AI generated ontologies and their impact on the quality of real-world data

Alexander Büsser, Exploris Health**
Matteo Togninalli, Isomorphic Labs*

February 12th, 2025

* Work conducted while at Visium
** Work conducted while at Idorisa

Is Dridorexant the only sleep medication to improve daytime functioning?

The Challenge of Clinical Trials

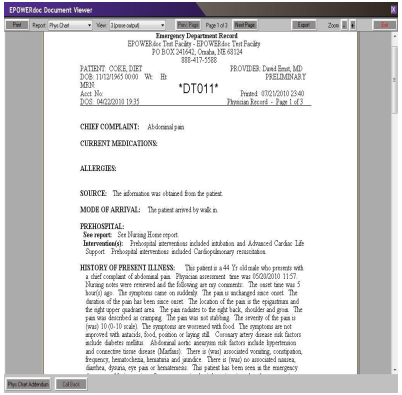
- **Clinical trials are the gold standard** for evaluating safety and efficacy.
- **Challenges:**
  - Expensive and time-consuming (mean cost of cardiovascular phase 3 trials: 157M USD[1])
  - Conducted in controlled settings, limiting generalizability, operationally complex
  - Slow to adapt to real-world clinical practice needs
- **Growing interest in Real-World Evidence (RWE):**
  - Uses EHRs, insurance claims, and other clinical data
  - Can supplement or replace some aspects of traditional trials

1: https://publichealth.jhu.edu/2018/cost-of-clinical-trials-for-new-drug-FDA-approval-are-fraction-of-total-tab

Benzodiazepine is associated with (statistically-) significant improvement in daytime functioning

Quality of Disease Representation is a key Barrier to Reproducibility

Physician Notes



EPOWERdoc Document Viewer

Print Report: Phys Chart View: 3 (prose output) Prev. Page Page 1 of 3 Next Page Export Zoom - + Exit

Emergency Department Record
EPOWERdoc Test Facility - EPOWERdoc Test Facility
PO BOX 241642, Omaha, NE 68124
888-417-5588

PATIENT: COKE, DIET PROVIDER: David Ernst, MD
DOB: 11/12/1965 00:00 Wt: Ht: PRELIMINARY
MRN:
DT011
Acct. No: Printed: 07/21/2010 23:40
DOS: 04/22/2010 19:35 Physician Record - Page 1 of 3

CHIEF COMPLAINT: Abdominal pain

CURRENT MEDICATIONS:

ALLERGIES:

SOURCE: The information was obtained from the patient.

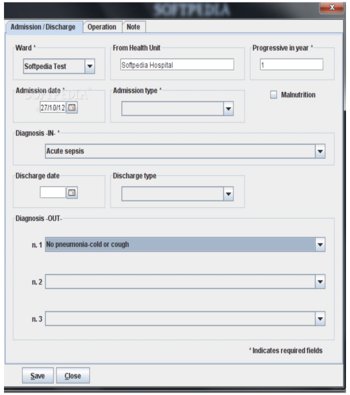
MODE OF ARRIVAL: The patient arrived by walk in.

PREHOSPITAL:
See report: See Nursing Home report.
Intervention(s): Prehospital interventions included intubation and Advanced Cardiac Life Support. Prehospital interventions included Cardiopulmonary resuscitation.

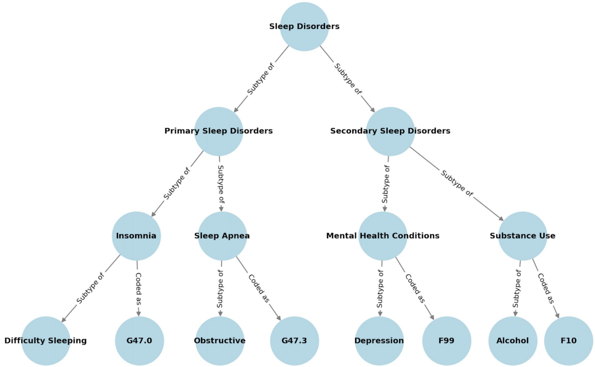
HISTORY OF PRESENT ILLNESS: This patient is a 44 Yr old male who presents with a chief complaint of abdominal pain. Physician assessment time was 05/20/2010 11:57. Nursing notes were reviewed and the following are my comments:. The onset time was 5 hour(s) ago. The symptoms came on suddenly. The pain is unchanged since onset. The duration of the pain has been since onset. The location of the pain is the epigastrium and the right upper quadrant area. The pain radiates to the right back, shoulder and groin. The pain was described as cramping. The pain was not stabbing. The severity of the pain is (was) 10 (0-10 scale). The symptoms are worsened with food. The symptoms are not improved with antacids, food, position or laying still. Coronary artery disease risk factors include diabetes mellitus. Abdominal aortic aneurysm risk factors include hypertension and connective tissue disease (Marfans). There is (was) associated vomiting, constipation, frequency, hematochezia, hematuria and jaundice. There is (was) no associated nausea, diarrhea, dysuria, eye pain or hematemesis. This patient has been seen in the emergency

