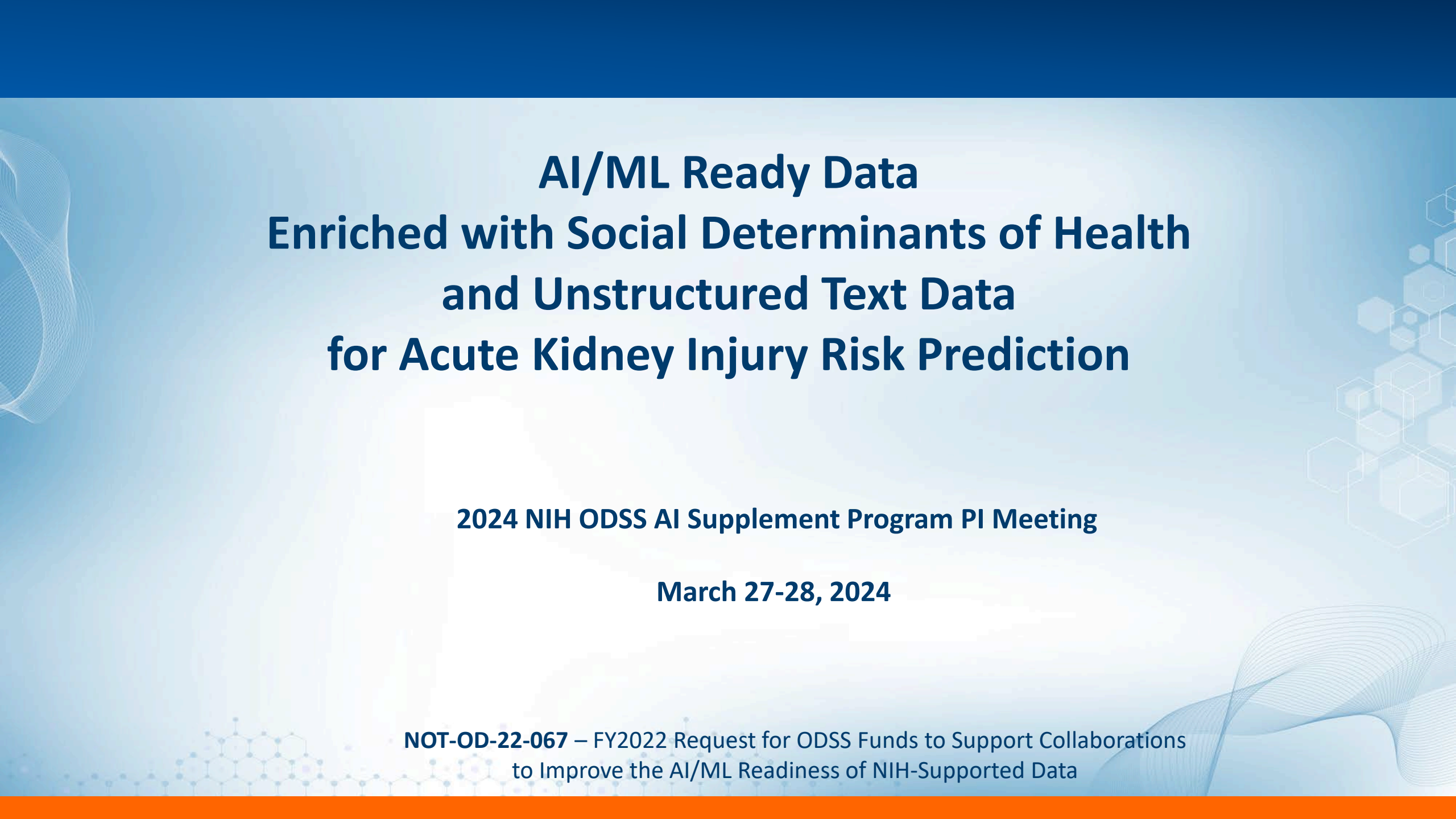


## **Breakout Session 8: Track A**

# **AI /ML Ready Data Enriched with Social Determinants of Health and Unstructured Text Data for Acute Kidney Injury Risk Prediction**

Dr. Tezcan Ozrazgat Baslanti

*Research Associate Professor, University of Florida*



# **AI/ML Ready Data Enriched with Social Determinants of Health and Unstructured Text Data for Acute Kidney Injury Risk Prediction**

**2024 NIH ODSS AI Supplement Program PI Meeting**

**March 27-28, 2024**

**NOT-OD-22-067** – FY2022 Request for ODSS Funds to Support Collaborations  
to Improve the AI/ML Readiness of NIH-Supported Data

## Background:

- Acute Kidney Injury (AKI) is a prevalent major health problem.<sup>1</sup>
- AKI is associated with an increased risk of poor short-term and long-term outcomes.<sup>2</sup>
- Drug-associated AKI (D-AKI) is the third to fifth leading cause of AKI.<sup>3,4</sup>

## Goal of Parent Project:

- To assess the effectiveness of a clinical surveillance system augmented with real-time predictive analytics to support a pharmacist-led intervention to reduce the progression and complications of drug-associated acute kidney injury.

