NLP SUMMIT HEALTHCARE

A Path Towards an Automated and Inclusive Clinical Research

Azadeh Mobasher
Principal Data Scientist
Genentech

Background

- Quality healthcare relies on many sources of information, however majority of data is unstructured and difficult to analyze

Traditional Medicine

- Millions of people are taking medications that will **not** help them [1]
- **Top ten** highest-grossing drugs in the United States help 1 in 4 of the people who take them [1]
- There are drugs that are **harmful** to certain ethnic groups because of the **bias** towards white western participants in classical clinical trials [1]
- **Doubling** time of medical knowledge is merely **73 days** today compared to **50 years** in 1950 [2]

445,812
ClinicalTrials.gov

35 million+ citations
PubMed

1.2 billion+ unique non-identified patients
IQVIA

4.5 million+ medical and healthcare concepts
UMLS

Precision Medicine

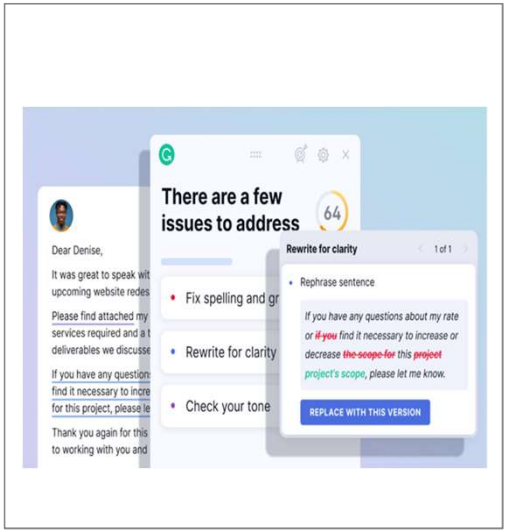
- Delivering individual patient level medicine, by considering genomics & omics, lifestyle, preferences, health history, medical records, compliance and exogenous factors [1]

NLP SUMMIT

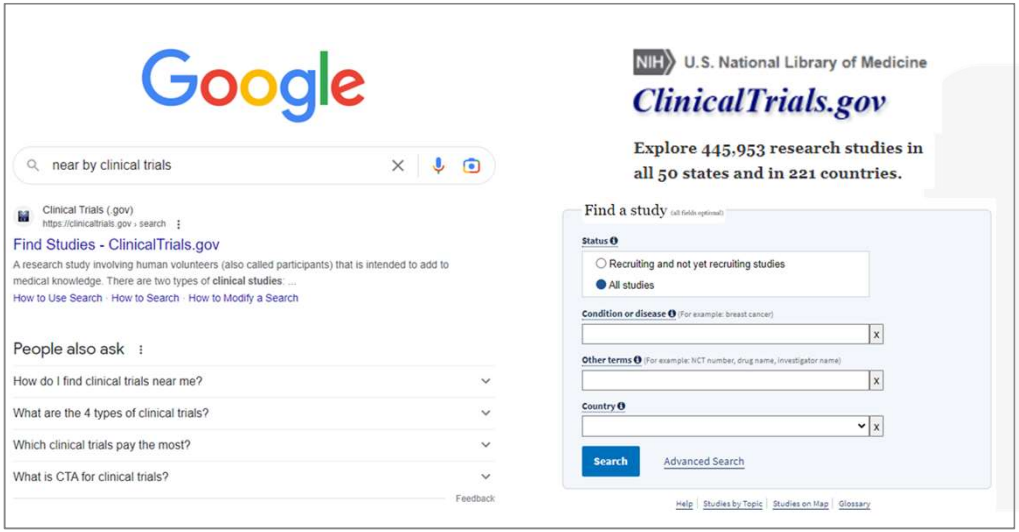
[1] Personalized medicine: Time for one-person trials
[2] How will artificial intelligence change medical training?

