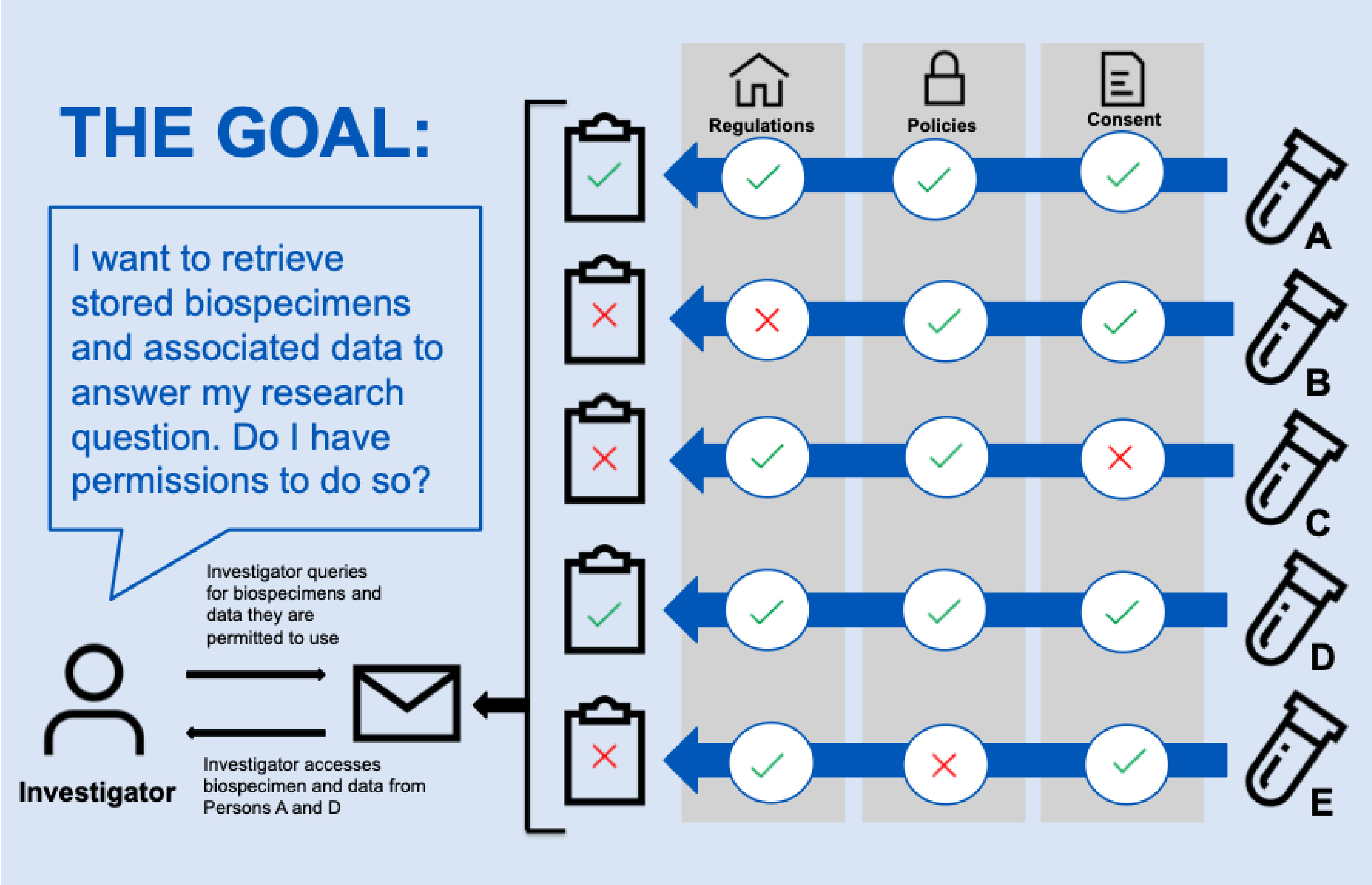
Evaluation of Semantic Models for Representing Biospecimen and Data Sharing Permissions in Biobank Consent Forms

Taylor B. Harrison, PhD[1,2], Emma Sandstrom[1], Elizabeth E. Umberfield, PhD, RN, Michelle L. McGowan, PhD[1,2], Cui Tao, PhD[1]

