



Rensselaer

why not change the world?®



IBM

leidos

AN ONTOLOGY-ENABLED APPROACH FOR USER-CENTERED AND KNOWLEDGE-ENABLED EXPLANATIONS OF AI SYSTEMS

Shruthi Chari| PhD Candidate, RPI| 6/27/2024

Committee: Prof. Deborah L. McGuinness (Advisor), Prof. Oshani Seneviratne (Co-Advisor), Prof. James A. Hendler, Dr. Pablo Meyer, Dr. Prithwish Chakraborty

Portions of this research were partially supported by IBM Research AI through the AI Horizons Network.

Portions of this research were supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

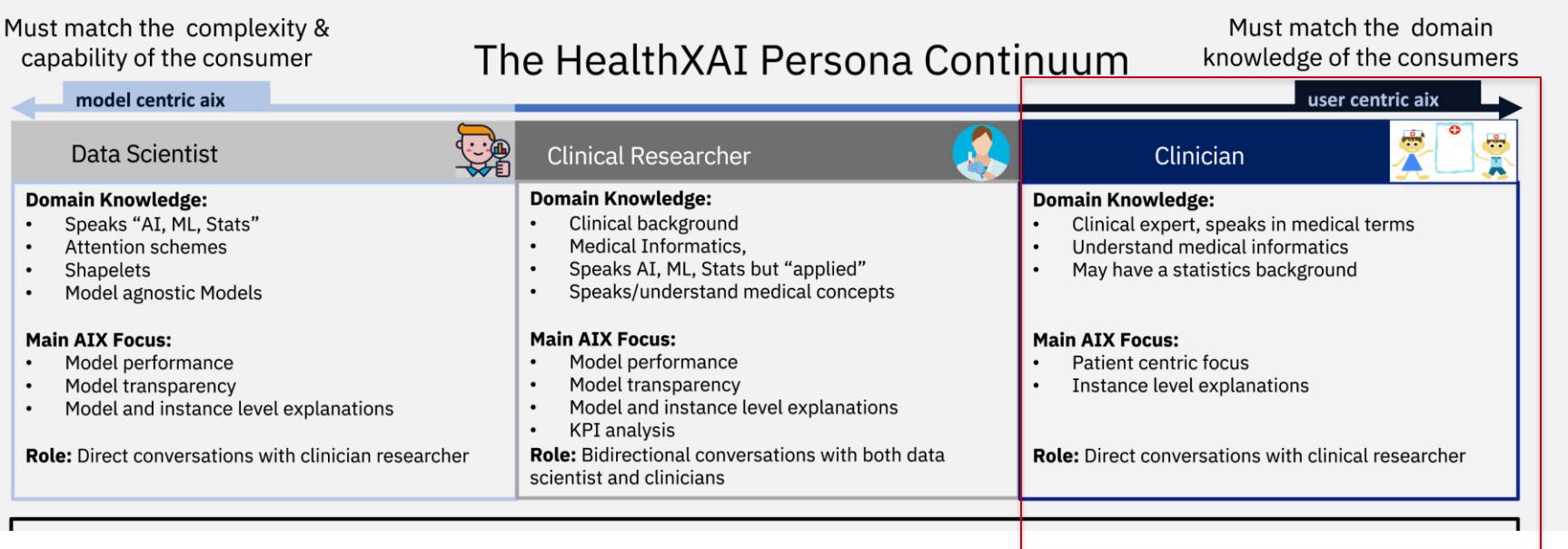
Motivation

- Current era: Several **black-box machine learning methods** that have a lot of **model explanation approaches**.
- Why are explanations important? [2, 3]
 - Trust, Education and Debug
- What are the gaps and opportunities?
 - Position statements emphasize a **need for user-centered explainability** [1, 2, 3, 4]
 - **Users are diverse** and hence require **different explanations** to trust AI systems [1, 4]
 - Lack of a **unified framework** to generate different explanation types in various use cases and that leverage **knowledge and context** as dimensions.

1. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.
2. Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288).
3. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.
4. Dey, S., Chakraborty, P., Kwon, B. C., Dhurandhar, A., Ghalwash, M., Saiz, F. J. S., ... & Meyer, P. (2022). Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, 3(5), 100493.



End-user spectrum [1]



1. Dey, S., Chakraborty, P., Kwon, B. C., Dhurandhar, A., Ghalwash, M., Saiz, F. J. S., ... & Meyer, P. (2022). Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, 3(5), 100493.

Definition of Explanations We Proposed and Adhere to

Explanations are:

``an account of the system, its workings, *implicit and explicit knowledge* used in its reasoning processes and the *specific decision*, that is *sensitive* to the *end-user's understanding, context, and current needs.*” [1]

1. Chari S., Seneviratne O., Gruen DM., McGuinness DL. “Foundations of Explainable Knowledge-Enabled Systems.” In Ilaria Tiddi, Freddy Lecue, Pascal Hitzler (eds.), Knowledge Graphs for eXplainable AI -- Foundations, Applications and Challenges. Studies on the Semantic Web, pp 23 - 48; 2020

Research Questions around Explainable AI (XAI)

1. How can we **formally represent explanations** with support for **interacting AI systems, additional data sources, and along different dimensions?**
 - Design an ***Explanation Ontology (EO)***, a general-purpose semantic representation that can represent fifteen different literature-derived explanation types via their system-, interface- and user- related components.
2. How **useful and feasible** are user-centered explanations for **clinical settings?**
 - Design and implement a ***clinical question-answering (QA) system to provide contextual explanations***, to address questions from clinical practice guidelines to help clinicians interpret risk prediction scores and their post-hoc explanations in a comorbidity risk prediction setting.
3. Is it feasible to **combine explanations** from **multiple data sources** and **AI methods?**
 - Design a general-purpose framework, ***MetaExplainer***, capable of providing ***multiple explanations from various perspectives*** to an end-user question in various use cases.



Rensselaer IBM

Explanation Ontology

Chari, S., Seneviratne, O., Ghalwash, M., Shirai, S., Gruen, D. M., Meyer, P., ... & McGuinness, D. L. Explanation Ontology: A general-purpose, semantic representation for supporting user-centered explanations. *Semantic Web*, (2023 - Preprint), 1-31.

Chari, S., Seneviratne, O., Gruen, D. M., Foreman, M. A., Das, A. K., & McGuinness, D. L. (2020, November). Explanation ontology: a model of explanations for user-centered AI. In *International Semantic Web Conference* (pp. 228-243). Cham: Springer International Publishing.

Key Points

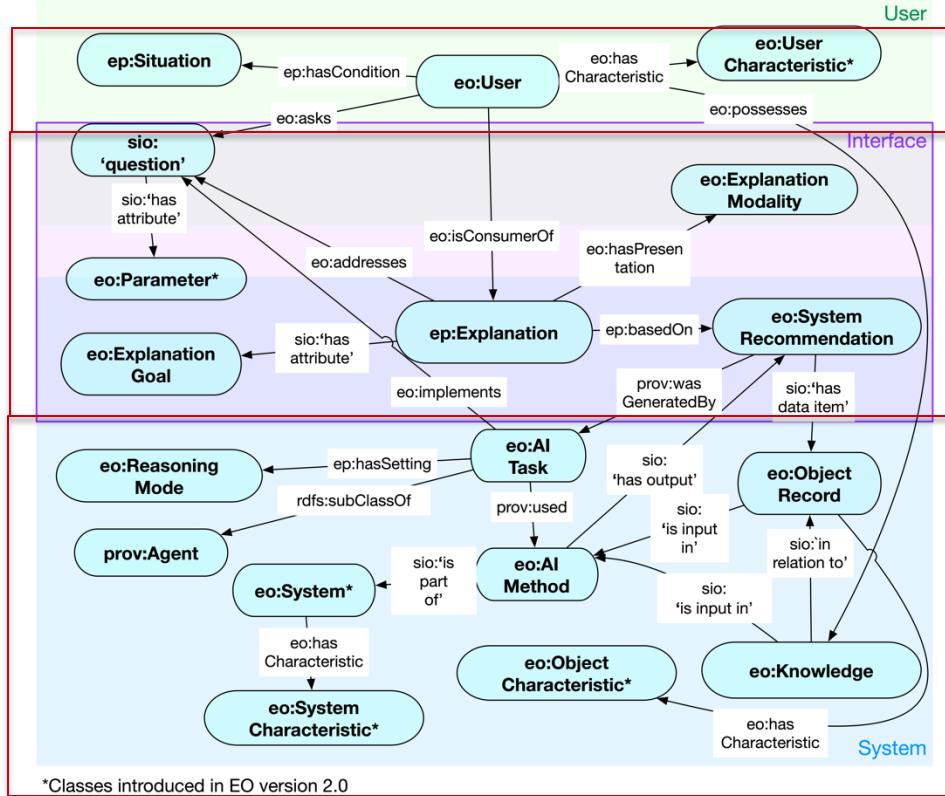
Addresses the Q1: How can we **formally represent explanations** with support for interacting AI systems, additional data sources, and along different dimensions?

We have designed the Explanation Ontology, to represent user-centered explanations from their supporting components.

In the EO:

- We capture system-, interface- and user- dependencies of explanations.
- We currently represent fifteen user-centered and literature-derived explanation types.
- We have represented explanations using EO in a few exemplar use cases.

Concept Map of the Explanation Ontology (EO)



Key

Class



Connecting property

Prefixes

eo	Explanation Ontology
ep	Explanation Patterns Ontology
sio	SemanticScience Integrated Ontology
prov	Prov-O

15 Literature-Derived Explanation Types

Explanation Type	Explanation Type
Case Based	Impact*
Contextual	Rationale*
Contrastive	Responsibility*
Counterfactual	Safety and Performance*
Data*	Scientific
Everyday	Simulation Based
Fairness*	Statistical
	Trace Based

Preferred by Clinicians [1]

Within the EO, each explanation type has a:

- Definition
- Equivalence
- Condition
- Prototypical Question they Address

Preferred by Clinicians [1]

*Introduced from Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 593.

1. Gruen, D. M., Chari, S., Foreman, M. A., Seneviratne, O., Richesson, R., Das, A. K., & McGuinness, D. L. (2021). Designing for ai explainability in clinical context. AAAI Workshop on Trustworthy AI.

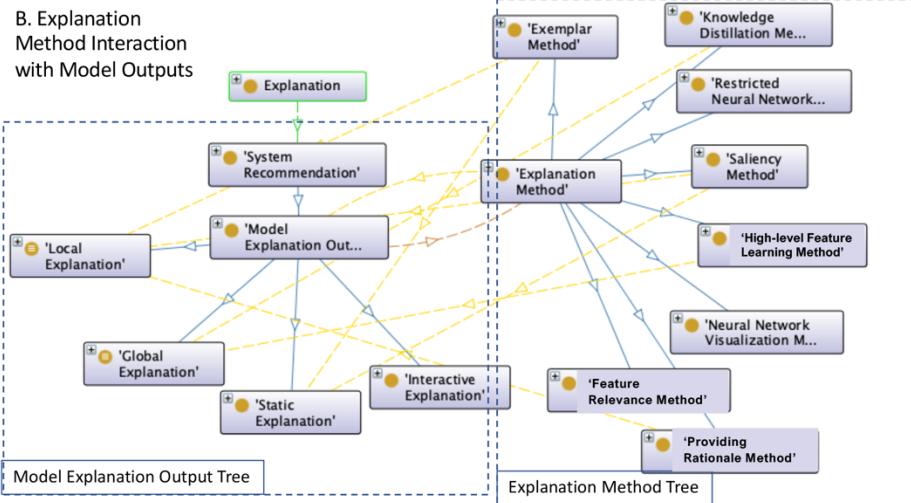
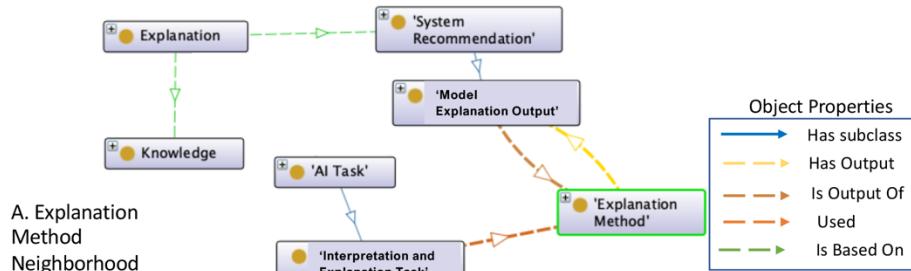


Use Cases

Use Case	Example	Explanation Type Inferred
Drug Recommendation	Why Drug B over Drug A?	Contrastive
Food Recommendation	Why should I eat spiced cauliflower soup? Why creamed broccoli soup over tomato soup?	Contextual and Contrastive
Proactive Retention*	What is the retention action outcome for this employee?	Rationale
Health Survey Analysis*	Who are the most representative patients in this questionnaire? Which questionnaires have the highest number of most representative patients?	Case Based and Contextual
Medical Expenditure*	What are the rules for expenditure prediction? What are patterns for high-cost patients?	Data
Credit Approval*	What are the rules for credit approval? What are some representative customers for credit? What factors if present and if absent contribute most to credit approval?	Data, Case Based and Contrastive

*AIX 360 use cases: <https://aix360.mybluemix.net>

EO V2 – Support for more Explainer Methods; V3 – Support for Explanation Modalities and Metrics