Natural Language Processing

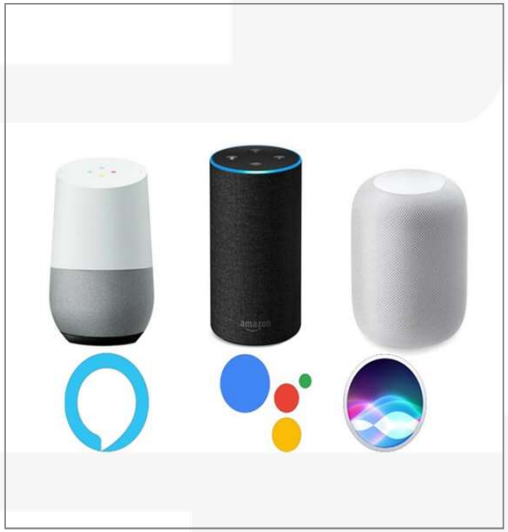
- Natural Language Processing (NLP) describes the interaction between **human language** and **computers**



Spelling and Grammar Checks



Search Engines



Digital Voice Assistants

NLP SUMMIT

Clinical trials patient matching

- **Essential**: Annual market over $46 billion
  - Clinical trials play important roles in drug development but often suffer from inaccurate, insufficient and expensive patient recruitment

- **Time consuming**
  - 50% of trials **delayed**, 25% of cancer trials **failed** due to enrollment
  - 37% of sites **fail** to meet their recruitment targets

- **High costs**
  - Average recruitment cost: $6000 to $7500 per patient

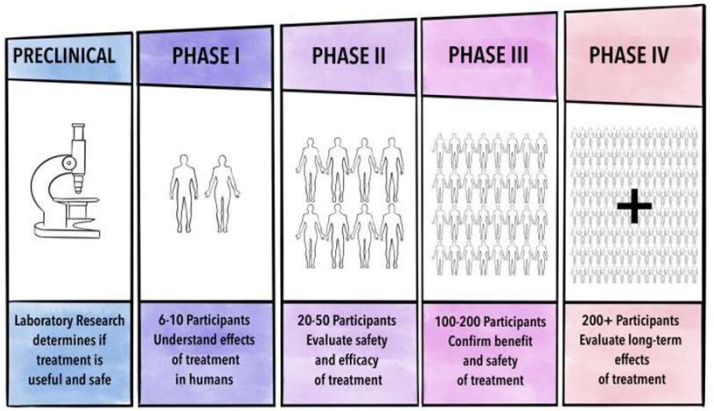
- **Underrepresentation**
  - 13.4% of the U.S. population is Black versus only 5% of trial participants
  - 18% of the US population is Latinx versus only 1% of trial participants

NLP SUMMIT

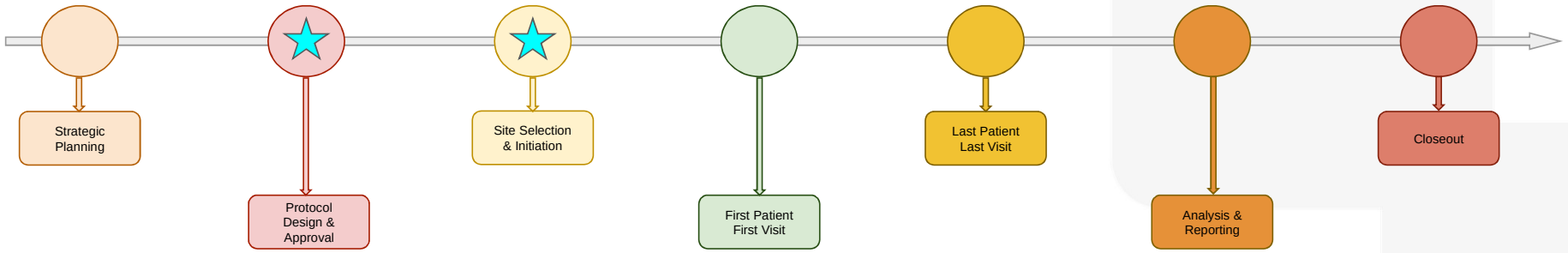
Clinical trials process

- **Data-driven practices** to identify PIs and select sites
  - Collaboration with data owners to locate principal investigators and select sites
  - Manual and rule-based with focus on centralized site selection, takes at least 2 to 4 weeks per study protocol

