minnesota cancer alliance summit

2026 | *the power of collaboration*

February 26, 2026

McNamara Alumni Center

Minneapolis, MN

Breakout Session #3

Memorial Hall

February 25, 2026

McNamara Alumni Center

Artificial Intelligence in Cancer Screening, Navigation, Early Detection, and Therapeutic Decision Making

David Perdue, MD, MSPH

~~*Ali Khammanivong, PhD, MS*~~

Rui Zhang, PhD, FAMCI, FAMIA, FIAHSI

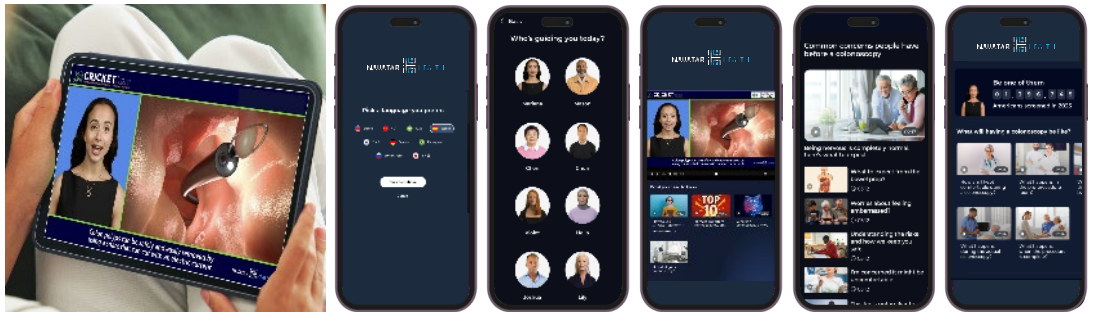
Financial Disclosure Statement

- This session is not eligible for continuing education hours due to unmitigable relevant financial disclosures for David Perdue, MD, MPH and Ali Khammanivong, PhD, MS.

NAVATAR HEALTH

"We make healthcare make sense....
for everyone"

Making AI Make Sense: Artificial Intelligence in Cancer Prevention and Early Detection



David G. Perdue MD MSPH
Founder and Principal
Navatar Health

2025 MN CUP FINALIST

MINNESOTA CUP SEMIFINALIST

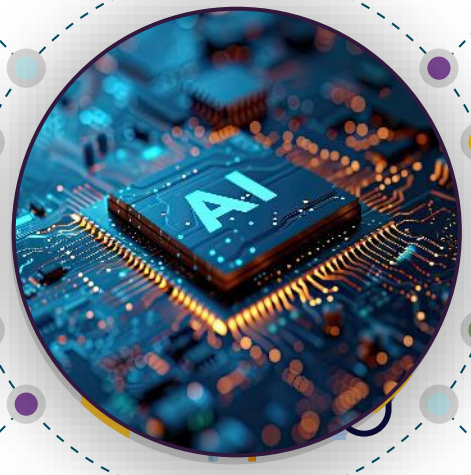
BETA Cohort Spring 2025

Because Every 1% Counts

www.NavatarHealth.com

Info@NavatarHealth.com

Applications Of Artificial Intelligence In Medicine



01
Disease diagnosis
Accurate illness identification through symptom analysis and data.

02
Drug discovery
Speeds research, predicts properties, finds potential treatments.

03
Medical imaging analysis
Improves diagnostics, detects anomalies in scans effectively.

04
Personalized treatment plans
Tailors therapies, utilizes patient data, genetic information.

05
Virtual health assistants
Provides care, schedules, offers medical guidance remotely.

06
Remote monitoring systems
Tracks health data, alerts changes, enhances remote care.

07
Precision surgery
Guides surgeons, boosts accuracy, minimizes invasiveness,.

08
Health informatics
Manages records, analyzes data, enhances healthcare delivery.

09
Predictive analytics
Forecasts outcomes, identifies risks, optimizes care plans effectively.

10
Medical research
Speeds discoveries, analyzes data, informs treatments efficiently.

Overview

Part 1

- Health literacy as the hidden barrier to better care
- Brief overview of the Navatar platform and our use of Generative AI

Part 2

- Artificial intelligence demystified
- Using AI to engage, educate, and guide patients
- Keeping humans at the center of care

David G. Perdue MD MSPH
Boarded Gastroenterologist
Founder and Principal
Navatar Health

University of Minnesota
Assistant Professor 2005-2008
Endoscopy Director 2005-2008

American Indian Cancer Foundation
Founder/ Board Member 2008
Chief Medical Officer 2008-2013

MNGI Digestive Health
Chief Medical Officer 2017-2022
Board of Directors 2013-2017
Director of Quality 2017-2022

Navatar Health
Founder Principal January 2025-

Over **30 Peer-Reviewed Publications**, Mostly on Colorectal Cancer and Screening Knowledge, Attitudes, Beliefs and Barriers

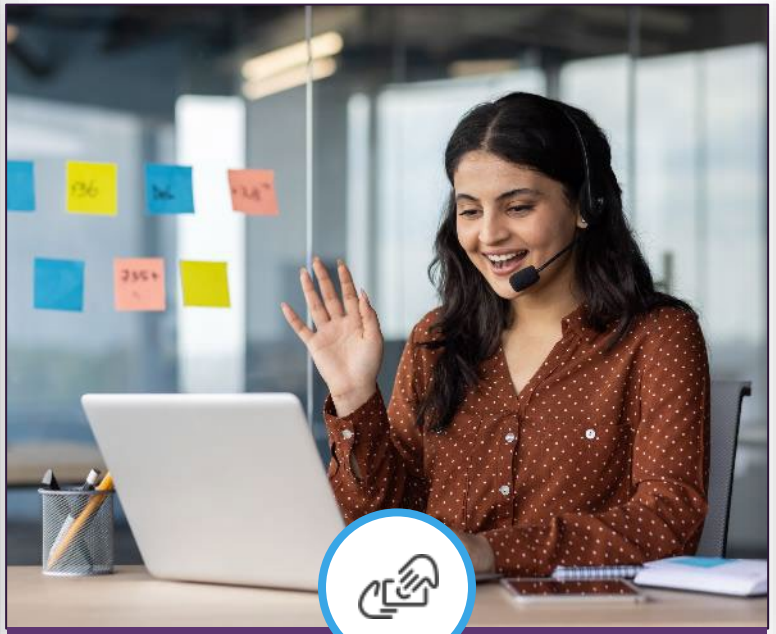
American Indian Cancer Foundation

Minnesota Intertribal Colorectal Cancer Council

NAVATAR HEALTH

Our Community Caregivers Are Irreplaceable

Relationship Based



Cultural & lived experience

Trust cannot be automated

There are only ~**18 CHWs per 100,000** U.S. residents. -US Bureau of Labor Statistics 2023

25% of U.S. cancer patients do not have access to patient navigators. -Commission on Cancer 2024

NAVATAR HEALTH

The U.S. Healthcare Spending—

The U.S. Spends $4.5 Trillion Annually On Healthcare

That's **40%** of the Global health budget being spent on 4% of the worlds population,
And **17%** of our GDP (**30%** more GDP than 2nd place Switzerland)

Yet We Are LAST Among Peer Nations in:

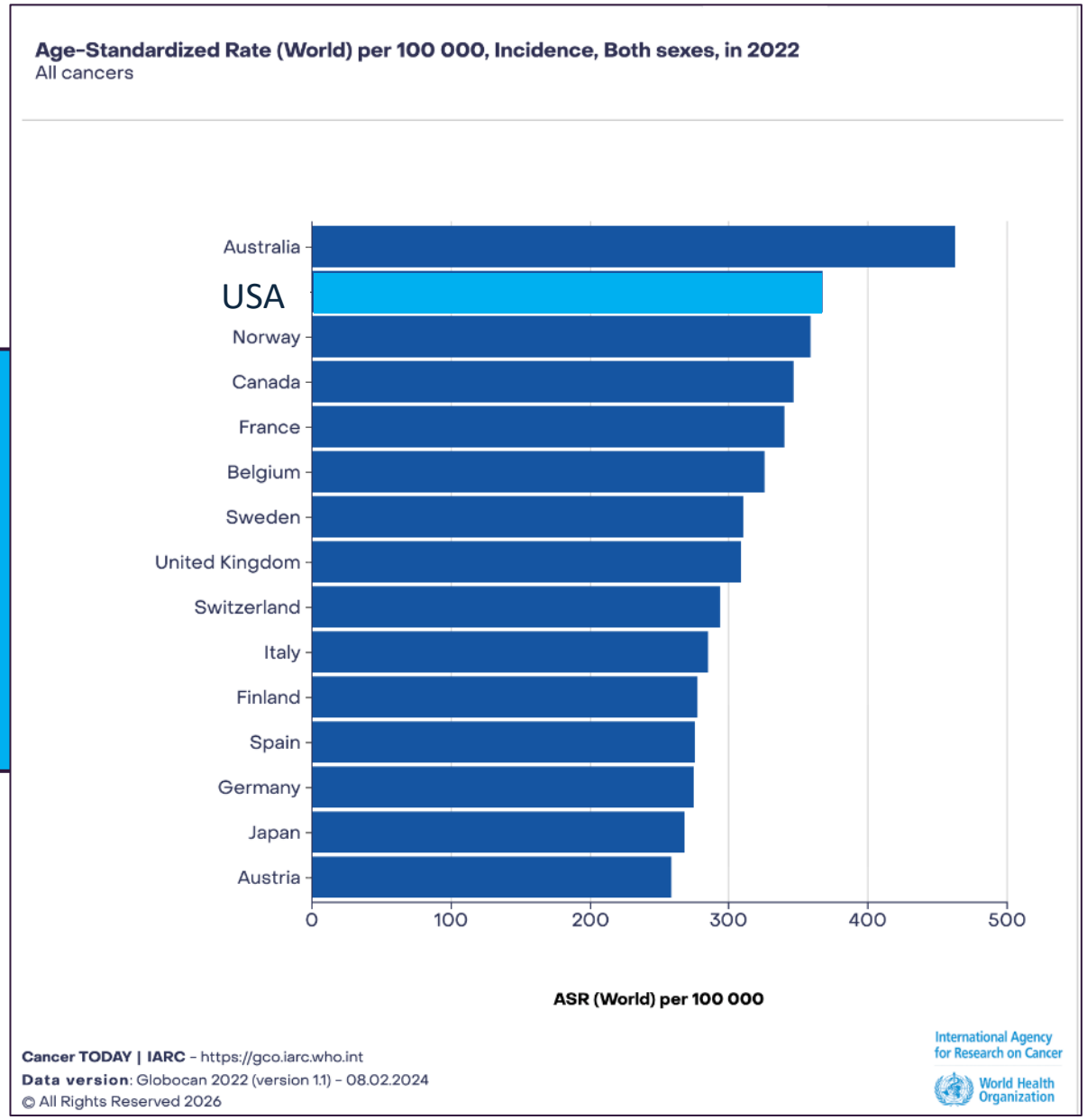
- Chronic Disease Prevalence
- Life Expectancy
- Preventable Hospitalizations
- Preventable Deaths

NAVATAR HEALTH

The U.S. Lags in Cancer Incidence

Cancer Incidence U.S. Versus Peer Nations

-IARC 2022



NAVATAR HEALTH

And Low Health Literacy is a Major Villain

Only 12% of Americans Have Adequate Health Literacy

- Centers for Disease Control

Health Literacy: The ability to understand and act on health information

54% of All Americans
Read below the 6th grade level

1 in 5 Americans
Speak a language other than English at home

66% with Low Health Literacy were born in the US

48% with low health literacy have a college degree

Only 2.1%
of patient-facing health information meets AMA literacy guidelines

-J of Patient Exper, 2021

The American Medical Association recommends patient information be written at the 6th grade reading level

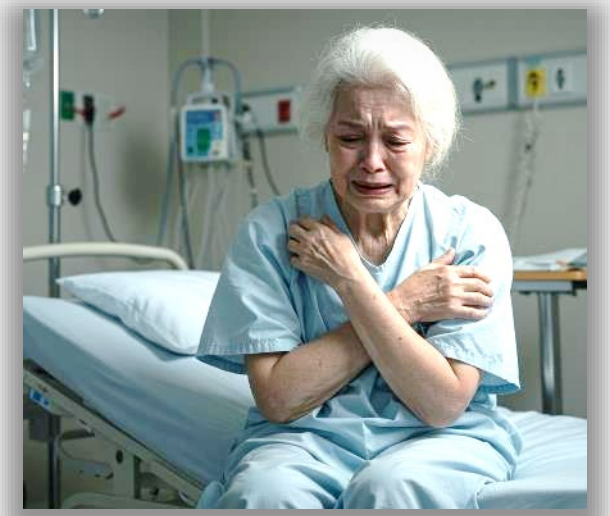
Healthcare assumes comprehension. Too many patients are left guessing.

Key Point: Great care can't deliver results if patients can't understand what to do.

Poor Health Literacy Accounts For:

32% of Hospitalizations

-J Med Internet Res 2019



34% of Readmissions

-J Am Geriatr Soc, 2021

130% Higher ER Usage

-Acad Emerg Med, 2019

Improvement could save Medicare $25.4 B /year

- United Health Group, 2020

Low Health Literacy Accounts For 17% of U.S. Personal Healthcare Expenditures*. (~$646 Billion/ Year) - *Milken Institute, 2022*

*PHCE totals $3.8 Trillion per year and represents total health spending minus administrative costs, public health programs and insurance operations.