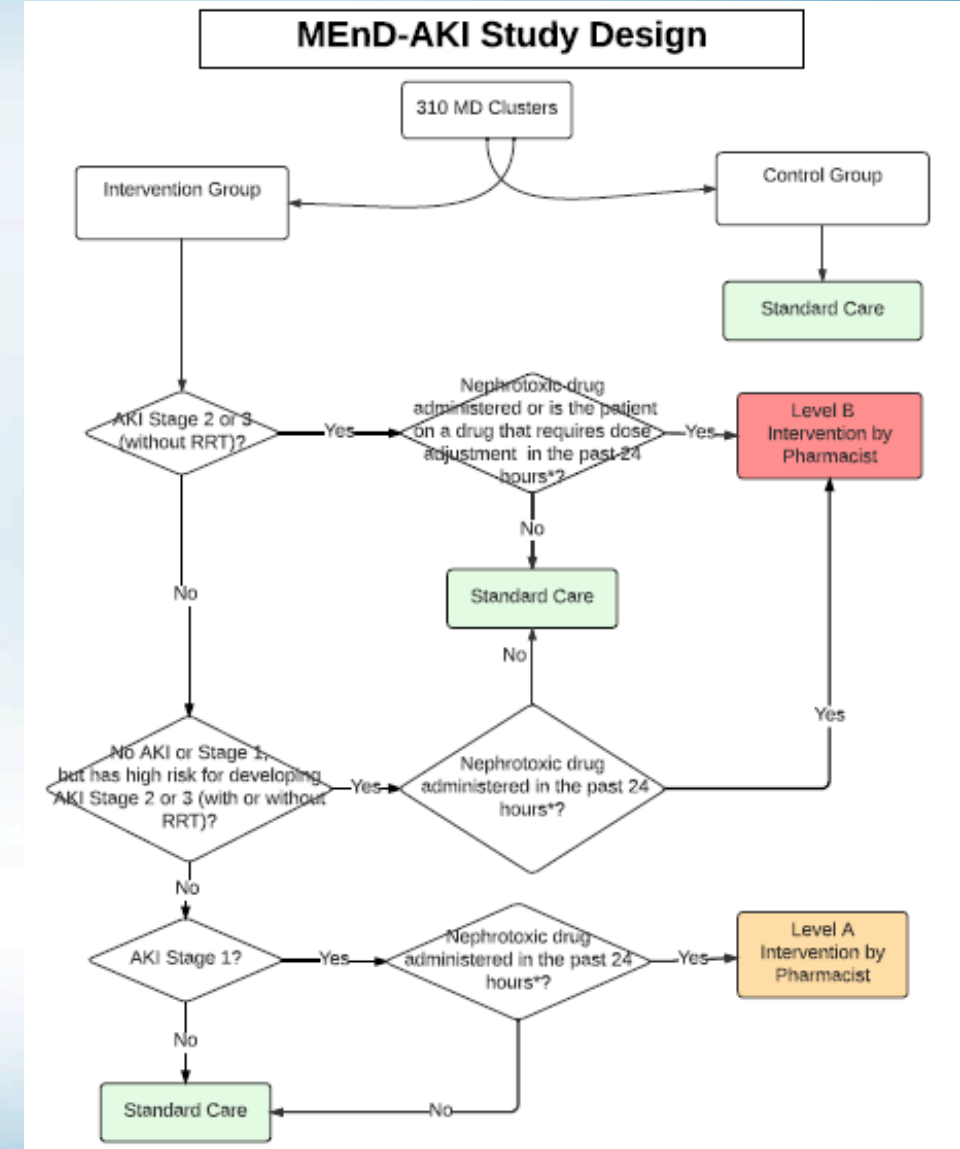


Figure 1. Study design

<sup>1</sup> Ronco C, et al. PMID: 31777389.

<sup>2</sup> Lameire N, et al. PMID: 23727171.

<sup>3</sup> Pazhayattil GS, et al. PMID: 25540591.

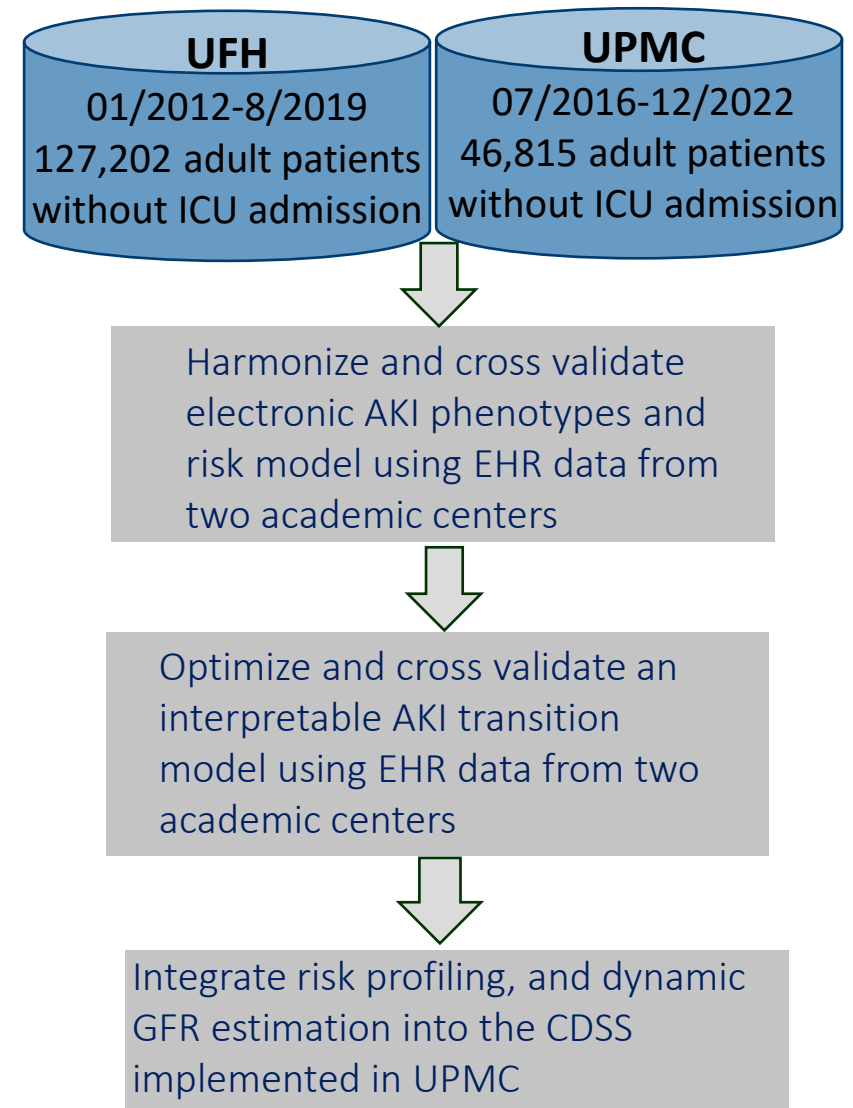
<sup>4</sup> Mehta R, et al. PMID: 25853333.

## Background:

- Acute Kidney Injury (AKI) is a prevalent major health problem.<sup>1</sup>
- AKI is associated with an increased risk of poor short-term and long-term outcomes.<sup>2</sup>
- Drug-associated AKI (D-AKI) is the third to fifth leading cause of AKI.<sup>3,4</sup>

## Goal of Parent Project:

- To assess the effectiveness of a clinical surveillance system augmented with real-time predictive analytics to support a pharmacist-led intervention to reduce the progression and complications of drug-associated acute kidney injury.



**Figure 2.** Aim 1 tasks

<sup>1</sup> Ronco C, et al. PMID: 31777389.

<sup>2</sup> Lameire N, et al. PMID: 23727171.

<sup>3</sup> Pazhayattil GS, et al. PMID: 25540591.

<sup>4</sup> Mehta R, et al. PMID: 25853333.

arXiv > cs > arXiv:2402.04209

Computer Science > Machine Learning

[Submitted on 6 Feb 2024]

## Acute kidney injury prediction for non-critical care patients: a retrospective external and internal validation study

Esra Adiyeye, Yuanfang Ren, Benjamin Shickel, Matthew M. Ruppert, Ziyuan Guan, Sandra L. Kane-Gill, Raghavan Murugan, Nabihah Amatullah, Britney A. Stottlmyer, Tiffany L. Tran, Dan Ricketts, Christopher M Horvat, Parisa Rashidi, Azra Bihorac, Tezcan Ozrazgat-Baslanti

**Background:** Acute kidney injury (AKI), the decline of kidney excretory function, occurs in up to 18% of hospitalized admissions. Progression of AKI may lead to irreversible kidney damage. **Methods:** This retrospective cohort study includes adult patients admitted to a non-intensive care unit at the University of Pittsburgh Medical Center (UPMC) (n = 46,815) and University of Florida Health (UFH) (n = 127,202). We developed and compared deep learning and conventional machine learning models to predict progression to Stage 2 or higher AKI within the next 48 hours. We trained local models for each site (UFH Model trained on UFH, UPMC Model trained on UPMC) and a separate model with a development cohort of patients from both sites (UFH-UPMC Model). We internally and externally validated the models on each site and performed subgroup analyses across sex and race. **Results:** Stage 2 or higher AKI occurred in 3% (n=3,257) and 8% (n=2,296) of UFH and UPMC patients, respectively. Area under the receiver operating curve values (AUROC) for the UFH test cohort ranged between 0.77 (UPMC Model) and 0.81 (UFH Model), while AUROC values ranged between 0.79 (UFH Model) and 0.83 (UPMC Model) for the UPMC test cohort. UFH-UPMC Model achieved an AUROC of 0.81 (95% confidence interval [CI] [0.80, 0.83]) for UFH and 0.82 (95% CI [0.81,0.84]) for UPMC test cohorts; an area under the precision recall curve values (AUPRC) of 0.6 (95% CI, [0.05, 0.06]) for UFH and 0.13 (95% CI, [0.11,0.15]) for UPMC test cohorts. Kinetic estimated glomerular filtration rate, nephrotoxic drug burden and blood urea nitrogen remained the top three features with the highest influence across the models and health centers. **Conclusion:** Locally developed models displayed marginally reduced discrimination when tested on another institution, while the top set of influencing features remained the same across the models and sites.