- **Patient-trial matching**: finding qualified patients given structured EHR and unstructured eligibility criteria



www.gene.vision



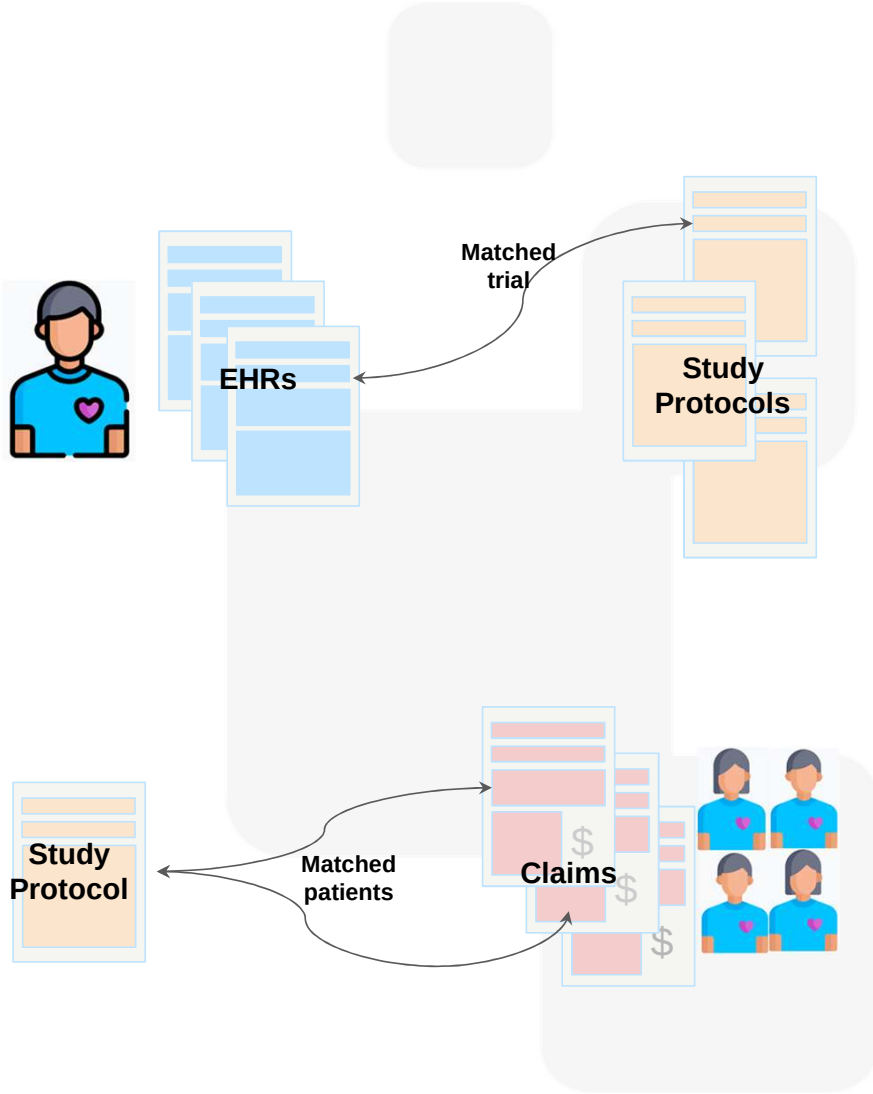
Where automated patient-trial matching can help to inform study design and empower targeted outreach programs

NLP SUMMIT

Patient-trial matching

- Finding matching trials for a patient
  - Given a patient electronic medical records, identify one or multiple trials that a patient is eligible

