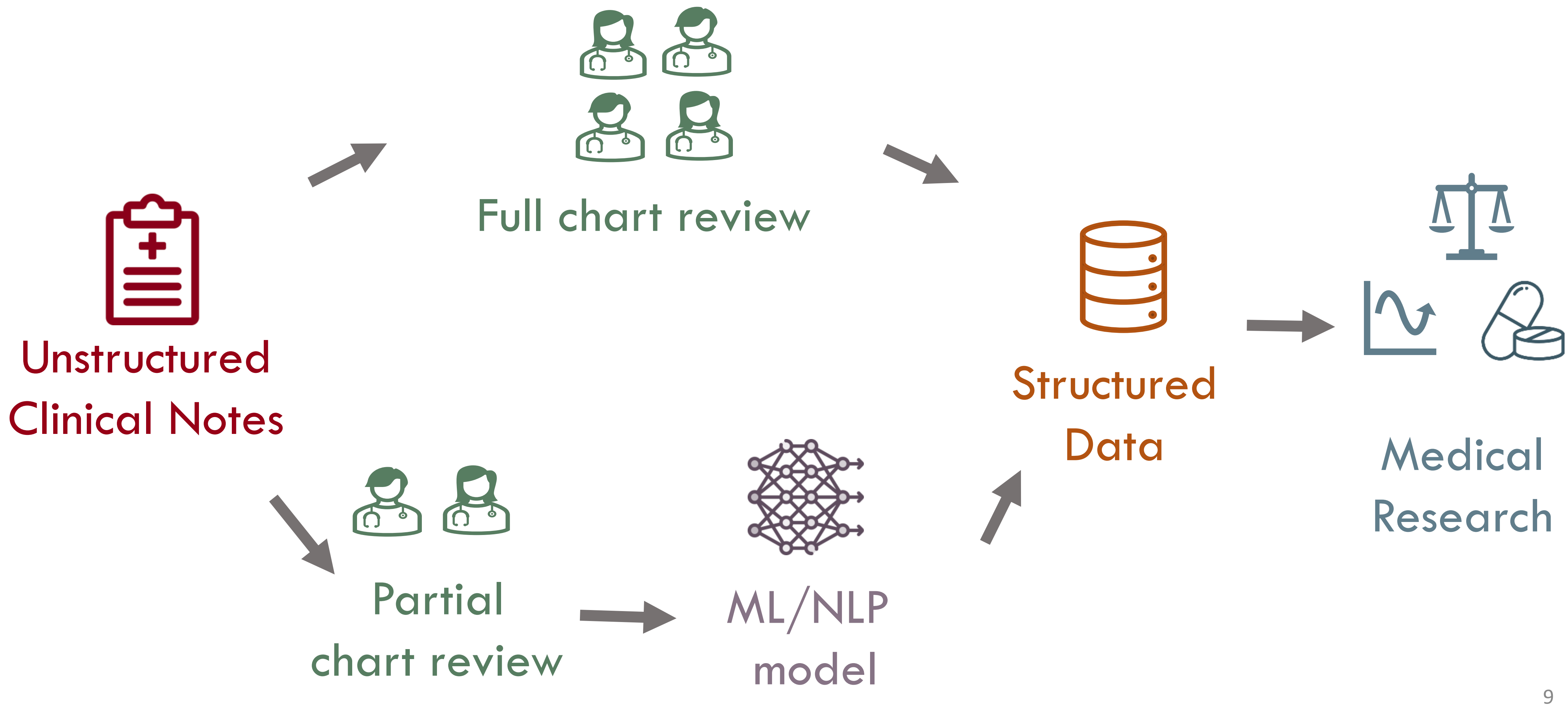


A daunting task



x1000s

Status quo for information extraction



Status quo for information extraction

Clinical Notes → Partial chart review → ML/NLP model → Structured Data → Medical Research

	Variable	# of Training Data
Agrawal, Adams, Nussbaum, Birnbaum. Machine Learning for Health (**ML4H**) NeurIPS Workshop, 2018.	Start/stop dates for oral medications	6,000+
Birnbaum, Nussbaum, Seidl-Rathkopf, Agrawal, et al. arXiv, 2020.	Binary metastasis	17,000+
Alkaitis, Agrawal, Riely, Razavi, Sontag. JCO Clinical Cancer Informatics, 2021.	Binary reason for stopping treatment	8,000+ and 1500+

Can recreate survival analyses achieved by full chart review

The partial chart review is still a huge bottleneck:

Variable + setting specific

Large amount of annotation time

Difficult to share across institutions

Other Uses

Central problem in EHRs (and in health data) is **information extraction**. How do we extract semi-structured insights from clinical data, that is:

- Customized to each use case
- Accurate
- Trustworthy, with provenance back to the original text
- Fast
- Cheaper
- …

This is useful **across healthcare:**

- Real world evidence
- Clinical trial matching

Other Uses

Information extraction is a core problem **across all of healthcare.**

Clinical trial matching
Given clinical trial criteria, how can we find patients that are eligible?

Transfers and continuity
How can we concisely summarize a patient's history for a new doctor?

Quality of care
How do we ensure that patients are receiving high-quality care across institutions?

Coding & billing
How can a hospital efficiently and accurately bill for the care delivered?

Patient understanding
How can we enable patients to understand their own medical record?

Decision support
How can we aid clinicians to administer the best possible care?

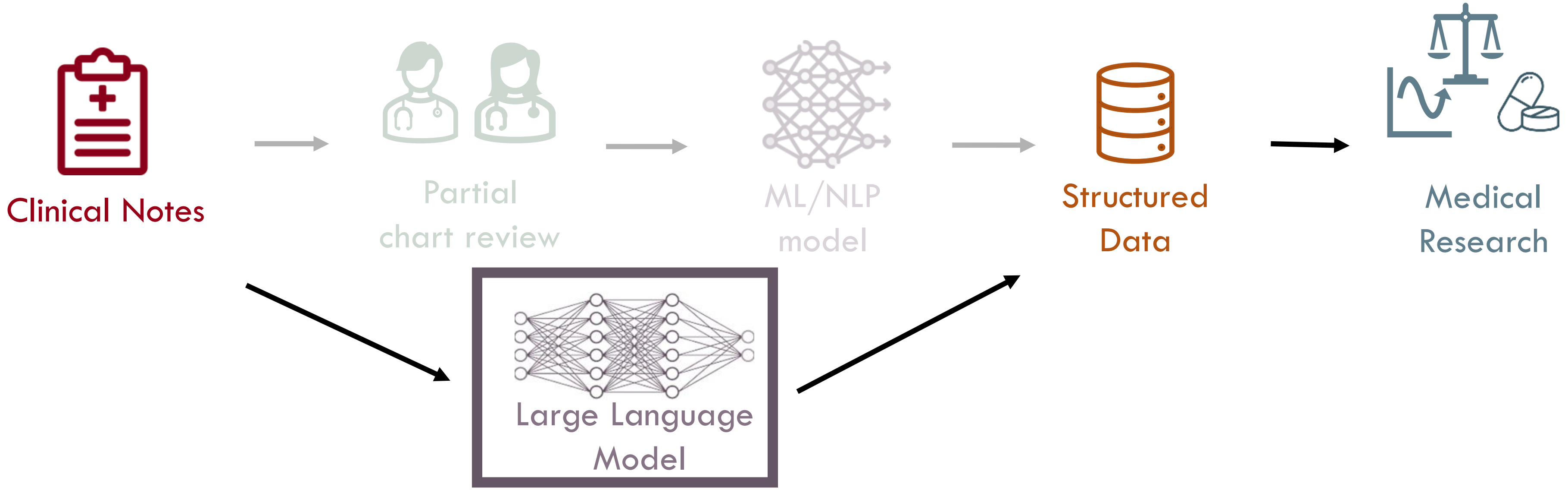
Trustworthy ML **for healthcare.**

- Accuracy is paramount – "good enough" doesn't cut it.
    - Long tail in clinical data (across subspecialties, patients, providers, presentations, …)
    - Context is key, "d/c" could mean *discharge* in an ED note but *discontinue* in a medication list.
- Provenance/justification is key – need to point back to the source to explain every decision.
- Humans need to be in the loop, but clinical expertise != ML expertise.

Outline

- **How can we leverage large language models to help in healthcare information extraction?**
- How can we incentivize cleaner clinical documentation?
- How can human-AI teams contribute?

Large Language Models for Clinical Text

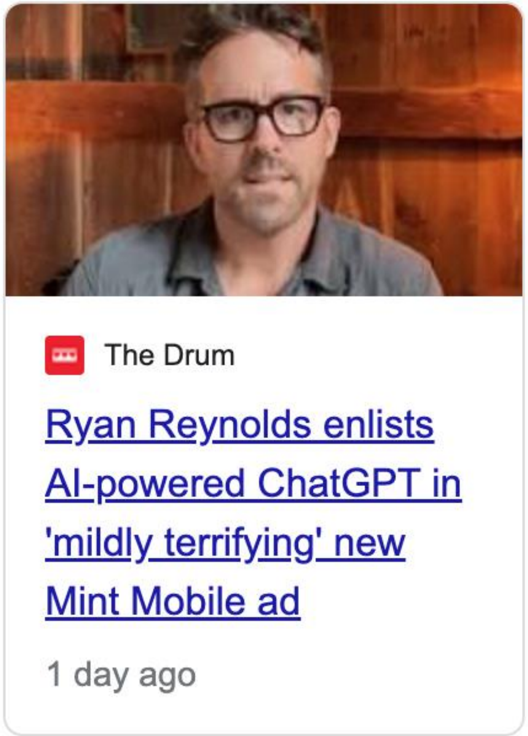


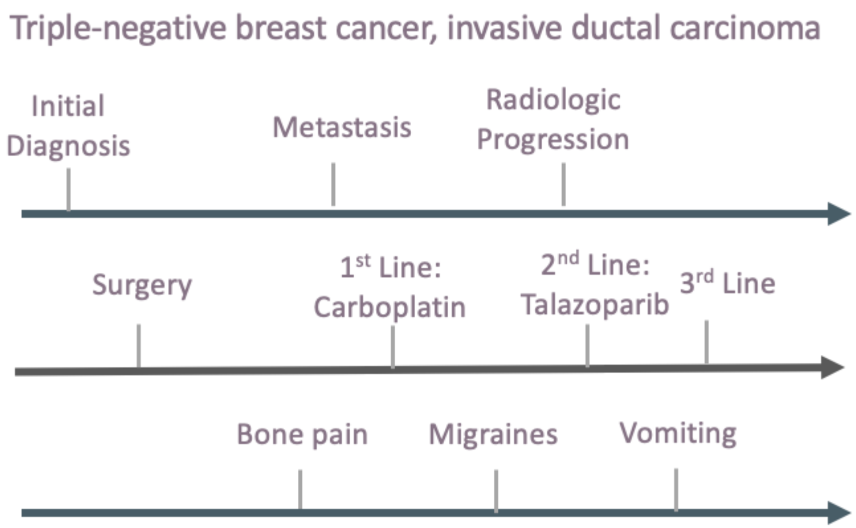
Large Language Models are Few-Shot Clinical Information Extractors

*Empirical Methods in Natural Language Processing (**EMNLP**), 2022.*

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, David Sontag

Can large language models help us structure clinical data?





Challenge #1: Clinical Text Availability

Most existing labeled data sets are under **data use agreements** and can't be sent over APIs directly, without special agreements

Benchmarking with existing publicly labels could suffer from ***label leakage***

Creation of Benchmark Datasets

We re-annotate the existing publicly available CASI dataset to release **three new** few-shot extraction **datasets**:

- Clinical coreference resolution
- Medication + status classification
- Medication + attribute relation extraction

Each contains 5 examples for development (e.g. prompt design) and 100 examples for test

**Moon et al,* " A sense inventory for clinical abbreviations…"

Challenge #2: Obtaining structured, evidence-backed output

Goal: List medications, and their reason, dosage, and frequency, as available.