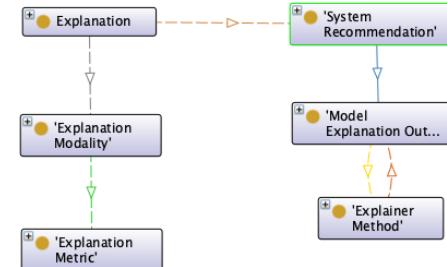


C. Examples of User-centric Explanation Dependencies on Model Explanations

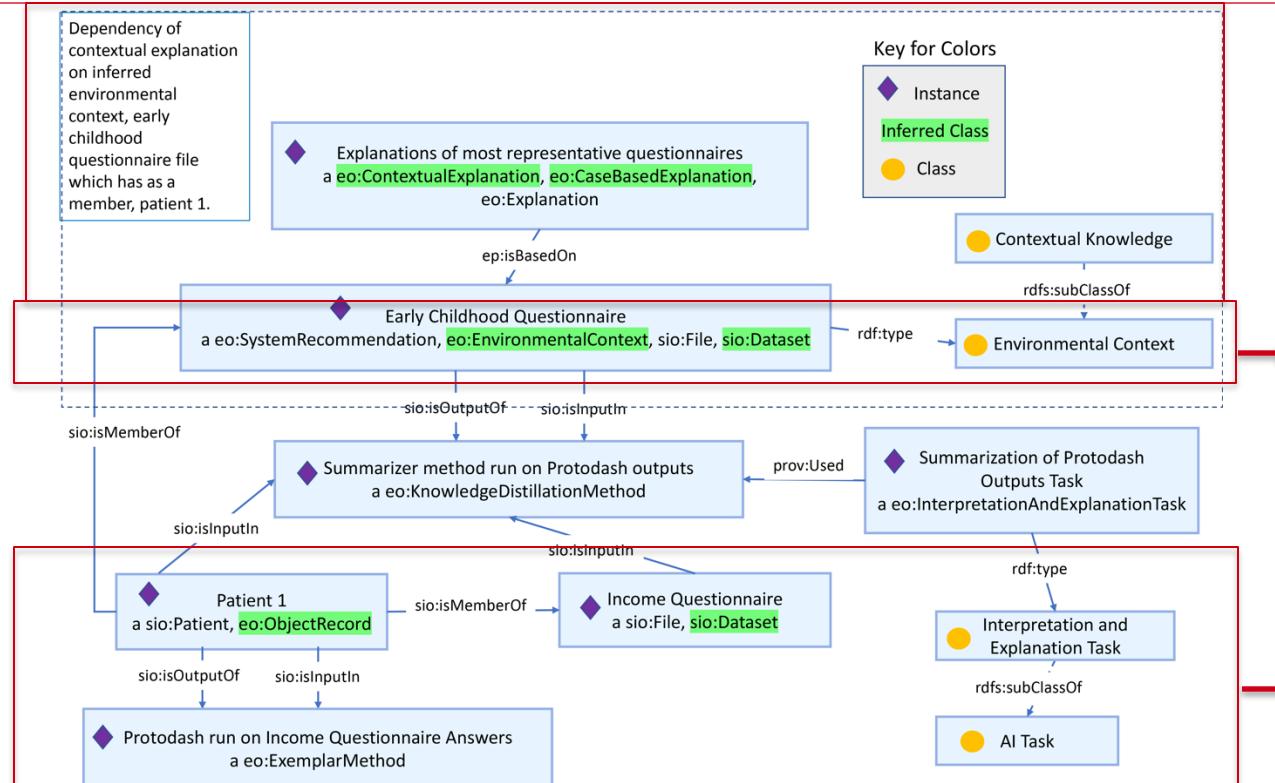
Class: 'Case based explanation'
EquivalentTo:
 'is based on' some
 ('Model Explanation Output' **and**
 ('is output of' **some**
 ('Exemplar Method' **or** 'High level feature learning
 method' **or** 'Feature Relevance Method')))

Class: 'Data explanation'
EquivalentTo:
 'is based on' some
 ('Model Explanation Output' **and**
 ('is output of' **some**
 ('Exemplar Method' **or** 'High level feature learning
 method' **or** 'Feature Relevance Method')))

D. Support for Explanation Modalities and Metrics in V3



EO KG in Health Survey Analysis Use Case



Model Explanation Outputs from [1]

Questionnaire	Prototypes representative of Income
0	Early Childhood
1	Physical Functioning
2	Acculturation
3	Disability
4	Physical Activity
5	Smoking - Secondhand Smoke Exposure
6	Cognitive Functioning
7	Sleep Disorders
8	Diabetes

	8	132	690	1475	2449	2912	3899
Respondent sequence number	73565.00	73689.00	74247.00	75032.00	76006.00	76469.00	77456.00
Income from wages/salaries	1.00	1.00	2.00	1.00	1.00	1.00	1.00
Income from self employment	2.00	2.00	2.00	2.00	1.00	2.00	2.00
Income from Social Security or RR	2.00	2.00	2.00	2.00	2.00	2.00	1.00
Income from other disability pension	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Income from retirement/survivor pension	2.00	2.00	2.00	2.00	2.00	1.00	2.00
Income from Supplemental Security Income	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Income from state/county cash assistance	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Income from interest/dividends or rental	2.00	1.00	2.00	1.00	2.00	2.00	2.00

1. <https://nbviewer.org/github/IBM/AIX360/blob/master/examples/tutorials/CDC.ipynb#study1>



Evaluation Strategies [1]

1. Evolution-based – *what are the gain in edits from V1 to V2 of EO?*
2. Task-based - *what can the intended user do with EO?*
3. Application-based – *what do EO's representations enable?*

1. Amith, M., He, Z., Bian, J., Lossio-Ventura, J.A., Tao, C.: Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *Journal of biomedical informatics* 80, 1–13 (2018)

Access Links for EO

Website: <https://tetherless-world.github.io/explanation-ontology/index>

Github: <https://github.com/tetherless-world/explanation-ontology>





Contextualizing Model Explanations via a Knowledge-Augmented Question-Answering (QA) Method

Chari, S., Acharya, P., Gruen, D. M., Zhang, O., Eyigoz, E. K., Ghalwash, M., ... & McGuinness, D. L. (2023). Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes. *Artificial Intelligence in Medicine*, 137, 102498.

Key Points

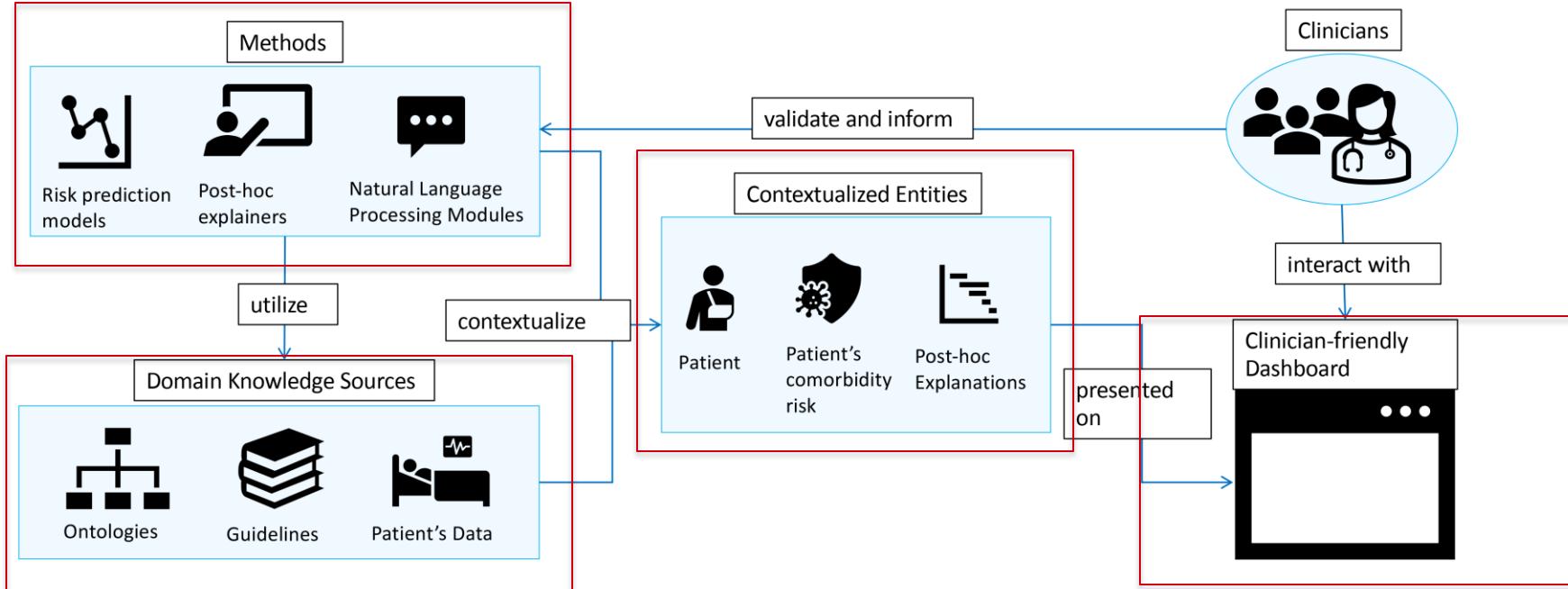
1.Addresses the Q2: How **useful and feasible** are user-centered explanations for **clinical settings**?

We have designed a method to generate explanations to contextualize ML risk predictions and feature importances.

Our method is:

- An end-end pipeline to **explain risk predictions** and populate these explanations on a dashboard.
- We setup the explanation generation task as a **question-answering method** to extract relevant sentences from medical guidelines.
- We assess the feasibility and usefulness of providing such explanations both from **qualitative as well as quantitative** standpoints.

Contextualizing Entities of Interest in a Comorbidity Risk Prediction Setting [Inspired by need for context – 1, 2]

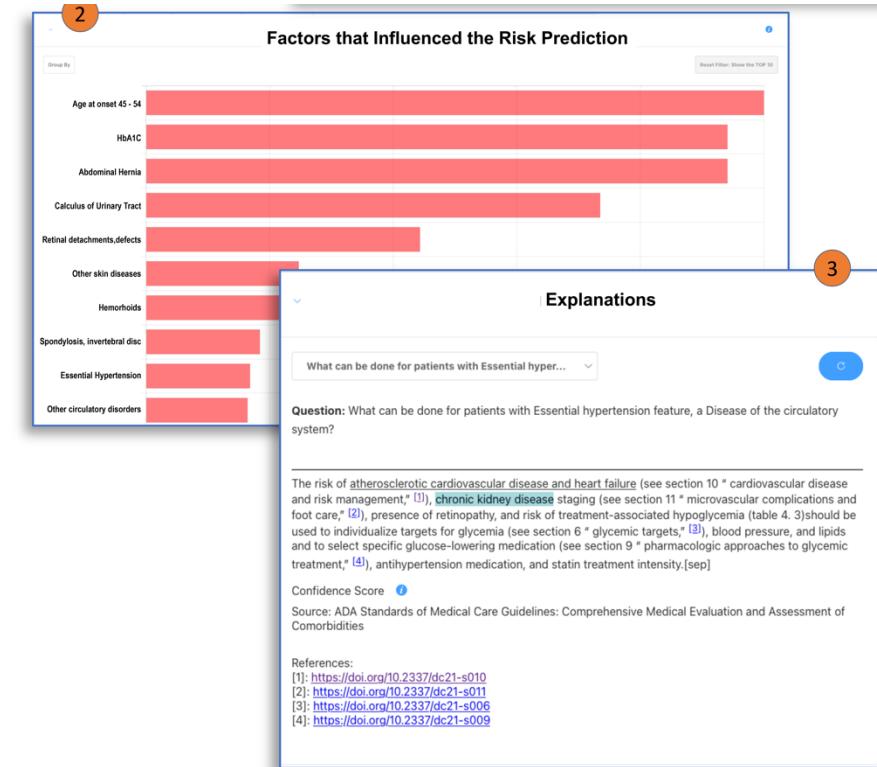
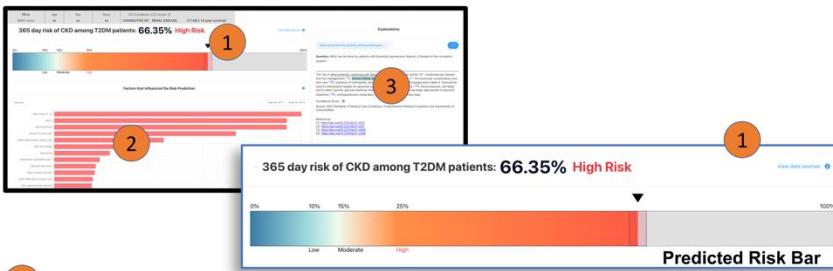


1. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750.
2. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019, October). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference* (pp. 359-380). PMLR



Displaying the QA Contextual Explanations Alongside Predictions and Patient Data

The Full Risk Prediction Dashboard



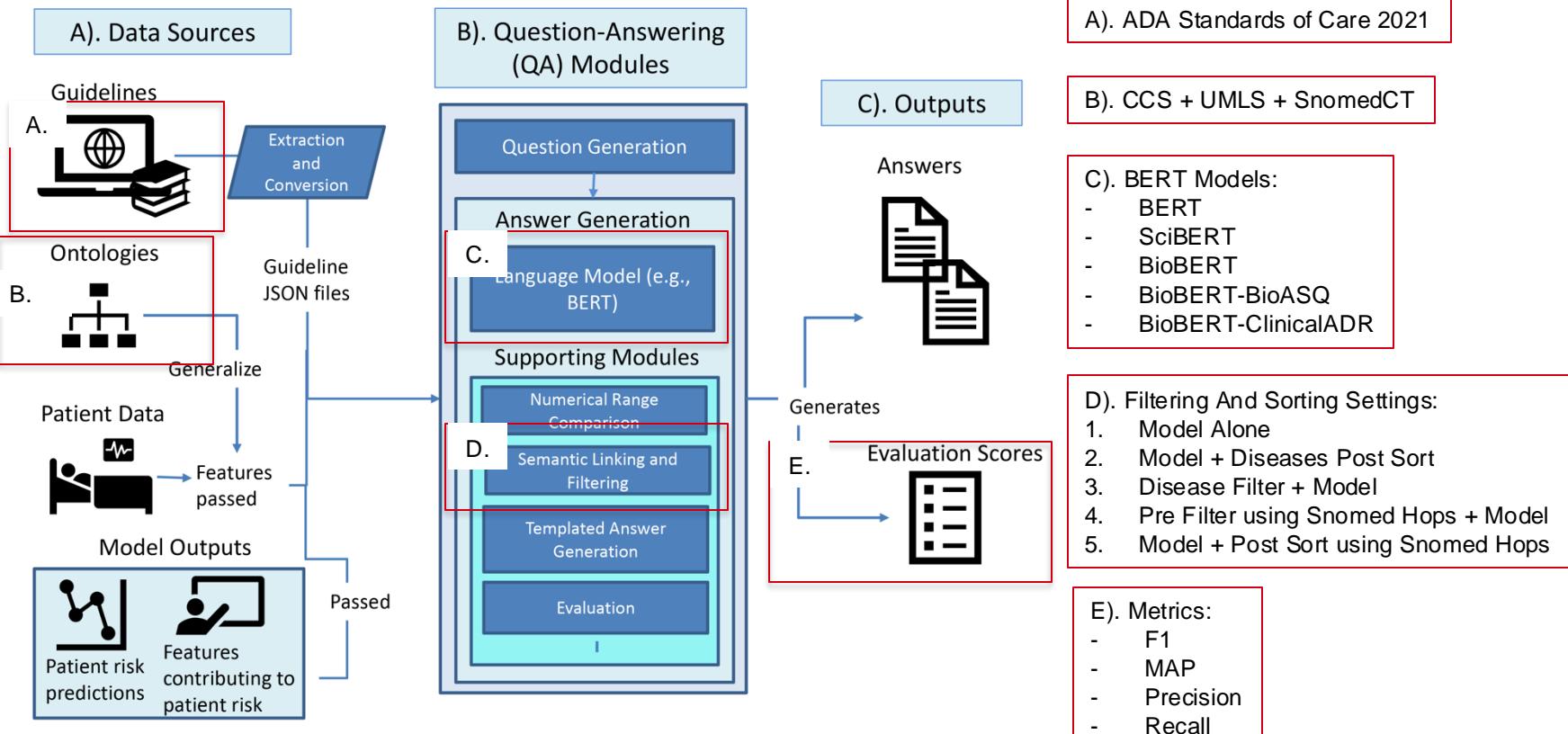
Example Answers

Question Type	Sample Question	Answer
1. Patient's T2DM summary	What is the patient's A1C value? What are their most frequent diagnoses codes?	Patient's A1C is A. Their most frequent diagnosis codes are essential hypertension, septicemia, etc.
2. Patient's risk summary	How does the predicted risk of the patient compare against the population?	The predicted risk of chronic kidney disease the patient is X %. The population averages for the same condition are as follows: For Medicare patients: Y % For patients with Charlson Comorbidity Index (CCI) score of 3 : Z %
3. Features contributing to patient's CKD risk	What can be done for Essential Hypertension?	10.3 For patients with diabetes and hypertension, blood pressure targets should be individualized through a shared decision-making process that addresses cardiovascular risk, potential adverse effects of antihypertensive medications, and patient preferences. C