NAVATAR HEALTH

Patients Want Video To Help Them Understand Their Care

60%

Use Social Media for Health Information

-Health Information National Trends Survey, USA 2022

Much of this content is unvetted, inconsistent, or misleading.

50%

Prefer Video-Based Health Education

-Health Information National Trends Survey, USA 2022

LLMs write at the 11th-12th grade level and are not designed for patient education

Key Point: Video is now where patients go to understand their health — for better or worse

Imagine if Every Patient Had…

A Familiar, Trusted Health Guide—

An Expert You Understand

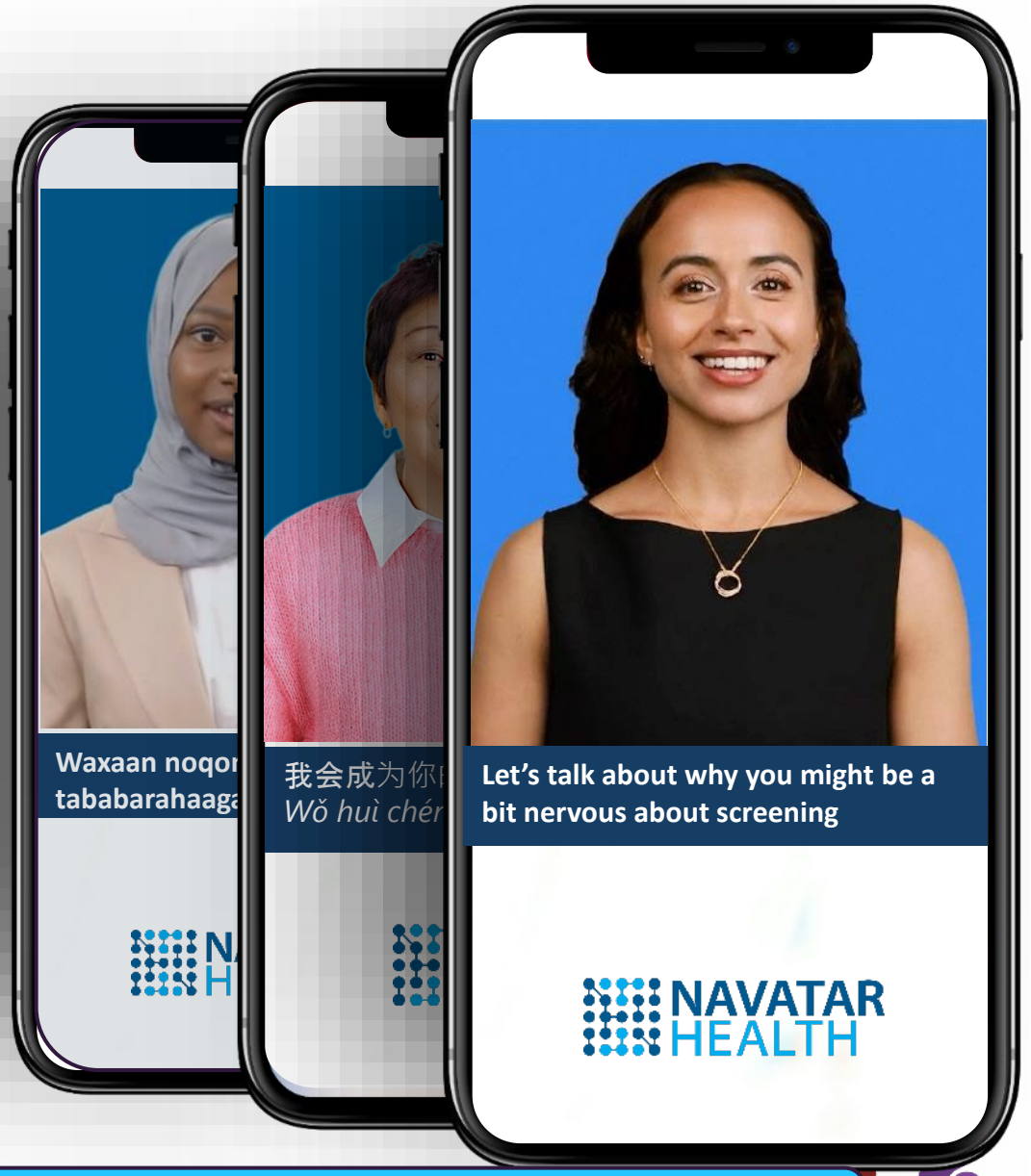
Who speaks your language and explains healthcare clearly, so you can follow-through

Supports You Can Rely On

Answers questions, guides decisions, and stays with you every step.

Always Available

On demand when ever help is needed,

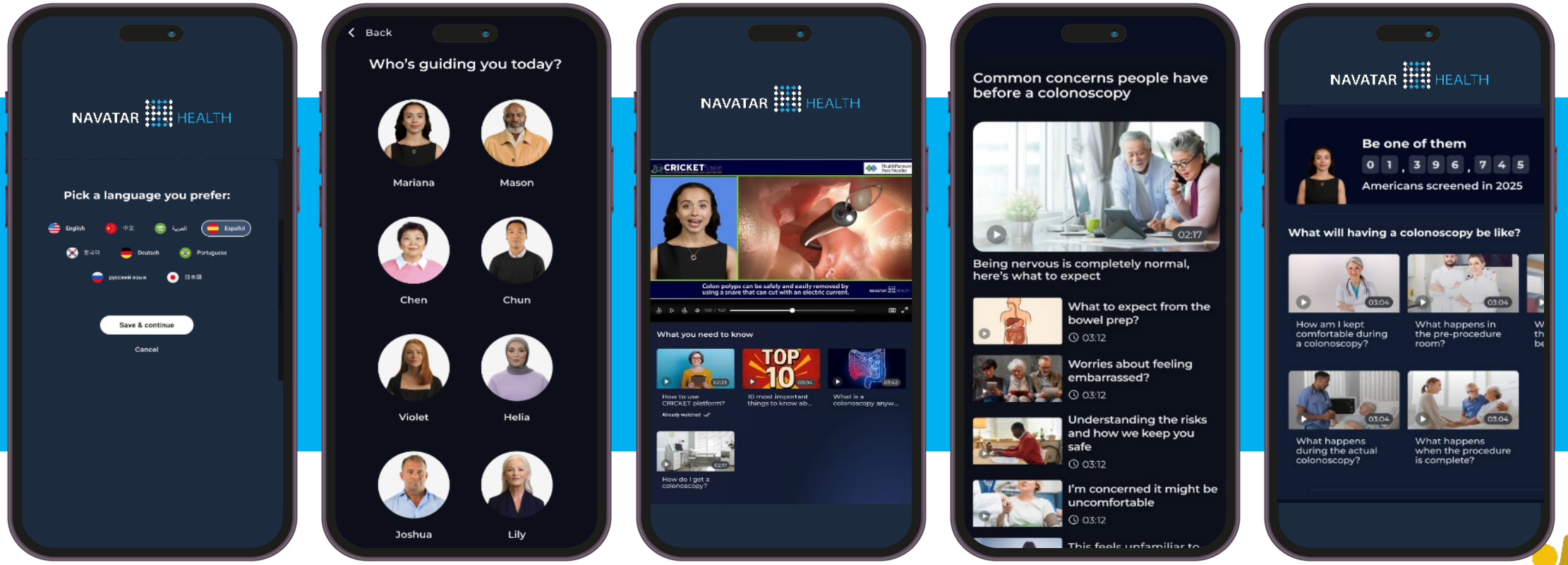


Culturally and linguistically matched AI health guides

Engage patients with human-like AI health guides

Educate with evidence-based, guideline-consistent content

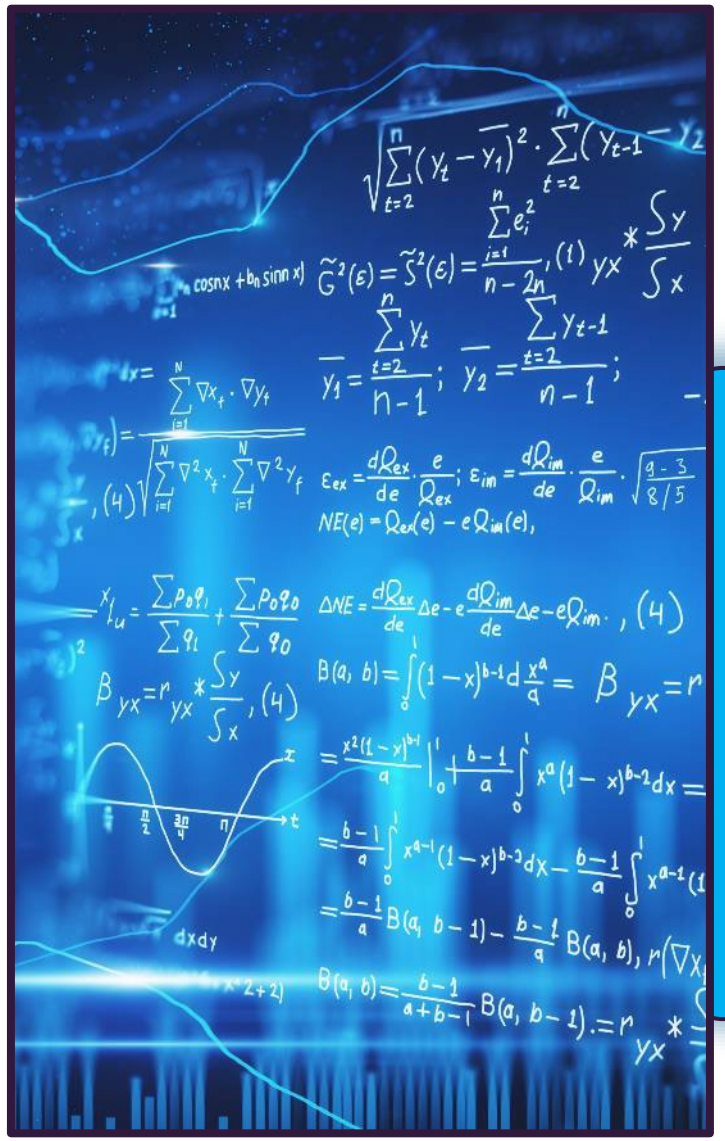
Support and Guide patients 24/7, across languages, devices and care journeys



Screening • Diagnosis • Follow-up • Prevention • Chronic Care

Video demonstration

https://vimeo.com/1146652156/8a28941f42?fl=ip&fe=ec



How Does Artificial intelligence. Work Anyway?

- Its NOT human intelligence:
- Trained on HUGE Datasets
- Breaks data into small pieces (tokens).
- Calculate how tokens are related and how probable they are to follow each other.
- Gets REALLY good at predicting the next word

It feels like it's thinking…But its just math!

Media Generative AI

Image/Video Generative AI

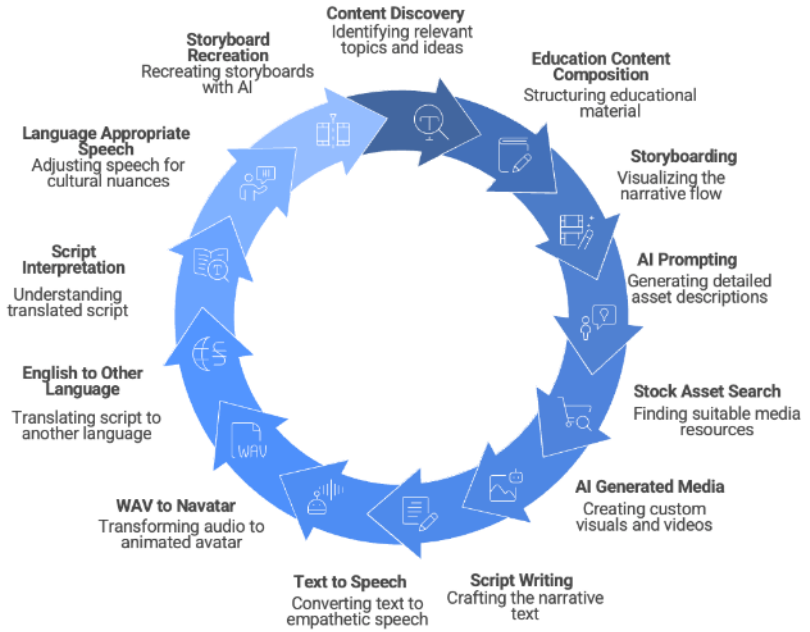
- Learns visual patterns from millions of images
- Starts with random noise and reshapes it step by step
- Predicts what each part of the image should look like
- Builds a new image that matches the prompt

Sound Generative AI

- Learns patterns from speech, music, and other audio
- Predicts small pieces of sound over time
- Builds audio step by step to match tone and style
- Generates new voices or music rather than replaying recordings