Subjects: **Machine Learning** (cs.LG); Artificial Intelligence (cs.AI)

Cite as: arXiv:2402.04209 [cs.LG]

(or arXiv:2402.04209v1 [cs.LG] for this version)

<https://doi.org/10.48550/arXiv.2402.04209>

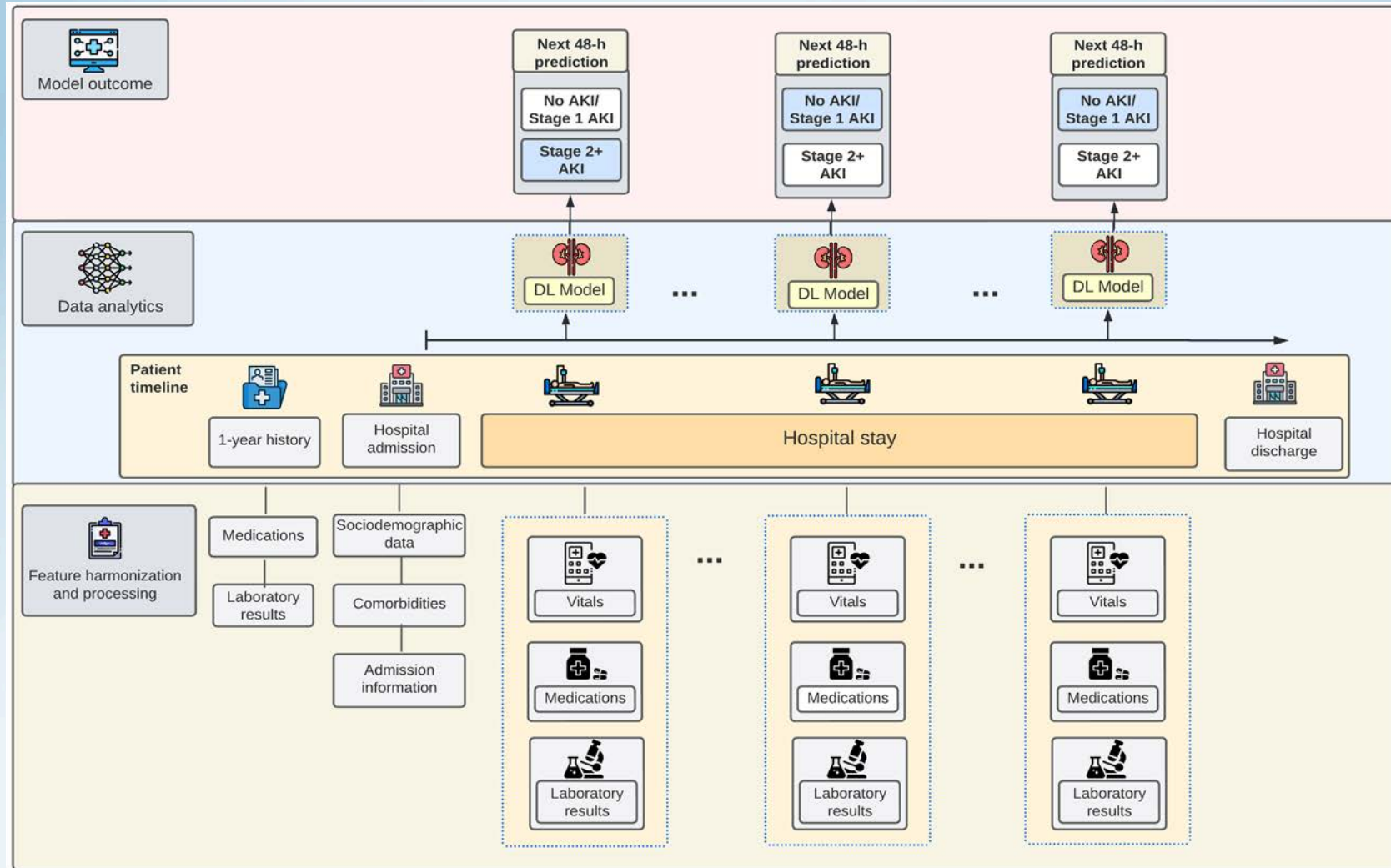


Figure 3. Overview of analytic framework

Table 1. Classification performance metrics for each model on test cohorts

	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	AUROC (95% CI)
<b>UFH Model</b>				
	0.81	0.66	0.81	0.81
UFH Test Cohort	(0.81, 0.81)	(0.63, 0.68)	(0.81, 0.81)	(0.79, 0.82)
	0.78	0.64	0.77	0.79
UPMC Test Cohort	(0.77, 0.78)	(0.62, 0.67)	(0.77, 0.78)	(0.78, 0.8)
<b>UPMC Model</b>				
	0.76	0.63	0.76	0.77
UFH Test Cohort	(0.76, 0.76)	(0.61, 0.66)	(0.76, 0.76)	(0.75, 0.78)
	0.82	0.69	0.82	0.83
UPMC Test Cohort	(0.82, 0.82)	(0.66, 0.71)	(0.82, 0.82)	(0.82, 0.84)
<b>UFH-UPMC Model</b>				
	0.77	0.72	0.77	0.81
UFH Test Cohort	(0.77, 0.77)	(0.7, 0.75)	(0.77, 0.77)	(0.8, 0.83)
	0.76	0.73	0.76	0.82
UPMC Test Cohort	(0.75, 0.76)	(0.71, 0.76)	(0.75, 0.76)	(0.81, 0.84)

CI: Confidence Interval; AUROC: area under the receiver operating characteristic curve. UFH Model was trained on UF development dataset, UPMC Model was trained on UPMC development dataset and UFH-UPMC Model was trained on combination of UFH and UPMC development datasets.

## Background

- Social determinants of health (SDOH) are important drivers of health inequities and disparities, and are responsible for between 30% and 50% of health outcomes.
- There is wealth of information contained in routine clinical notes that improves performance of prediction models for AKI.
- Our ongoing parent project is not yet utilizing SDOH and clinical notes that carry important information about patient health status and access to health care.
- **The proposed supplement project will develop integration, standardization, and processing tools and pipelines to create AI/ML ready data using SDOH and unstructured clinical text data.**



Figure 4. SDOH graphic.<sup>5</sup>

**Chief Complaint:**  
"swelling of tongue and difficulty breathing and swallowing"

**History of Present Illness:**  
77 y o woman in NAD with a h/o CAD, DM2, asthma and HTN on altace for 8 years awoke from sleep around 2:30 am this morning of a sore throat and swelling of tongue. She came immediately to the ED b/c she was having difficulty swallowing and some trouble breathing due to obstruction caused by the swelling. She has never had a similar reaction ever before and she did not have any associated SOB, chest pain, itching, or nausea. She has not noticed any rashes, and has been afebrile. She says that she feels like it is swollen down in her esophagus as well. In the ED she was given 25mg benadryl IV, 125mg solumedrol IV and pepcid 20 mg IV. This has helped the swelling some but her throat still hurts and it hurts to swallow. Nothing else was able to relieve the pain and nothing make it worse though she has not tried to drink any fluids because of trouble swallowing. She denies any recent travel, recent exposure to unusual plants or animals or other allergens. She has not started any new medications, has not used any new lotions or perfumes and has not eaten any unusual foods. Patient has not taken any of her oral medications today.