Example Answers

Question Type	Sample Question	Answer
4. Patient's lab values	What should be done for this patient, whose A1C levels are greater than 10 ?	The early introduction of insulin should be considered if there is evidence of ongoing catabolism (weight loss), if symptoms of hyperglycemia are present, or when A1C levels are greater than 10% [86 mmol/mol] or blood glucose levels greater than or equal to 300 mg/dL [16.7 mmol/L] are very high.
5. Patient's medication list	What do the guidelines state about the GLP-1 RA drug the patient is taking?	Meta-analyses of the trials reported to date suggest that GLP-1 receptor agonists and SGLT2 inhibitors reduce risk of atherosclerotic major adverse cardiovascular events to a comparable degree in patients with type 2 diabetes and established ASCVD (185).

Guideline Question-Answering (QA) Method



Quantitative Evaluation of QA - Disease Questions

Best performing models:

model	bleu	P@1	P@5	map	f1	recall
BERT	0.117	0.468	0.382	0.390	0.213	0.241
BioBERT	0.116	0.431	0.339	0.346	0.200	0.238
BioBERT-BioASQ	0.132	0.383	0.329	0.332	0.217	0.281
BioClinicalBERT-ADR	0.125	0.368	0.317	0.316	0.205	0.259
SciBERT	0.165	0.461	0.349	0.364	0.261	0.354

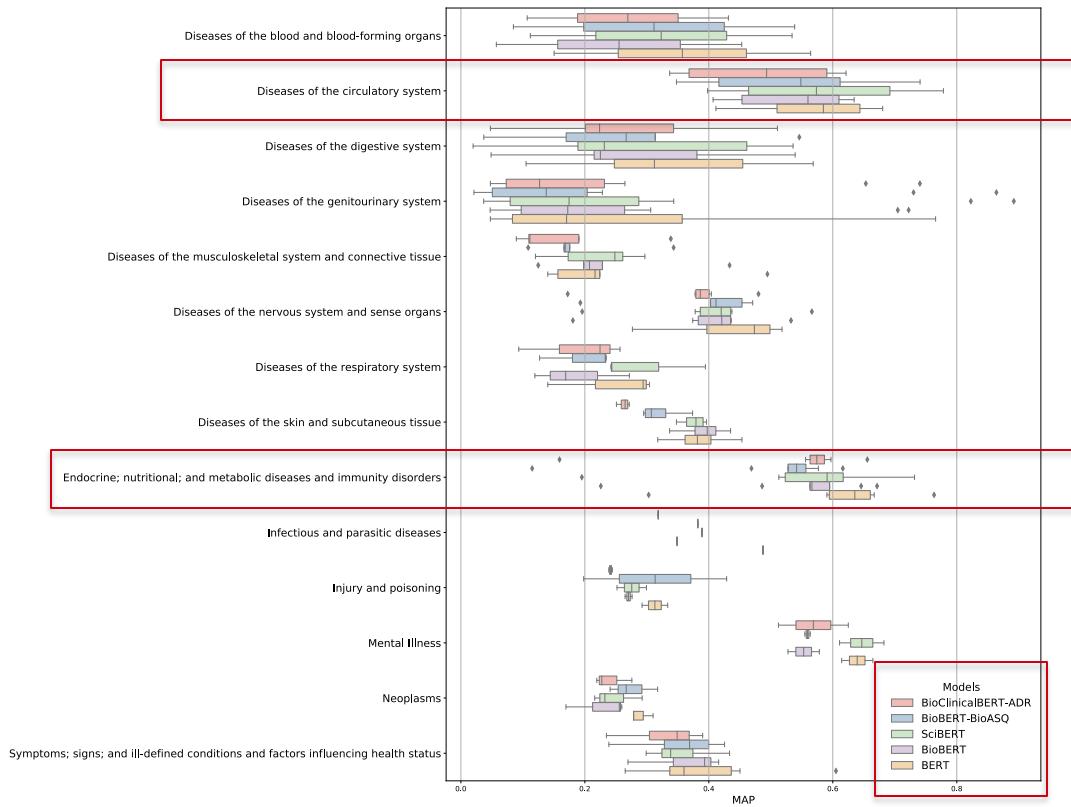
Best performing models with KAs:

model	bleu	P@1	P@5	map	f1	recall
BERT-KA	0.075	0.467	0.419	0.438	0.169	0.186
BioBERT-KA	0.127	0.434	0.348	0.353	0.215	0.254
BioBERT-BioASQ-KA	0.141	0.458	0.362	0.369	0.237	0.280
BioClinicalBERT-ADR-KA	0.121	0.406	0.321	0.330	0.202	0.242
SciBERT-KA	0.192	0.473	0.341	0.375	0.291	0.405

Models downloaded from: https://huggingface.co/models?pipeline_tag=question-answering&sort=trending&search=Bio
KA used Python libraries

<https://github.com/AnthonyMrios/pymetamap>; <https://owlready2.readthedocs.io/en/latest/pymedtermino2.html>

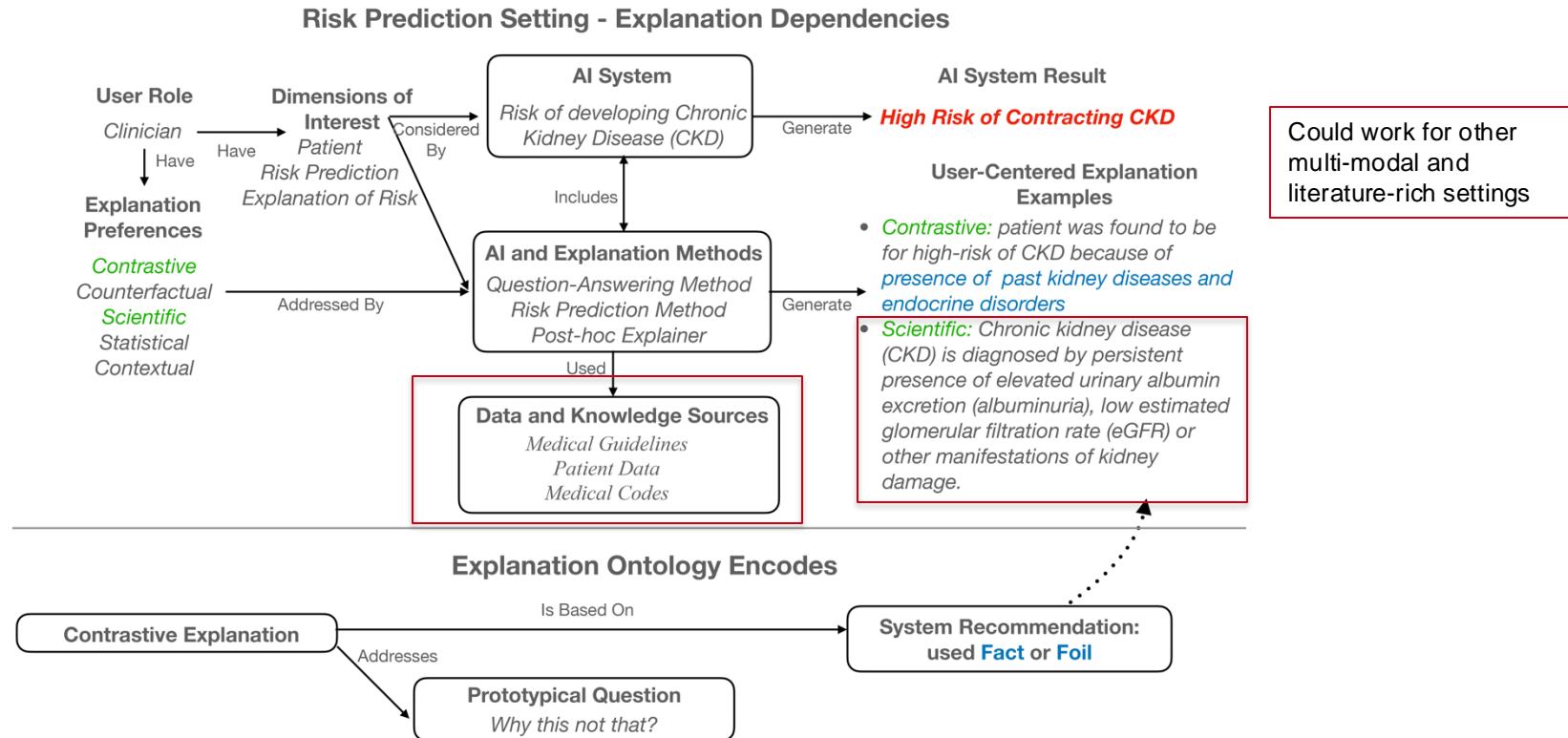
Quantitative Evaluation – Model Differences by Disease Group



Themes and Sub-themes from Expert Panel Sessions

Theme	Sub-Theme
Clinical Value of Explanations and Contextualizations	Value of contextual information around: CKD Risk, Patient features, Patient's Diabetes
Highlighting Actionability	Highlight Actionable and Modifiable Factors, Highlight the Impact of CKD risk prediction on Treatment Decisions for Diabetes and other conditions, Suggest Specific Actions to Reduce CKD risk
Connections to Patient Data	Connections to Patient's Clinical Indicators, Need for Information on Related Diagnoses, Connections to Patient's History
Connections to External Medical Domain Knowledge	Links to: Medication Databases Published Articles Support familiar categorizations

Opportunity for Multiple Explanations in A Risk Prediction Setting





MetaExplainer: A Framework to generate Natural-language Explanations

Github Repository: <https://github.com/tetherless-world/metaexplainer/tree/main>



Key Points

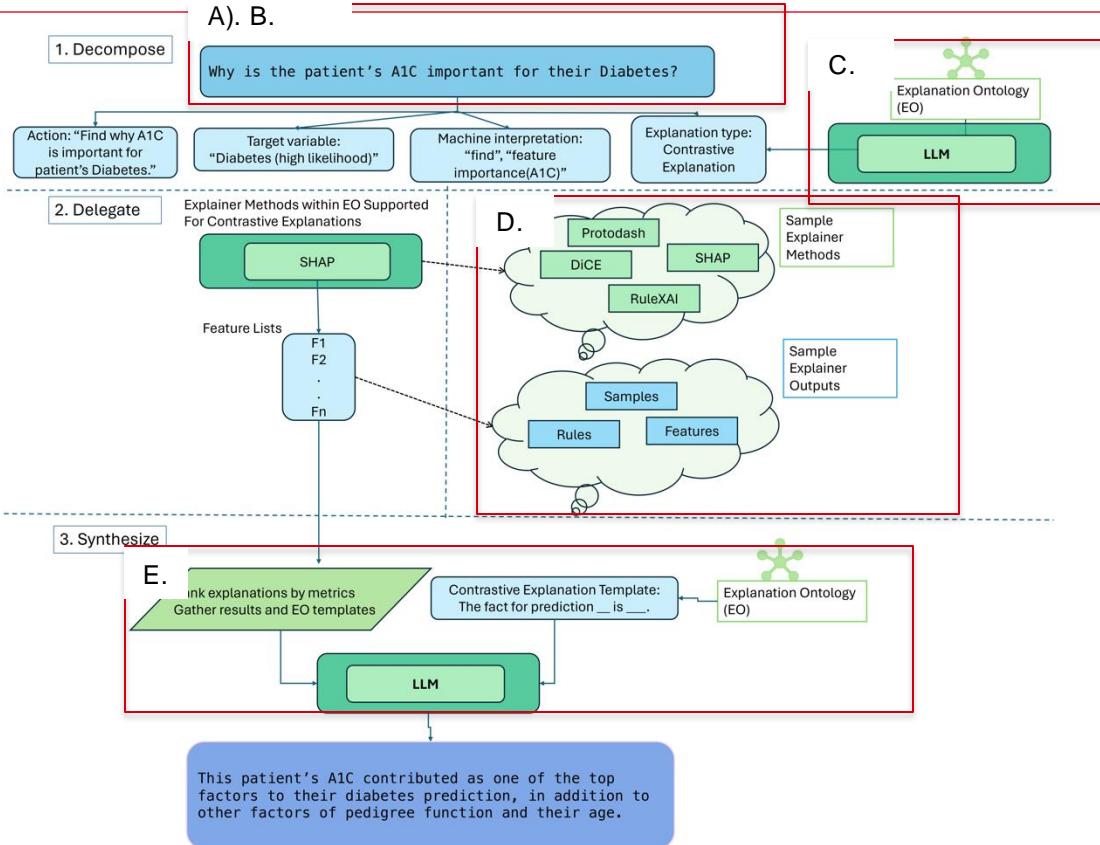
Addresses Q 3). Is it feasible to **combine explanations** from **multiple data sources** and **AI methods**?