Input: "[…] 500mg of metformin b.i.d. […]"

Expected completion: *"Medication: metformin Dosage: 500mg Frequency: b.i.d."*

Issue #1: Narrative format

Reality: *"The medication taken is metformin for the reason of diabetes at a dosage of 500mg…"*

Issue #2: Hallucinations

Encouraging quoted structured output

Naive approach:

Zero-shot prompt:

```
Input: 500 mg of metformin b.i.d.
Prompt: Label medications. Include dosage, reason, …
The medication taken is metformin for…
```

Complex post-processing (resolver) of LM output →

```
"Metformin": {reason: "diabetes",
              dosage: "500mg",
              frequency: "b.i.d."}
```

Encouraging quoted structured output

Our approach:

One-shot quoted example + guidance:

```
Input: He takes ibuprofen daily […].
Prompt: Label medications. Include dosage, reason, …
-medication: “statin”, frequency: “daily”
Input: 500 mg of metformin b.i.d.[…].
Prompt: Label medications. Include dosage, reason, …
-medication:“metformin”, dosage: “500mg”, “frequency”
```

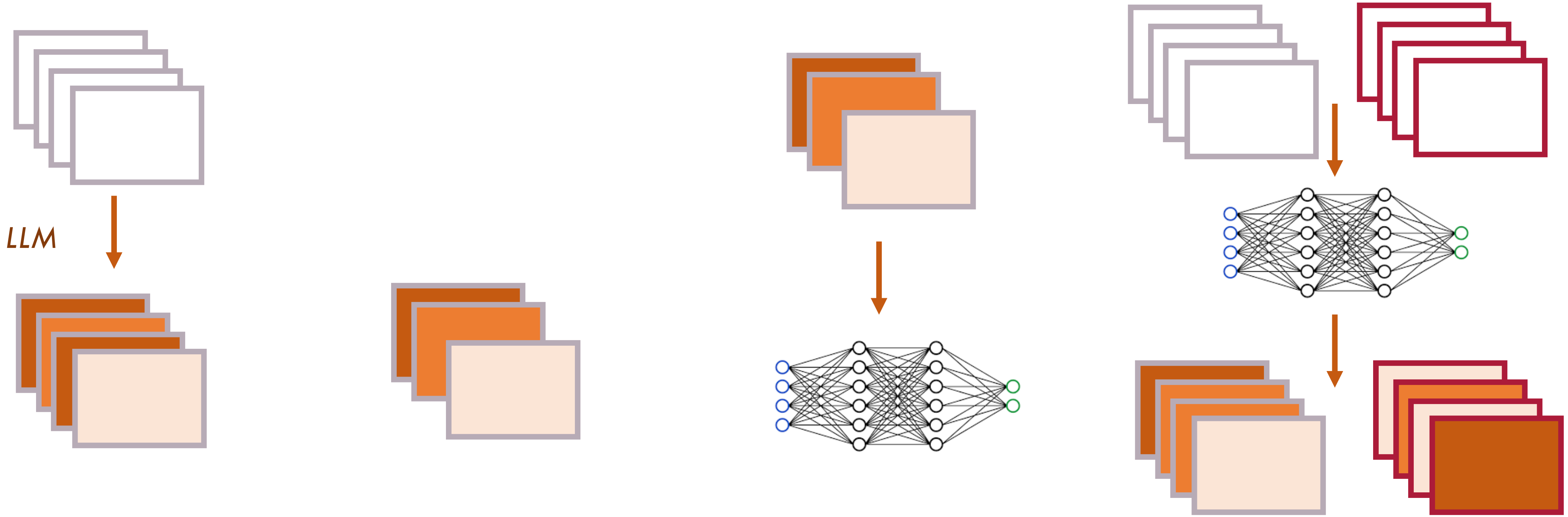
Minimal post-processing (resolver) of LM output →

```
“Metformin”: {dosage: “500mg”,
              frequency: “b.i.d.”}
```

Challenge #3: Deployability

- HIPAA compliance*
- Unwieldy size of models
- Model sensitivity to prompt wording
- Model miscalibration and overconfidence
    - When available

Treating LLM Outputs as Weak Labels



Step 1: Get LLM outputs on publicly available data

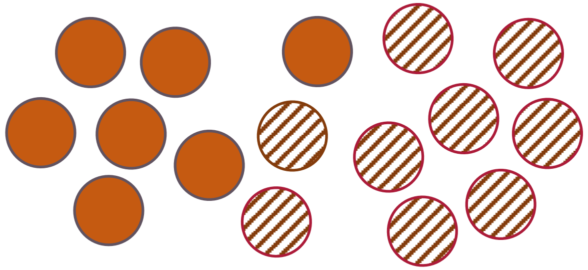
Step 2: Identify confident outputs*

Step 3: Train smaller model on confident outputs

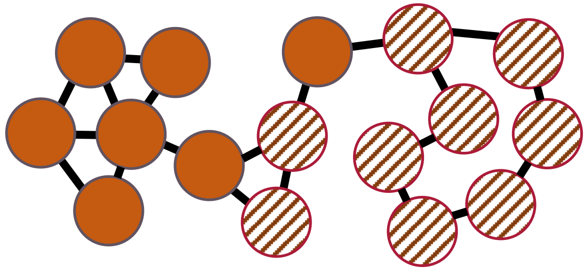
Step 4: Run smaller model on same or new data sets

Selection of confident outputs

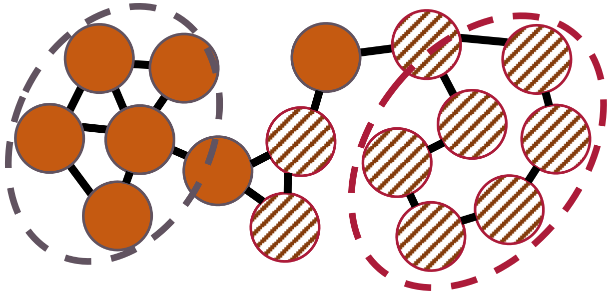
Deep models are often wildly overconfident and miscalibrated –
how can we determine when to trust their outputs?



1. Embed examples x with $\phi(x)$



2. Make K-Nearest Neighbors graph in ϕ



3. Select examples from the most homogeneous regions

Lang and Agrawal et al., ICML 2022, Lang et al, NeurIPS 2022

Selection of confident outputs

We use the cut statistic to define ``most homogeneous regions''

Test statistic for node u:

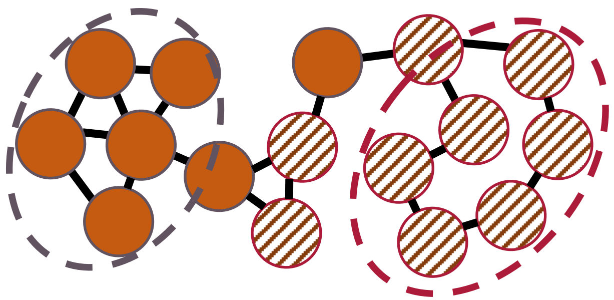
$$\frac{J_u - \mu_u}{\sigma}$$

$\sum_{v \in N(u)} w_{uv} I_{uv}$

(Weighted) sum of alike neighbors

$(1 - \hat{P}_{\hat{y}_u}) \sum_{v \in N(u)} w_{uv}$

Expected (weighted) sum of alike neighbors, if normal



3. Select examples from the most homogeneous regions

Lang and Agrawal et al., ICML 2022, Lang et al, NeurIPS 2022

Results: Clinical Acronym Disambiguation

Input: Clinical Text Snippet + Overloaded Acronym
Output: Multiple-choice Expansion of Acronym

Algorithm	CASI Acc.	CASI Macro F1
Random	0.31	0.23
Most Common	0.79	0.28
BERT (from Adams et al. (2020))	0.42	0.23
ELMo (from Adams et al. (2020))	0.55	0.38
LMC (from Adams et al. (2020))	0.71	0.51
GPT-3 edit + R: 0-shot	0.86	0.69
GPT-3 edit + R + weak sup	**0.90**	**0.76**

Zero-shot LM baseline trained on MIMIC data

OpenAI Engines: *`text-davinci-edit-001`*

Results: Clinical Acronym Disambiguation

Input: Clinical Text Snippet + Overloaded Acronym
Output: Multiple-choice Expansion of Acronym

Algorithm	**CASI Acc.**	**CASI Macro F1**	**MIMIC Accuracy**	**MIMIC Macro F1**
Random	0.31	0.23	0.32	0.28
Most Common	0.79	0.28	0.51	0.23
BERT (from Adams et al. (2020))	0.42	0.23	0.40	0.33
ELMo (from Adams et al. (2020))	0.55	0.38	0.58	0.53
LMC (from Adams et al. (2020))	0.71	0.51	0.74	**0.69**
GPT-3 edit + R: 0-shot	0.86	0.69	*	*
GPT-3 edit + R + weak sup	**0.90**	**0.76**	**0.78**	**0.69**

Example: Medication Information Parsing

Input: Clinical text snippet
Output: Medications, dosage, route, frequency, reason, duration

Baseline supervised on different clinical dataset

Subtask	Algorithm	Medication	Dosage	Route	Frequency	Reason	Duration
Token-level	PubMedBERT + CRF (Sup.)	0.82	0.92	0.77	0.76	0.35	**0.57**
	GPT-3 + R: 1-shot	**0.85**	0.92	**0.87**	**0.91**	**0.38**	0.52

OpenAI Engines: `text-davinci-edit-001`

Bonus: what might these models be learning from?

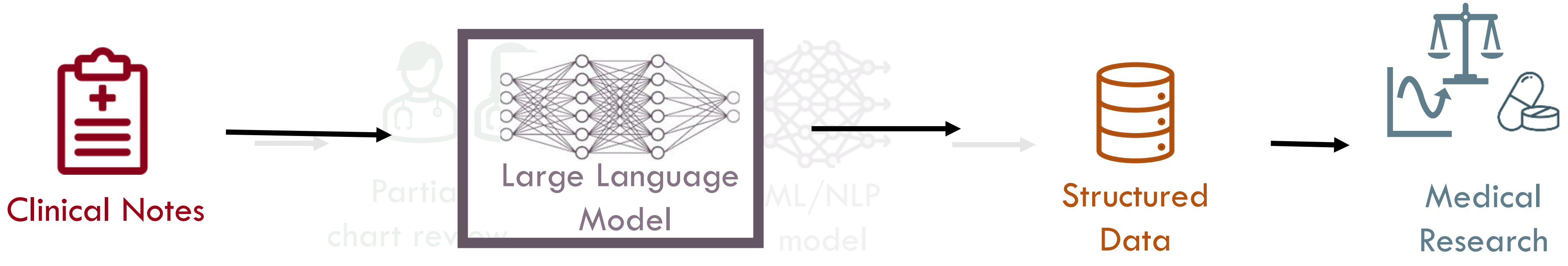
We classified sources of colloquial clinical jargon ("fx", "fracture") in a subset of Common Crawl data

Source	Median %
Research Articles	16%
Patient Health Resources	15%
Commercial Health	14%
Clinician Forums	13%
Patient Blogs + Forums	6%