Figure 5. Example clinical note.



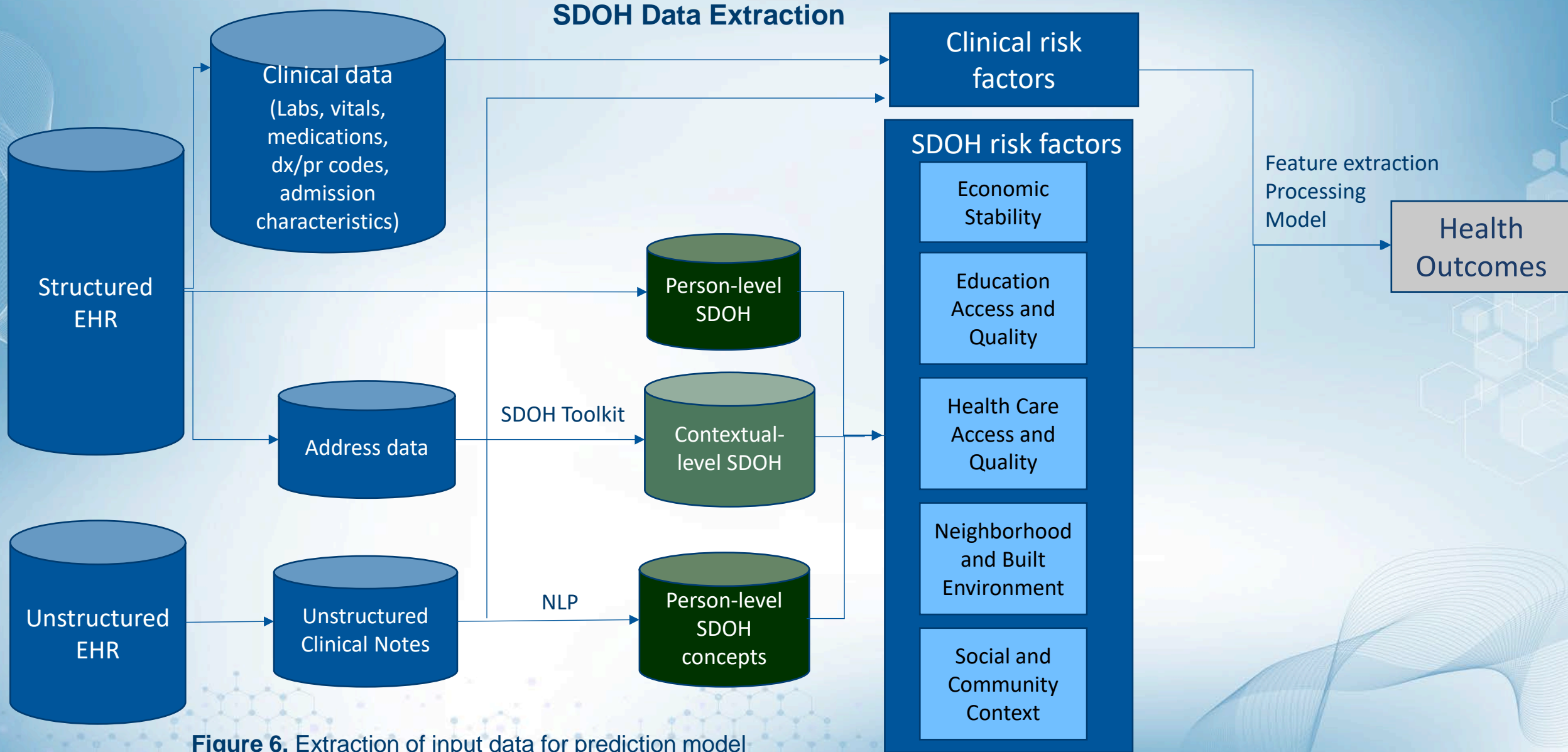


Figure 6. Extraction of input data for prediction model

## Aims of the Administrative Supplement Proposal

### **Aim 1: Preparation of AI/ML ready social determinants of health (SDOH) data.**

We will develop and assess tools for

- a) extracting, cleaning, imputing, preprocessing and representing data for various exposures contributing to a patient's SDOH exposome,
- b) integration of SDOH data to databases for them to be used in AKI risk model development and validation.

### **Aim 2: Preparation of multimodal AI/ML ready data that includes unstructured text data.**

We will develop and assess tools for

- a) extracting, cleaning, preprocessing and representing unstructured text data,
- b) integration of clinical data and unstructured text data to prepare multimodal AI/ML ready datasets at UF.

## SDOH Database

The SDOH Database gathers information from various sources across five main domains: economic stability, education access and quality, health care access and quality, neighborhood and built environment, and social and community context.

**Table 2.** SDOH datasets characters.

Dataset Name	Data Source Organization	Domains Covered	Numbers of Variables	Spatial scale	Temporal scale
<b>Agency for Healthcare Research and Quality (AHRQ)</b>	Agency for Healthcare Research and Quality (AHRQ)	Economic Context   Social Context   Healthcare Context   Physical Infrastructure   Education Context	329	Census Tract (FIPS-11)	1-year
<b>Neighborhood Atlas: Area Deprivation Index (ADI)</b>	Health Resources & Services Administration (HRSA)	Economic Context	2	Zip-9 and Census tract (FIPS-11)	1-year
<b>COVID-19 Reported Patient Impact and Hospital Capacity by Facility</b>	U.S. Department of Health & Human Services Hospital Utilization	Healthcare Context	128	County	Cross-sectional
<b>Socio-demographic: American Community Survey (ACS)</b>	United States Census Bureau	Social Context	56	Census Block Group (FIPS-12)	5-years
<b>Social Capital : County Business Patterns (CBP)</b>	United States Census Bureau	Economic Context	23	5-digit zip	1-year
<b>Crime and Safety</b>	The Uniform Crime Reporting (UCR)	Physical Infrastructure	84	County	1-year
<b>Food Environment Atlas</b>	US Department of Agriculture (USDA)	Economic Context	293	County	1-year
<b>Food Access: Food Access Research Atlas (FARA)</b>	US Department of Agriculture (USDA)	Economic Context	65	Census Tract (FIPS-11)	1-year
<b>National Walkability Index</b>	United States Environment Protection Agency (EPA)	Physical Infrastructure	114	Census Block Group (FIPS-12) / Census Tract (FIPS-11)	Cross sectional

## SDOH Linkage Tool:

We have developed a geographic information mapping tool to accurately link each patient's residential address to SDOH database, utilizing Federal Information Processing Standards codes and latitude and longitude coordinates, as appropriate.