Generative AI At Navatar Health

Agentic Prototyping Engine



Rapidly prototypes content for any medical topic or use case

Transcreates* content for different languages and ethnicities

Agentic Transcreation Engine

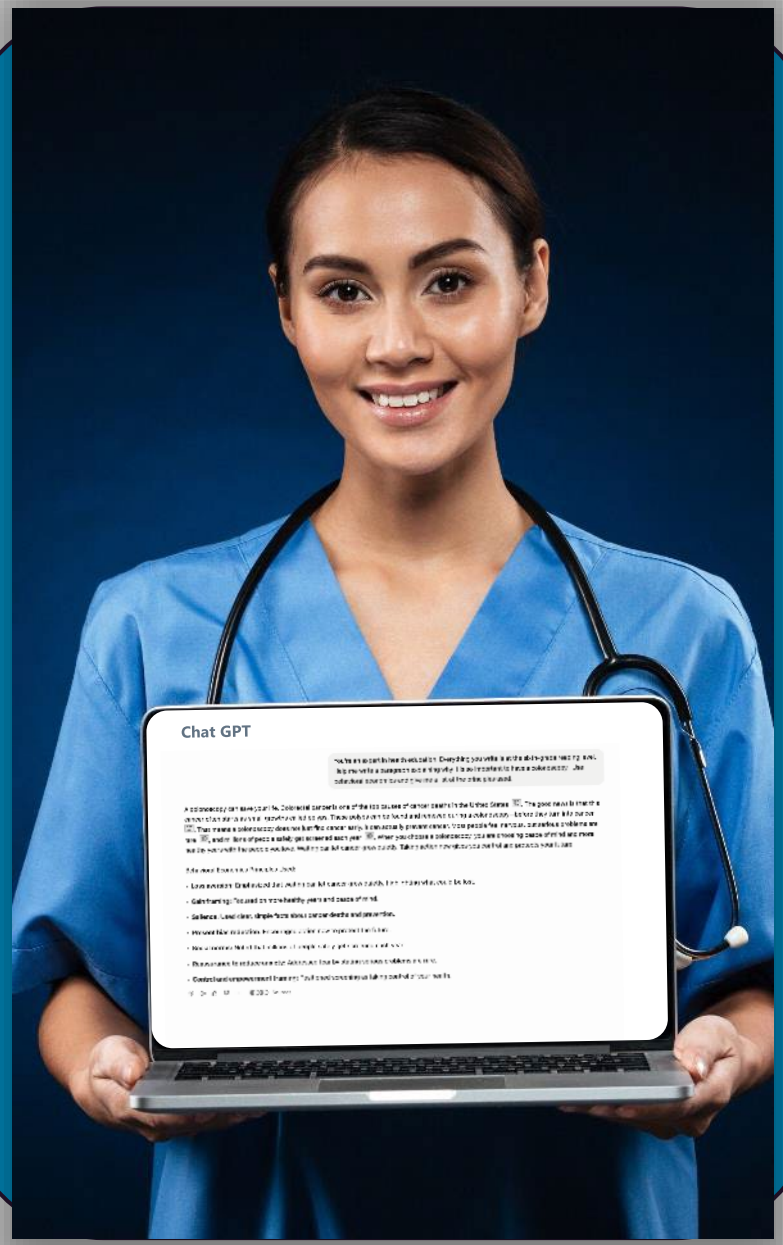
***Transcreation** is the science of redesigning a message so it aligns with a population's language, culture, and decision-making patterns.

Agentic Rapid Prototype Engine
Test and iterate new journeys fast

Agentic Transcreation Engine
Adapt content across language & culture

Realtime Conversational Navatars
Interactive, patient-facing guidance

Key Point: Navatar's modular AI approach lets us innovate, adapt, and outperform—now and in the future.



10 Practical Uses of AI for Navigators & CHWs

- Write and rewrite pretty much anything
- Draft outreach message templates
- Build FAQ sheets
- Translate & simplify education materials
- Identify and remove medical jargon
- Convert guidelines into plain language
- Role-play hesitant patient scenarios
- Generate media-based teaching materials
- Draft report narratives

The goal is simple: **use AI to reduce busy work so you can focus on judgment, creativity, empathy, and impact**

Every Hour Invest Learning Saves You Hundreds

AI Proficiency Requires New Skills

- Make **Prompting** Your Superpower
- Learn "**Custom AI Assistants**"
  - Allow customized output
    - Chat GPT: Custom GPTs
    - Claude: Projects
    - Gemini: Gems

Start with this Prompt:
You are an AI expert. Give me 15 of your best tips for prompting.

Learning Sources

YouTube – Prompt Engineering for Beginners
Search: "How to get better answers from AI"
https://www.youtube.com

LearnPrompting.org
Free, structured guide with hands-on examples
https://learnprompting.org

Google Prompting Essentials (Grow with Google)
Beginner-friendly, practical short course
https://grow.google/prompting-essentials

DeepLearning.AI Short Courses
Foundational AI + prompting courses
https://www.deeplearning.ai/short-courses

Coursera, Udemy
Have multiple courses

Prompting: Clarity in, Clarity Out

Simple Framework:

WHO- do you want the AI to be?
(what expertise or skills should it have)

WHAT- do you want to it to do?
(write, summarize, explain, design…)

WHAT- are you going to give it to help it
(an example, an outline, a source…?)

HOW- you want it outputted.
(for Word, as bulleted list…etc)

WHO: You are a MCHES-certified, health educator very experienced in working with Spanish-speaking immigrant populations

WHAT: Using the outline to compose text for a patient handout. Compose both an English and a Spanish version with numbered paragraphs for easy comparison. Write at the sixth-grade reading level and keep sentences less than 12 words.

WHAT: I'm attaching an outline of a handout on home cervical cancer screening.

HOW: Avoid markdown formatting. Output as plain text formatted for direct pasting into Microsoft Word with normal paragraph spacing. Avoid icons. Try to keep this to ~ 600-words.

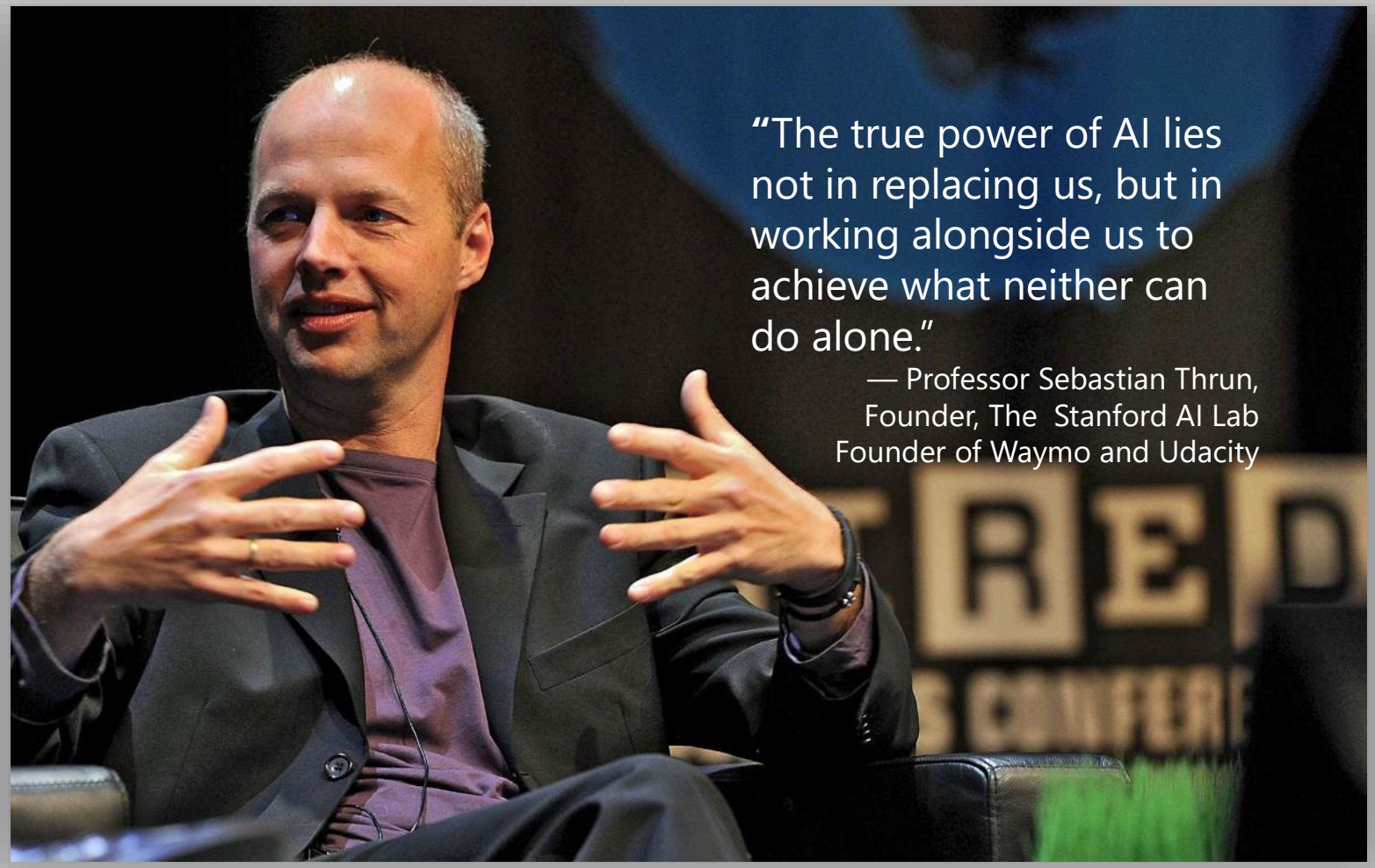
Hallucinations

AI Making Stuff Up:
Training Flaws/Training Data Flaws/Poor Prompts

'Human in the Loop' Is Critical

- **AI Just Drafts**
  AI generates initial content, data analysis, or recommendations.
- **Humans Decide**
  Human experts review, validate, and make the final decisions.
- **Verify Sources**
  Humans ensure the accuracy and reliability of information used or generated by AI.
- **Check Recency**
  Humans confirm the timeliness and relevance of the data and AI outputs.
- **Use Professional Judgment**
  Human expertise, critical thinking, and ethical considerations are indispensable for effective and responsible AI deployment.

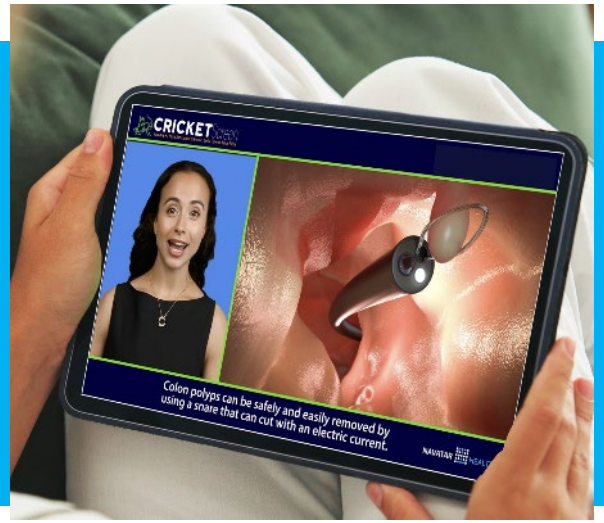
A.I. is Not Something to Fear
It is Something to Embrace and Learn

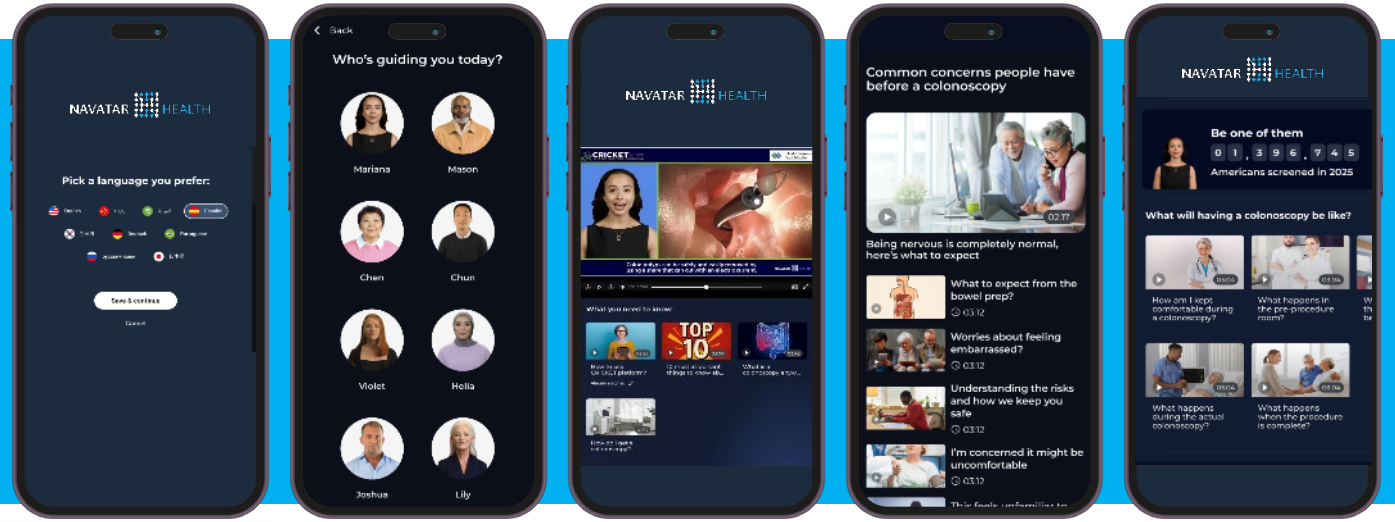


NAVATAR HEALTH

“We Make Healthcare Make Sense… for Everyone”