43% of mentions for qhs + bedtime

41% of mentions for carbo + carboplatin

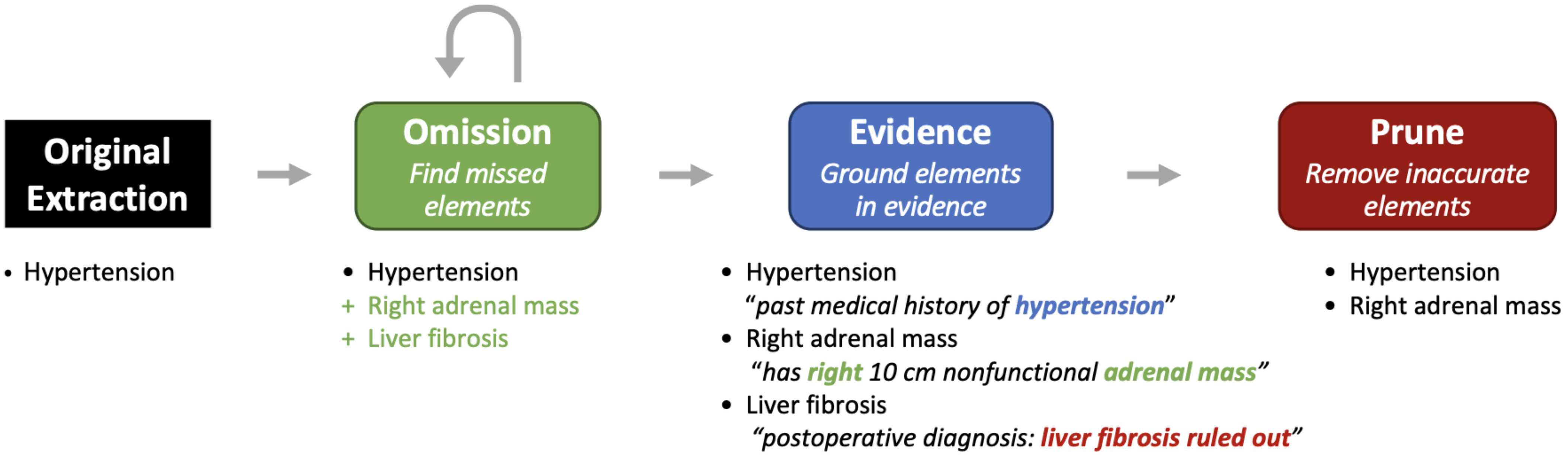
Takeaways: Large Language Models



The reasoning capabilities of and medical knowledge within LLMs could transform clinical information extraction

We developed further techniques to boost model performance, as naïve application of these models is insufficient

Follow-up: Increasing Reliability



Self-verification Improves Few-Shot Clinical Information Extraction

Zelalem Gero et al, IMLH 2023.

Case Study: LLMs for clinical trial matching

Core problem: how do we match *patients* to *trials?*

NIH U.S. National Library of Medicine

ClinicalTrials.gov

Find Studies ▾ About Studies ▾ Submit Studies ▾ Resources ▾ About Site ▾ PRS Login

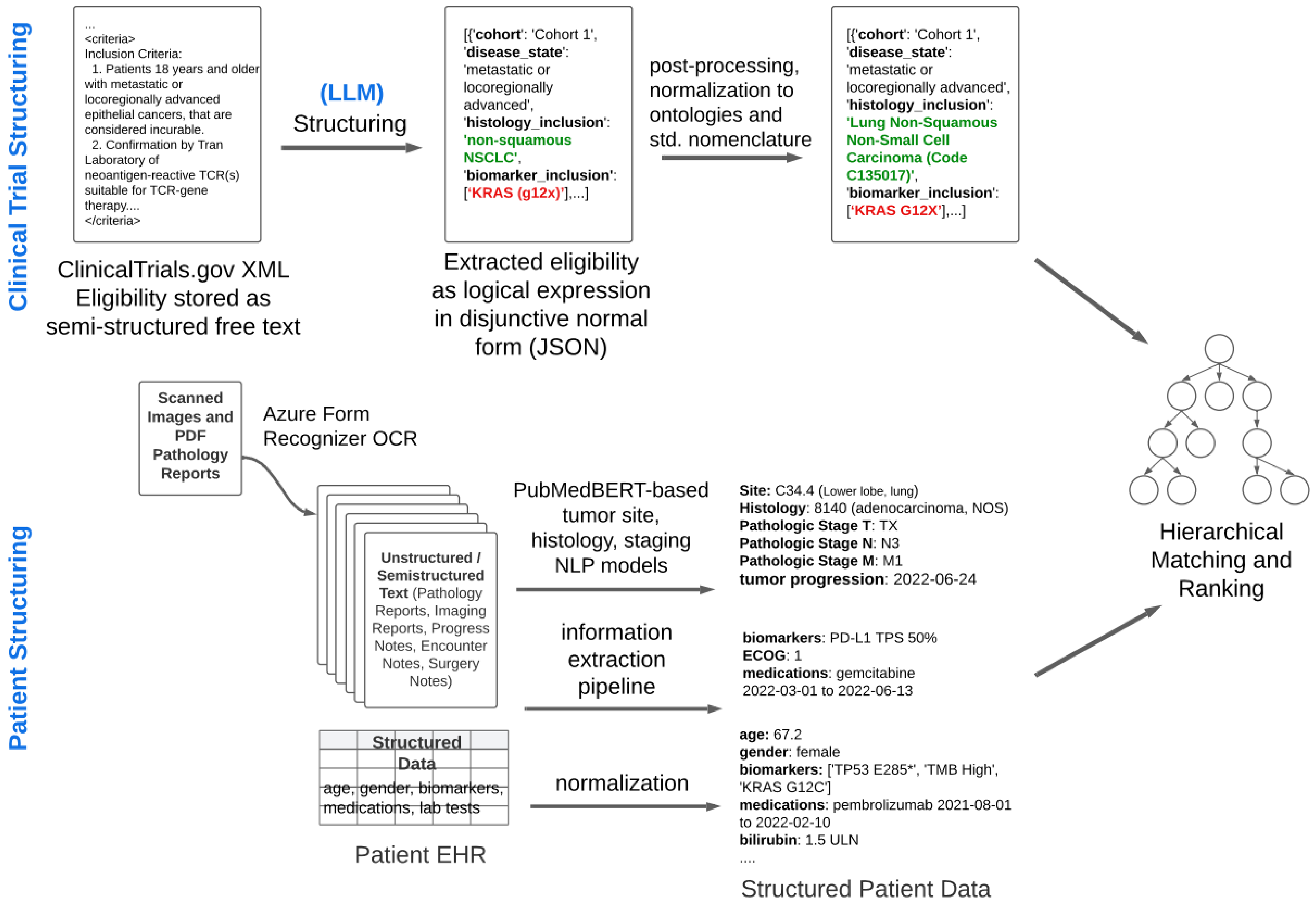
Eligibility Criteria Go to ▾

Inclusion Criteria:

-Histologically or cytologically confirmed high-grade neuroendocrine tumor that has progressed on first line therapy, excluding small cell lung cancer (SCLC). High grade includes any neuroendocrine neoplasm with a Ki-67 of >=20% or with mitotic count of more than 20 mitoses per high power field or any poorly differentiated neoplasm or any neoplasm lacking these that is deemed high grade by pathology consensus, based on other markers (necrosis or IHC demonstrating p53 or RB mutation).

Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology

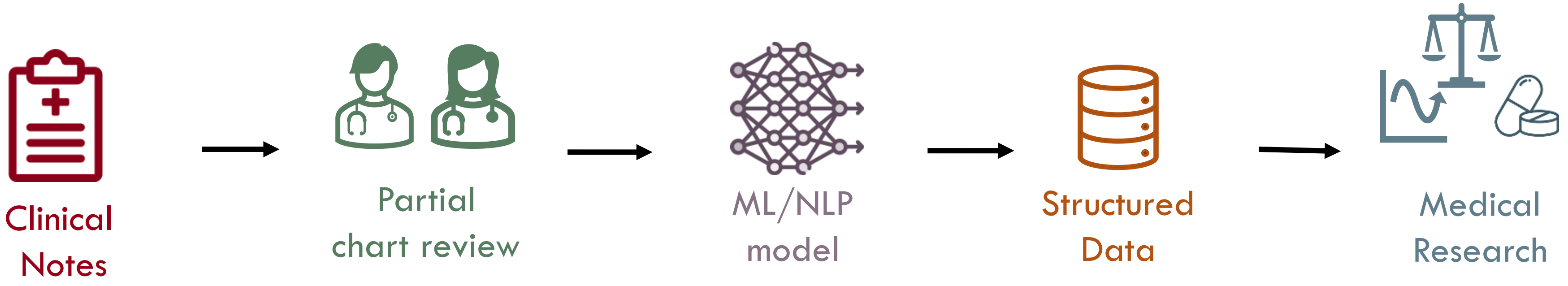
Cliff Wong et al, MLHC 2023.


Clinical Trial Structuring
...
<criteria>
Inclusion Criteria:
1. Patients 18 years and older with metastatic or locoregionally advanced epithelial cancers, that are considered incurable.
2. Confirmation by Tran Laboratory of neoantigen-reactive TCR(s) suitable for TCR-gene therapy....
</criteria>
ClinicalTrials.gov XML Eligibility stored as semi-structured free text
(LLM) Structuring
[{'cohort': 'Cohort 1', 'disease_state': 'metastatic or locoregionally advanced', 'histology_inclusion': 'non-squamous NSCLC', 'biomarker_inclusion': ['KRAS (g12x)'],...]
Extracted eligibility as logical expression in disjunctive normal form (JSON)
post-processing, normalization to ontologies and std. nomenclature
[{'cohort': 'Cohort 1', 'disease_state': 'metastatic or locoregionally advanced', 'histology_inclusion': 'Lung Non-Squamous Non-Small Cell Carcinoma (Code C135017)', 'biomarker_inclusion': ['KRAS G12X'],...]
Hierarchical Matching and Ranking
Patient Structuring
Scanned Images and PDF Pathology Reports
Azure Form Recognizer OCR
Unstructured / Semistructured Text (Pathology Reports, Imaging Reports, Progress Notes, Encounter Notes, Surgery Notes)
Structured Data
age, gender, biomarkers, medications, lab tests
Patient EHR
PubMedBERT-based tumor site, histology, staging NLP models
information extraction pipeline
normalization
Site: C34.4 (Lower lobe, lung)
Histology: 8140 (adenocarcinoma, NOS)
Pathologic Stage T: TX
Pathologic Stage N: N3
Pathologic Stage M: M1
tumor progression: 2022-06-24
biomarkers: PD-L1 TPS 50%
ECOG: 1
medications: gemcitabine 2022-03-01 to 2022-06-13
age: 67.2
gender: female
biomarkers: ['TP53 E285*', 'TMB High', 'KRAS G12C']
medications: pembrolizumab 2021-08-01 to 2022-02-10
bilirubin: 1.5 ULN
....
Structured Patient Data

Outline

- How can we leverage large language models?
- **How can we incentivize cleaner clinical documentation?**
- How can human-AI teams contribute?

Re-imagining clinical documentation



Fast, Structured Clinical Documentation via Contextual Autocomplete

*Machine Learning for Healthcare (**MLHC**), 2020*

Divya Gopinath, Monica Agrawal, Luke Murray, Steven Horng, David Karger, David Sontag

MedKnowts: Unified Documentation and Information Retrieval for EHRs

*User Interface and Software Technology (**UIST**), 2021*

Luke Murray, Divya Gopinath, Monica Agrawal, Steven Horng, David Sontag, David Karger

Conceptualizing ML for Dynamic Information Retrieval of EHR notes

*Machine Learning for Healthcare (**MLHC**), 2023*

Sharon Jiang, Shannon Shen, Monica Agrawal, Barbara Lam, Nicholas Kurtzman, Steven Horng, David Karger, David Sontag,

Re-imagining clinical documentation

What if we could collect high-quality clinical data *at the point of care*, without increasing physician burnout?