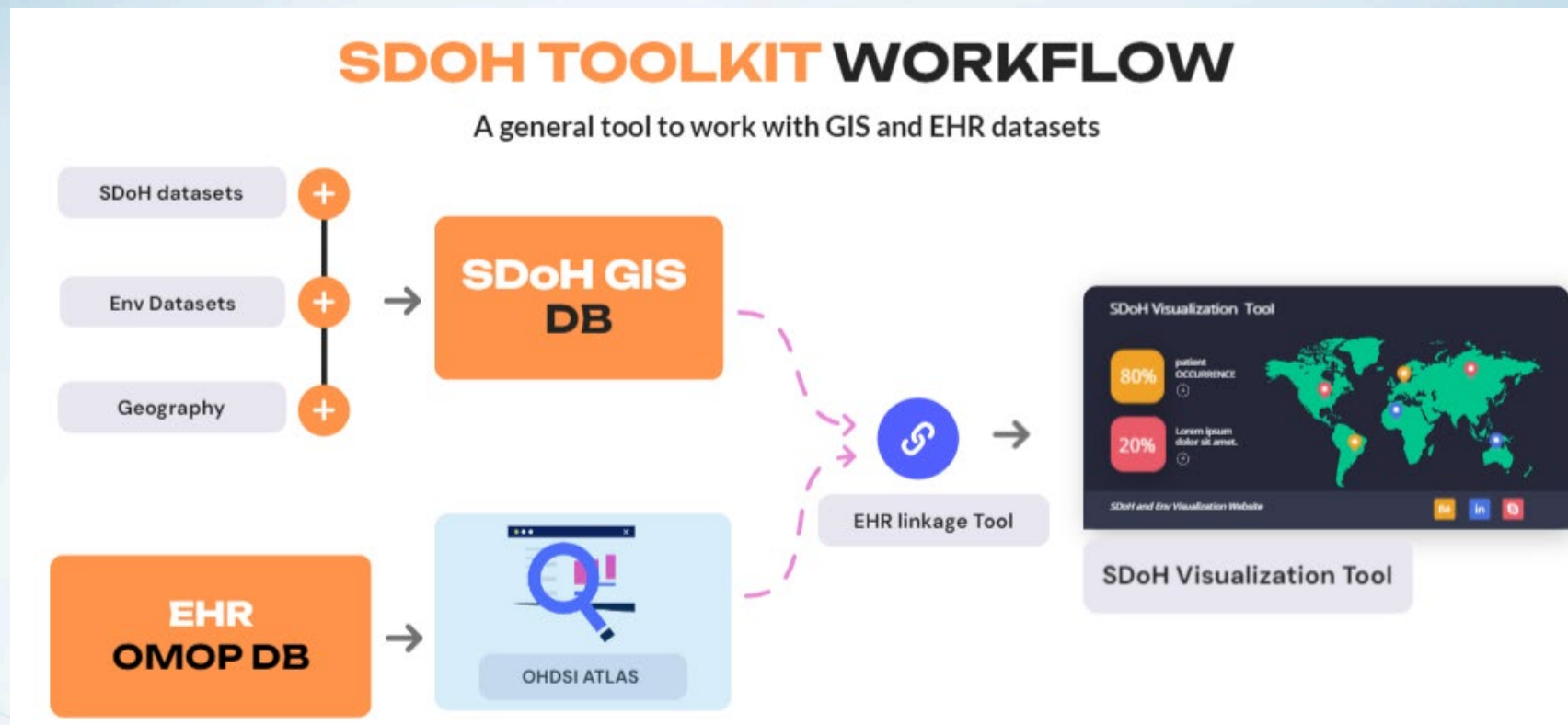
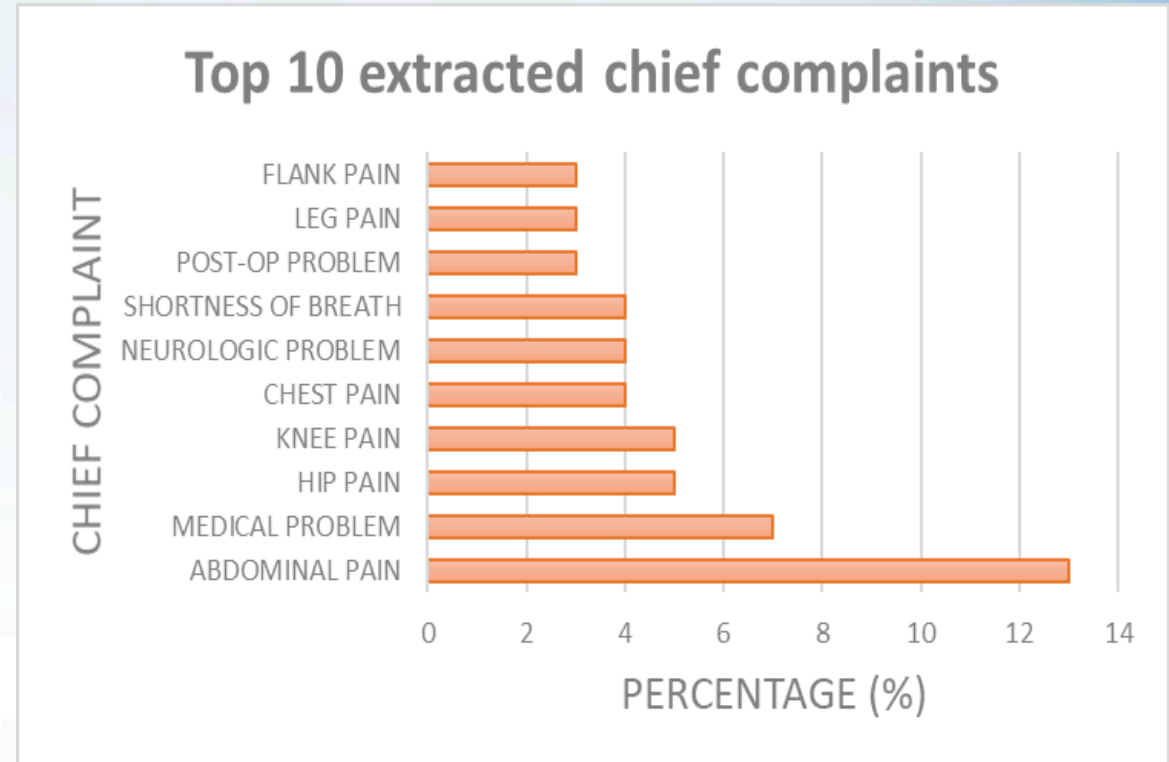
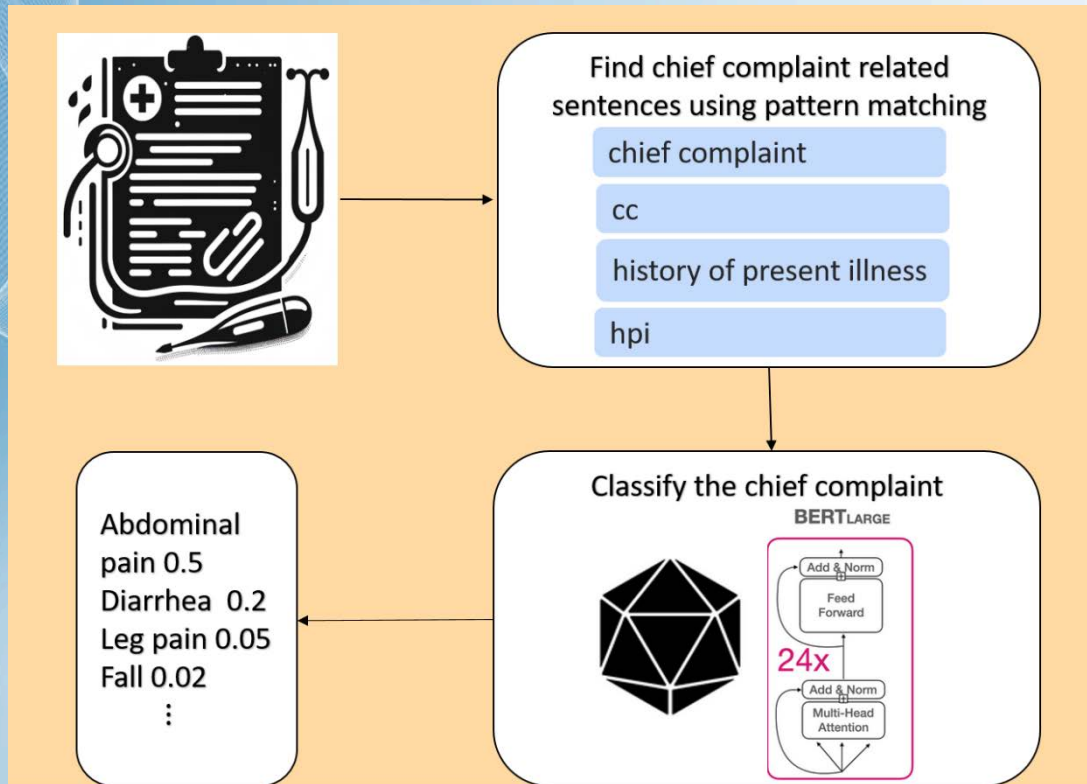