Better understanding drives better outcomes, lower costs, and scalable impact





2025 MN CUP FINALIST

BETA Cohort Spring 2025

Because Every 1% Counts

www.NavatarHealth.com

Info@NavatarHealth.com

Division of Computational Health Sciences
Department of Surgery
University of Minnesota
Driven to Discover®

Masonic Cancer Center
University of Minnesota
Comprehensive Cancer Center designated by the National Cancer Institute

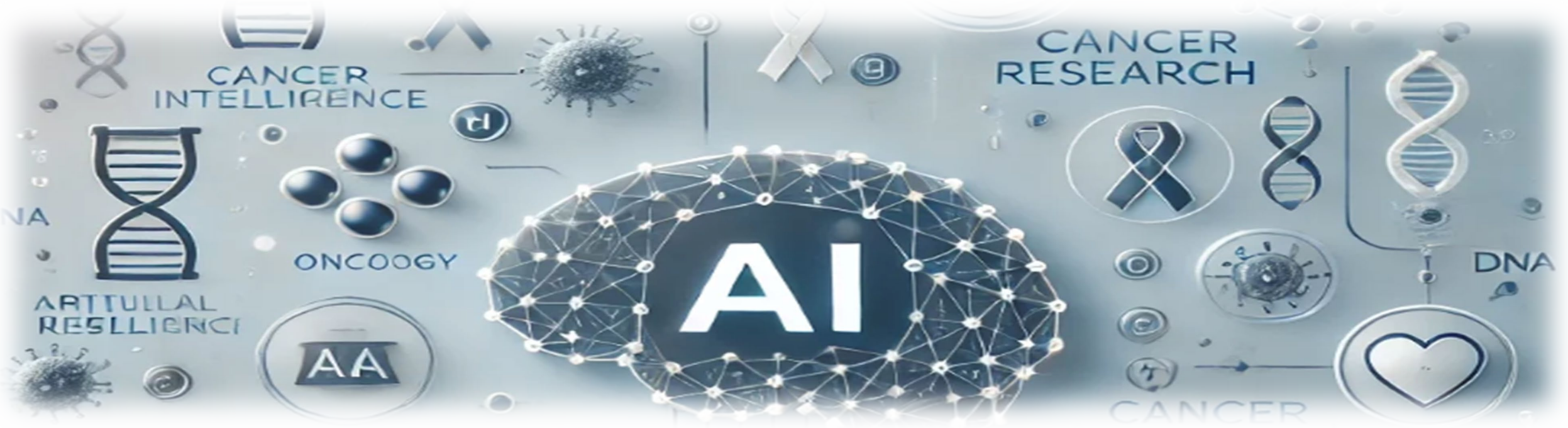
Center for Learning Health System Sciences
University of Minnesota
Driven to Discover®

Data Science and AI Hub

AI and LLM for Cancer Diagnosis and Treatment Risk Prediction

Rui Zhang, PhD, FACMI, FAMIA, FIAHSI

Professor and Founding Chief, Division of Computational Health Sciences
Chair, Data Science & AI for Healthcare (AID-H), Data Science Initiative
Director, Natural Language Processing/Information Extraction program
Associate Director, Center for Learning Health System Sciences
Member, Masonic Cancer Center
University of Minnesota, Twin Cities



- **CancerBERT: cancer phenotyping algorithm**
- **CancerLLM: domain specific LLM**
- **AI generalizability across institutions**
- **Treatment associated cardiotoxicity risk prediction**
- **All of Us data for oncology research**
- **Multimodal learning for cancer diagnosis**

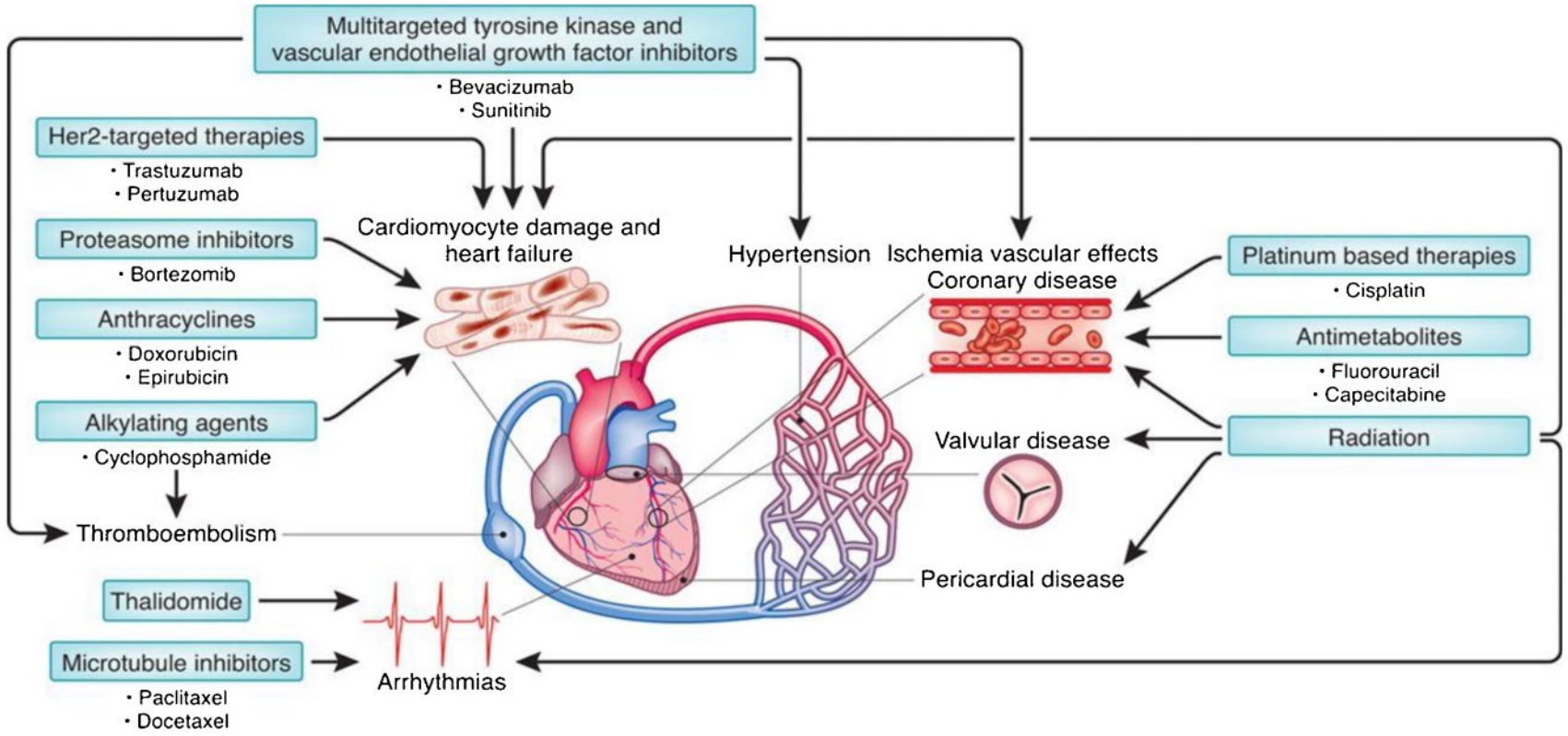
NIH NATIONAL CANCER INSTITUTE
1R01CA287413
(PI: Sun/Zhang/Cui)

NIH National Institute on Minority Health and Health Disparities
1R21MD019134
(PI: Zhang/Hou/Wang)

FDA
U01FD008720
(PI: Zhang/Liu)

Breast Cancer Treatment Associated Cardiotoxicity

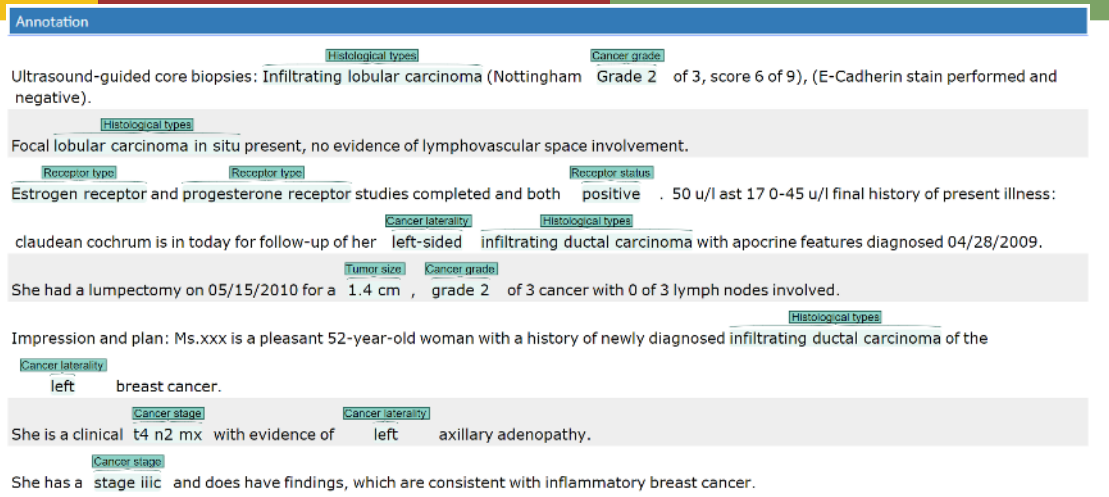
- Breast cancer has the second highest death rate in the U.S
- Cardiotoxicity is a significant problem associated with breast cancer treatments.
- It is one of the leading causes of death for breast cancer patients, ranging from 7.4% to 13.3%
- Different cancer treatments induce the cardiotoxicity in different mechanisms

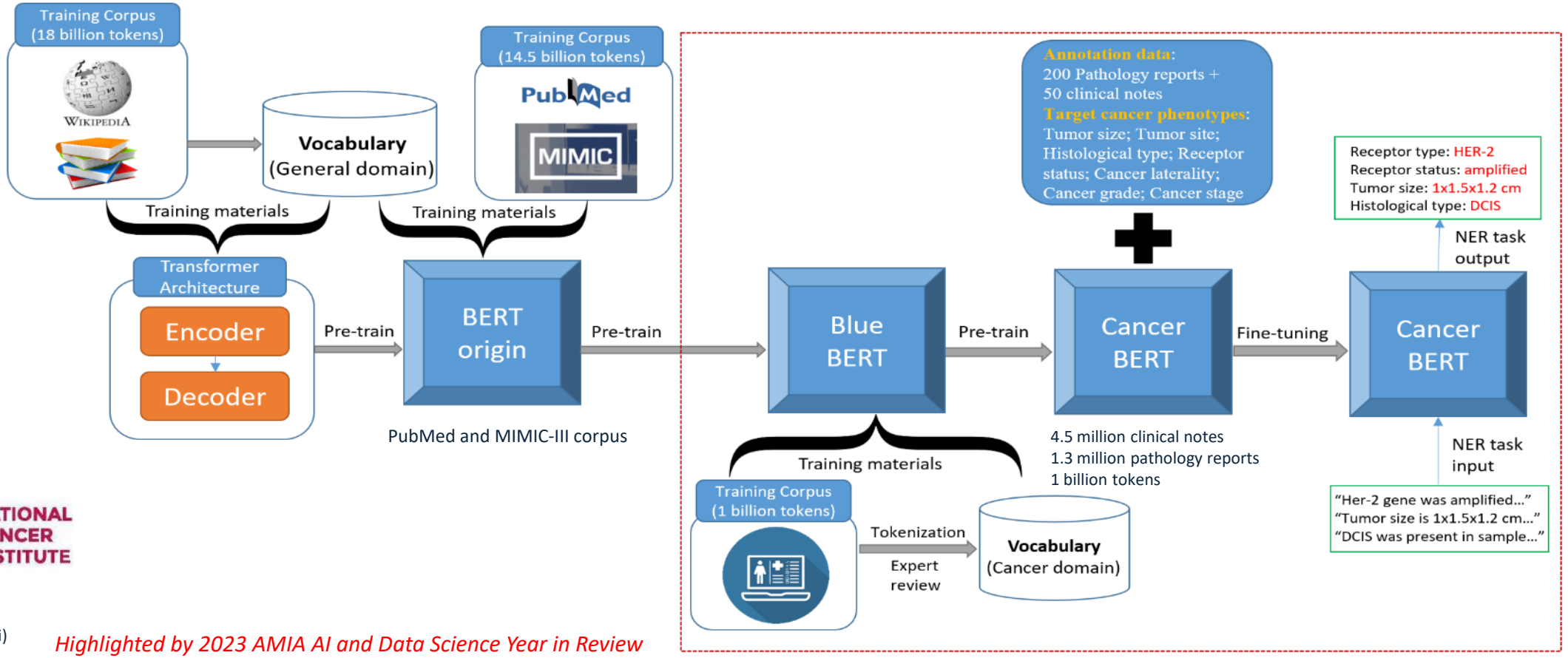


CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, Rui Zhang ✉

Journal of the American Medical Informatics Association, Volume 29, Issue 7, July 2022, Pages 1208–1216, https://doi.org/10.1093/jamia/ocac040





NIH NATIONAL CANCER INSTITUTE

1R01CA287413
(PI: Zhang/Sun/Cui)