Fast, Structured Clinical Documentation via Contextual Autocomplete

*Machine Learning for Healthcare (**MLHC**), 2020*

Divya Gopinath, Monica Agrawal, Luke Murray, Steven Horng, David Karger, David Sontag

MedKnowts: Unified Documentation and Information Retrieval for EHRs

*User Interface and Software Technology (**UIST**), 2021*

Luke Murray, Divya Gopinath, Monica Agrawal, Steven Horng, David Sontag, David Karger

Conceptualizing ML for Dynamic Information Retrieval of EHR notes

*Machine Learning for Healthcare (**MLHC**), 2023*

Sharon Jiang, Shannon Shen, Monica Agrawal, Barbara Lam, Nicholas Kurtzman, Steven Horng, David Karger, David Sontag,

EHRs have usability issues

WHY DOCTORS HATE THEIR COMPUTERS

Digitization promises to make medical care easier and more efficient. But are screens coming between doctors and patients?

By Atul Gawande

November 5, 2018

Issue #1: Time for Data Entry

Issue #2: Time for Information Retrieval

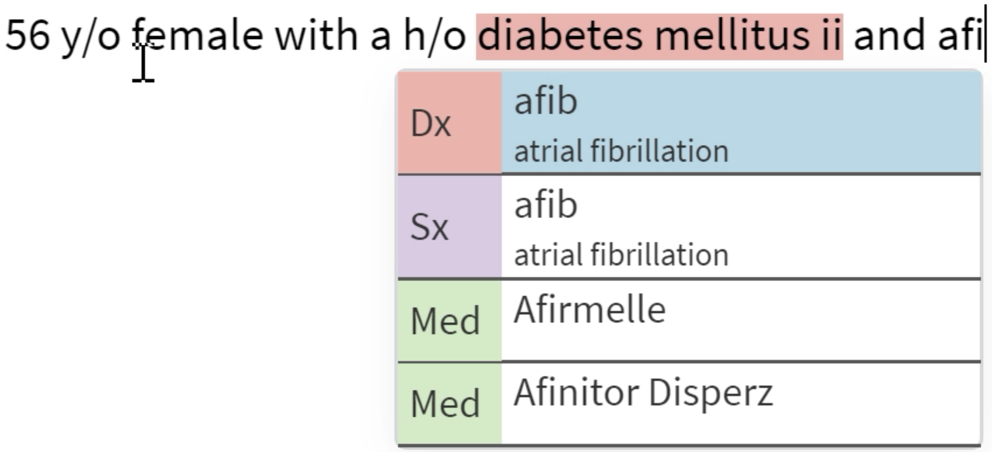
Challenge of Data Entry

Linguistic Characteristics of Medical Notes

Many of the entries on the medical records are in the form of notes which are neither complete sentences nor single word entries, but linguistic strings of an intermediate type, which we will hereafter call fragments. Fragments are a compressed type of linguistic material resulting from various transformations which have the effect of making linguistic strings shorter by reducing or deleting material. The writer of these stretches of material must make his entries brief, in order to save time and effort, but also make them informative and unambiguous. For this reason the deleted material has to be easily recover-

Anderson et al , *Grammatical Compression in Notes and Records, ACL 1975*

Solution: Streamlining Data Entry

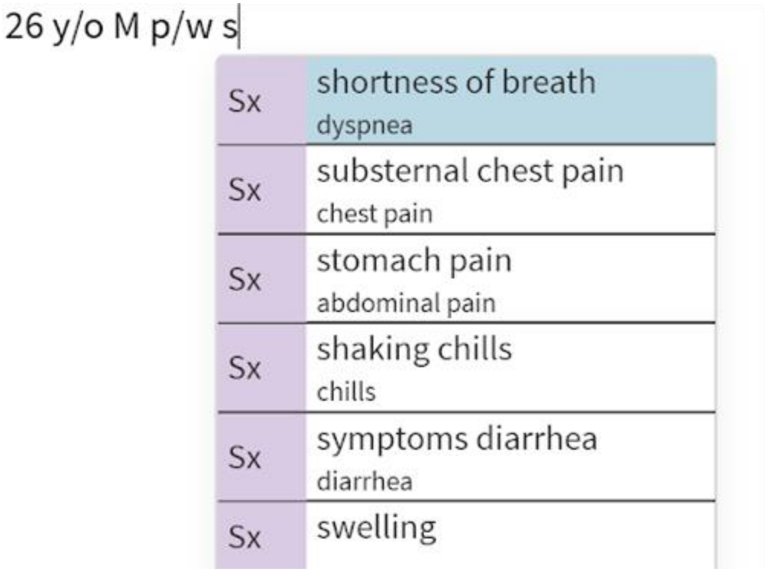


Contextual autocomplete

- Personalized to each patient
- Automatically normalizes concepts to ontologies as the note is being written
- Decreases documentation burden with fewer keystrokes

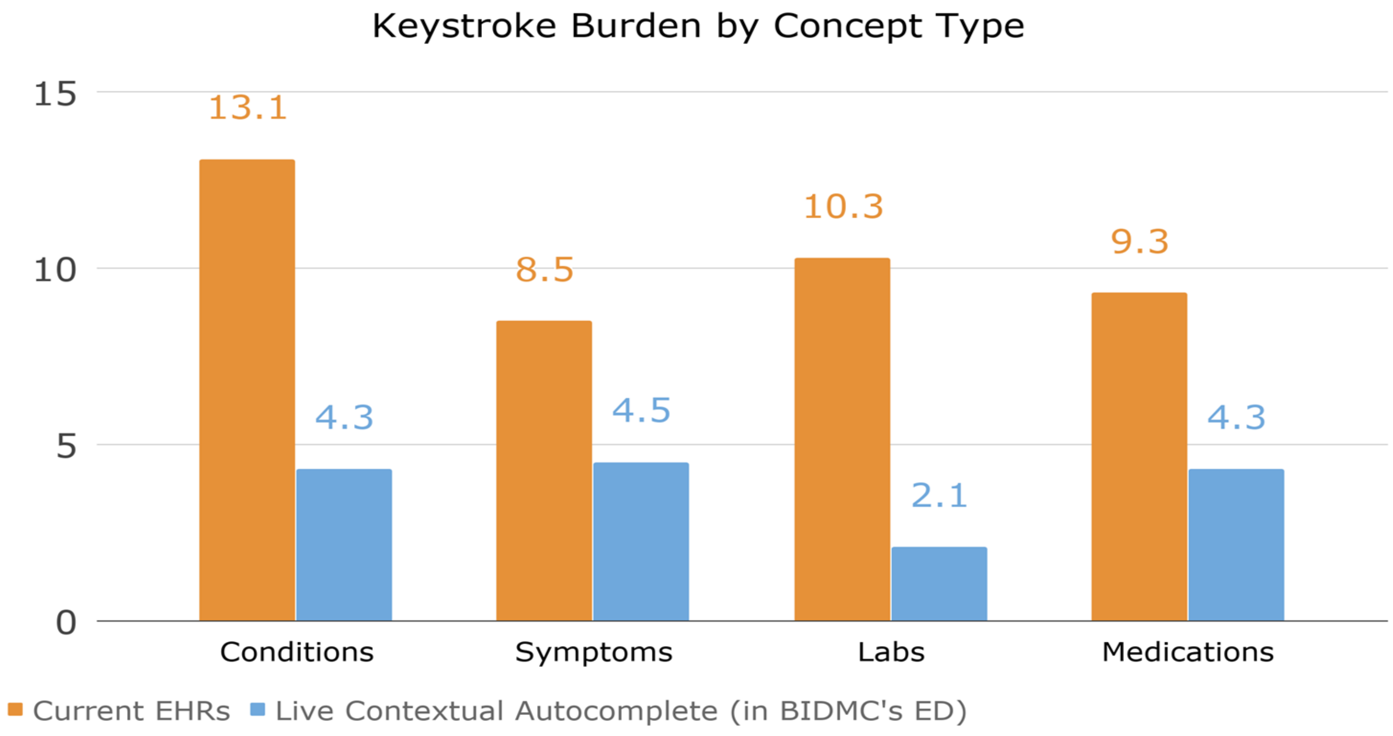
Sources of supervision

Use available information from a given patient to predict concepts that will be documented in a clinical note.



- (0) Prior notes (EHR)
- (1) Triage assessment
- (2) Chief complaint
- (3) Nurse's Notes
- (4) Doctor's Notes (our focus)

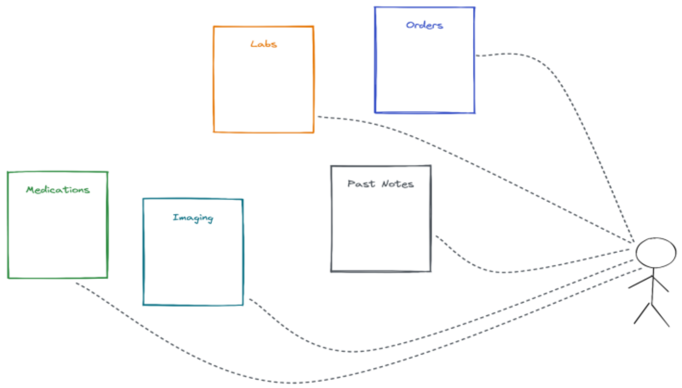
We dramatically reduced the **keystroke burden** of data entry in a **live setting.**