[1]Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA; [2]Biomedical Ethics Research Program, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA

OBJECTIVE



METHODS

ODRL
Open Digital Rights Language developed by the World Wide Web Consortium's (W3C)

FHIR
The HL7 Fast Healthcare Interoperability Resources (used by Epic & Oracle Health)

ICO
The Informed Consent Ontology (Community-based, initiated in 2012 at the University of Michigan)

DUO
The Data Use Ontology, endorsed by the Global Alliance for Genomics & Health (GA4GH)

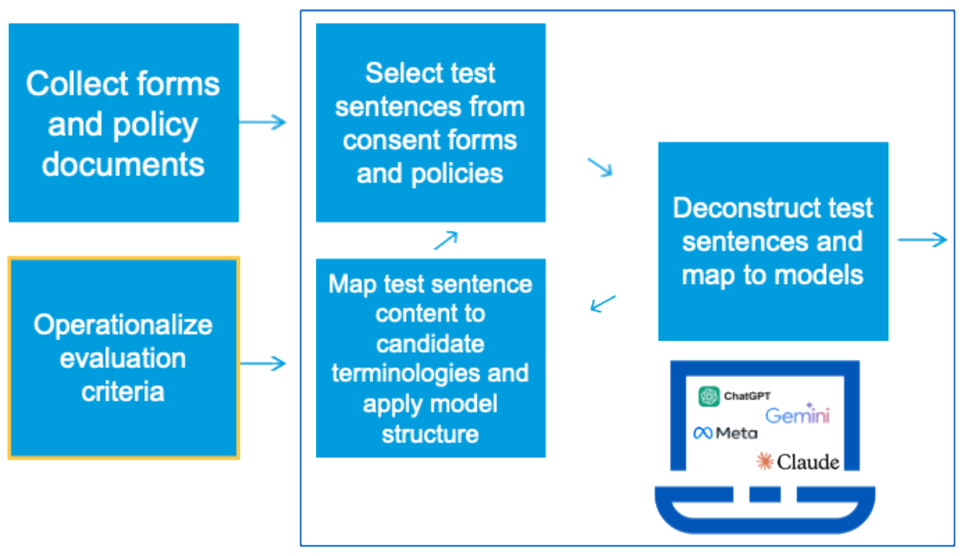
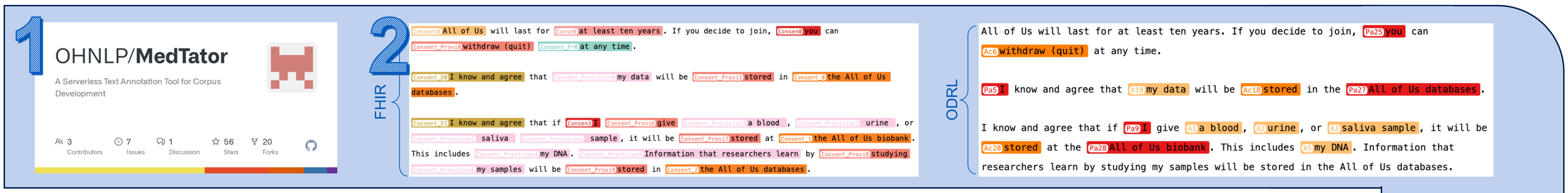


Table 1. Criteria used to evaluate each information model (ODRL, FHIR, ICO, DUO).

Evaluation Criteria	Conceptual Definition	Competency Question	Operational Definition
Consistency	When fitting test sentences into each model, contradictions or faulty meanings do not manifest.	Is the meaning of the permission or constraint in the test sentence expressible in the given model?	The number of expressible sentences/ total sentences.
Completeness	All concepts within the set of test sentences are accommodated by the model.	Are each of the individual concepts in the test sentence expressible in the given model?	The number of expressible concepts/ total concepts.
Conciseness	The model is free from irrelevant or redundant classes/concepts.	Are there classes/concepts in the model which are not used to express the test sentences from consent forms?	The number of unused classes/ concepts/total classes
Interoperability	The model is linked to controlled terminologies as response lists.	Does the model have recommended response lists (i.e.. value sets) that are standardized and conform to FAIR principles?	List of controlled terminologies that can be used to express concepts in the model.



"**Forward mapping**" is the annotation of real-world biobank consent forms using minimum language classified by selected information models.