Highlighted by 2023 AMIA AI and Data Science Year in Review

Entity type	BiLSTM-CRF	BERT-large Origin	BlueBERT (PubMed+MIMIC III)	BioBERT (PubMed)	character-BERT (Medical)	CancerBERT$_{OrigVoc}$ (EHRs corpus)	CancerBERT$_{CustVoc_997}$ (EHRs corpus)	CancerBERT$_{CustVoc_397}$ (EHRs corpus)
Hormone Receptor type	0.953 (0.957)	0.976 (0.985)	0.979 (0.984)	0.982 (0.987)	0.972 (0.983)	**0.984** (**0.988**)	0.979 (0.985)	0.982 (0.985)
Hormone Receptor status	0.856 (0.856)	0.846 (0.846)	0.885 (0.885)	0.859 (0.859)	0.851 (0.851)	**0.901*** (**0.901***)	0.887 (0.887)	0.891 (0.891)
Tumor size	0.664 (0.709)	0.663 (0.767)	0.781 (0.819)	**0.785** (0.821)	0.674 (0.684)	0.765 (0.813)	0.784 (0.824)	0.781 (**0.827***)
Tumor site	0.562 (0.771)	0.696 (0.769)	0.711 (0797)	**0.749*** (0.799)	0.688 (0.762)	0.733 (0.792)	0.715 (0.787)	0.727 (**0.824***)
Tumor grade	0.910 (0.910)	0.857 (0.857)	0.891 (0.891)	0.886 (0.886)	0.833 (0.833)	0.891 (0.891)	0.898 (0.898)	**0.915* (0.915***)
Tumor laterality	0.935 (0.935)	0.926 (0.926)	0.931 (0.931)	0.943 (0.943)	0.934 (0.934)	0.939 (0.939)	0.947 (0.947)	**0.953* (0.953***)
Cancer stage	0.908 (0.908)	0.804 (0.804)	0.870 (0.870)	0.869 (0.869)	**0.907 (0.907)**	0.870 (0.870)	0.885 (0.885)	0.898 (0.898)
Histological type	**0.885*** (0.938)	0.823 (0.918)	0.843 (0.922)	0.855 (0.934)	0.861 (**0.943***)	0.849 (0.922)	0.862 (0.937)	0.862 (0.938)
Macro average	0.834 (0.873)	0.824 (0.859)	0.862 (0.887)	0.868 (0.889)	0.840 (0.862)	0.867 (0.889)	0.871 (0.896)	**0.876* (0.904***)
Micro average	0.876 (0.905)	0.873 (0.907)	0.898 (0.921)	0.904 (0.926)	0.883 (0.906)	0.903 (0.925)	0.906 (0.930)	**0.909*** (**0.933***)

*Note: The scores were averaged scores based on 10 runs, * indicates statistically higher than other methods (CI: 0.95).*

> Comput Struct Biotechnol J. 2023 Aug 22;22:32-40. doi: 10.1016/j.csbj.2023.08.018. eCollection 2023.

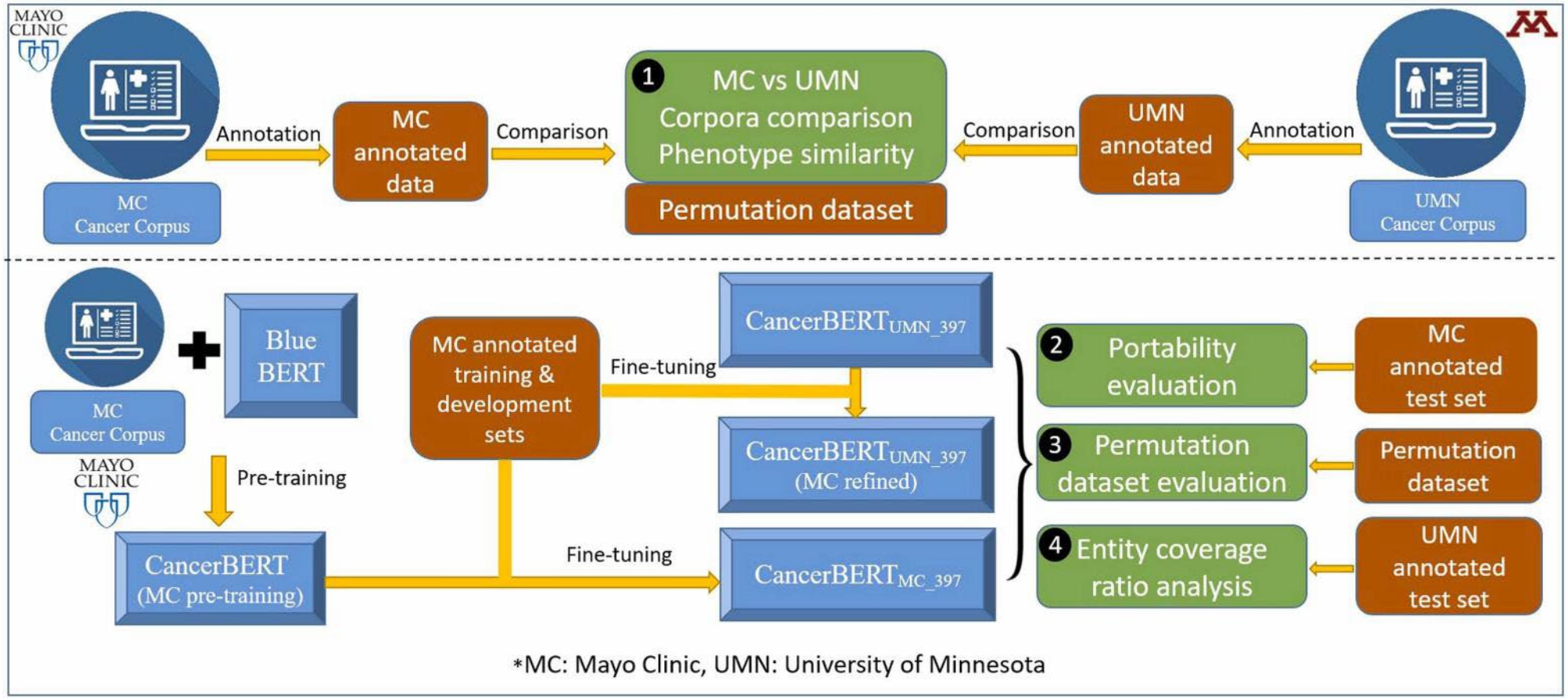
A cross-institutional evaluation on breast cancer phenotyping NLP algorithms on electronic health records

Sicheng Zhou [1], Nan Wang [2], Liwei Wang [3], Ju Sun [4], Anne Blaes [5], Hongfang Liu [3], Rui Zhang [6]

Affiliations + expand

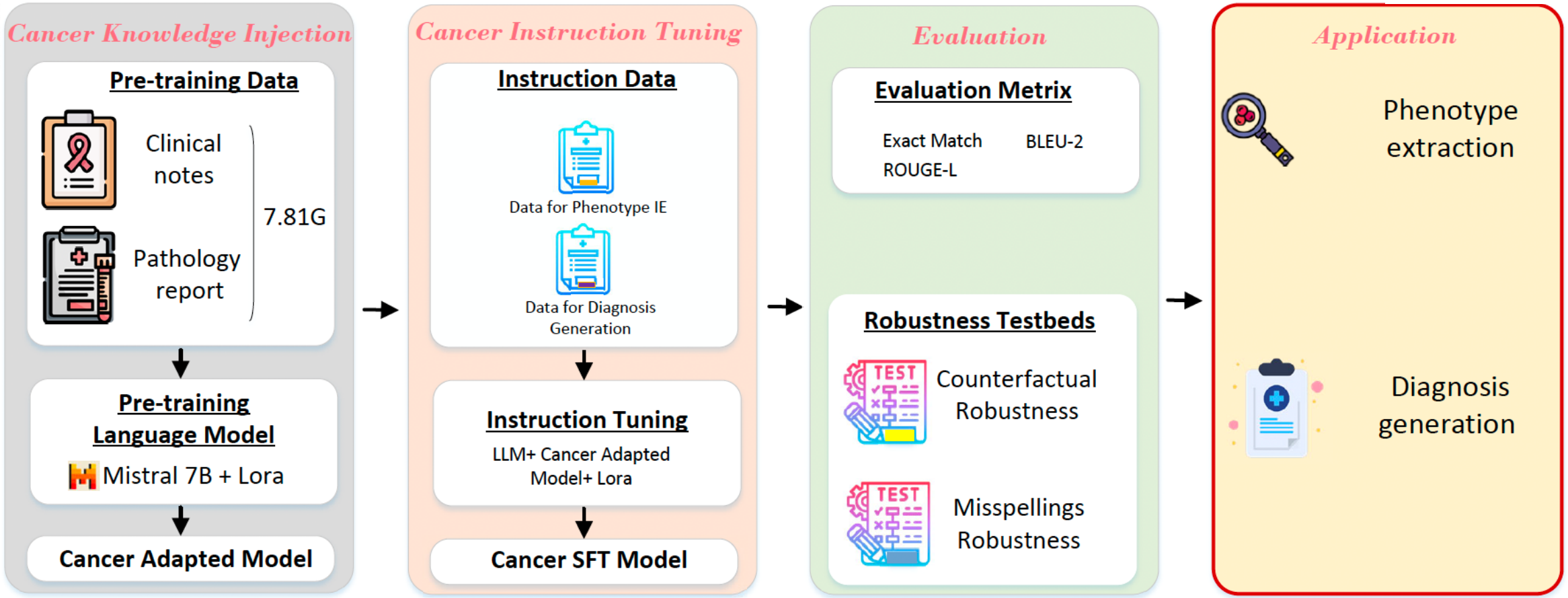
PMID: 37680211 PMCID: PMC10480628 DOI: 10.1016/j.csbj.2023.08.018

- Evaluated the generalizability of BERT-based models across two healthcare systems from different perspectives
- Evaluated the impact of corpus heterogeneity on NLP models' generalizability
- Compared two strategies for transferring models between clinical institutes, i.e., i) direct transfer vs ii) continuous fine-tuning



- **Construct permutation test set**
  - **Identify unique entities from Mayo Clinic that not appear in UMN data**
  - **Replace the entities in UMN data with the randomly sampled entities from Mayo Clinic**
- **The permutation dataset could simulate variations in the data**
- **Provides a better sense of generalizability of models when encounter new data**

CancerLLM: A Large Language Model in Cancer Domain



NPJ Digital Medicine, in press. 2026

Phenotyping Extraction

By providing sentences from cancer pathology reports or cancer clinical notes, the model needs to extract eight specific entities: hormone receptor type, hormone receptor status, tumor size, tumor site, cancer grade, histological type, tumor laterality, and cancer stage.

	question
1	What is the tumor size in the given context?
2	What is the histological type in the given context?
3	Please identify the receptors mentioned in the provided context.
4	What is the receptor type in the given context?
5	Please identify the value of tumor laterality in the provided context.
6	What is the stage of cancer in the given context?
7	Please describe the tumor location in the given context
8	What is the grade of cancer in the given context?

Table 1: Question types

	Exact Match			BLEU-2			ROUGE-L			F1
Approach	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Average
PMC LLaMA 7B	88.61	88.61	88.61	90.36	90.36	90.36	93.26	93.26	93.26	90.74
Medalpaca 7B	89.28	89.28	89.28	91.27	91.27	91.27	93.89	93.89	93.89	91.48
LLAMA-2 7B	89.18	89.18	89.18	90.83	90.83	90.83	93.53	93.53	93.53	91.18
Mistral 1*7B	89.47	89.47	89.47	91.30	91.30	91.30	94.19	94.19	94.19	91.65
Mixtral 8*7B	90.23	90.23	90.23	92.05	92.05	92.05	94.50	94.50	94.50	92.26
Bio-Mistral 7B	88.90	88.90	88.90	91.05	91.05	91.05	93.79	93.79	93.79	91.24
LLama3 8B	89.94	89.94	89.94	91.75	91.75	91.75	94.34	94.34	94.34	92.01
Qwen-7B	88.90	88.90	88.90	90.67	90.67	90.67	93.57	93.57	93.57	91.05
Deepseek 8B	87.19	87.19	87.19	89.12	89.12	89.12	92.24	92.24	92.24	89.52
MedLLaMA 13B	88.80	88.80	88.80	90.87	90.87	90.87	93.31	93.31	93.31	90.99
PMC LLaMA 13B	87.95	87.95	87.95	89.64	89.64	89.64	92.58	92.58	92.58	90.06
Medalpaca 13B	88.61	88.61	88.61	90.37	90.37	90.37	92.94	92.94	92.94	90.64
LLaMA2 13B	89.85	89.85	89.85	91.54	91.54	91.54	94.21	94.21	94.21	91.86
LLaMA2 70B	90.04	90.04	90.04	91.62	91.62	91.62	93.98	93.98	93.98	91.88
Llama3-OpenBioLLM-70B	88.33	88.33	88.33	90.02	90.02	90.02	93.15	93.15	93.15	90.50
ClinicalCamel-70B	**92.02**	**92.02**	**92.02**	**93.62**	**93.62**	**93.62**	**95.52**	**95.52**	**95.52**	**93.72**
CancerLLM 7B(Ours)	89.37	89.37	89.37	91.98	91.98	91.98	93.98	93.98	93.98	91.78

Cancer Diagnosis Generation

By giving the information from cancer clinical notes, which includes the 1) reason for visit, 2) treatment site, 3) subjective information, 4) nursing Review of Systems (ROS), 5) objective observations, and 6) laboratory test results, the model is expected to generate the correct cancer diagnosis.