Challenge of Information Retrieval

Doctors have to manually synthesize past data into data driven narratives

- Past Labs
- Past Medications
- Relevant Notes
- Relevant Imaging

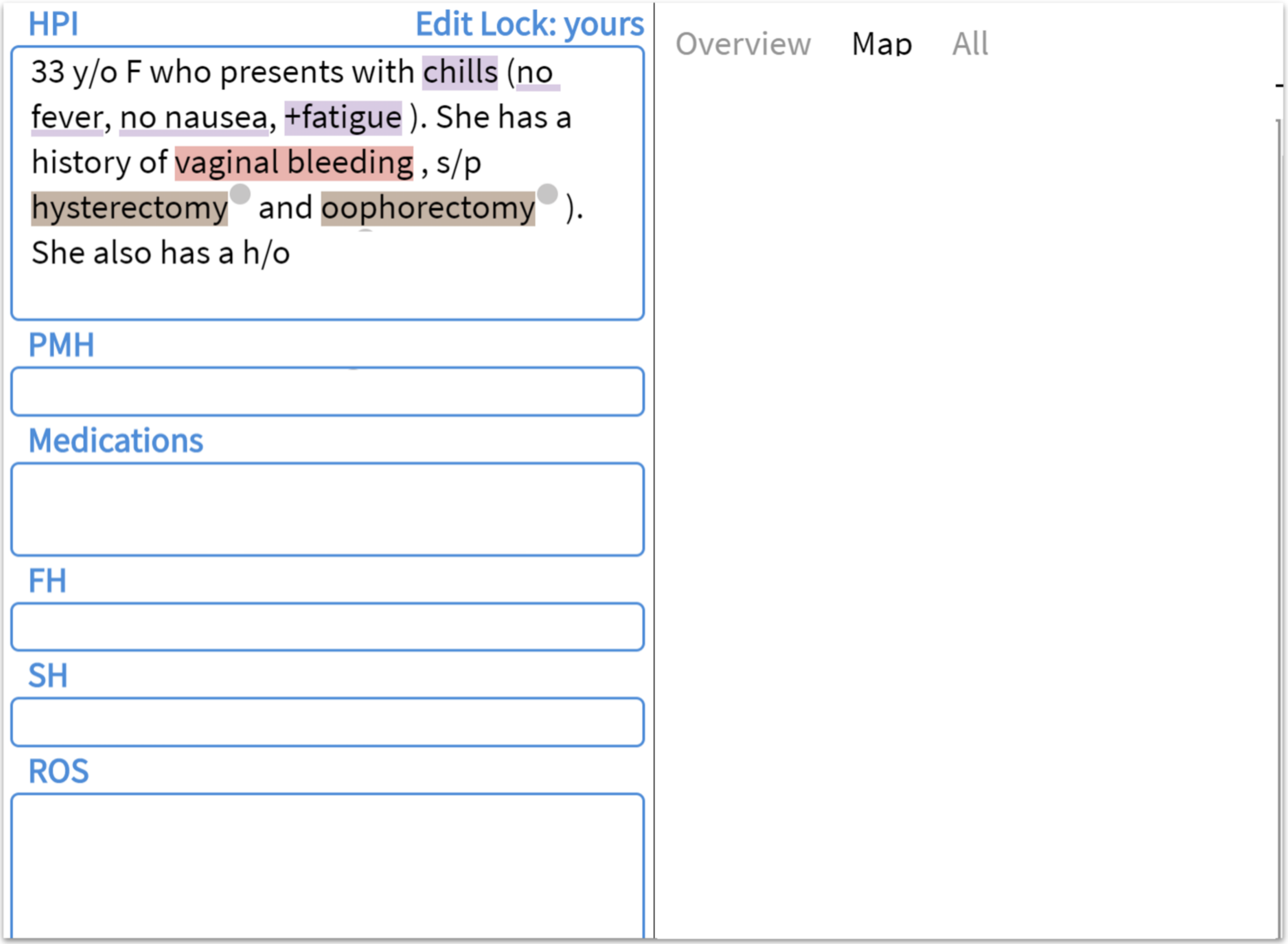


Death By 1,000 Clicks: Where Electronic Health Records Went Wrong

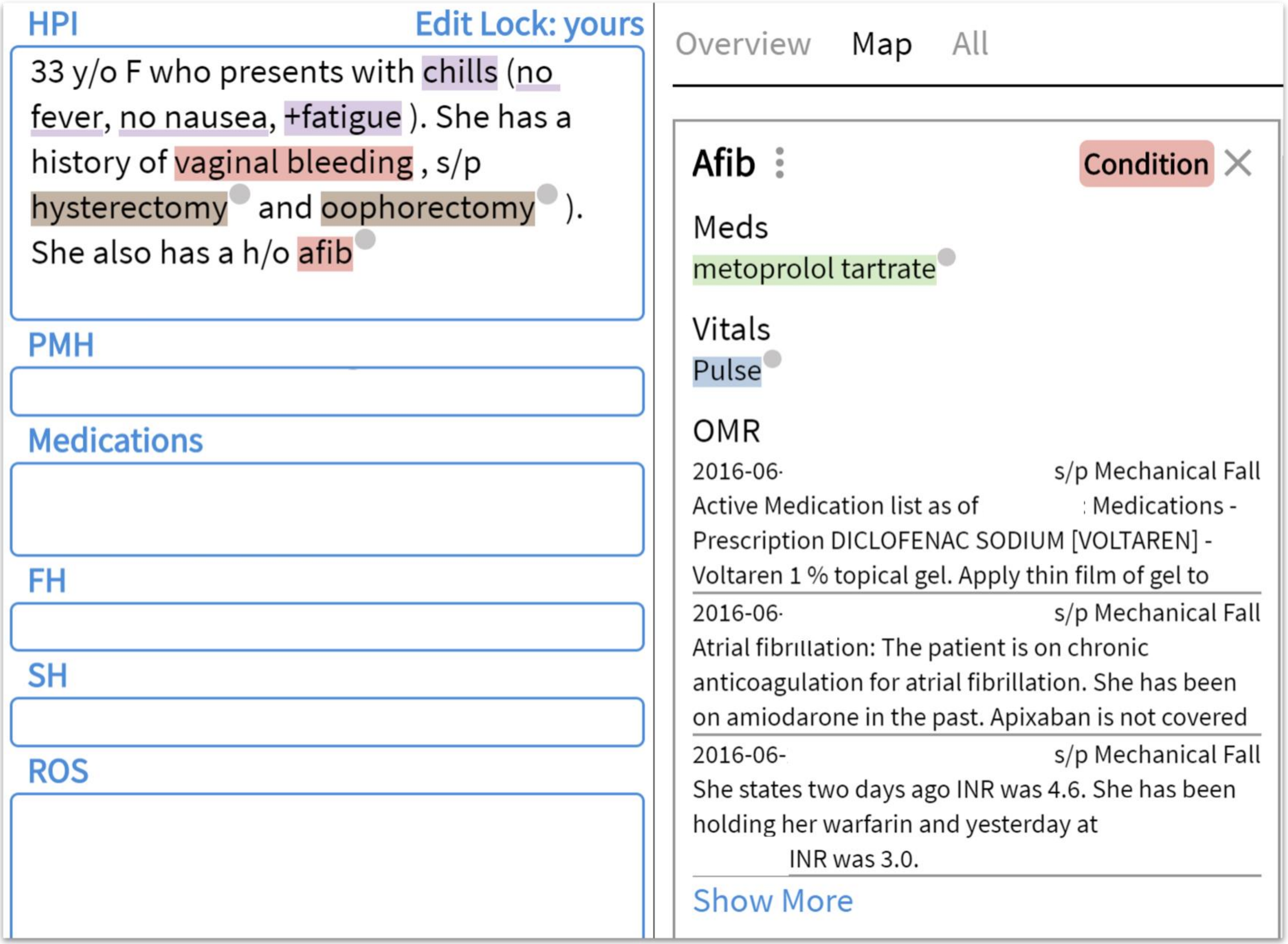
The U.S. government claimed that turning American medical charts into electronic records would make health care better, safer and cheaper. Ten years and $36 billion later, the system is an unholy mess. Inside a digital revolution that took a bad turn.

By Fred Schulte and Erika Fry, Fortune • MARCH 18, 2019

Solution: Streamlining Information Retrieval



Solution: Streamlining Information Retrieval



Filling in Redundant Information

HPI

Edit Lock: yours

93 y/o F p/w nonproductive cough , fever , nausea , but no chills . She has a history of an oophorectomy and type 2 diabetes . She has mild hypertension and is on Coumadin to treat this.

PMH

Medications

FH

SH

ROS

Filling in Redundant Information

HPI

Edit Lock: yours

93 y/o F p/w nonproductive cough , fever , nausea , but no chills . She has a history of an oophorectomy and type 2 diabetes . She has mild hypertension and is on Coumadin to treat this.

PMH

oophorectomy, type 2 diabetes, hypertension

Medications

Coumadin

FH

SH

ROS

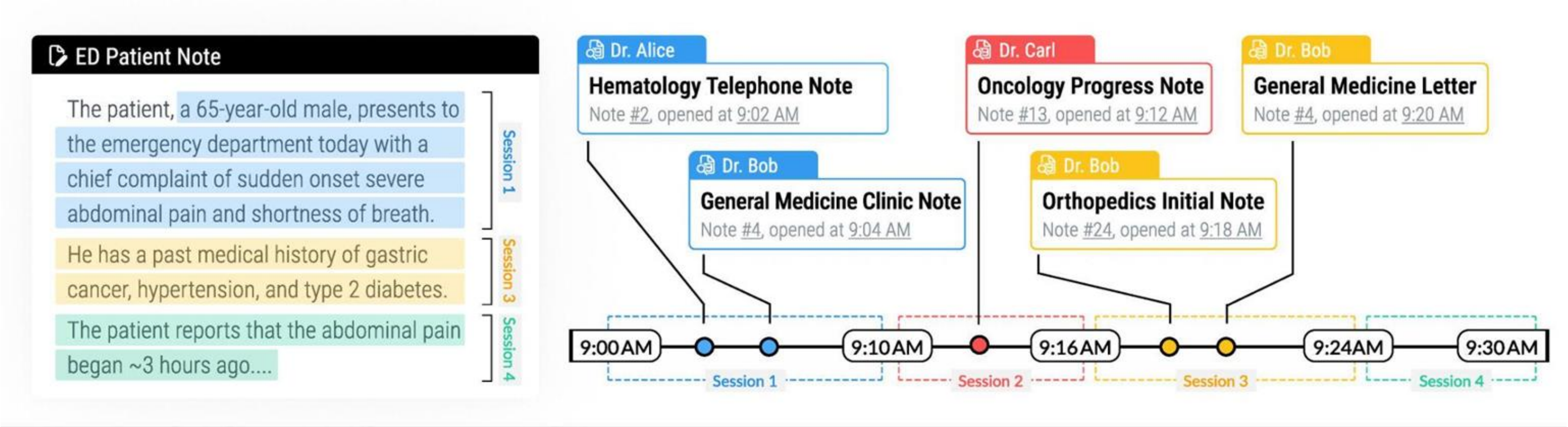
Constitutional:, fever, nausea, no chills
Head / Eyes: No diplopia
ENT: no earache
Resp:, nonproductive cough
Cards: No chest pain
Abd: No abdominal pain

Deployment + Evaluation

- We designed MedKnowts in a year-long iterative prototyping process with a clinician and their scribes across 1185 patients.
- We evaluated MedKnowts in a month-long deployment with four scribes across 234 patients.
- In a user questionnaire:
  - Would use frequently – median 5/5
  - Quick learning curve – median 5/5
  - Easy to use – median 4.5/5

Newer direction: leveraging EHR audit logs

We can use EHR audit logs to characterize the note writing process



We can also use the signal from those audit logs to learn how to auto-surface notes (AUC of 0.963).

With the advent of LLMs, what changes?

Bootstrapping/zero-shot performance at information extraction is **significantly better than before**, but still some critical gaps:

LLMs still struggle with the long tail:

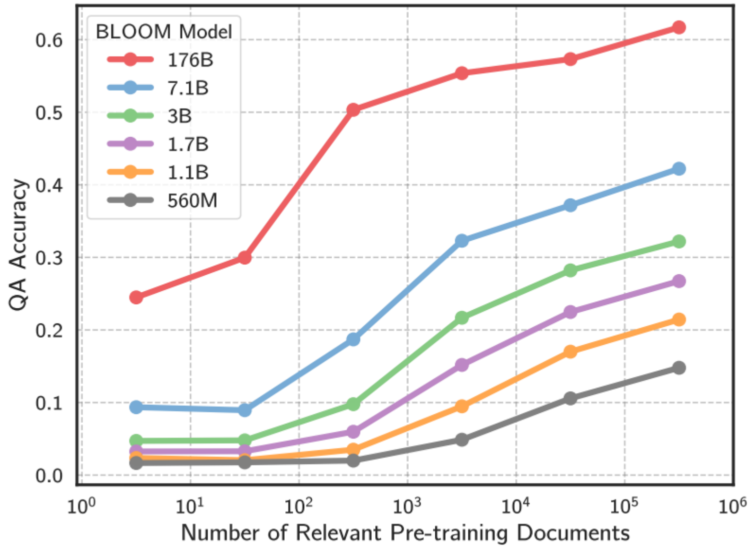


Figure 1. Language models struggle to capture the long-tail of information on the web. Above, we plot accuracy for the BLOOM model family on TriviaQA as a function of how many documents in the model's pre-training data are relevant to each question.