We have designed the MetaExplainer framework; to respond to a large variety of user questions along different explanation types.

Different stages are:

- **Decompose:** Convert user question into machine-interpretable components.
- **Delegate:** Parse the components to identify and run explainer methods.
- **Synthesis:** Combine results of explainer methods along templates for identified explanation type.

MetaExplainer Framework – add EO graph next slide



A). PIMA Indian Diabetes Dataset

B). Generated using GPT-3.5 Turbo

C). Fine-tuned Decompose Models

- Llama2
- Llama3

Metrics: Classification Accuracy, F1s

D). Explanation Types

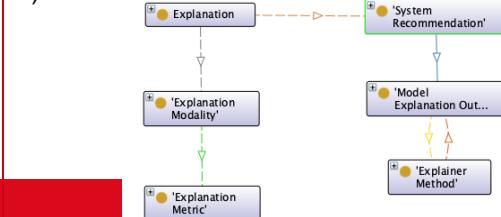
1. Case-based (*Protodash*)
2. Contrastive (*SHAP*)
3. Counterfactual (*DiCE*)
4. Data (*Protodash*)
5. Rationale (*RuleXAI*)

Metrics: Dependent on Explanation Modality

E). RAG over Explainer Result Dataframes – LlamaIndex and GPT-3.5 Turbo

Metrics: Relevance, Precision

F). EO is used for:



Example Results

User question: Why did the model classify a 40-year-old female with a BMI of 30 and a Diabetes Pedigree Function of 0.1 as not having Diabetes?

Decompose Output from LLama3 FT model

Question: Why did the model classify a 40-year-old female with a BMI of 30 and a Diabetes Pedigree Function of 0.1 as not having Diabetes?

Machine interpretation: "Why did the model predict that the patient with (age=40, sex=Female, BMI=30, Diabetes Pedigree Function=0.1, Diabetes = no) does not have Diabetes?"

Action,['with']

Feature groups: "[{'Age': '40', 'Sex': 'Female', 'BMI': '30', 'Diabetes Pedigree Function': '0.1', 'Diabetes': 'no'}]"

Explanation type: Rationale Explanation

Delegate Output from RuleXAI [1]

"IF BMI = (-inf, 27.45) THEN label = {0}"

"IF DiabetesPedigreeFunction = (-inf, 0.39) AND Pregnancies = (-inf, 9.5) THEN label = {0}"

"IF DiabetesPedigreeFunction = <0.39, inf) THEN label = {1}"

"IF Pregnancies = <9.5, inf) THEN label = {1}"

1. Macha, Dawid, et al. "RuleXAI—A package for rule-based explanations of machine learning model." *SoftwareX* 20 (2022): 101209.

Example Results

User question: Why did the model classify a 40-year-old female with a BMI of 30 and a Diabetes Pedigree Function of 0.1 as not having Diabetes?

Synthesis Output from LlamaIndex using GPT-3.5 Turbo on Result Dataframes

Explanation of Subset matched within Dataset:

Based on the query results, there are no full matches in the dataset based on the specified feature group of Age: 40, Sex: Female, BMI: 30, and Diabetes Pedigree Function: 0.1.

However, the dataset summary shows that there are 10 entries with an average age of 40, average BMI of 33.48, and an average Diabetes Pedigree Function value of 0.3568. The outcome variable has a mean value of 0.4, indicating that there is a mix of positive and negative outcomes in the dataset.

Explanation of Rules:

The data provided includes information on rules for labeling based on BMI values. The rule states that if the BMI falls within the range of negative infinity to 27.45, the label assigned is 0. This decision is based on the specific criteria set within the machine learning model. The mean, standard deviation, and quartile values are also provided for further context. This explanation helps users understand the rationale behind the decision-making process and allows them to assess the validity of the model's reasoning.

Results per Stage – Every stage of the MetaExplainer can be evaluated

Generation results for Decompose from LLama3

I: Performance Metrics for Text Fields from Llama3

Field	F1 (%)	Precision (%)	Recall (%)
F1 Exact Match scores on text fields			
Machine Interpretation	59.06	55.91	62.58
Action	57.48	50.00	67.60
Likelihood	81.46	84.34	78.77
F1 Levenshtein scores on text fields			
Machine Interpretation	18.87	18.87	18.87
Action	19.23	19.61	18.87
Likelihood	81.13	81.13	81.13
Exact match on text fields			
Machine Interpretation		29.53	
Action		28.74	
Likelihood		40.73	

Classification Models run on chosen Dataset

› 5.11: Model Performance Metrics

Model	Precision	Recall	F1	Sensitivity	Specificity
Logistic Regression	0.77	0.77	0.77	0.61	0.86
Decision Tree	0.73	0.73	0.73	0.63	0.79
Random Forest	0.75	0.75	0.75	0.63	0.82

Evaluation metrics [1] on Delegate Explainer Outputs

Table 5.12: Summary of Metrics for Explanation Modalities

Metric	Mean values	Modality	Explainer Method
Average rule length	2.39	Rules	RuleXAI
Fidelity	0.31	Rules	RuleXAI
Non representativeness	0.026	Samples	Protodash and DiCE
Diversity	340.96	Samples	Protodash and DiCE
Faithfulness	0.71	Features	SHAP
Monotonicity	0.095	Features	SHAP

Evaluation of RAG outputs from Synthesis

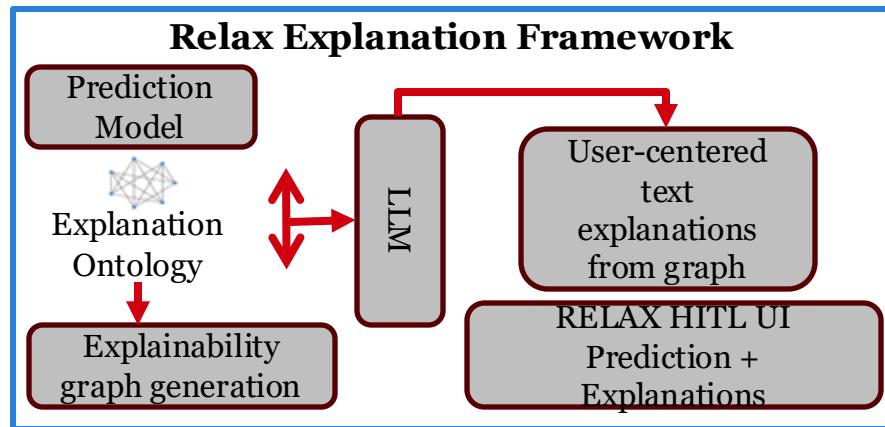
Table 5.13: Results of RAG metrics [5] for small-set of natural-language explanations generated by the Synthesis stage.

Metric	Value
Faithfulness	0.071
Answer relevance	0.68
Context-utilization	0.92

1. Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 593.



The RELAX Explainability Approach for Authorship Attribution



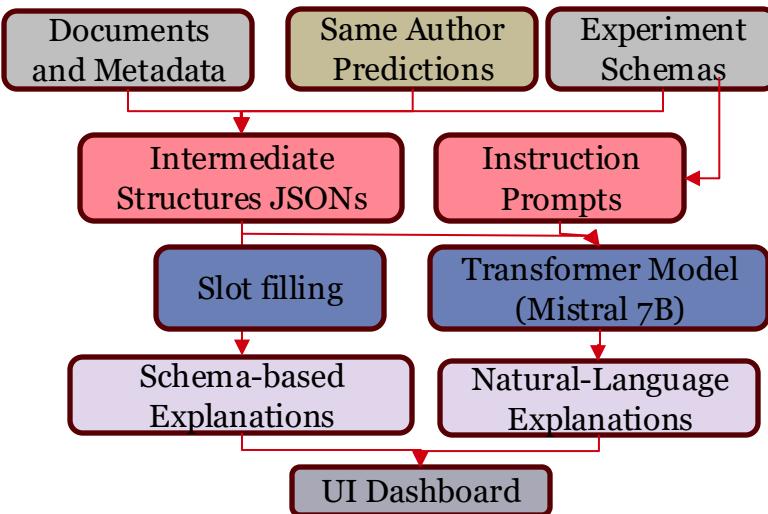
Current Progress

- Zero shot LLM-generated explanations from features and authorship model predictions
- Prototyped two interpretable authorship attribution models
- Explainability experiment (right & next slide) focused end-user trust as mediated by the explanation.

Next Steps

- Develop ML explanation graphs using Explanation Ontology (EO)
- Develop method for prompt generation from EO-based explanation graphs

Explaining Same Author Prediction for TA2 with zero-shot LLM



Innovations

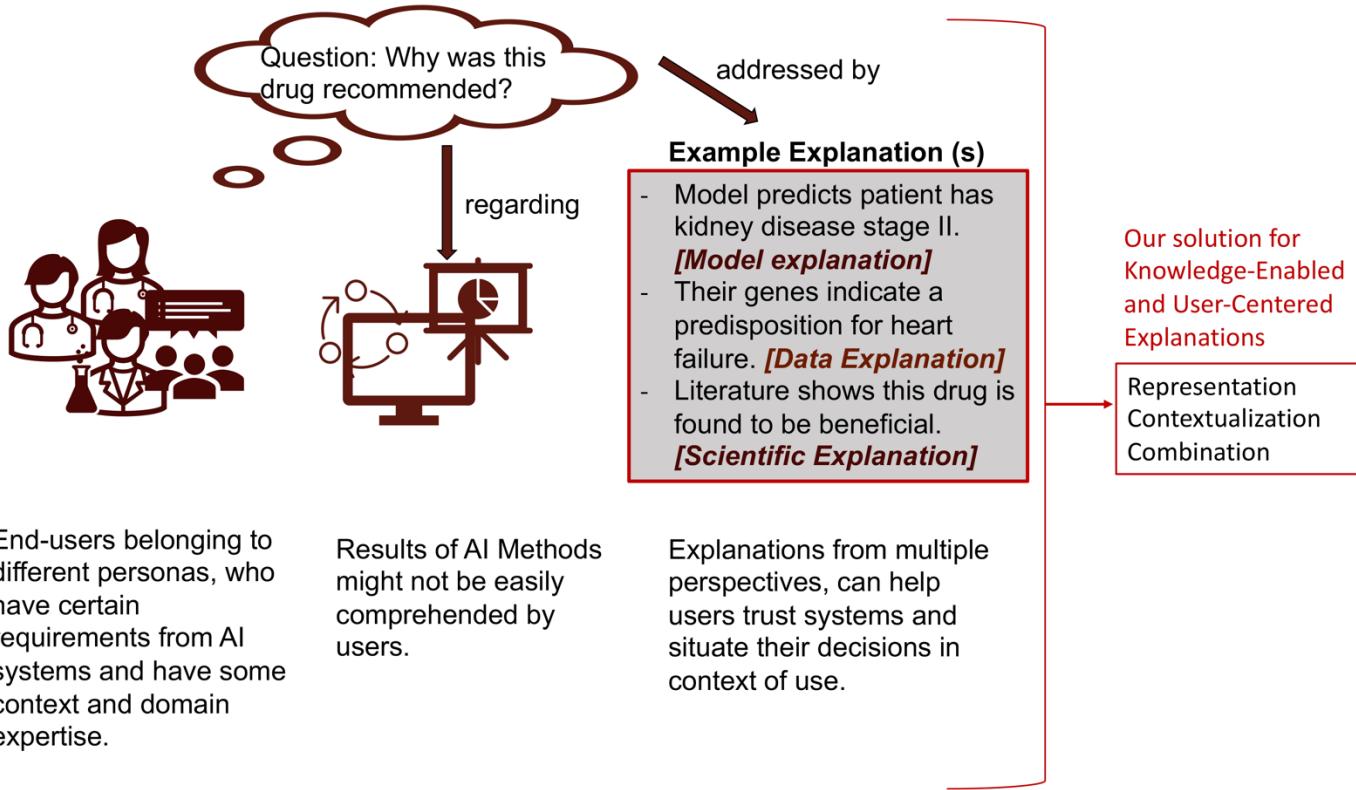
1. Developed user-centric explanation generator to improve end-user system comprehension and trust
2. Explanations pull features with most authorship predictive power and presents them for users

Conclusion



Rensselaer

An Ontology-Enabled Approach For User-Centered and Knowledge-Enabled Explanations of AI Systems



Limitations and Future Work

- User-centered aspects of Explainability and Challenges
 - Domain experts and users time is hard to come by. However, there are methods that can be employed like active and reinforcement learning to leverage user feedback.
- Grounding Explanations in Domain Knowledge and Context
 - Need to setup pipelines to extract insights from knowledge sources in each domain – but this is easier with LLMs
- Disagreement and Overload Problem of Explanations
 - Explanations can convey different signals and they might conflict / complement each other – a combination of similarity and metric-based techniques can be used to resolve these conflicts.
- Metrics for Explainable AI
 - There are lot of subjective metrics for explanations, however, there are fewer objective metrics – there is scope for innovative approaches here.



Conclusion

- Presented three contributions to support user-centered explanations by:
 - a general-purpose semantic representation for diverse explanation types,
 - a method to support contextual explanations to interpret AI system results, and
 - a framework to respond to a wide-variety of user questions via explanations,
- Two contributions are evaluated in clinical settings of real-world importance but have applications beyond these domains.
- Open-sourced the semantic representation and MetaExplainer and invite the community to provide feedback and try it out.
- Combine NLP, Knowledge Representation and Explainable ML methods to improve decision-making for clinicians using AI systems.

Why is XAI still important? In the foundational model era, there are exciting opportunities to extract insights from domain literature and multi-modal and -omics data (e.g., clinician notes, scans, drug components) for goals such as precision medicine and XAI can be a big part of helping domain experts use these AI systems.

Collaborators

I thank my amazing collaborators over the years from RPI, IBM and LEIDOS.