	Exact Match			BLEU-2			ROUGE-L			F1
Approach	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Average
PMC LLaMA 7B	47.04	47.04	47.04	56.41	56.41	56.41	66.98	66.98	66.98	56.81
Medalpaca 7B	41.75	41.75	41.75	50.16	50.16	50.16	62.07	62.07	62.07	51.33
LLAMA-2 7B	33.18	33.18	33.18	42.80	42.80	42.80	55.09	55.09	55.09	43.96
Mistral 1*7B	45.40	45.40	45.40	54.29	54.29	54.29	65.52	65.52	65.52	55.07
Mixtral 8*7B	51.32	51.32	51.32	59.35	59.35	59.35	69.70	69.70	69.70	60.12
Bio-Mistral 7B	62.26	62.26	62.26	68.40	68.40	68.40	76.02	76.02	76.02	68.89
LLama3 8B	51.60	51.60	51.60	60.13	60.13	60.13	69.97	69.97	69.97	60.57
Qwen-7B	76.66	76.66	76.66	80.59	80.59	80.59	85.46	85.46	85.46	80.90
Deepseek 8B	34.82	34.82	34.82	44.98	44.98	44.98	56.97	56.97	56.97	45.59
MedLLaMA 13B	39.02	39.02	39.02	48.98	48.98	48.98	60.63	60.63	60.63	49.54
PMC LLaMA 13B	54.97	54.97	54.97	62.56	62.56	62.56	71.06	71.06	71.06	62.86
Medalpaca 13B	40.66	40.66	40.66	50.58	50.58	50.58	62.40	62.40	62.40	51.21
LLaMA2 13B	45.76	45.76	45.76	54.70	54.70	54.70	65.80	65.80	65.80	55.42
LLaMA2 70B	50.23	50.23	50.23	59.63	59.63	59.63	69.25	69.25	69.25	59.70
Llama3-OpenBioLLM-70B	54.42	54.42	54.42	62.36	62.36	62.36	71.65	71.65	71.65	62.81
ClinicalCamel-70B	54.60	54.60	54.60	63.34	63.34	63.34	72.73	72.73	72.73	63.55
CANCERLLM 7B(Ours)	**83.50**	**83.50**	**83.50**	**86.60**	**86.60**	**86.60**	**90.34**	**90.34**	**90.34**	**86.81**

Efficiency and Retrieval-augmented Generation

	Phenotype Extraction			Diagnosis Generation		
LLMs	F1	Time	Used Memory	F1	Time	Used Memory
Bio-Mistral 7B	91.24	1:06:55	5,746 MB	68.89	1:07:45	5,802 MB
Mistral 1*7B	91.65	57:05	5,598 MB	55.07	1:07:49	5,680 MB
Mixtral 8*7B	92.26	2:01:27	25,086 MB	60.12	2:16:14	25,166 MB
PMC LLaMA 13B	90.06	1:08:59	8,208 MB	62,86	1:19:52	9,208 MB
LLaMA2 13B	91.86	1:08:43	8,204 MB	55.42	1:24:17	9,254 MB
ClinicalCamel-70B	93.72	2:50:16	37,716 MB	63.55	3:05:37	37,67 MB
CANCERLLM(Ours)	91.78	1:14:12	5,550 MB	86.81	1:26:33	5,768 MB

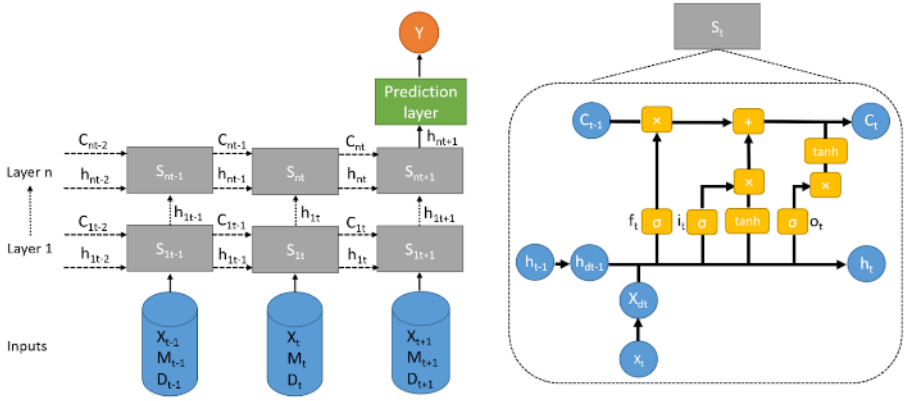
		Exact Match			BLEU-2			ROUGE-L			F1
Task	Retriever	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Average
Phenotype Extraction	Random	89.47	89.47	89.47	91.30	91.30	91.30	93.91	93.91	93.91	91.55
	Medcpt	87.95	87.95	87.95	90.18	90.18	90.18	92.97	92.97	92.97	90.36
	Contriever	88.14	88.14	88.14	90.62	90.62	90.62	94.04	94.04	94.04	90.93
	SGPT	83.02	83.02	83.02	85.67	85.67	85.67	89.93	89.93	89.93	86.20
	Specter2	88.61	88.61	88.61	90.81	90.81	90.81	93.67	93.67	93.67	91.03
	No-retriever	89.37	89.37	89.37	91.98	91.98	91.98	93.98	93.98	93.98	91.78
Diagnosis Generation	Random	43.30	43.30	43.30	52.90	52.90	52.90	63.63	63.63	63.63	53.27
	Medcpt	58.34	58.34	58.34	65.99	65.99	65.99	73.54	73.54	73.54	65.95
	Contriever	82.50	82.50	82.50	85.15	85.15	85.15	88.80	88.80	88.80	85.48
	SGPT	43.94	43.94	43.94	52.97	52.97	52.97	63.75	63.75	63.75	53.55
	Specter2	85.78	85.78	85.78	89.09	89.09	89.09	92.49	92.49	92.49	89.12
	No-retriever	83.50	83.50	83.50	86.60	86.60	86.60	90.34	90.34	90.34	86.81

Risk prediction of cardiotoxicity

- The ICD-9 and ICD-10 codes were used to identify patients diagnosed with breast cancer between 2011-2020.
- Unstructured features from clinical texts (Extract by CancerBERT)
  - Receptor status: HER2, ER, PR
  - Cancer stage: Overall stage and TNM stage
  - Histological type: e.g., invasive carcinoma, metastatic carcinoma
  - Laterality: left or right
- Index date:
  - defined as the first date of any breast cancer treatment: chemotherapy, radiation or targeted therapy
  - the patients with heart diseases before index date were excluded
- Prescriptions of cardiovascular medications and heart diseases (outcomes) were extracted from the follow-up period (after the index date)
- All other variables, the longitudinal observations before index data were summarized as the value before and closest to the index date
- Missing values for continuous variables were imputed either using average values or normal values

Category	Variable
Outcomes	congestive heart failure (CHF), coronary artery disease (CAD), cardiomyopathy (CM), myocardial infarction (MI)
Vitals	systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI)
Labs	high-density lipoprotein (HDL), low-density lipoprotein (LDL), hemoglobin A1c (Hba1c), troponin, triglyceride, abnormal blood pressure, abnormal blood lipid
Pre-conditions	Hyperlipidemia, diabetes, hypertension
Cardiovascular related medications	Insulin, Metformin, Statins, ACE inhibitor, Angiotensin II receptor antagonists, Antihypertensive combinations, Vasodilators, Antiarrhythmic, Beta blockers, Calcium blockers
Cancer treatments	Radiation therapy, Chemotherapy, Targeted therapy
Demographics	Age
Cancer phenotypes (NLP features)	Receptor status, cancer stage, histological type, laterality

iScience, 2024 (27),7:110329.



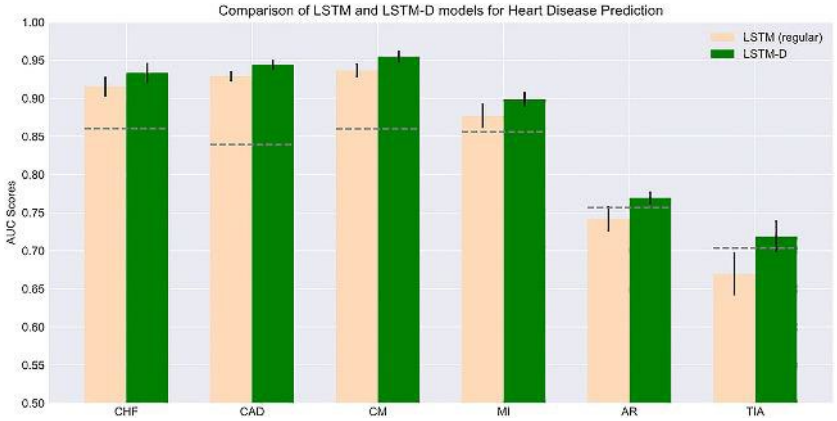
(C)

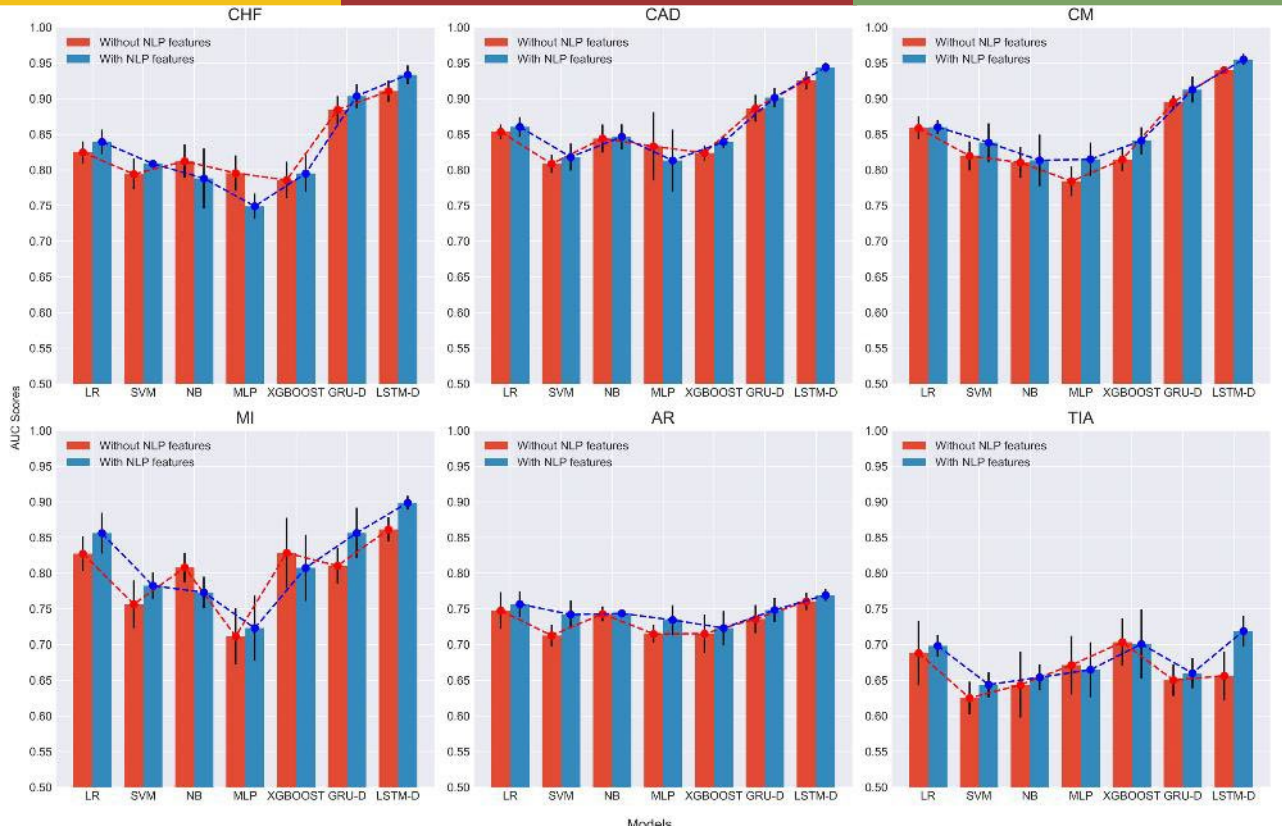
$$X = \begin{bmatrix} 3.5 & 2.6 & NA & 2.5 & 3.0 & NA & 4.1 & 2.4 \\ 115 & NA & 122 & 119 & NA & 131 & 128 & NA \end{bmatrix}$$