LLMs can be "distracted" by irrelevant information in ways that traditional methods may not be:

Original Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now?
Modified Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, the age of Claire's father is 3 times of Jessica's age.* How old is Jessica now?
Standard Answer 24

Takeaways: Re-imagining documentation

Via a redesign of the EHR, it is possible to simultaneously:

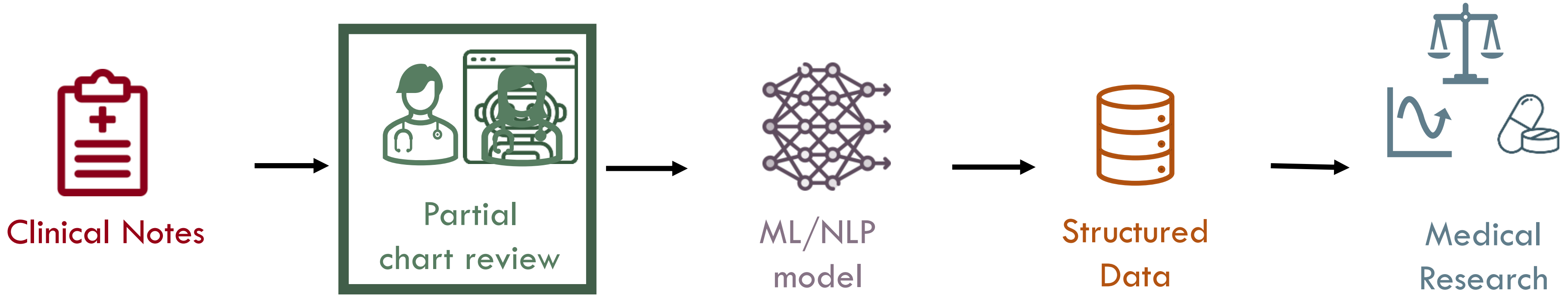
- Obtain cleaner data as a natural byproduct
- Reduce physician workload

These features can be integrated into live workflows via careful *opt-in* design

Outline

- How can we leverage large language models?
- How can we incentivize cleaner clinical documentation?
- **How can human-AI teams contribute?**

Human-AI Teams for Clinical Annotation



Assessing the Impact of Automated Suggestions on Decision Making

*Conference on Human Factors in Computing Systems (**CHI**), 2021.*

Ariel Levy*, Monica Agrawal*, Arvind Satyanarayan, David Sontag

Clinical concept recognition

“Pt given carbo ia for her TNBC. Will dc.”

Clinical concept recognition

Patient? Prothrombin time?

Carbodome?

Intra-arterial?

D/C current?

discontinue?

" **Pt** given **carbo** **ia** for her **TNBC**. Will **dc**. "

Physical therapist?

Carboplatin?

Intra-articular?

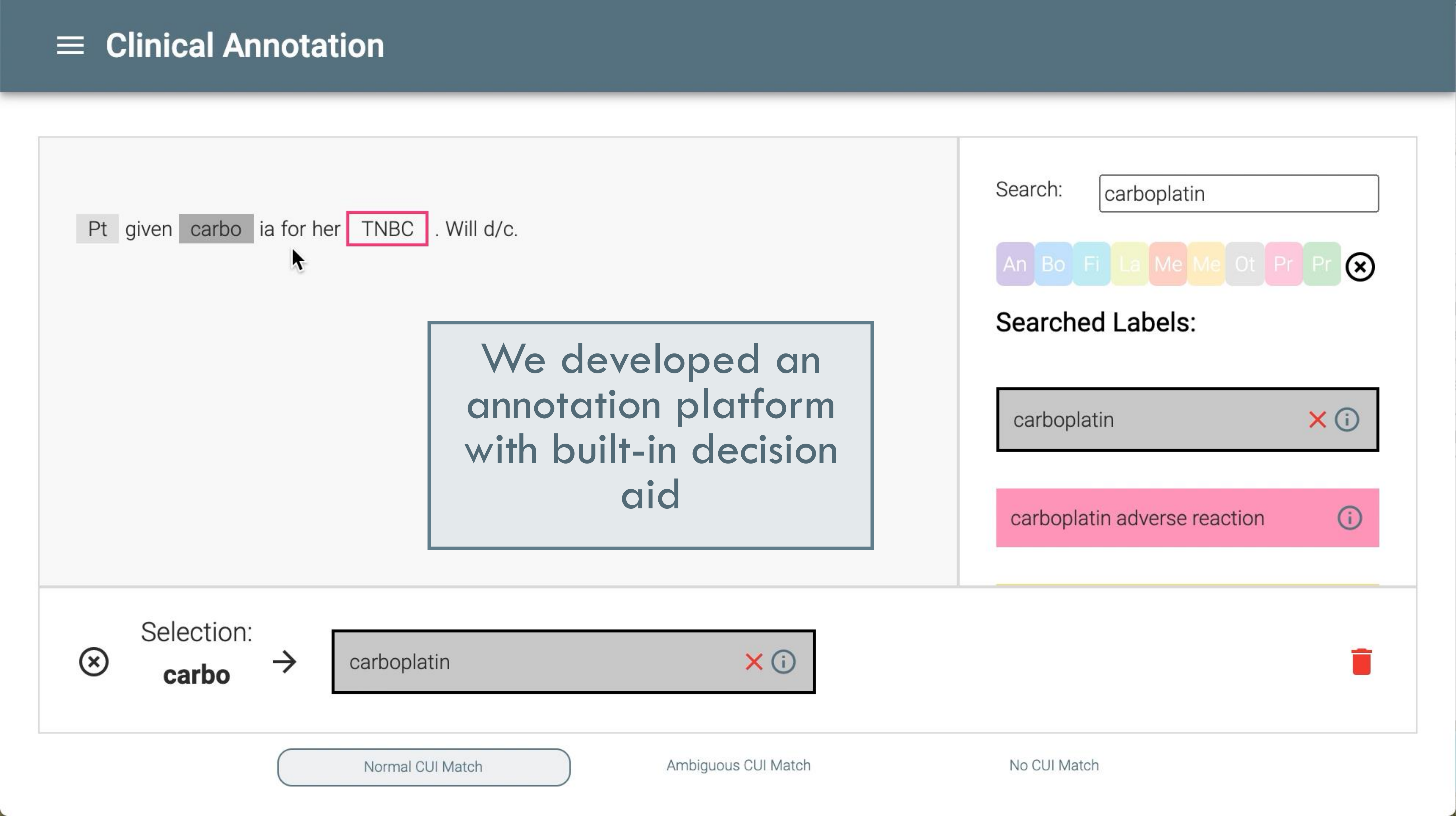
discharge?

Doctor of Chiropractic?