Committee Members



Prof. Deborah
McGuinness, RPI



Prof. Oshani
Seneviratne, RPI



Prof. James A.
Hendler, RPI



Dr. Prithwish
Chakraborty, Amazon



Dr. Pablo Meyer
IBM



Dr. Henrique Santos,
RPI



Dr. Jamie P.
McCusker, RPI



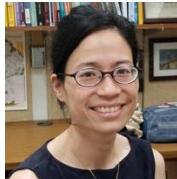
Prof. Kristin P.
Bennett, RPI



Dr. Daniel M. Gruen,
RPI



Dr. Amar K. Das,
Guardant Health



Dr. Ching-Hua
Chen, IBM



Morgan
Foreman, IBM



Dr. Mohamed
Ghalwash, IBM



Dr. Elif Eyigoz,
IBM



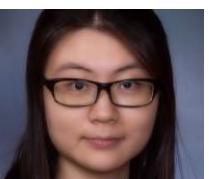
Dr. Noelia
Creaghe, LEIDOS



Dr. Alex Rosenfield,
LEIDOS



Dr. Sola Shirai, IBM



Dr. Miao Qi,
Neptec OS



Jade Franklin, RPI



Dr. Nneka Agu,
Amazon



Dr. Neha
Keshan, RPI



Dr. Sabbir
Rashid, RPI



Prasant Acharya,
Epic



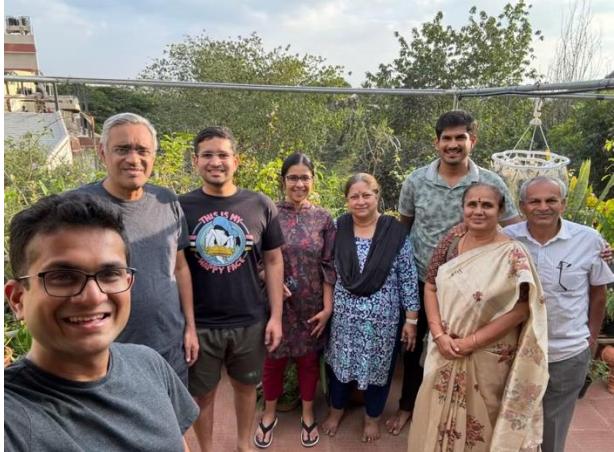
Olivia Zhang,
Hercules Pharmaceuticals



Rensselaer

An Ontology-Enabled Approach For User-Centered and Knowledge-Enabled Explanations of AI Systems

Special thanks to my Family, Friends and all the Activities!



Thank You. Questions?

Contact: charis@rpi.edu



Rensselaer

why not change the world?®

References



Rensselaer

An Ontology-Enabled Approach For User-Centered and Knowledge-Enabled Explanations of AI Systems

Publications

1. S Chari, P Acharya, DM Gruen, O Zhang, EK Eyigoz, M Ghalwash, O Seneviratne, FS Saiz, P Meyers, P Chakraborty, DL McGuinness; "Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes"; *Artificial Intell. Medicine J.*, 102498; 2023
2. S Chari, O Seneviratne, M Ghalwash, S Shirai, DM Gruen, P Meyers, P Chakraborty, DL McGuinness; "Explanation Ontology: A General-Purpose, Semantic Representation for Supporting User-Centered Explanations"; *Semantic Web J.*, In Press, 2023
3. S Chari, P Chakraborty, M Ghalwash, O Seneviratne, EK Eyigoz, DM Gruen, FS Saiz, CH Chen, PM Rojas, DL McGuinness; "Leveraging Clinical Context for User-Centered Explainability: A Diabetes Use Case"; *KDD Applied Data Science in Healthcare (DSHealth) Workshop*; 2021 [**Best workshop paper**]
4. DM Gruen., S Chari, MA Foreman, O Seneviratne, R Richesson, AK Das, and DL McGuinness. "Designing for AI explainability in clinical context." *Trustworthy AI Workshop*, AAAI, 2021.
5. S Chari, O Seneviratne, DM Gruen, M Foreman, AK Das, DL McGuinness. "Explanation Ontology: A Model for User-Centric Explainable AI." *In International Semantic Web Conference* (pp. 228 - 243); 2020 [**Best paper award**]
6. S Chari, O Seneviratne, DM Gruen, DL McGuinness. "Foundations of Explainable Knowledge-Enabled Systems." In Ilaria Tiddi, Freddy Lecue, Pascal Hitzler (eds.), *Knowledge Graphs for eXplainable AI -- Foundations, Applications and Challenges. Studies on the Semantic Web*, pp 23 - 48; 2020
7. S Chari, O Seneviratne, DM Gruen, DL McGuinness. "Directions for Explainable Knowledge-Enabled Systems." In Ilaria Tiddi, Freddy Lecue, Pascal Hitzler (eds.), *Knowledge Graphs for eXplainable AI -- Foundations, Applications and Challenges. Studies on the Semantic Web*, pp 245 - 261; 2020

Posters

1. O Zhang, S Chari, FS Saiz, DM Gruen, P Acharya, O Seneviratne, P Meyer, DL McGuinness, P Chakraborty; "Understanding Clinician Workflows to Design AI Risk Prediction Models"; AMIA Inf. Summ., Apr 2022
2. S Chari, O Zhang, P Acharya, FS Saiz, M Ghalwash, EK Eyigoz, O Seneviratne, DM Gruen, PM Rojas, P Chakraborty, DL McGuinness; "Towards Providing Clinical Context for a Diabetes Risk-Prediction Use Case via User-centered Explainability"; MCBK Symp., Jul 2021
3. S Chari, P Chakraborty, O Seneviratne, M Ghalwash, DM Gruen, D Sow, DL McGuinness; "Towards Clinically Relevant Explanations for Type-2 Diabetes Risk Prediction with the Explanation Ontology"; AMIA Symp. Poster 2021
4. S Chari, O Seneviratne, DM Gruen, M Foreman, AK Das, DL McGuinness; "Explanation Ontology in Action: A Clinical Use-Case"; Posters and Demo Track, 19th International Semantic Web Conference 2020.
5. S Chari, O Seneviratne, DM Gruen, M Foreman, AK Das, DL McGuinness. "Supporting User-centric explanation types for clinical reasoning." In MCBK, 2020.



References used in Slides

1. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.
2. Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288).
3. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.
4. Dey, S., Chakraborty, P., Kwon, B. C., Dhurandhar, A., Ghalwash, M., Saiz, F. J. S., ... & Meyer, P. (2022). Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, 3(5), 100493.
5. Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 593.
6. Amith, M., He, Z., Bian, J., Lossio-Ventura, J.A., Tao, C.: Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *Journal of biomedical informatics* 80, 1–13 (2018)
7. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750.
8. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019, October). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference* (pp. 359-380). PMLR.
9. Groopman, E. E., Marasa, M., Cameron-Christie, S., Petrovski, S., Aggarwal, V. S., Milo-Rasouly, H., ... & Gharavi, A. G. (2019). Diagnostic utility of exome sequencing for kidney disease. *New England Journal of Medicine*, 380(2), 142-151.
10. Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *international journal of Computer Applications*, 68(13), 13-18.
11. Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. *Information*, 11(9), 421.

Supplementary Slides

EO Extra



Rensselaer

An Ontology-Enabled Approach For User-Centered and Knowledge-Enabled Explanations of AI Systems

45

12/13/2024

Contrastive Explanation Representation in EO – Why this not that?

Description: Contrastive Explanation

Equivalent To +

- 'is based on' some ('System Recommendation' and (used some Foil))
- Explanation and (wasGeneratedBy some 'Abductive Task')
- 'is based on' some ('System Recommendation' and (used some Fact))
- ('is based on' some ('System Recommendation' and (used some Fact))) and ('is based on' some ('System Recommendation' and (used some Foil)))

?

@

X

O

?

@

X

O

?

@

X

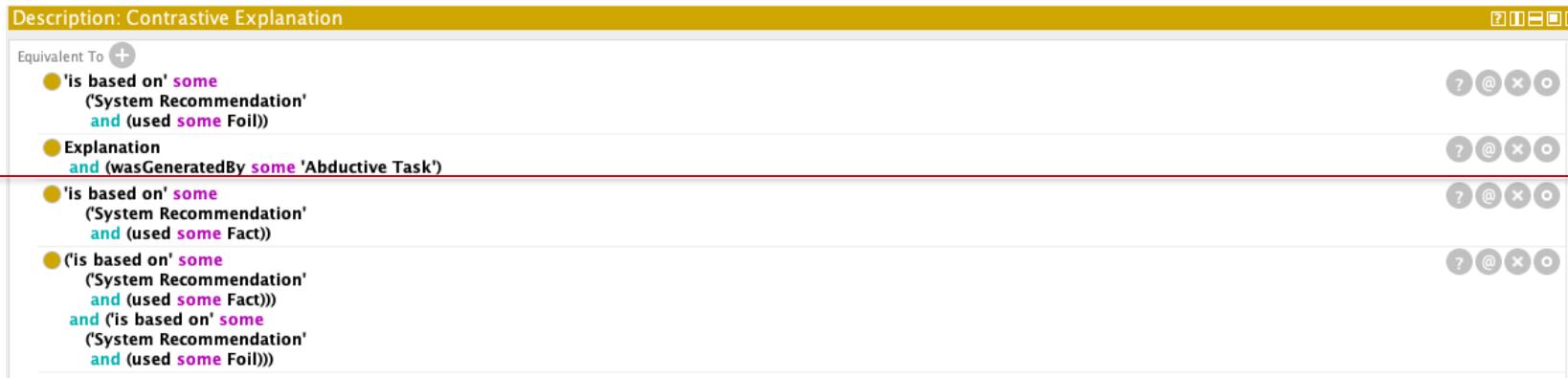
O

?

@

X

O

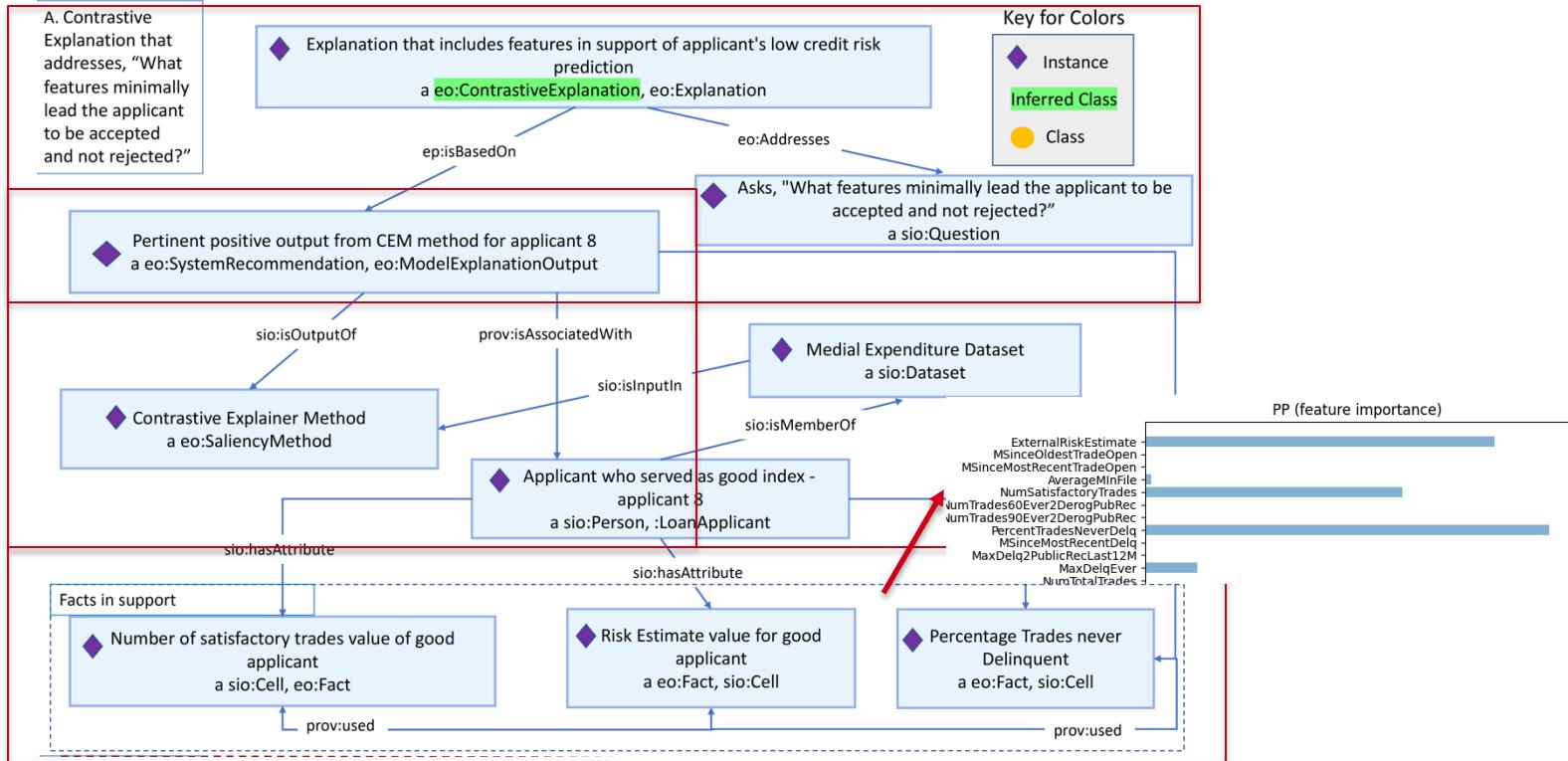


Develop Methods to Combine Explanations in a Multi-modality setting

- **Setting:** Stage Prediction of Patient's Chronic Kidney Disease
- **Data Modalities Involved in Prediction:**
 - Patient Diagnosis Codes
 - Patient CT Scans
 - Patient Genes
 - Genes found to be important for CKD from literature [1]
- **Multi-step Explanation Comparison Approach:**
 - Compare the set of diseases the patient has and the ones they are pre-disposed due to their genes [use semantic linkages] -> only compare relevant diseases
 - Compare similarity of answers from literature for disease sets and actions on them
 - Combination of medical guidelines and published literature on Pubmed
 - Expand on text similarity metrics [2, 3] to compute similarity of literature-derived scientific explanations.

1. Groopman, E. E., Marasa, M., Cameron-Christie, S., Petrovski, S., Aggarwal, V. S., Milo-Rasouly, H., ... & Gharavi, A. G. (2019). Diagnostic utility of exome sequencing for kidney disease. *New England Journal of Medicine*, 380(2), 142-151.
2. Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *international journal of Computer Applications*, 68(13), 13-18.
3. Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. *Information*, 11(9), 421.

EO KG in Credit Approval Use Case



Why was it classified as Contrastive?