★ Finding eligible patients for a trial
  - Given a clinical trial study protocol, identify as many as eligible patients and consequently corresponding providers
  - Claims data contains anonymized patients information
  - Outreach campaigns do not use patient data



NLP SUMMIT

Current challenges

- Manual and rule-based with high focus on centralized site location
- Time consuming and resource intensive to convert natural language in clinical trials study protocols to query language
- Inflexible due to inadequate rule coverage

Providers of potential eligible patients

Data sources (EHR, Claims)

Data owners

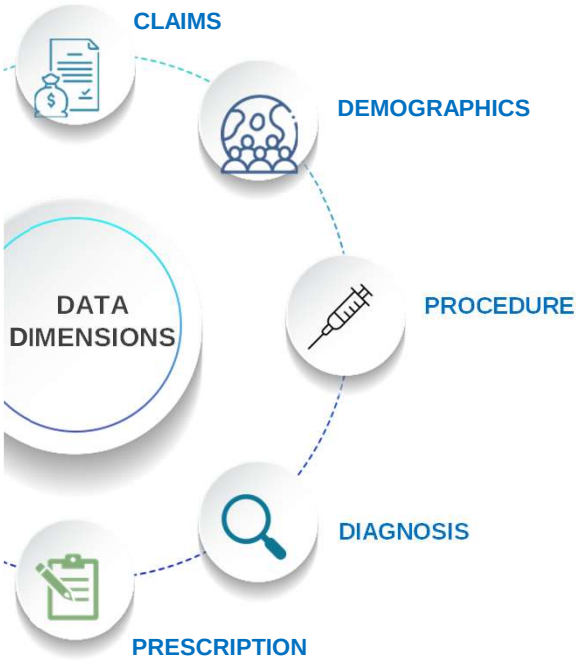
Clinical Trials Study Protocol

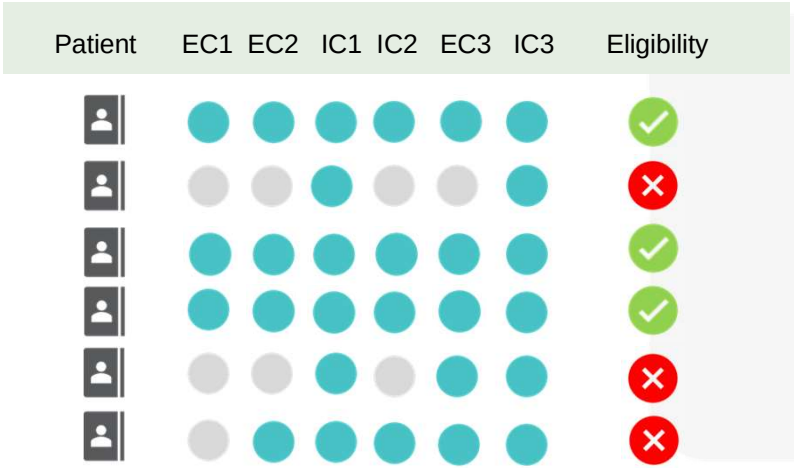
Clinical Operations Team Member

NLP SUMMIT

An automated method

- Empowering outreach programs
- Enable faster design and site selection cycles
- Inform clinical trial study design and its impact on potential patient pool population