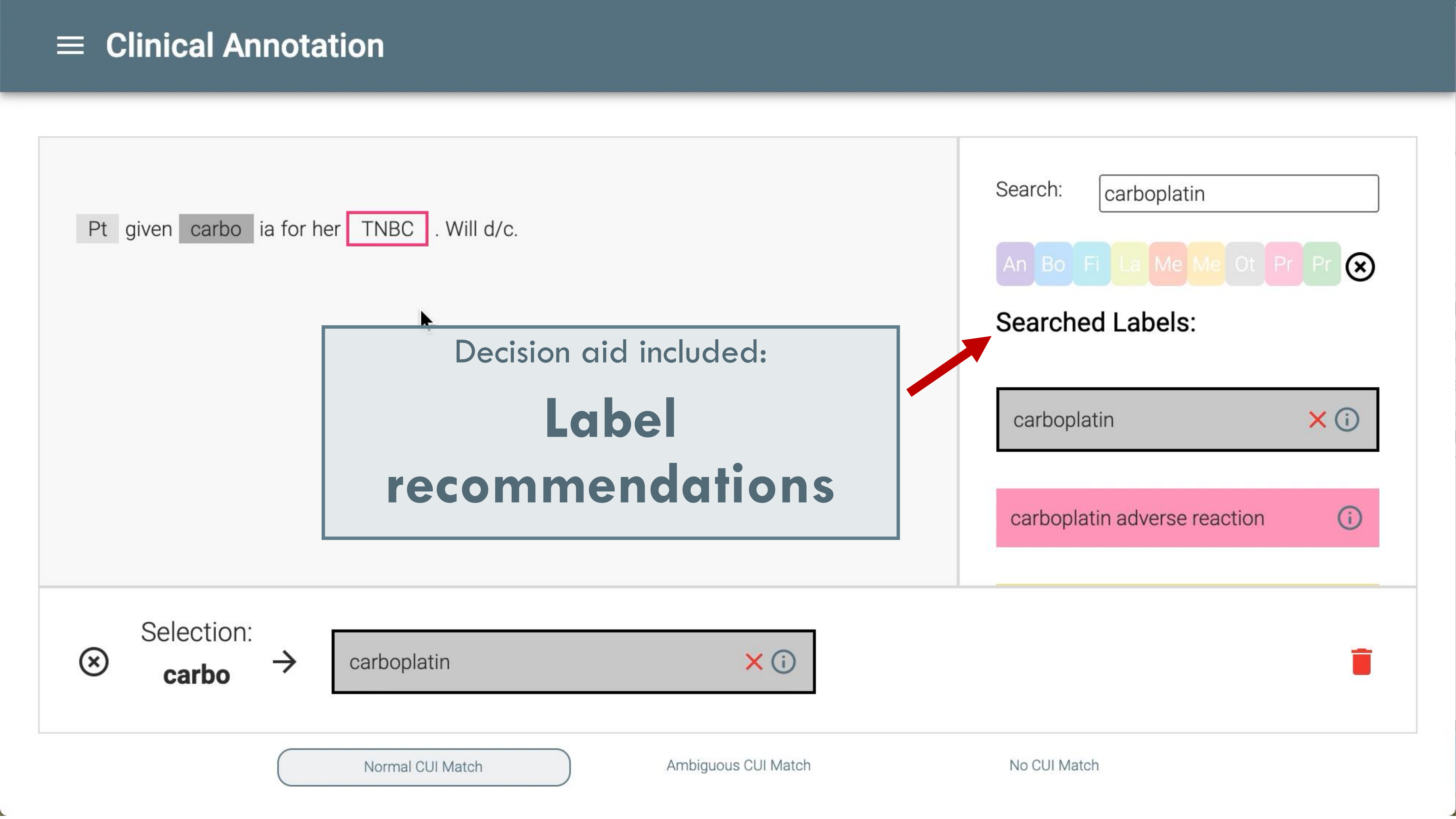
Clinical concept recognition

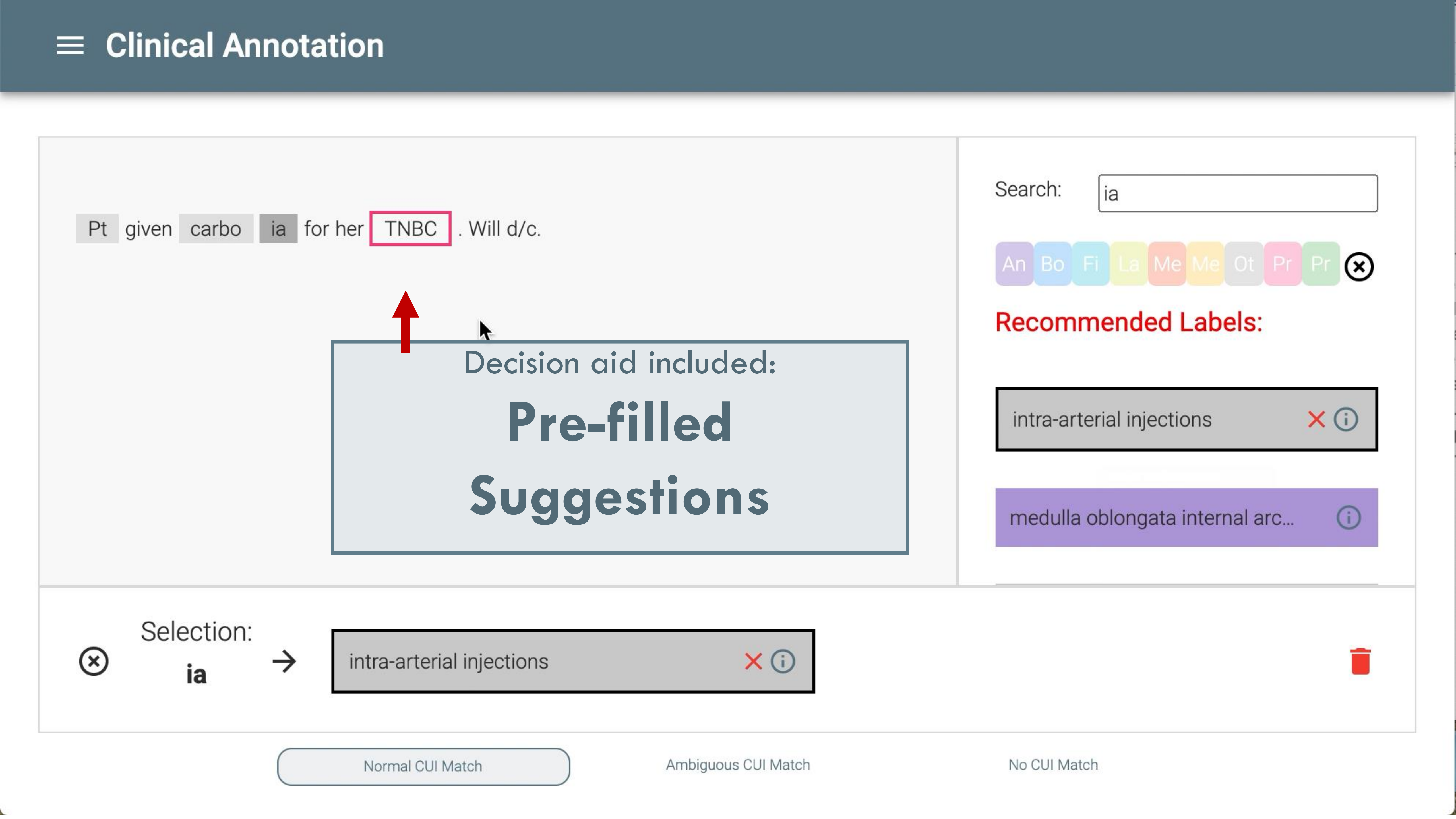
Difficulties include the many labels and the large label space (over 400,000 concepts)

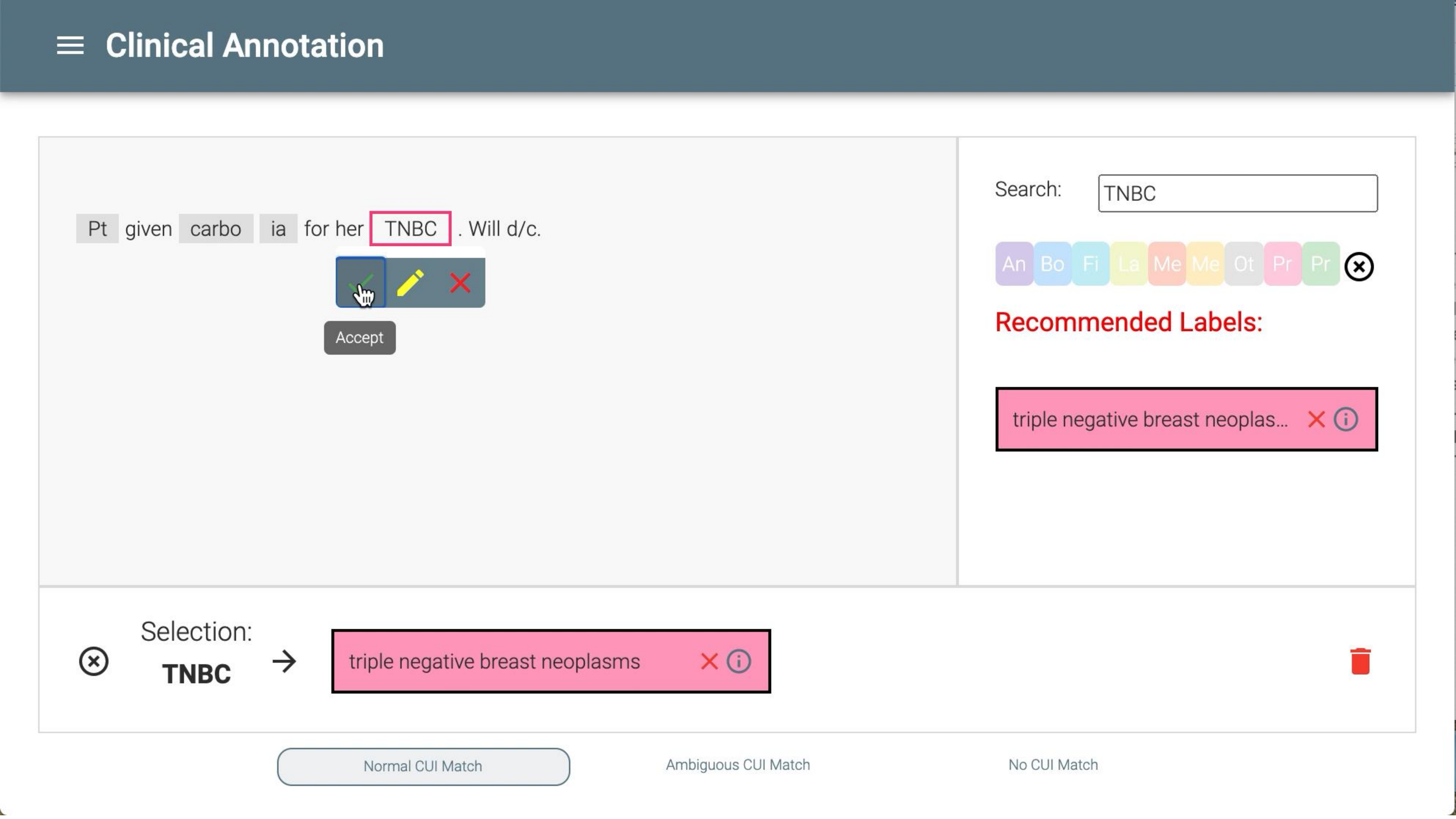
Patient (C0030705)		Carboplatin (C0079083)	Intra-arterial (C1561451)			Triple-neg. breast cancer (C3539878)		Discontinue (C1706472)
" Pt	given	carbo	ia	for	her	TNBC.	Will	dc. "



Clinical Annotation
Pt given carbo ia for her TNBC . Will d/c.
Search:
carboplatin
An Bo Fi La Me Me Ot Pr Pr
Searched Labels:
carboplatin
carboplatin adverse reaction
We developed an annotation platform with built-in decision aid
Selection:
carbo
carboplatin
Normal CUI Match
Ambiguous CUI Match
No CUI Match


Clinical Annotation
Pt given carbo ia for her TNBC . Will d/c.
Search:
carboplatin
An Bo Fi La Me Me Ot Pr Pr
Searched Labels:
carboplatin
carboplatin adverse reaction
Decision aid included:
Label recommendations
Selection:
carbo
carboplatin
Normal CUI Match
Ambiguous CUI Match
No CUI Match


Clinical Annotation
Pt given carbo ia for her TNBC . Will d/c.
Decision aid included:
Pre-filled Suggestions
Search:
ia
An Bo Fi La Me Me Ot Pr Pr
Recommended Labels:
intra-arterial injections
medulla oblongata internal arc...
Selection:
ia
intra-arterial injections
Normal CUI Match
Ambiguous CUI Match
No CUI Match


Clinical Annotation
Pt given carbo ia for her TNBC . Will d/c.
Accept
Search:
TNBC
An Bo Fi La Me Me Ot Pr Pr
Recommended Labels:
triple negative breast neoplas...
Selection:
TNBC
triple negative breast neoplasms
Normal CUI Match
Ambiguous CUI Match
No CUI Match

Example Impact

One Clinician Is All You Need–Cardiac Magnetic Resonance Imaging Measurement Extraction: Deep Learning Algorithm Development

Pulkit Singh[1]; Julian Haimovich[2, 3, 4]; Christopher Reeder[1]; Shaan Khurshid[2, 3, 5]; Emily S Lau[3, 4]; Jonathan W Cunningham[4, 6]; Anthony Philippakis[1, 7]; Christopher D Anderson[8, 9, 10]; Jennifer E Ho[4, 11]; Steven A Lubitz[2, 3, 4, 5]; Puneet Batra[1]

BROAD INSTITUTE

JMIR Publications
Advancing Digital Health & Open Science

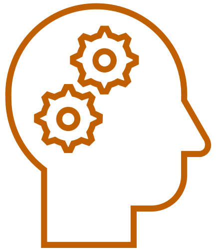
Goal: Extraction of 21 Measurements from Cardiac MRI Reports

Macro F1 score: 0.957

Clinician labeling time: ~11 hours for all training data

Due to the ever-growing presence of automated decision aid,
we build on past work to ask:

How does domain expertise mediate the influence of decision aid?



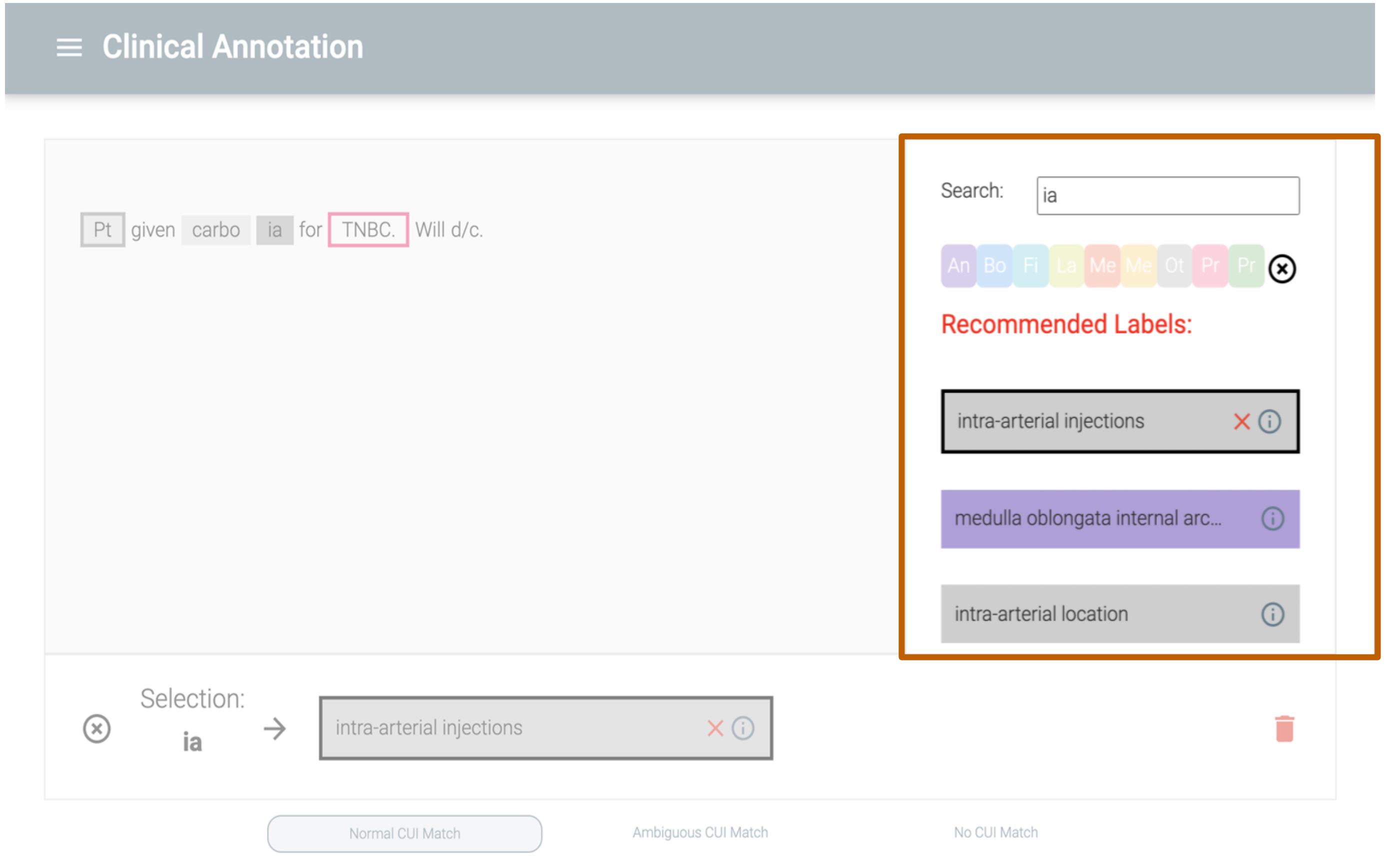
- In a task with a complex decision space
- Using objective measures of trust and agency
- Over an extended period of use to factor in fatigue