Figure 7. SDOH Toolkit

## Extracting medical concepts chief complaints and home medications

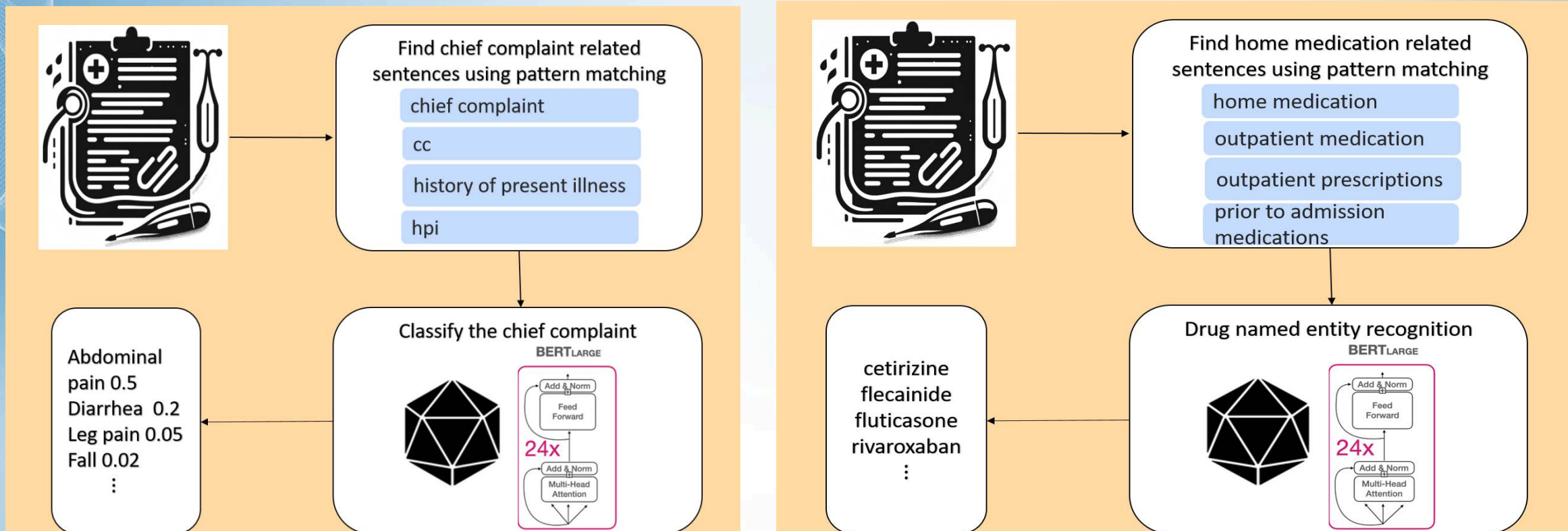
We have developed a pipeline to extract and standardize the chief complaints and home medications using pattern matching and a pre-trained BERT model.



**Figure 8.** Pipelines for extracting medical concepts for chief complaints (left) and home medications (right).

## Extracting medical concepts chief complaints and home medications

We have developed a pipeline to extract and standardize the chief complaints and home medications using pattern matching and a pre-trained BERT model.



**Figure 8.** Pipelines for extracting medical concepts for chief complaints (left) and home medications (right).

## Extracting SDOH: literature review of existing tools (\* Code is available)

- **2 studies using lexicon and regular expressions approach:** Bejan et al. 2018, Dorr et al. 2019.
  - Homelessness, adverse childhood experiences, chronic stress, social isolation, financial insecurity
- **5 studies using rule-based toolkit and platform (i.e., Moonstone, CLEVER) approach:** Oreskovic et al. 2017, Navathe et al. 2018, Conway et al. 2019, Bucher et al. 2019\*, Morrow et al. 2022\*
  - Anxiety, depression, tobacco use, alcohol abuse, drug abuse, housing instability, fall risk, poor social support, marital status, job instability, justice, social connections, detoxification, military sexual trauma, access to lethal means, food insecurity
- **4 studies using rule-based and paired with traditional ML and NLP system (i.e., CLAMP, cTAKES) approach:** Chilman et al. 2021, Rawat et al. 2022\*, Shah-Mohammadi et al. 2022, Hatef et al. 2021
  - Occupation, community and social context, economic stability, physical environment, health system, education, food, mental distress, social distress, legal distress, medical distress, family distress, housing issues, insurance status, neighborhood characteristics
- **6 studies using traditional supervised ML approach:** Feller et al. 2020, Ahsan et al. 2021\*, Kessler et al. 2023, Feller et al. 2018, Rouillard et al. 2022, Teng and Wilcox 2022
  - Alcohol use, sexual orientation, homelessness, substance use, sexual history, HIV status, drug use, housing status, transportation needs, housing insecurity, food insecurity, financial insecurity, employment/income insecurity, insurance insecurity, and poor social support
- **11 studies using deep learning, particular BERT approach:** Yu et al. 2022\*, Yu et al. 2021\*, Bashir et al. 2022, Mitra et al. 2021, Mitra et al. 2023, Richie et al. 2023, Zhao and Rios 2023, Newman-Griffis and Fosler-Lussier 2021, Han et al. 2022, Lituiev et al. 2022\*, Lybarger et al. 2023\*
  - Relationship status, social status, family history, employment status, race/ethnicity, gender, social history, sexual orientation, diet, alcohol, smoking, housing insecurity, social isolation, illicit drug use, violence, transition of care, food insecurity, physical activity, marital status, education, occupation, ethnicity, language, financial constraint

## Extracting SDOH: tool exploration and comparison of extraction

### Comparison of extraction rate from 24,652 H&P notes

Table 3. SDOH outcomes by Lituiev et al.2022.