Explanation 1 Display laconic explanation

Explanation for: 'Explanation that includes features in support of applicant's low credit risk prediction' Type 'Contrastive Explanation'

- 1) 'Explanation that includes features in support of applicant's low credit risk prediction' is based on 'Pertinent positive output from CEM method for applicant 8'
- 2) 'is based on' some ('System Recommendation' and (used some Fact) SubClassOf 'Contrastive Explanation'
- 3) 'Number of satisfactory trades value of good applicant' Type Fact
- 4) 'Pertinent positive output from CEM method for applicant 8' used 'Number of satisfactory trades value of good applicant'
- 5) 'Pertinent positive output from CEM method for applicant 8' Type 'System Recommendation'

In ALL other justifications [?](#)

In NO other justifications [?](#)

In NO other justifications [?](#)

In NO other justifications [?](#)

In ALL other justifications [?](#)

Explanation 2 Display laconic explanation

Explanation for: 'Explanation that includes features in support of applicant's low credit risk prediction' Type 'Contrastive Explanation'

- 1) 'Explanation that includes features in support of applicant's low credit risk prediction' is based on 'Pertinent positive output from CEM method for applicant 8'
- 2) 'is based on' some ('System Recommendation' and (used some Fact) SubClassOf 'Contrastive Explanation'
- 3) 'Risk Estimate value for good applicant' Type Fact
- 4) 'Pertinent positive output from CEM Inferred 1 for applicant 8' used 'Risk Estimate value for good applicant'
- 5) 'Pertinent positive output from CEM method for applicant 8' Type 'System Recommendation'

In ALL other justifications [?](#)

In NO other justifications [?](#)

In NO other justifications [?](#)

In NO other justifications [?](#)

In ALL other justifications [?](#)



Evolution-based Evaluation [1]

Additions to EO from V1.0 to V2.0, enabled us to:

- Capture more **granular representations** of 'AI methods' and their interactions with the explanation **types**, and support more ways to generate explanations.
- Introduce **characteristics at various strategic attributes** that contribute to explanations (e.g., at the system, user, and object classes), which provide the flexibility to define characteristics at multiple levels and allow for better considering explanation types through the restriction of equivalent classes.
- Include more of the **contributing attributes of the explanation ecosystem** itself (such as capturing the 'system' in which the 'AI methods' are run), which helps maintain better provenance of the infrastructure contributing to the explanations.

[1] Amith, M., He, Z., Bian, J., Lossio-Ventura, J. A., & Tao, C. (2018). Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *Journal of biomedical informatics*, 80, 1-13.

Task-based Evaluation

Kendall, Elisa F., and Deborah L. McGuinness. "Ontology engineering." *Synthesis Lectures on the Semantic Web: Theory and Technology* 9.1 (2019): i-102.

Setting	Competency Question	Answer	SPARQL Query length	Property Restrictions accessed?	Inference Required?	Filter Statements
System Design	Q1. Which AI model(s) is/are capable of generating this explanation type (e.g. trace-based)?	Knowledge-based systems, Machine learning model: decision trees	8	Yes	No	No
System Design	Q2. What example questions have been identified for counterfactual explanations?	What other factors about the patient does the system know of? What if the major problem was a fasting plasma glucose?	4	No	No	No
System Design	Q3. What are the components of a scientific explanation?	Generated by an AI Task, Based on recommendation, and based on evidence from study or basis from scientific method	2	Yes	No	No
System Analysis	Q4. Given the system has ranked specific recommendations by comparing different medications, what explanations can be provided for that recommendation?	Contrastive explanation	8	Yes	No	No
System Analysis	Q5. Which explanation type best suits the user question asking about numerical evidence, and how does a system generate such an answer?	Explanation type: statistical; System: run 'Inductive' AI task with 'Clustering' method to generate numerical evidence	18	Yes	No	No
System Analysis	Q6. What is the context for data collection and application of the contextual explanation, say for example from the health survey analysis use case?	Explanation type: contextual Environmental context: 'Early childhood questionnaire' in a US location	5	No	Yes	No

Application-based Evaluation

Example questions that can be asked of our use case knowledge graphs that are modeled using our EO.

Use Case	Competency Question	Answer	SPARQL Query Length	Property Restrictions Accessed?	Inference Required?	Filter Statements
Food Recommendation	Q1. What explanation types are supported?	Contextual and Contrastive	2	No	Yes	No
Food Recommendation	Q2. Why should I eat spiced cauliflower soup?	Cauliflower is in season.	5	No	Yes	Yes
Proactive Retention	Q2. What is the retention action outcome for employee 1?	Employee 1 is likely to remain in the same organization.	4	No	No	Yes
Health Survey Analysis	Q3. Who are the most representative patients in the income questionnaire?	Patient 1 and 2.	3	No	Yes	No
Health Survey Analysis	Q5. Which questionnaire did patient 1 answer?	Income, early childhood and social determiners.	3	No	Yes	Yes
Medical Expenditure	Q6.What are the rules for high-cost expenditure?	Individuals are in poor health, have limitations in physical functioning and are on health insurance coverage.	4	No	Yes	Yes
Credit Approval	Q7. What factors contribute most to a loan applicants credit approval?	Facts: Number of satisfactory trades and risk estimate value	7	No	Yes	No



SPARQL Query for Retrieving Facts Supporting Contrastive Explanation

SPARQL Query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX eo: <https://purl.org/heals/eo#>
PREFIX ep: <http://linkedu.eu/dedalo/explanationPattern.owl#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX ceo: <https://purl.org/heals/eo_creditapprovalusecase#>
```