67% of biobanks report significant underutilization of their biospecimens, with primary reasons being underdeveloped and unharmonized data management technologies.[1]

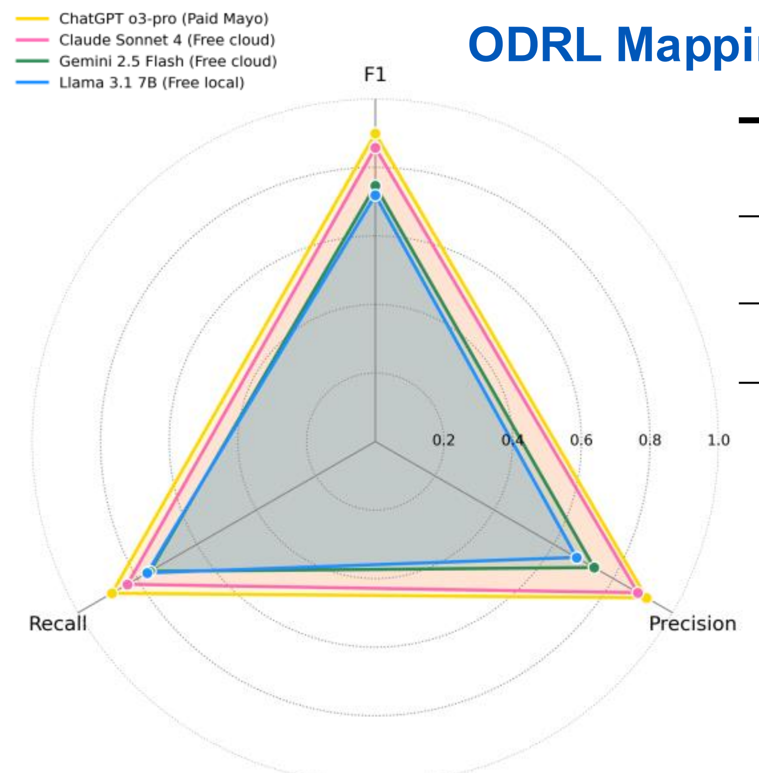
"**Backward mapping**" turns annotated text and information model labels into sentences that preserve the original meaning.

3 Permission Sentence (ODRL)

Party assigner: I	Action: stored
Asset: my data	Constraint: All of Us Databases

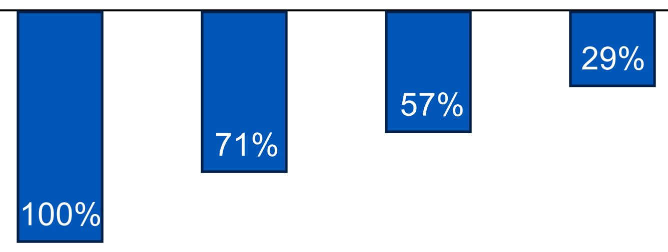
4 I know and agree that my data will be stored in the All of Us databases.

RESULTS



ODRL Mapping

	ChatGPT o3-pro (Paid Mayo)	Claude Sonnet 4 (Free cloud)	Gemini 2.5 Flash (Free cloud)	Llama 3.1 7B (Free local)
F1	**0.899**	0.857	0.746	0.719
Precision	**0.912**	0.882	0.735	0.677
Recall	**0.886**	0.833	0.758	0.767



Preliminary results of forward and backward mapping (n = 7) using zero-shot in-context learning prompts. Forward mapping performance is reported in the table and backward mapping accuracy is reported in the corresponding bar plot.

CONCLUSIONS

Our aim was to evaluate four established machine-interpretable models for expressing permissions, restrictions, and obligations for biospecimen and data sharing and reuse. Our preliminary results suggest that LLMs like o3-pro may help optimize biorepository resource usage, advance science, and improve health outcomes while simultaneously honoring specimen contributors' autonomous choices.

REFERENCE

1. iSpecimen. Widespread Underuse of Specimens Concerns Biobanks. ISBER Annual Meeting & Exhibits [Internet]. Dallas; 2018 [cited 2022 Sep 26]; Available from: https://www.ispecimen.com/news_items/widespread-underuse-specimens-concerns-biobanks-says-new-research/

Let's connect & collaborate: harrison.taylor@mayo.edu