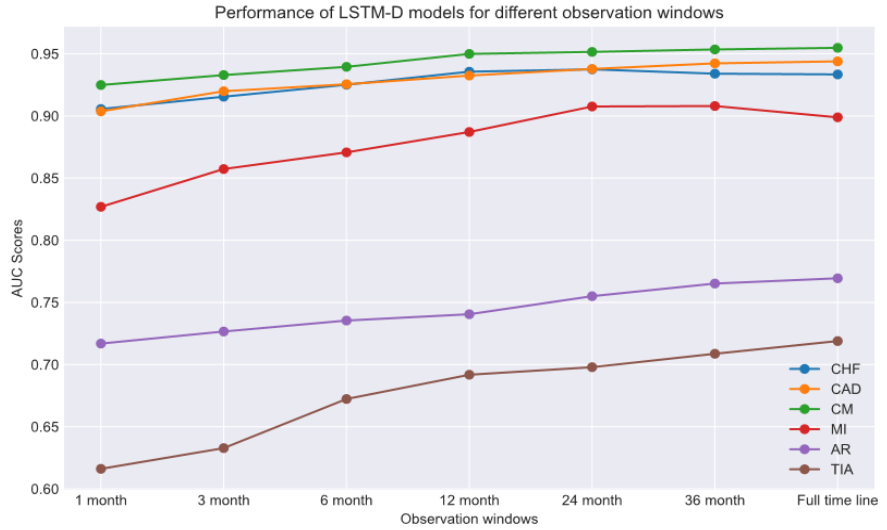
$$M = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & 0.1 & 0.22 & 0.2 & 0.15 & 0.26 & 0.12 & 0.14 \\ 0 & 0.1 & 0.12 & 0.2 & 0.35 & 0.11 & 0.12 & 0.26 \end{bmatrix}$$

Time points (days) = [0 10 22 42 57 68 80 94]







- X: time series feature matrix
- M: mask matrix indicates positions of missing values
- D: delta time depend on time interval between missing values
- Calculated decayed X and h (hidden states) as input for normal LSTM

JOURNAL ARTICLE EDITOR'S CHOICE FEATURED

Multi-modality risk prediction of cardiovascular diseases for breast cancer cohort in the *All of Us* Research Program FREE

Han Yang, MSE, Sicheng Zhou, MS, Zexi Rao, MS, Chen Zhao, MS, Erjia Cui, PhD, Chetan Shenoy, MD, Anne H Blaes, MD, Nishitha Paidimukkala, PhD, Jinhua Wang, PhD, Jue Hou, PhD ✉, Rui Zhang, PhD ✉

Journal of the American Medical Informatics Association, Volume 31, Issue 12, December 2024, Pages 2800–2810, https://doi.org/10.1093/jamia/ocae199

All of Us RESEARCH PROGRAM

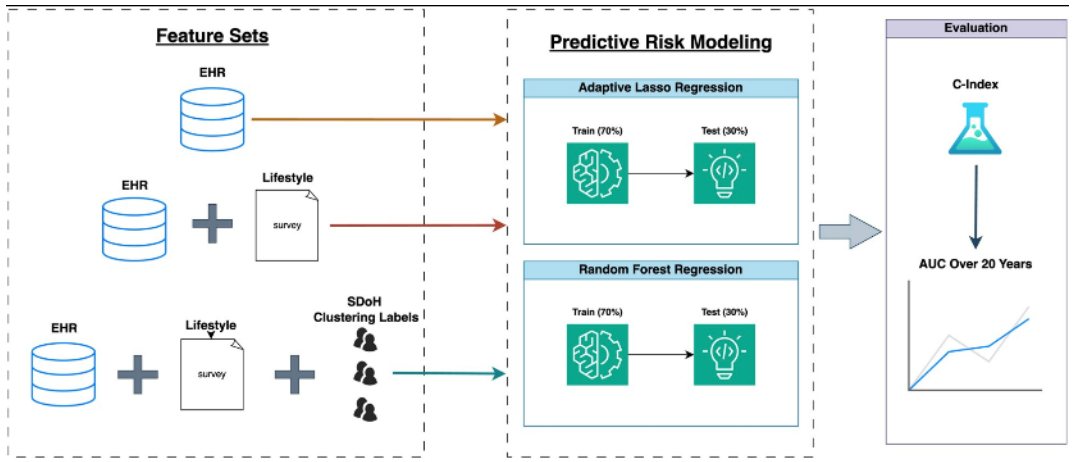
837,000+ Participants

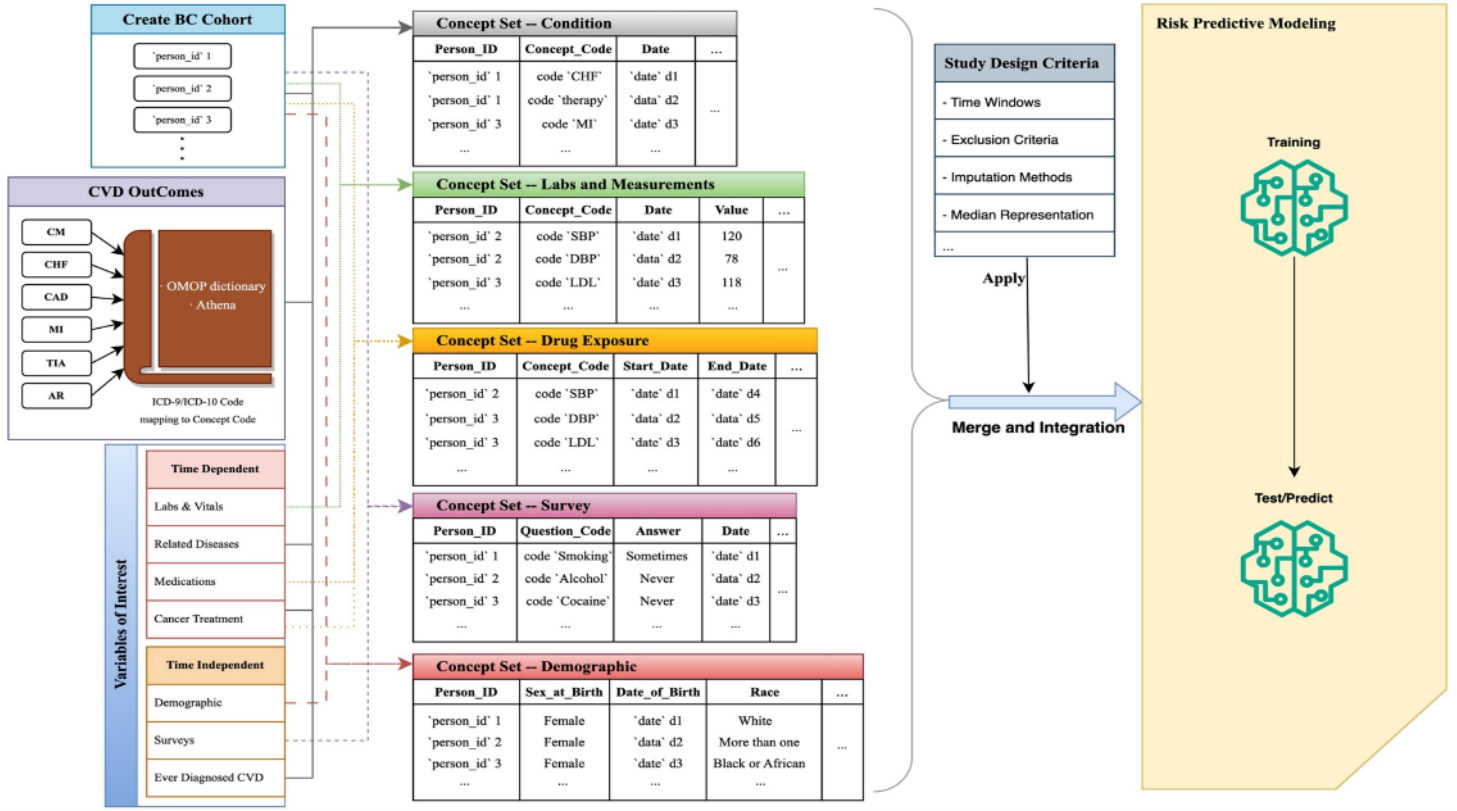
453,000+ Electronic Health Records

589,000+ Biosamples Received

The *All of Us* Research Program is a historic effort to collect and study data from one million or more people living in the United States. The goal of the program is better health for all of us.

- Integrating EHR, Lifestyle, SDoH for prediction
- Leveraging clustering method to group SDoH variables





National Institute on Minority Health and Health Disparities

1R21MD019134
(PI: Zhang/Hou/Wang)

Comprehensive list of variables collected from *All of Us* under different categories.

	Category	Variables
Time-dependent variables	Labs	High-density lipoprotein (HDL), Low-density lipoprotein (LDL), Hemoglobin A1c (HbA1c), Triglyceride
	Diseases	Type-2 diabetes, Hypertension
	Vitals	Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Body mass index (BMI), Weight, Height
	Cardiovascular related medications	Insulin, Metformin, Statins, ACE inhibitor, Angiotensin II receptor antagonists, Antihypertensive combinations, Beta blockers, Calcium blockers, Vasodilators, Antiarrhythmic, Diuretics
	Cancer treatments	Radiation therapy, Chemotherapy, Targeted therapy
	Demographics (part I)	Age
	Surveys	Lifestyle, Social Determinant of Health*
Time-independent variables	Ever diagnosed CVD or not	Cardiomyopathy (CM), Congestive heart failure (CHF), Coronary artery disease (CAD), Myocardial infarction (MI), Transient ischemic attack/stroke (TIA), and Arrhythmias (AR)
	Demographics (part II)	Sex at birth, Race, and Ethnicity

Models	Feature sets	CVD outcomes					
		CM	CHF	CAD	MI	TIA	AR
Adaptive Lasso	EHR	58.9	71.5	65	64.4	66.6	67.4
	EHR+Lifestyle	58.8	62.8	65	71.2	68	59.3
	EHR+Lifestyle+SDoH	58.4	72	67.1	74.5	67.1	66.2
Random Forest	EHR	62.3	72.8	75.2	68.2	64.6	61.7
	EHR+Lifestyle	63.8	72.8	74.9	66.7	64.5	64.5
	EHR+Lifestyle+SDoH	63.8	73.6	73.3	70.9	60.8	65.6

Adding lifestyle and SDoH data generally improved model performance.

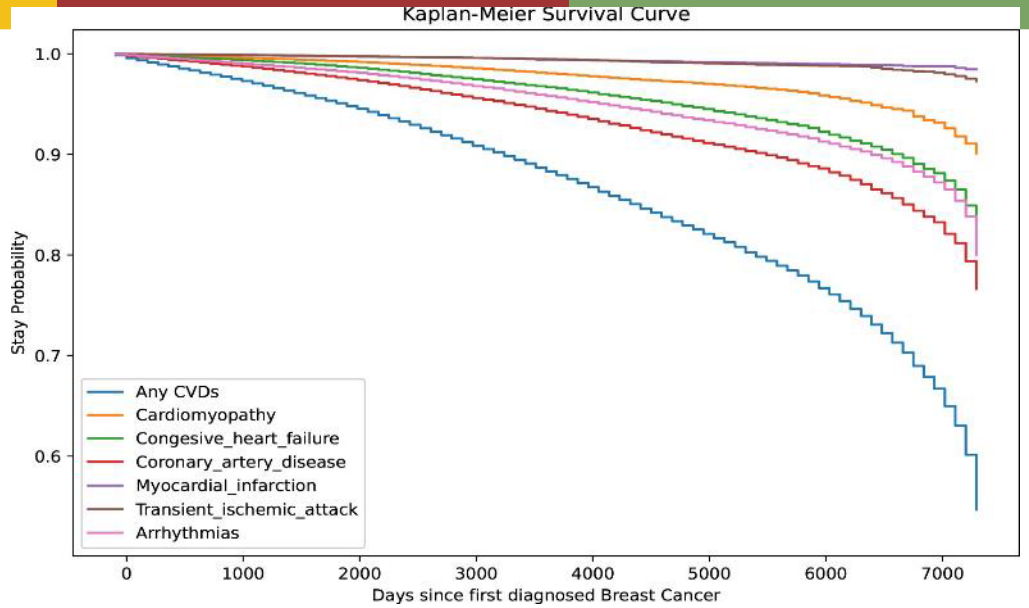


Table 2. The descriptive statistics of Breast Cancer cohort.