Lituiev et al., 2022		
SDOH concept	# Concepts	Rate (%)
Marital or partnership status	14,937	61
Housing	9,470	38
Depression	6,538	27
Anxiety	5,817	24
Social isolation	5,587	23
Transportation	277	1.1
Insurance status	273	1.1
Food	0	0
Financial strain	0	0
Pain scores	0	0

Table 4. SDOH outcomes by SODA (Yu et al. 2022)

Yu et al., 2022, SODA		
SDOH concept	# Concepts	Rate (%)
Tobacco use	23,441	95
Alcohol use	23,346	95
Gender	22,886	93
Drug use	22,391	91
Sexual activity	20,178	82
Marital status	18,767	76
Occupation	14,253	58
Living supply	9,095	37
Partner	5,880	24
Living condition	3,825	16
Race	3,126	13
Abuse (physical or mental)	2,443	10
Education	1,569	6
Physical activity	1,252	5
Employment status	813	3
Transportation	180	0.7
Ethnicity	56	0.2
Language	43	0.2
Financial constraint	16	0
SDOH ICD	3	0
Social cohesion	2	0



## Adding SDOH and unstructured clinical notes into AKI risk prediction model

- **Outcome:** AKI stage 2 or higher in the first 96 hours of hospital admission. Prevalence of stage 2 or higher AKI was **3%**.
- **Patients:** 24,579 non-ICU adult patients admitted to UF Health between 2016 and 2017.  
Development: N=20,949; Validation: N=3,630.
- **Model:** XGBoost (Extreme Gradient Boosting) model.
- **Features:**
  - EHR data: demographics, comorbidities, medications and laboratory measurements in the past one year, reference creatinine, and eGFR (estimated glomerular filtration rate).
  - Contextual-level SDOH: 67 contextual-SDOH variables (eg. area deprivation index (ADI), social vulnerability index (SVI))
  - Chief complaints from clinical notes
  - Home medications from clinical notes: derived features include number of home medication, accumulated nephrotoxicity, and indicators of having specific drugs
  - SDOH from clinical notes: Housing, social isolation, depression, anxiety, and living supply.

### Results:

Models	ALL	FEMALE	MALE	AFRICAN-AMERICAN	NON AFRICAN-AMERICAN	AGE >=65	AGE<65
EHR	0.74 (0.69, 0.79)	0.77 (0.69, 0.83)	0.70 (0.61, 0.79)	0.74 (0.64, 0.83)	0.74 (0.67, 0.79)	0.71 (0.61, 0.79)	0.75 (0.67, 0.82)
EHR + contextual SDOH	0.71 (0.65, 0.76)	0.75 (0.67, 0.82)	0.66 (0.55, 0.75)	0.71 (0.61, 0.81)	0.71 (0.63, 0.77)	0.66 (0.55, 0.76)	0.73 (0.64, 0.8)
EHR + SDOH from clinical notes	0.73 (0.68, 0.79)	0.76 (0.67, 0.83)	0.70 (0.6, 0.79)	0.75 (0.65, 0.83)	0.73 (0.66, 0.79)	0.75 (0.68, 0.82)	0.75 (0.68, 0.82)
EHR + chief complaints	0.73 (0.68, 0.79)	0.76 (0.68, 0.83)	0.70 (0.6, 0.79)	0.76 (0.66, 0.86)	0.73 (0.66, 0.79)	0.72 (0.61, 0.81)	0.74 (0.66, 0.81)
EHR + home medications	0.74 (0.68, 0.79)	0.75 (0.67, 0.82)	0.71 (0.62, 0.8)	0.74 (0.65, 0.84)	0.73 (0.67, 0.79)	0.71 (0.6, 0.8)	0.75 (0.68, 0.81)