Study Overview

- 18 clinicians from 9 institutions
- Study Novelties
  - Joint study of agency (what to label?) and trust (how to label?) using objective measures
  - Large space of 400k+ labels
  - ~8 hours of annotation per user
- Two stages
  - Stage 1: Label Recommendations
  - Stage 2: Pre-filled Suggestions

Stage 1: Label Recommendations

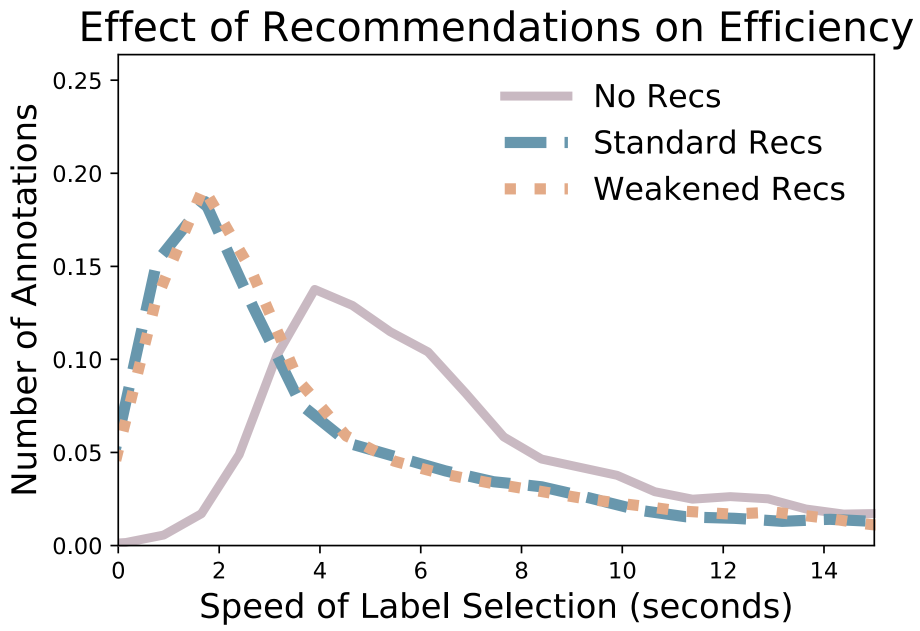
We analyze accuracy, speed, and search behavior, particularly where recommendations are inaccurate



How does user behavior shift?

Users with full recommendations created **more annotations** (average of 12%) than those without any ($p<0.02$)

The **median time** to choose a label **halves** with recommendations: from 6 seconds to 3 seconds ($p<0.05$)



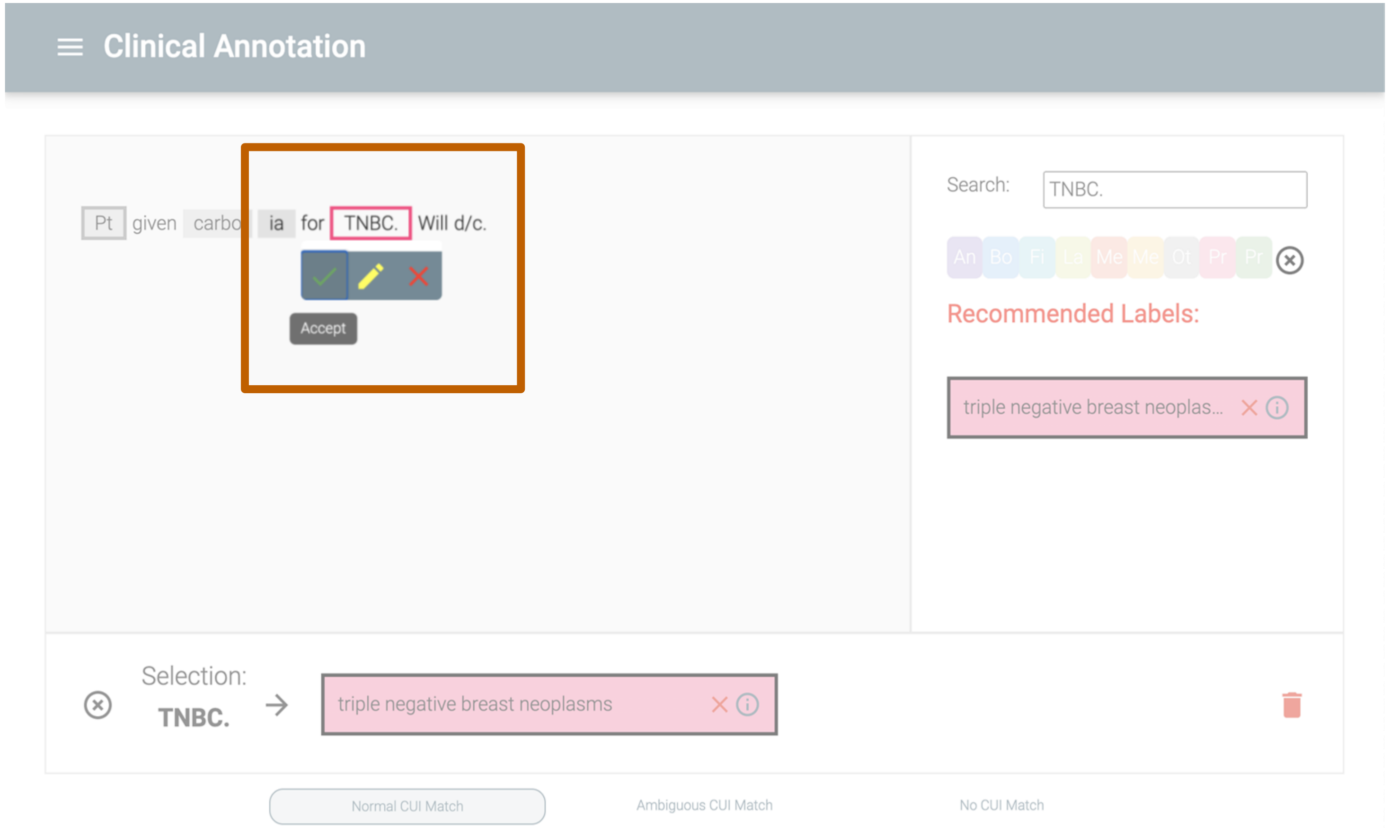
Do users search when needed, or misplace trust?

Yes, they generally search when the recommendation algorithm truly doesn't surface the correct answer.

However, they search **less often when they may *expect* the correct answer to be surfaced.**

Stage 2: Pre-filled suggestions

We analyze **accuracy** and **speed**, particularly where suggestions are inaccurate, and **additional annotation behavior**



Do users react appropriately?

Mostly. They:

- accept 99% of correct labels+spans
- accept only 17% of incorrect labels
- accept 33% of incorrect spans

Overall, they tend to have higher performance than users without pre-filled suggestions.

Do users react appropriately?

There was large user variability in accepting of incorrect labels and spans – **not correlated with their prior task performance**

Providing label confidence made no discernable difference.

What about agency for creating new annotations?

Users experience **loss of agency** in creating the new nontrivial annotations that don't come pre-filled:
they made 12% fewer than in Stage 1

No such drop was observed in users without pre-filled suggestions, making the loss significant
($p<0.01$)

What about agency for creating new annotations?

This loss of agency went **unnoticed** by users.

What about agency for creating new annotations?

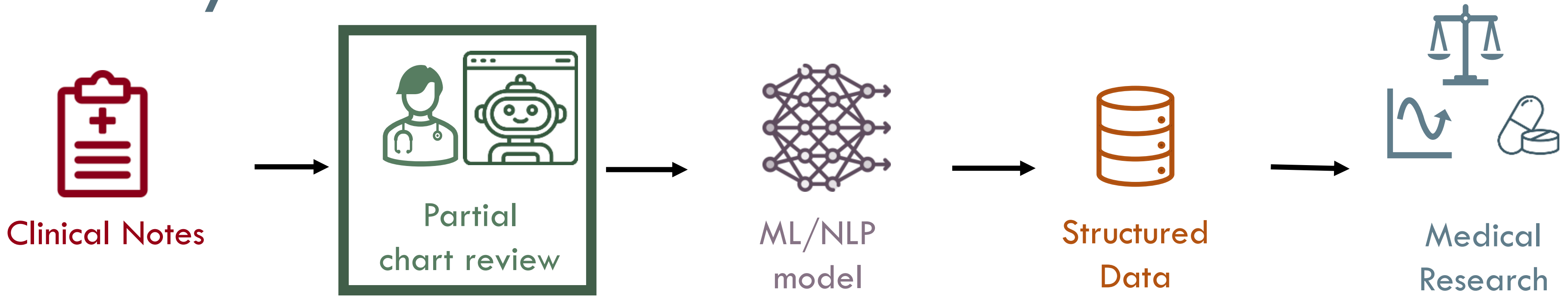
"I made sure to double check if there were parts that were not annotated."

"Made it easier to scan the remaining unmarked parts for words to annotate."

This loss of agency went **unnoticed** by users.

"[Pre-filled annotations] freed up mental bandwidth to spend more energy on unmarked text."

Takeaways: Human-AI Teams



- With appropriate mental models, users properly modulated trust and mediated model errors.
- Users lost agency without noticing, highlighting the importance of objective measures.
- Both UIs and ML systems should consider such effects in their design

Conclusion

A holy grail in ML for healthcare is **information extraction.** This would solve fundamental challenges **across healthcare.**

Core takeaways:

1. LLMs are getting us **much closer** to making ML-augmented information extraction possible, but has many challenges that need to be addressed, particularly for healthcare data (long tail, data availability, security & compliance, explainability/trust, etc.)
2. Rather than applying LLMs as a post-hoc bandaid to extract insights from clinical data, the true gamechanger is **collecting clean data at the point-of-care,** incentivized by ML-driven information retrieval.
3. ML for healthcare is a very **human problem** – we need to design human-centered systems that understand the impact of introducing ML into workflows.

CS329T: Projects & Datasets

Dataset	Description
Clinical Trial Matching	All FDA clinical trial eligibility criteria are freely available online.
Medical Information Mart for Intensive Care (MIMIC)	Vast dataset of de-identified structured & unstructured clinical data across ICU and ED.
PMC Patients	Patient summaries extracted from PubMed case reports; 167k+ patients.
Adverse Drug Event Corpus	Extracts all adverse drug events (ADEs) from a set of clinical notes.
Synthetic note generation	As in here, generate synthetic notes

Any questions?

Leverage large language models.

Incentivize cleaner clinical documentation

Quantify the impact of human-AI teams

Beyond the talk: Reach out to us at divya@layerhealth.com / monica@layerhealth.com