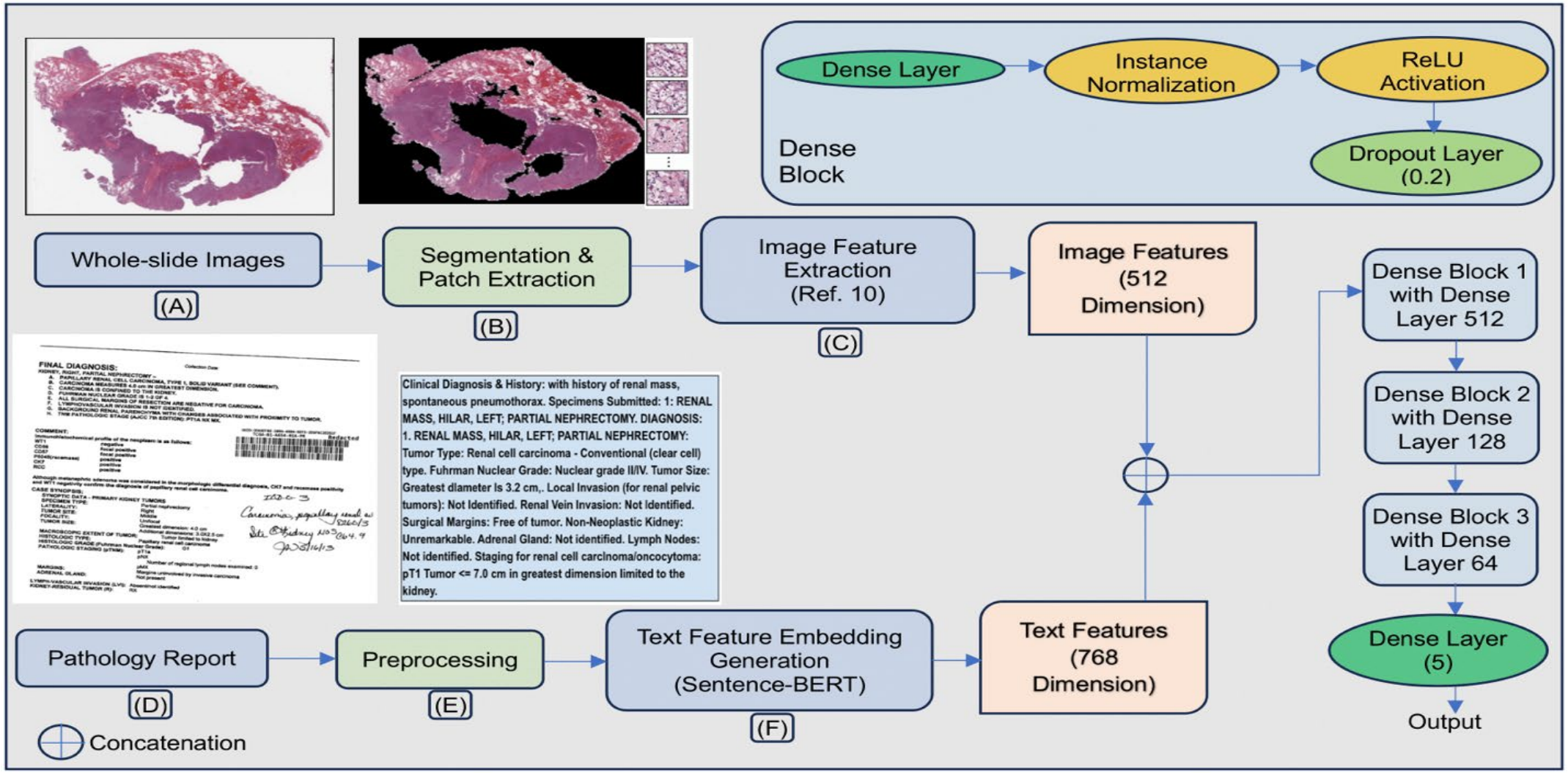
Variables	Total BC patients (n = 9377)	BC having CVDs (n = 3079)
Age (median) when first diagnosed BC	58	61
Sex at Birth = Female (%)	9062 (96.64)	2938 (95.42)
Race = White (%)	6368 (67.91)	2070 (67.23)
Race = Black or African American (%)	1182 (12.61)	462 (15.00)
Race = Asian (%)	225 (2.40)	45 (1.46)
Race = More than one population (%)	124 (1.32)	38 (1.23)
Ethnicity = Hispanic or Latino (%)	1192 (12.71)	351 (11.40)
BMI (median) when first diagnosed BC	39.9	39.41
Received radiotherapy after first diagnosed BC (%)	141 (47.32)*	52 (44.07)*
Received chemotherapy after first diagnosed BC (%)	345 (5.05)*	113 (4.25)*
Received targeted therapy after first diagnosed BC (%)	217 (3.18)*	61 (2.29)*
Lifestyle (Smoking=Yes)	3591 (38.30)	1347 (43.75)
Lifestyle (Alcohol=Yes)	8412 (89.71)	2729 (88.63)
CVD outcome, CM (%)	529 (5.64)	529 (17.18)
CVD outcome, CHF (%)	967 (10.31)	967 (31.41)
CVD outcome, CAD (%)	1758 (18.75)	1758 (57.10)
CVD outcome, MI (%)	343 (3.66)	343 (11.14)
CVD outcome, TIA (%)	299 (3.19)	299 (9.71)
CVD outcome, AR (%)	1484 (15.83)	1484 (48.20)

Asterisk (*) denotes the percentage is calculated after dropping the missing value.

Multimodal data fusion to improve cancer diagnosis

- Whole slide images (WSIs) are essential for cancer diagnosis but are challenging to analyze due to their gigapixel resolution.
- Automated AI solutions can expedite diagnosis, but precise annotations and model training remain difficult.
- Multi-modal approaches integrating WSIs with clinical data, such as pathology reports, show promise in improving diagnostic accuracy and reducing variability
- We developed the MPath-Net, a multimodal system integrating WSIs and pathology reports to enhance cancer subtype classification accuracy and reduce diagnostic variability

MPath-Net overview

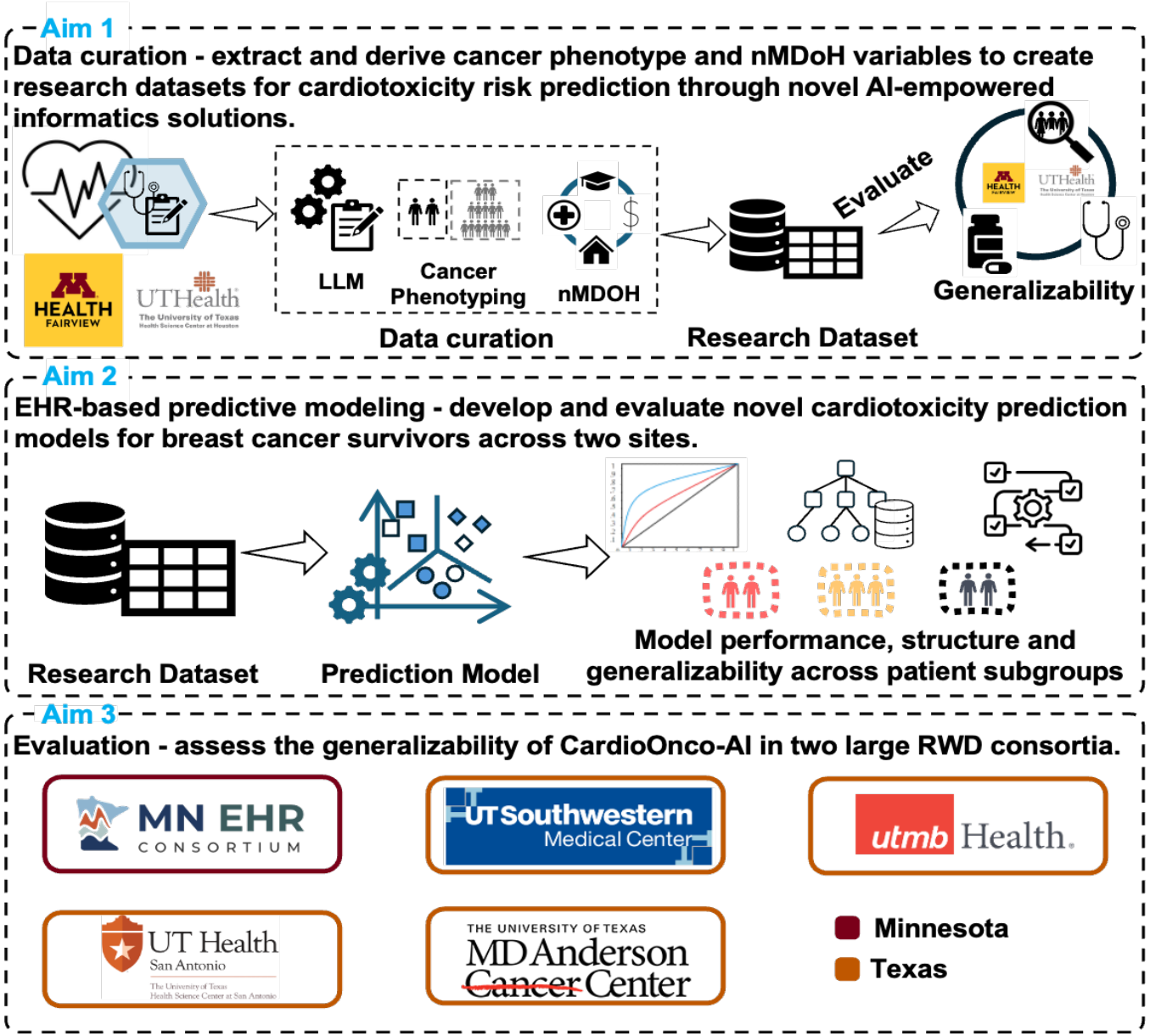


Block diagram of our proposed end-to-end multi-modal WSI classification model MPath-Net, (A) A typical WSI; (B) Segmentation and patch extraction from WSI; (C) Image feature extraction; (D) A typical pathology report in pdf format; (E) Preprocessing of pathology report; (F) Text feature embedding generation using Sentence-BERT model

Comparing with multi-instance learning models

Method	Accuracy	Precision	Recall	F1-score	AUC
TransMIL	0.9215 (0.9154, 0.9276)	0.9099 (0.9021,0.9177)	0.9158 (0.9096,0.9221)	0.9102 (0.9029,0.9174)	0.9863 (0.9822,0.9904)
ACMIL	0.9229 (0.9163, 0.9295)	0.9192 (0.9120,0.9264)	0.9231 (0.9162,0.9300)	0.9165 (0.9094,0.9236)	0.9885 (0.9845,0.9925)
ABMIL	0.9034 (0.8956, 0.9113)	0.9048 (0.8967,0.9129)	0.8998 (0.8918,0.9079)	0.8982 (0.8906,0.9059)	0.9900 (0.9860,0.9939)
MaxMIL	0.8866 (0.8785, 0.8948)	0.9266 (0.9212,0.9319)	0.8892 (0.8801,0.8982)	0.8990 (0.8917,0.9063)	0.9907 (0.9867,0.9947)
MeanMIL	0.8248 (0.8152, 0.8345)	0.8595 (0.8508,0.8681)	0.8279 (0.8185,0.8373)	0.8266 (0.8167,0.8365)	0.9810 (0.9768,0.9851)
MPath-Net_V1 (SentenceBERT)	**0.9487 (0.9443, 0.9531)**	**0.9495 (0.9441,0.9548)**	**0.9445 (0.9396,0.9495)**	**0.9460 (0.9406,0.9514)**	**0.9902 (0.9861,0.9942)**

CardioOnco-AI: AI-Empowered Cardiotoxicity Risk Prediction Among Breast Cancer Survivors Using Multi-Site Real-World Data



U01FD008720 (PI: Zhang/Liu)

Co-I: Blaes, Ramu, Melton, Drawz, Sun

To develop and validate a scalable, generalizable, and explainable AI-powered informatics framework, **CardioOnco-AI**, which innovates on the curation and modeling of real-world data (RWD) for individualized prediction of cardiotoxicity among breast cancer survivors.

Acknowledgements

Extramural Funding

NIH National Cancer Institute
1R01CA287413
(PI: Zhang/Sun/Cui)

NIH National Institute on Aging
1R01AG078154 (PI: Zhang/Xu)
3R01AT009457-04S1 (PI: Zhang)

NIH National Center for Complementary and Integrative Health
U01AT012797 (PI: Kilicoglu/Zhang/Tao)

NIH National Center for Complementary and Integrative Health
2R01AT009457 (PI: Zhang, renewal)
1R01AT009457 (PI: Zhang)

NIH National Institute on Minority Health and Health Disparities
1R21MD019134
(PI: Zhang/Hou/Wang)

NIH National Institutes of Health Office of Dietary Supplements
3R01AT009457-03S1
(PI: Zhang)

NIH National Institute of Diabetes and Digestive and Kidney Diseases
R01DK115629 (PI: Ku//Zhang/Dudley)

CDC Centers for Disease Control and Prevention
Midwest Disease Modeling and Analytics Preparedness Center

FDA
U01FD008720
(PI: Zhang/Liu)

Department of Veterans Affairs
NSF
Firearm and suicide prevention (NLP lead: Zhang)

Medtronic

Cisco
FL for BiomedNLP
(PI: Zhang/Suni)

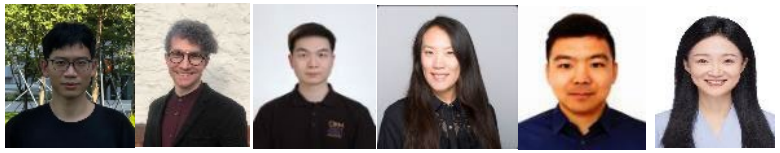
aws

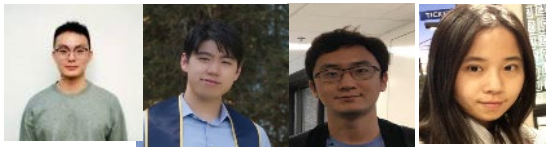
Minnesota Partnership for Biotechnology and Medical Genomics | University of Minnesota | Mayo Clinic

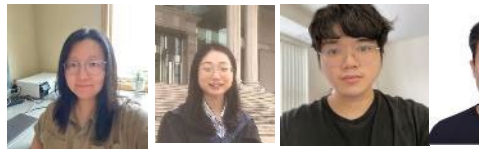
UMN NLP/IE program
(Director: Zhang; Associate Director: Melton)

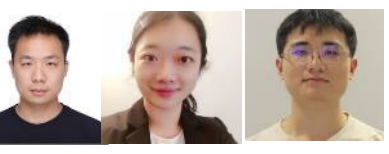
Alumni

Lab members









Email: ruizhang@umn.edu Lab: https://ruizhang.umn.edu/

Mayo D529 suit