**Table 5.** Performance outcomes across subgroups and models.

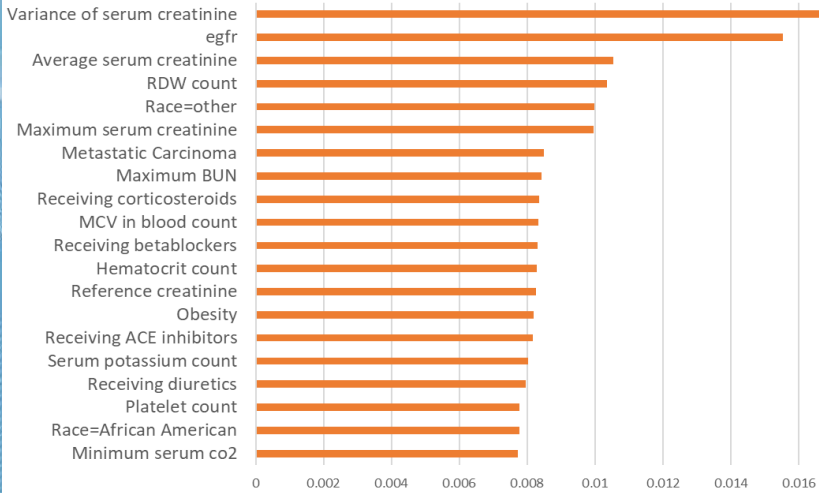


Figure 9. A) Feature importance for model with EHR only

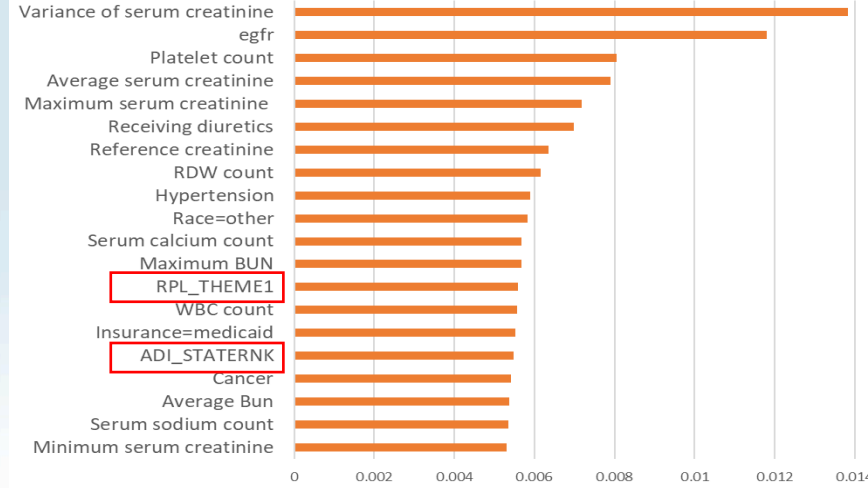


Figure 9. B) Feature importance for model with EHR and contextual SDOH

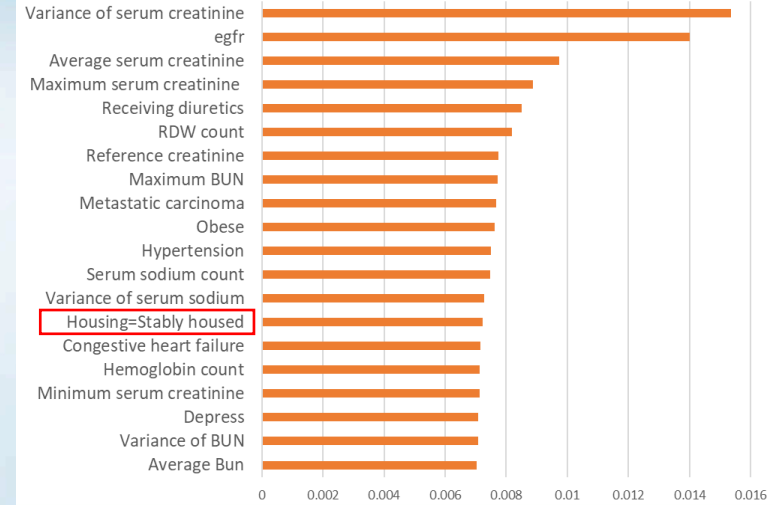


Figure 9. C) Feature importance for model with EHR and SDOH from clinical notes

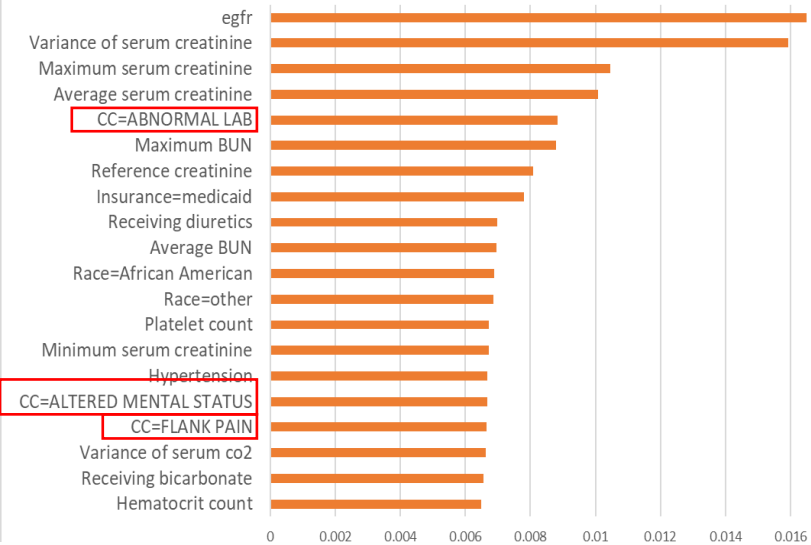


Figure 9. D) Feature importance for model with EHR and chief complaints

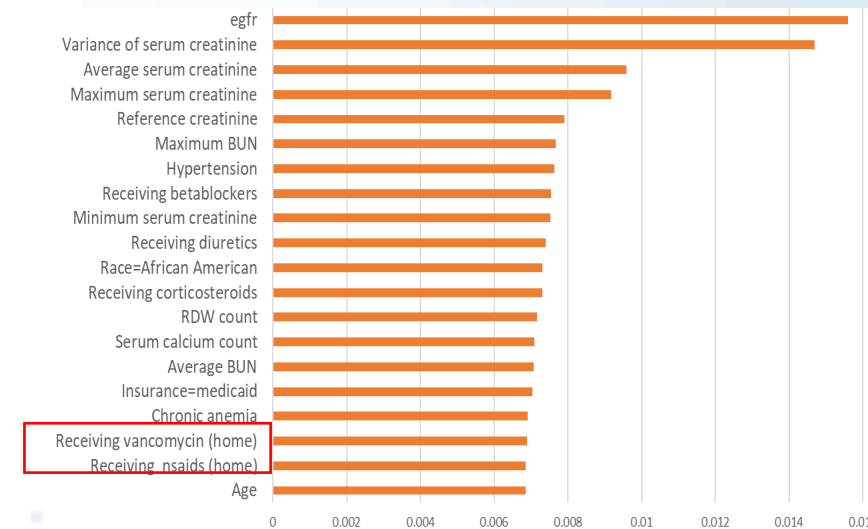


Figure 9. E) Feature importance for model with EHR and home medications

**Abbreviations.** eGFR, estimated glomerular filtration rate  
BUN, blood urea nitrogen  
MCV; mean corpuscular volume  
ACE: Angiotensin-converting enzyme

RPL\_THEME 1: Socioeconomic Status (sub-index of Social Vulnerability Index)  
ADI\_STATERNK: Area Deprivation Index State Ranking

## Challenges

- Difficulties with accurately assigning geographic coordinates to patient addresses due to errors and missingness
- Lack of gold standard terminologies and ontology for encoding SDoH
- Generalizability of the pre-trained models; lack of validation
- Lack of public annotated datasets and well-maintained software/code

## Future Plans

- Development of a data catalog and visualization tool to aid researchers in understanding and visualizing the data within a geographic context.
- Expedition of the annotation process of SDoH dataset
- Training of our model on our clinical notes leveraging the large language model technique
- Standardization of the extracted SDoH concepts and application to our AKI risk model

## Publications/Posters

1. Adiyeye E, Ren Y, Shickel B, Ruppert MM, Guan Z, Kane-Gill SL, Murugan R, Amatullah N, Stottlemeyer BA, Tran TL, Ricketts D, Horvat CM, Rashidi P, Bihorac A, Ozrazgat-Baslanti T. Acute kidney injury prediction for non-critical care patients: a retrospective external and internal validation study. arXiv preprint [arXiv:2402.04209](https://arxiv.org/abs/2402.04209). 2024 Feb 6.
2. Amatullah N, Stottlemeyer BA, Zerfas I, Stevens C, Ozrazgat-Baslanti T, Bihorac A, Kane-Gill SL. Challenges in Pharmacovigilance: Variability in the Criteria for Determining Drug-Associated Acute Kidney Injury in Retrospective, Observational Studies. *Nephron*. 2023;147(12):725-732. doi: 10.1159/000531916. Epub 2023 Aug 23. PMID: 37607496; PMCID: PMC10776175. DOI: [10.1159/000531916](https://doi.org/10.1159/000531916)
3. Adiyeye E, Ren Y, Ruppert MM, Shickel B, Kane-Gill SL, Murugan R, Rashidi P, Bihorac A, Ozrazgat-Baslanti T. A deep learning-based dynamic model for predicting acute kidney injury risk severity in postoperative patients. *Surgery*. 2023 Sep;174(3):709-714. doi: 10.1016/j.surg.2023.05.003. Epub 2023 Jun 13. PMID: 37316372; PMCID: PMC10683578. DOI: [10.1016/j.surg.2023.05.003](https://doi.org/10.1016/j.surg.2023.05.003)
4. Johnson C, Adiyeye E, Sai A, Guan Z, Ren Y, Bihorac A, Ozrazgat-Baslanti T. Impact of Social Determinants of Health on Perioperative Acute Kidney Injury – A Means to Improving AKI Prediction. University of Florida, College of Medicine Celebration of Research Day, Gainesville, FL. February 12, 2024. (Poster)
5. Vedala AS, Adiyeye E, Bible L, McKean J, Mohr A, Ozrazgat-Baslanti T, Bihorac A. . Impact of Social Determinants of Health on Post-operative Outcomes. University of Florida, College of Medicine Celebration of Research Day, Gainesville, FL. April, 2023. (Poster)
6. Adiyeye E, Fleeting C, Davidson A, Vedala SA, Johnson C, Uddin R, Newsom M, Bihorac A, Ozrazgat-Baslanti T. Review of association of contextual social determinants of health and critical care outcomes (in preparation, *Critical Care Medicine*)

# Thank you!

## Funding Support:

- R01 DK121730 NIH/NIDDK (University of Pittsburgh Subaward)
- ODSS Funds to Support Collaborations to Improve the AI/ML Readiness of NIH-Supported Data



Azra Bihorac, MD MS

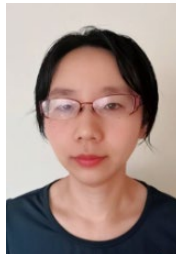


Sandra Kane-Gill, PharmD MS

## Administrative Supplement:



Tezcan Ozrazgat-Baslanti, PhD



Yuanfang Ren, PhD



Benjamin Shickel, PhD

## Acknowledgements:

Esra Adiyeye, PhD  
Sai Annanya Sree Vedala

**PRISMA<sup>P</sup>**

**UF** | Intelligent Clinical Care Center (IC<sub>3</sub>)  
UNIVERSITY of FLORIDA

<https://ic3.center.ufl.edu/>