```
SELECT ?subject ?sysRec ?applicant ?fact
```

```
WHERE {
?subject a eo:ContrastiveExplanation .
?subject rdfs:label ?s .
?subject ep:isBasedOn ?sysRec .
?sysRec prov:used ?fact .
?fact a eo:Fact .
```

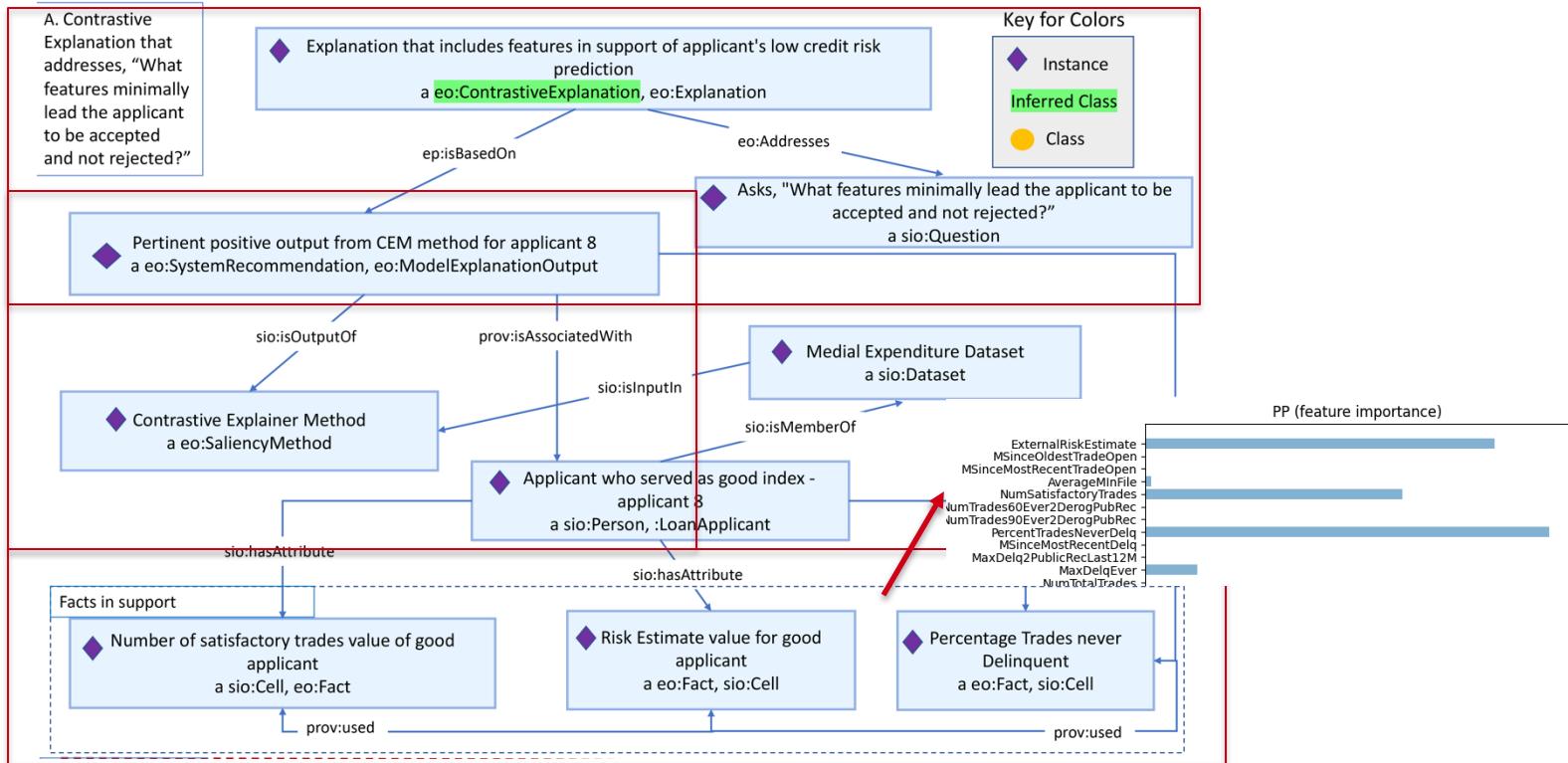
```
?sysRec sio:SIO_001403 ?applicant .
?applicant a ceo:LoanApplicant .
}
```

Query Results:

subject	sysRec
Explaination that includes features in support of applicant's loan Pertinent positive output from CEM method for applicant 8	Explaination that includes features in support of applicant's loan Pertinent positive output from CEM method for applicant 8

applicant	fact
Applicant who served as good index – applicant 8	Number of satisfactory trades value of good applicant
Applicant who served as good index – applicant 8	Risk Estimate value for good applicant

EO KG in Credit Approval Use Case – [Change to Medical Expenditure]



Why was it classified as Contrastive?

Explanation 1 Display laconic explanation

Explanation for: 'Explanation that includes features in support of applicant's low credit risk prediction' Type 'Contrastive Explanation'

- 1) 'Explanation that includes features in support of applicant's low credit risk prediction' is based on 'Pertinent positive output from CEM method for applicant 8'
- 2) 'is based on' some ('System Recommendation' and (used some Fact) SubClassOf 'Contrastive Explanation'
- 3) 'Number of satisfactory trades value of good applicant' Type Fact
- 4) 'Pertinent positive output from CEM method for applicant 8' used 'Number of satisfactory trades value of good applicant'
- 5) 'Pertinent positive output from CEM method for applicant 8' Type 'System Recommendation'

In ALL other justifications [?](#)
In NO other justifications [?](#)
In NO other justifications [?](#)
In NO other justifications [?](#)
In ALL other justifications [?](#)

Explanation 2 Display laconic explanation

Explanation for: 'Explanation that includes features in support of applicant's low credit risk prediction' Type 'Contrastive Explanation'

- 1) 'Explanation that includes features in support of applicant's low credit risk prediction' is based on 'Pertinent positive output from CEM method for applicant 8'
- 2) 'is based on' some ('System Recommendation' and (used some Fact) SubClassOf 'Contrastive Explanation'
- 3) 'Risk Estimate value for good applicant' Type Fact
- 4) 'Pertinent positive output from CEM Inferred 1 for applicant 8' used 'Risk Estimate value for good applicant'
- 5) 'Pertinent positive output from CEM method for applicant 8' Type 'System Recommendation'

In ALL other justifications [?](#)
In NO other justifications [?](#)
In NO other justifications [?](#)
In NO other justifications [?](#)
In ALL other justifications [?](#)

Guideline QA Extra

Contextual Explanation - Definition

Explanations that contain context, are often *explicit information* [1] to *characterize the situation* of (an) entity(ies), wherein “an *entity* is a person, place, or object that is *considered relevant* to the interaction between a user and an application”[2]

1. Dey, A. K., Abowd, G. D., & Wood, A. (1998, January). CyberDesk: A framework for providing self-integrating context-aware services. In *Proceedings of the 3rd international conference on Intelligent user interfaces* (pp. 47-54).
2. Lieberman, H., & Selker, T. (2000). Out of context: Computer systems that adapt to, and learn from, context. *IBM systems journal*, 39(3.4), 617-632.

Question Type	Sample Question	Answer
1. Patient's T2DM summary	What is the patient's A1C value? What are their most frequent diagnoses codes?	Patient's A1C is A. Their most frequent diagnosis codes are essential hypertension, septicemia, etc.
2. Patient's risk summary	How does the predicted risk of the patient compare against the population?	The predicted risk of chronic kidney disease the patient is X %. The population averages for the same condition are as follows: For Medicare patients: Y % For patients with Charlson Comorbidity Index (CCI) score of 3 : Z %
3. Features contributing to patient's CKD risk	What can be done for Essential Hypertension?	10.3 For patients with diabetes and hypertension, blood pressure targets should be individualized through a shared decision-making process that addresses cardiovascular risk, potential adverse effects of antihypertensive medications, and patient preferences. C

Question Types

Question Type	Contextualized Entity	Domain Knowledge Source
1 Patient's T2DM summary	Patient	Patient data
2 Patient's risk summary	Risk Prediction	Risk Prediction and population data
3 Features contributing to patient's CKD risk	Post-hoc Explanation	Feature importances and ADA guidelines
4 Patient's medication list	Patient and Risk Prediction	Patient Data and guidelines
5 Patient's lab values	Patient	Patient Data and guidelines

Knowledge Augmentation Settings for QA

Setting No.	Type
1	Model Alone
2	Model + Disease Post Sorting
3	Pre-Disease Sorting + Model
4	Model + Post Snomed Hop Sorting (3, 5, 8)
5	Pre Snomed Hop Sorting (3, 5, 8) + Model

Summary of patient cohort

Table 5

Summary (generated using Tableone library [45]) of 20 prototypical patients highlighting the demographic and diagnoses counts. We report the disease diagnoses by their higher-level disease groupings (e.g. for T2DM the higher-level code is endocrine, nutritional and metabolic disorders). We highlight the conditions that are most prevalent amongst the patients (>50%).

Feature	Overall counts (%)
Age at onset 45–54	4 (20.0)
Age at onset ≥55	15 (75.0)
Age at onset ≤44	1 (5.0)
SEX - FEMALE	7 (35.0)
Mood disorders	3 (15.0)
Diseases of the blood and blood-forming organs	3 (15.0)
Diseases of the circulatory system	17 (85.0)
Diseases of the digestive system	6 (30.0)
Diseases of the genitourinary system	9 (45.0)
Diseases of the musculoskeletal system and connective tissue	12 (60.0)
Diseases of the nervous system and sense organs	9 (45.0)
Diseases of the respiratory system	11 (55.0)
Diseases of the skin and subcutaneous tissue	7 (35.0)
Endocrine; nutritional; and metabolic diseases and immunity disorders	20 (100.0)
Infectious and parasitic diseases	10 (50.0)
Injury and poisoning	4 (20.0)
Mental illness	3 (15.0)
Neoplasms	6 (30.0)
Symptoms; signs; and ill-defined conditions and factors influencing health status	10 (50.0)



Sentences extracted from ADA 2021 Guidelines

Table 6

Coverage statistics from extracted content from the ADA Standards of Care - Diabetes Guidelines 2021. We report these statistics on the recommendations and discussion sentences we extracted across chapters.

Field	Count
Chapters	16
No. of sentences	2379
Tokens from BERT	118 350
Avg. BERT tokens per sentence	49
Metamap semantic types covered	116/126

Performance of QA framework on Drug and Lab Result Questions

Results on Drug Questions

Table 9

Performance of guideline QA with different knowledge augmentations of language model approaches reported at mean average precision (MAP), F1 and recall at top-10 answers and precision at top-1 and top-5 for 6 anti-diabetic drug feature questions. Best and second-best values for each column is highlighted in Green and Blue color, respectively. Language model (e.g. BERT) suffixed with KA represents the corresponding knowledge augmented model (e.g. BERT-KA).

Model	bleu	P@1	P@5	Map	f1	Recall
BERT	0.100	0.910	0.751	0.757	0.254	0.206
BioBERT	0.100	0.726	0.643	0.635	0.231	0.192
BioBERT-BioASQ	0.081	0.708	0.694	0.704	0.222	0.162
BioClinicalBERT-ADR	0.075	0.593	0.614	0.597	0.192	0.146
SciBERT	0.121	0.947	0.757	0.772	0.281	0.228
BERT-KA	0.099	0.900	0.863	0.821	0.281	0.213
BioBERT-KA	0.083	0.802	0.704	0.720	0.234	0.170
BioBERT-BioASQ-KA	0.117	0.711	0.725	0.716	0.272	0.221
BioClinicalBERT-ADR-KA	0.085	0.598	0.595	0.587	0.199	0.152
SciBERT-KA	0.128	0.912	0.823	0.794	0.298	0.232

Results on Lab Questions

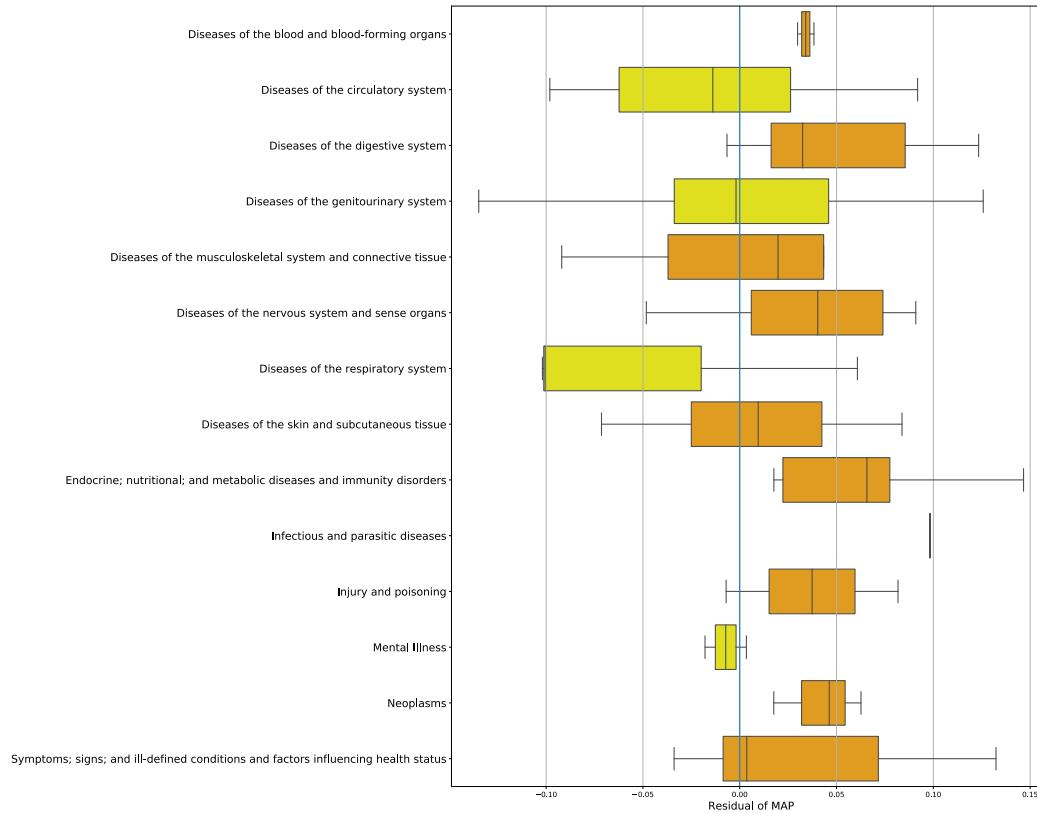
Table 10

Results of Guideline QA with rule augmentation of language model approaches for numerical comparisons reported for 9 questions across the 20 prototypical patients identified from our predicted high-risk chronic kidney disease cohort. The split of question variations is equal across the different numerical range comparison operators of lesser than, equal to and greater than.

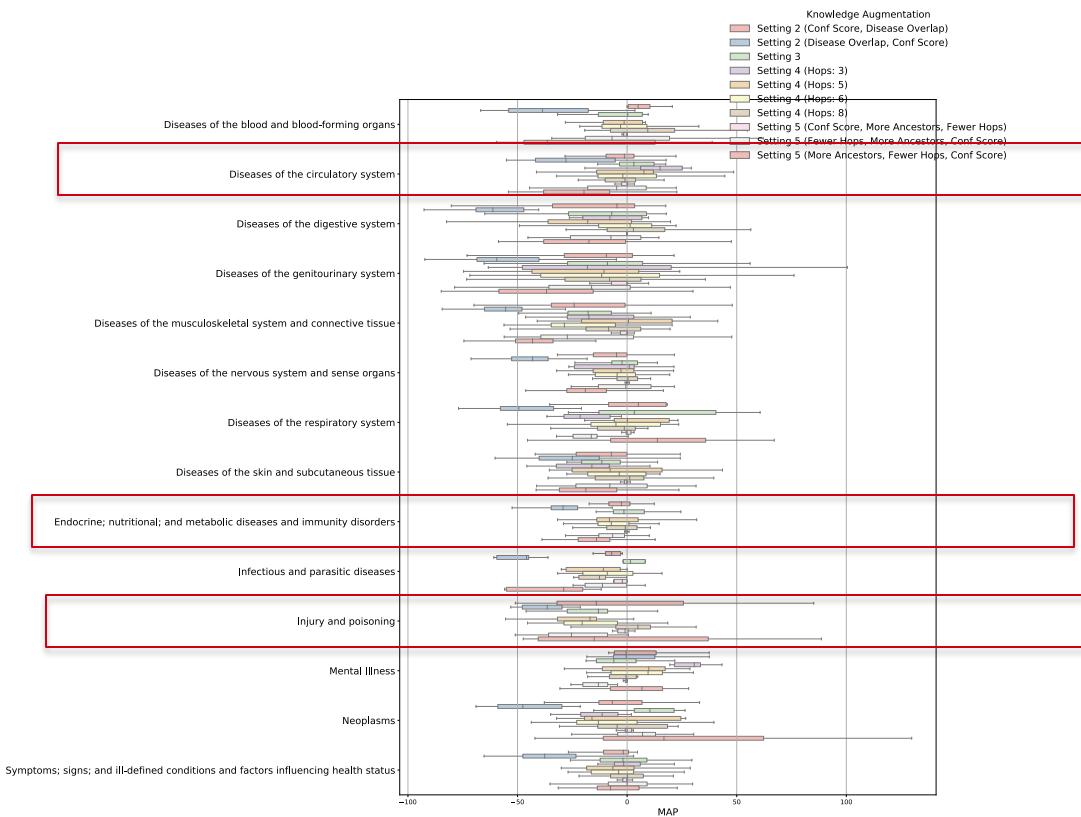
Comparison	Accuracy	TP	TN	FP	FN	Total
Overall	0.78	7	7	3	0	18
Lesser Than	0.84	2	3	1	0	6
Equal To	0.67	1	3	2	0	6
Greater Than	100	4	2	0	0	6

TP — True Positives, TN — True Negatives, FP — False Positives, FN — False Negatives. Accuracy computed as accuracy = (TP + TN)/Total.

Best performing models, BERT and SciBERT differences by Disease Groups



Differences between knowledge augmentation approaches



Qualitative Evaluation – Themes from Expert Panel Interviews

Theme (# of subthemes)	Sources	Current Coverage
1. Clinical Value of Explanations and Contextualizations (3)	Guidelines, Feature Importances, Medical Ontologies, Patient Data and Published Literature	✓ 3 / 3
2. Highlighting Actionability ** (3)	Guidelines, Published Literature	✓ 1 / 3
3. Connections to Patient Data (3)	History, Past Diagnoses and Clinical Indicators	✓ 1 / 3
4. Connections to External Medical Domain Knowledge (3)	Guidelines, Medical Ontologies Medication Databases, Published Literature and Familiar Categorizations	✓ 1 / 3

*Sources we currently support in our QA approach to provide context are shown in green

** Not currently supported by our methods

MetaExplainer Extra

Psuedocode of MetaExplainer

Algorithm 1 Metaexplainer

Require: Explanation Ontology (EO), Data Store (DS)

- 1: Explanation type - explainer graph, $G^1 = \{(t_i, em_j), \forall i \in N \text{ and } \forall j \in M\}$