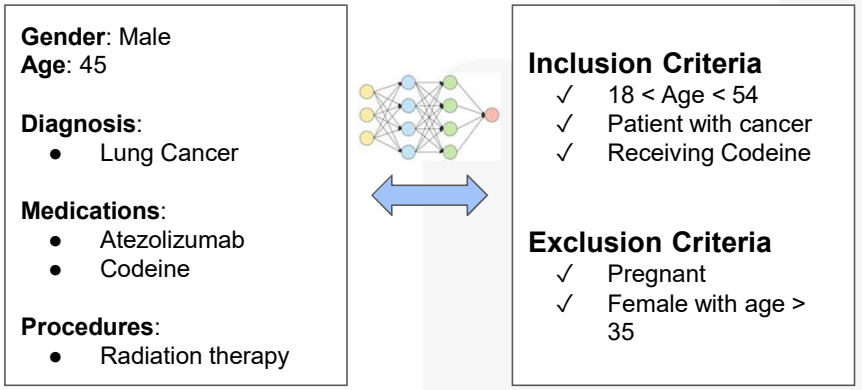




NLP SUMMIT

Problem description

- Structured Electronic Health Records data in IQVIA Claims
  - Procedures
  - Diagnosis
  - Drugs
- Clinical trials study protocols
  - Inclusion criteria
  - Exclusion criteria



NLP SUMMIT

Current methodologies

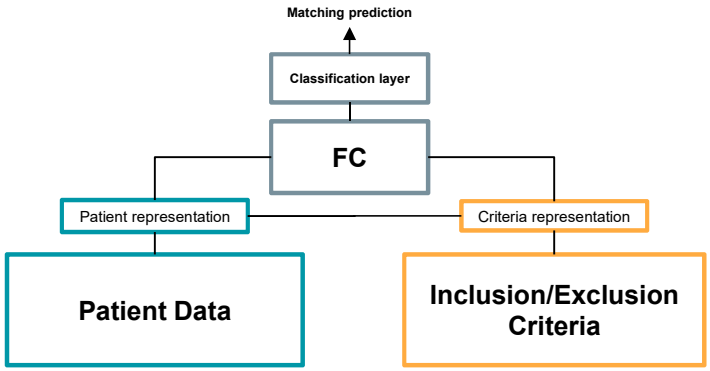
- Rule-based systems:
  - Extracting named entities and relations from inclusion/exclusion criteria and construct rules to identify patients
  - Requires: (1) large number of annotations or (2) train models to extract rules or (3) combined ML & rule-based methods
  - One of the state-of-the-art models: Criteria2Query
- Deep embedding based models:
  - Jointly embeds patient records and trial eligibility criteria in the same latent space and then formulate the problem as a binary classification task per criteria
  - Some models distinguish between inclusion and exclusion criteria
  - One of the state-of-the-art models: COMPOSE

NLP SUMMIT

Problem formulation

- Input data definition
  - Patient records
    - In claims data, each patient can be represented as a sequence of multivariate observations
    - Each patient can have many records across time
    - Each record will be represented by list of diagnosis, procedures and products
    - Each record also contains patient demographics information
  - Clinical trials
    - A list of inclusion criteria
    - A list of exclusion criteria
  - Multi-level descriptions of diagnosis, products and procedures