- 2: Data type - explainer graph, $G^2 = \{(d_k, em_j), \forall k \in K \text{ and } \forall j \in M\}$

3: Input

- 4: uq User Question

- 5: $rq = \text{Decompose}(uq)$

- 6: $eq = \text{Delegate}(rq)$

- 7: $E = \text{Synthesis}(eq)$

8: Output

- 9: rq list of questions reframed from uq

- 10: E list of explanations $\{E_e\}$ that answer uq where $E_e = \{\mathcal{E}_e, t_e, em_e, rq_e, uqe\}$
-



Metrics for Explanation Modalities

Explanation Modality	Metric	Definition
Representative Samples	Diversity	Diversity is the distance between example points. [1]
Representative Samples	Non-Representativeness	High non-representativeness, however, can also be indicative of factual inaccuracy. A highly diverse set of examples demonstrates the degree of integration of the explanation. [1]
Feature Importances	Faithfulness	Are “relevant” features truly relevant? [2]
Feature Importances	Monotonicity	The monotonicity for feature attributions a_i is defined as the Spearman’s correlation coefficient $p_S(a, e)$ [28]. $a = (\dots, a_i , \dots)$ is a vector containing the absolute values of the attributions. $e = (\dots, E(l(y^*, f_i); X_i x^*-i), \dots)$ contains the corresponding (estimated) expectations, as computed in Eq.(1). [1]
Rules	Fidelity	The approximation should correctly capture the black box model behavior in all parts of the feature space. [3]
Rules	Average Rule Length	The average number of antecedents, connected with the AND operator, of the rules contained in each ruleset Lakkaraju et al. (2016) ; Wu et al. (2018) . a_i represents the number of antecedents of the i th rule and $R = A $ the number of rules. [4]

1. A.-p. Nguyen and M.R. Martínez, On quantitative aspects of model interpretability, arXiv preprint arXiv:2007.07584 (2020).
2. D. Alvarez Melis and T. Jaakkola, Towards robust interpretability with self-explaining neural networks, Advances in neural information processing systems 31 (2018).
3. H. Lakkaraju, E. Kamar, R. Caruana and J. Leskovec, Interpretable & exploratory approximations of black box models, arXiv preprint arXiv:1707.01154 (2017).
4. Vilone, Giulia, and Luca Longo. "A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods." *Frontiers in artificial intelligence* 4 (2021): 717899.



Prompt for Generation of Questions of Different Explanation Types

Prompt to generate questions to populate question bank

Generate 20 questions for Data Explanation for the Diabetes domain. Data Explanation questions are of the form,

- What is the data?
- How has the data been used in a particular decision?
- How has the data been used to train the ML model?

Your features could be:

- Glucose
- Diabetes Pedigree Function
- Age
- Sex
- Blood Pressure

The feature values range from:

- Diabetes Pedigree Function: 0.1 - 1
- Age 20 - 80
- Sex: Female or Male
- Blood Pressure: 90 - 200

For each question, also generate a machine interpretation which includes:

- predicate logic translation of the question,
- action mentioned in the question
- explanation type
- and the target variable in the question and if there is a low / high likelihood.

Prompt for Instruction fine-tuning of Llama to Generate Machine Parses

Instruction Prompt Example for Fine-tuning LLama to decompose question *uq* into response

Instruction:

Pretend that you are a machine who is trying to understand what action to take on a question. Your job is to determine what type of explanation to provide and how to provide that explanation. If there exist methods to produce the explanation, you must detect what the methods are. If the author of the question expresses an explicit preference for a particular explanation type, you must detect it.

You will be given a question and its interpretation for each task. The interpretation will include the machine interpretation - which is a predicate logic translation of the question, action in the question, explanation type that can address a question of this type, target variable in the question which is being enquired about and a likelihood of that variable. The options will be denoted by:

- Question:
- Machine interpretation:
- Action:
- Explanation type:
- Target variable:

You MUST respond with a valid Machine interpretation, action, explanation type, and target variable.

Task:

Question: How does the gender distribution affect the Diabetes prediction outcome?

Response:

- **Machine interpretation:** Summary(Sex, Diabetes)
- **Action:** Investigating the relationship between gender and Diabetes prediction.
- **Explanation type:** Data Explanation
- **Target variable:** Diabetes prediction, low likelihood.



Training details for fine-tuning

- Temperature, random seed = 0 for GPT-3.5 Turbo
- Fine-tuning of Llama2 and Llama3
 - 80:20 split
 - batch_size = 15

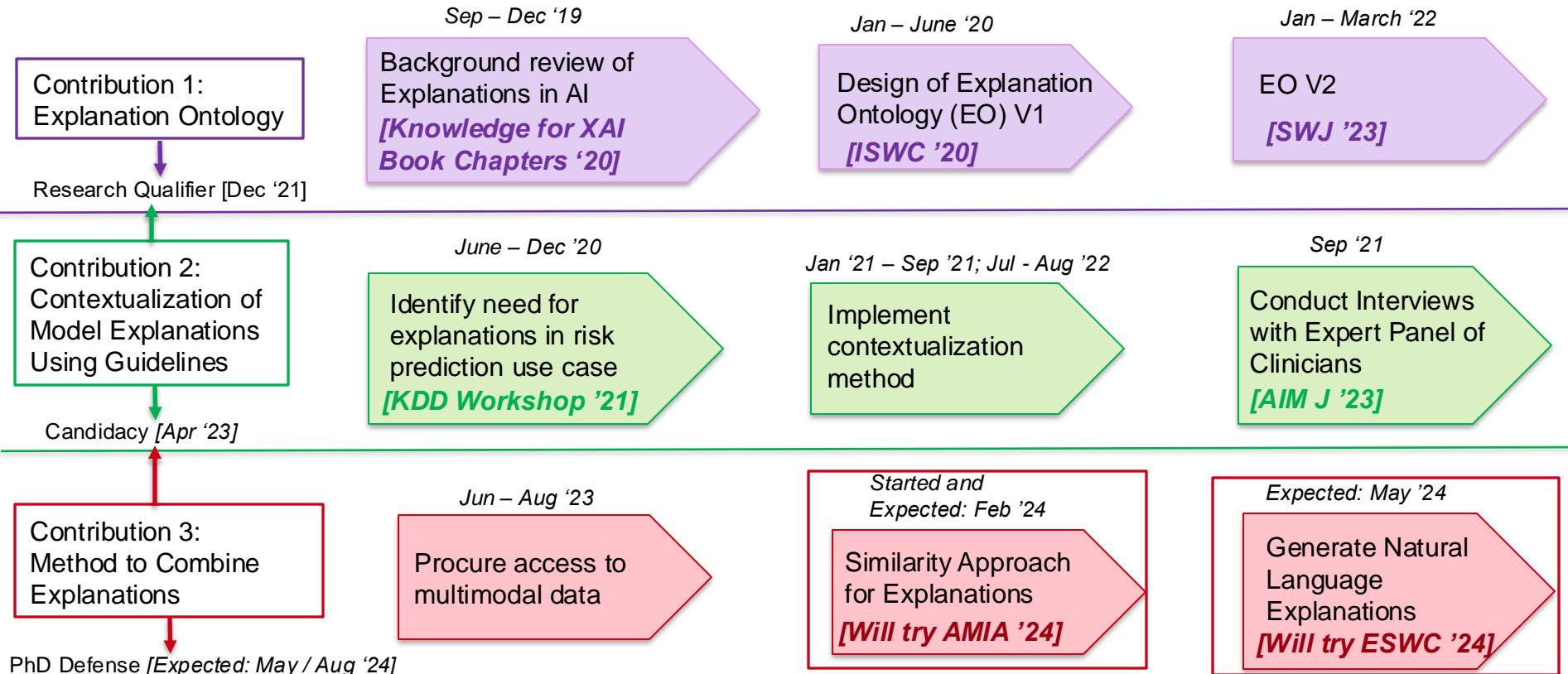
```
# Training Params
train_params = TrainingArguments(
    output_dir=codeconstants.OUTPUT_FOLDER + '/llm_results/' + self.refined_model_name + '_decompose_results_modified',
    num_train_epochs=12,
    per_device_train_batch_size= self.batch_size,
    gradient_accumulation_steps=1,
    optim="paged_adamw_32bit",
    save_steps=25,
    logging_steps=25,
    learning_rate=2e-4,
    weight_decay=0.001,
    fp16=False,
    bf16=False,
    max_grad_norm=0.3,
    max_steps=-1,
    warmup_ratio=0.03,
    group_by_length=True,
    lr_scheduler_type="constant",
    report_to="tensorboard"
)
```

Delegate parse statistics

Table 5.10: Summary of Delegate Parses

Parse stats for mode: Generated by fine-tuned LLama3 model	
Usable passes	221
Number of explanation types	
Explanation type	Count
Rationale Explanation	61
Data Explanation	55
Counterfactual Explanation	33
Contextual Explanation	28
Contrastive Explanation	25
Case Based Explanation	19
Length of unusable passes	58
Parse stats for mode: human-verified GPT generated	
Usable passes	279
Number of explanation types	
Explanation type	Count
Data Explanation	80
Case Based Explanation	60
Rationale Explanation	50
Contextual Explanation	35
Contrastive Explanation	29
Counterfactual Explanation	25
Length of unusable passes	0

Progress So Far and Remaining Tasks

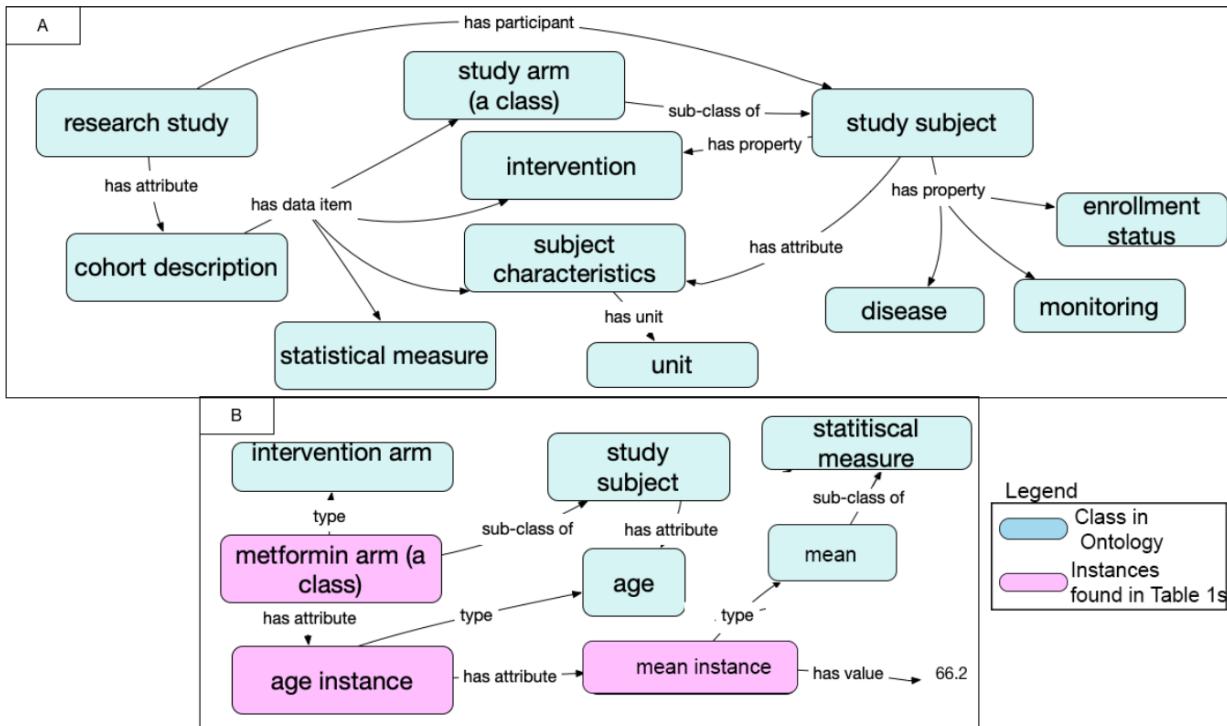


A Method to Represent Study Populations – mostly will shorten

Chari, S., Qi, M., Agu, N. N., Seneviratne, O., McCusker, J. P., Bennett, K. P., ... & McGuinness, D. L. (2019, October). Making study populations visible through knowledge graphs. In *International Semantic Web Conference* (pp. 53-68). Cham: Springer International Publishing.

Franklin, J. D., Chari, S., Foreman, M. A., Seneviratne, O., Gruen, D. M., McCusker, J. P., ... & McGuinness, D. L. (2020). Knowledge Extraction of Cohort Characteristics in Research Publications. In *AMIA Annual Symposium Proceedings* (Vol. 2020, p. 462). American Medical Informatics Association.

Study Cohort Ontology [1]



"Metformin is effective and safe, is inexpensive and widely available, and may reduce risk of cardiovascular events and death (89)" – from

1

89. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HA. 10-year follow-up of intensive glucose control in type 2 diabetes. *N Engl J Med* 2008;359:1577-1589
[Google Scholar](#) [Crossref](#) [PubMed](#)

Sample Table 1 from Clinical Trial PDF [2]

Sample characteristic ^a	Cases		Total (n=1708)	Controls Exposed (n=332)	Unexposed (n=1376)
	Cases (n=854)	Total (n=332)			
Diabetes (exposure) ^b	265 (31%)	332 (19%)	—	—	—
Demographics					
Male	325 (38%)	637 (37%)	73 (22%)	564 (41%)	63 (13%)
Age, years [mean (SD)]	69 (10.8)	63 (9.9)	64 (11.7)	—	—
18-40	34 (4%)	150 (9%)	27 (8%)	124 (9%)	—
41-60	111 (13%)	242 (14%)	50 (15%)	193 (14%)	—
61-80	589 (69%)	1244 (73%)	239 (72%)	1004 (73%)	—
81+	120 (14%)	72 (4%)	16 (5%)	55 (4%)	—
Education					
≤ High school	77 (9%)	165 (4%)	13 (4%)	151 (11%)	—
High school	325 (38%)	735 (43%)	116 (35%)	620 (45%)	—
Some college	367 (43%)	678 (40%)	170 (51%)	509 (37%)	—
≥ College	85 (10%)	130 (8%)	33 (10%)	96 (7%)	—
Insurance status					
Public	486 (57%)	926 (54%)	195 (59%)	729 (53%)	—
Private	248 (29%)	567 (33%)	100 (30%)	468 (34%)	—
None	120 (14%)	215 (13%)	37 (11%)	179 (13%)	—
Personal medical history					
CCI, median (min-max)	5 (0-15)	2 (0-10)	3 (0-10)	0 (0-7)	—
Heart failure	453 (53%)	404 (24%)	60 (18%)	344 (25%)	—
Atrial fibrillation	265 (31%)	238 (14%)	73 (22%)	165 (12%)	—
Hypertension	290 (34%)	375 (22%)	86 (26%)	289 (21%)	—
Pharmacologic agent use					
Sulfonylureas	538 (63%)	692 (40%)	183 (55%)	509 (37%)	—
Vasodilators	154 (18%)	195 (11%)	43 (13%)	151 (11%)	—
Diuretics	461 (54%)	728 (43%)	136 (41%)	592 (43%)	—
Beta blockers	239 (28%)	416 (24%)	113 (34%)	303 (22%)	—
Statins	325 (38%)	452 (27%)	123 (37%)	310 (24%)	—
NSAIDs	376 (44%)	731 (43%)	139 (42%)	592 (43%)	—

^aVariable distributions are reported as n (%) unless otherwise specified.

^bExposure distribution not reported for strata defined by exposure status.

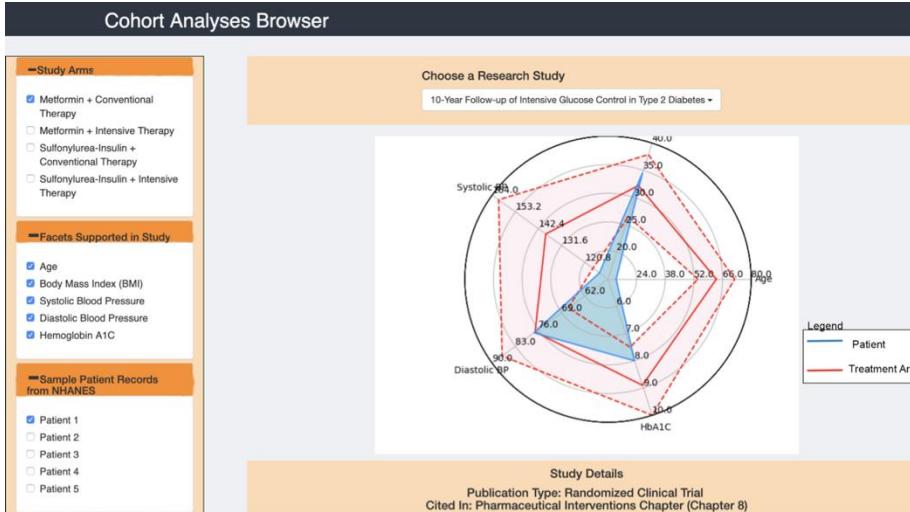
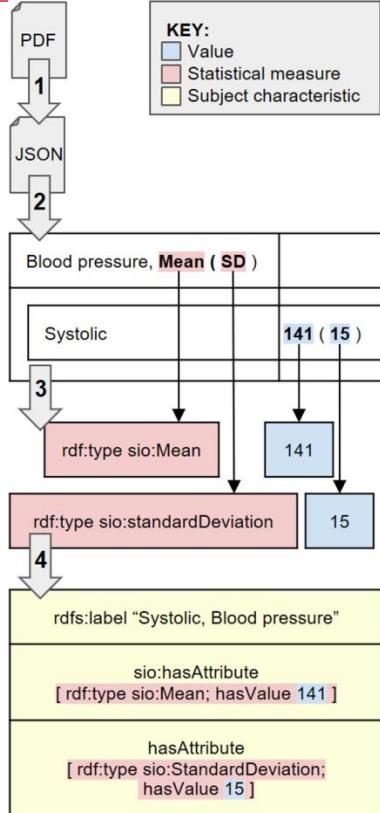
Abbreviations: CCI, Charlson Comorbidity Index; min, minimum; max, maximum; NSAID, non-steroidal anti-inflammatory drug; SD, standard deviation.

1. https://diabetesjournals.org/care/article/47/Supplement_1/S158/153955/9-Pharmacologic-Approaches-to-Glycemic-Treatment

2. Hayes-Larson, Eleanor, et al. "Who is in this study, anyway? Guidelines for a useful Table 1." *Journal of clinical epidemiology* 114 (2019): 125-132.



Study Cohort Extraction Pipeline and Analyses



- A). The knowledge extraction pipeline consists of four steps: 1) PDF-JSON conversion, 2) Creation of the tree table, 3) Annotating tokens with KG elements, and 4) Assembling a KG from annotated KG elements.
- B). Cohort Similarity Visualization that can compare patient to study





Rensselaer

why not change the world?®