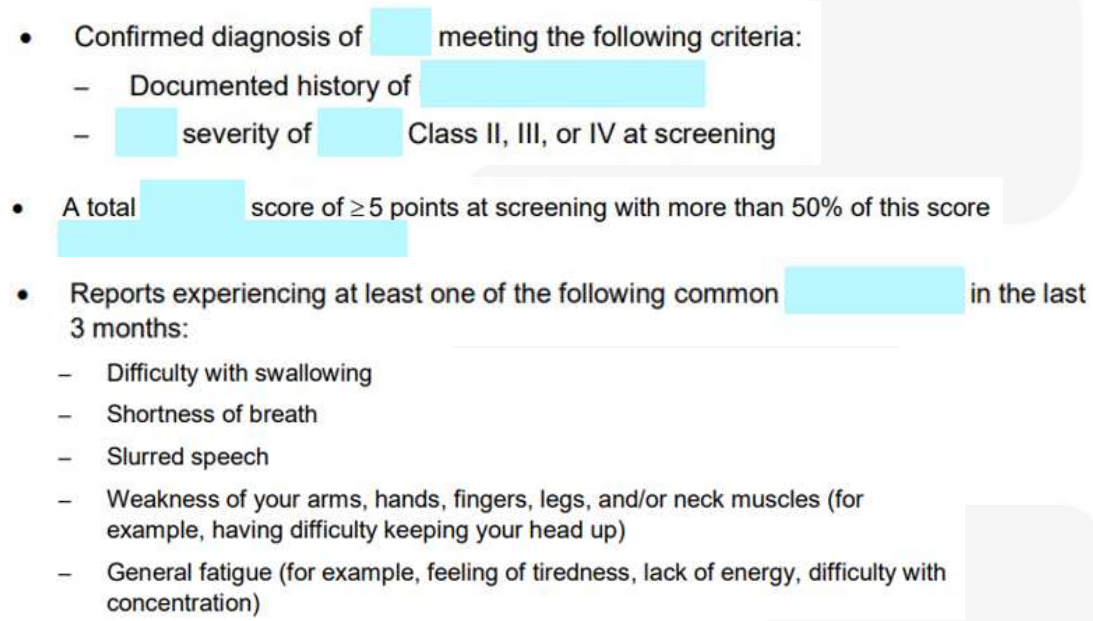
- Problem definition
  - Patient criteria matching
    - Multi-class classification task
    - Classify the matching results between patients and eligibility criteria into “**match**”, “**mismatch**” and “**unknown**”
  - Patient trial matching
    - A patient is a match only if the patient **match**es all inclusion criteria and **mismatch**es all exclusion criteria



Psuedo-Siamese Network

Data and modeling challenges

- Existing methodologies suffer from poor recall
- Training data is not available
  - Manual data annotation
  - Evaluate with clinical trials team
- Complex clinical trial criteria
  - Eligibility criteria often encode more general disease concepts (e.g. cardiovascular disease)
  - EHR represents patients conditions using more specific medical codes (e.g. peripheral vascular disease or PAD)
  - Severity or score calculations, multi-line criteria, etc.



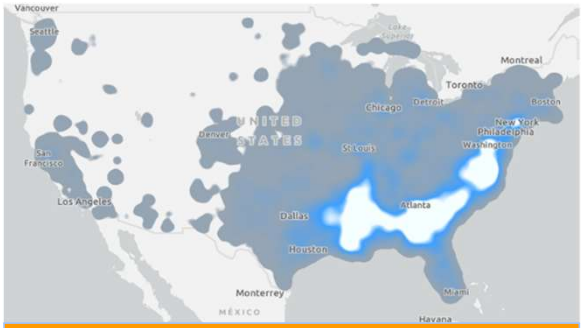
- Confirmed diagnosis of meeting the following criteria:
  - Documented history of
  - severity of Class II, III, or IV at screening
- A total score of ≥5 points at screening with more than 50% of this score
- Reports experiencing at least one of the following common in the last 3 months:
  - Difficulty with swallowing
  - Shortness of breath
  - Slurred speech
  - Weakness of your arms, hands, fingers, legs, and/or neck muscles (for example, having difficulty keeping your head up)
  - General fatigue (for example, feeling of tiredness, lack of energy, difficulty with concentration)

NLP SUMMIT

Collaboration – rare disease trial

- Identifying providers of potential patients accurately and empower targeted outreach campaigns
- Measuring accuracy of automated approaches and their impact is imperative.
- Empowering inclusive research for underrepresented patient populations

Sum of # potential patients of identified providers per zip code



Diversity Index w.r.p.t. specific ethnicity group

Sum of # potential patients of identified providers within 150 miles of selected sites

NLP SUMMIT

Conclusion

- Quality healthcare relies on many sources of information
- Healthcare professionals are baffled with **mountains** of **healthcare data** and **ever-expanding medical knowledge** that is making it increasingly hard to master
- Clinical operations teams are working tirelessly to advocate for **health equity** and **precision medicine**
- AI/NLP can help to empower **inclusive research** and facilitate **faster research cycles**

NLP SUMMIT

NLP SUMMIT

HEALTHCARE

Presented by

John Snow LABS

www.nlpsummit.org