Phys Chart Addendum Call Back

Claims

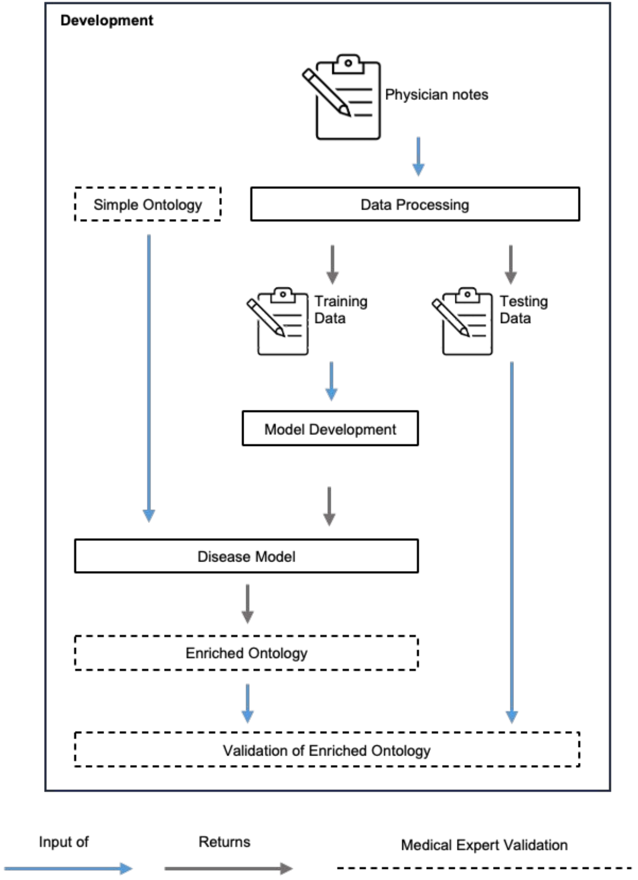


Disease Representation



Disease-Specific Medical Ontology Learning framework

Disease-Specific Medical Ontology Learning framework



Data

- Physician notes of 82,722 insomnia patients were used for this study from Amazing Charts LLC

Simple ontology

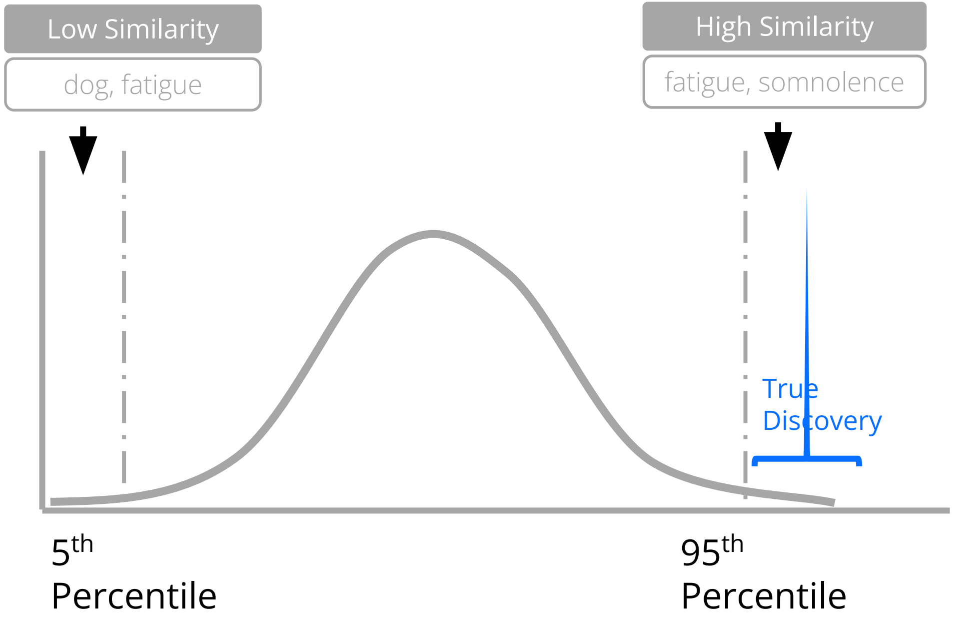
- The patient reported outcome measure Insomnia Daytime Symptoms and Impacts Questionnaire was used as the simple ontology

Model development end evaluation

- An ensemble of 8 word embeddings using word2vec was trained on different bootstrap samples to stabilize the concept extraction
- A evaluation metric "statistical power" was used to quantify the quality of the embedding

A Robust Evaluation for Medical Embeddings

Statistical power measures how well embeddings distinguish true medical relationships from random ones by testing their significance against a null distribution.



How statistical power works

1. Compute **cosine similarity** for known relationships.
2. Create a **null distribution** from **random medical concept pairs**.
3. Check if known relationships score **above the 95th percentile**.
4. **Power = % of true relationships detected above chance.**

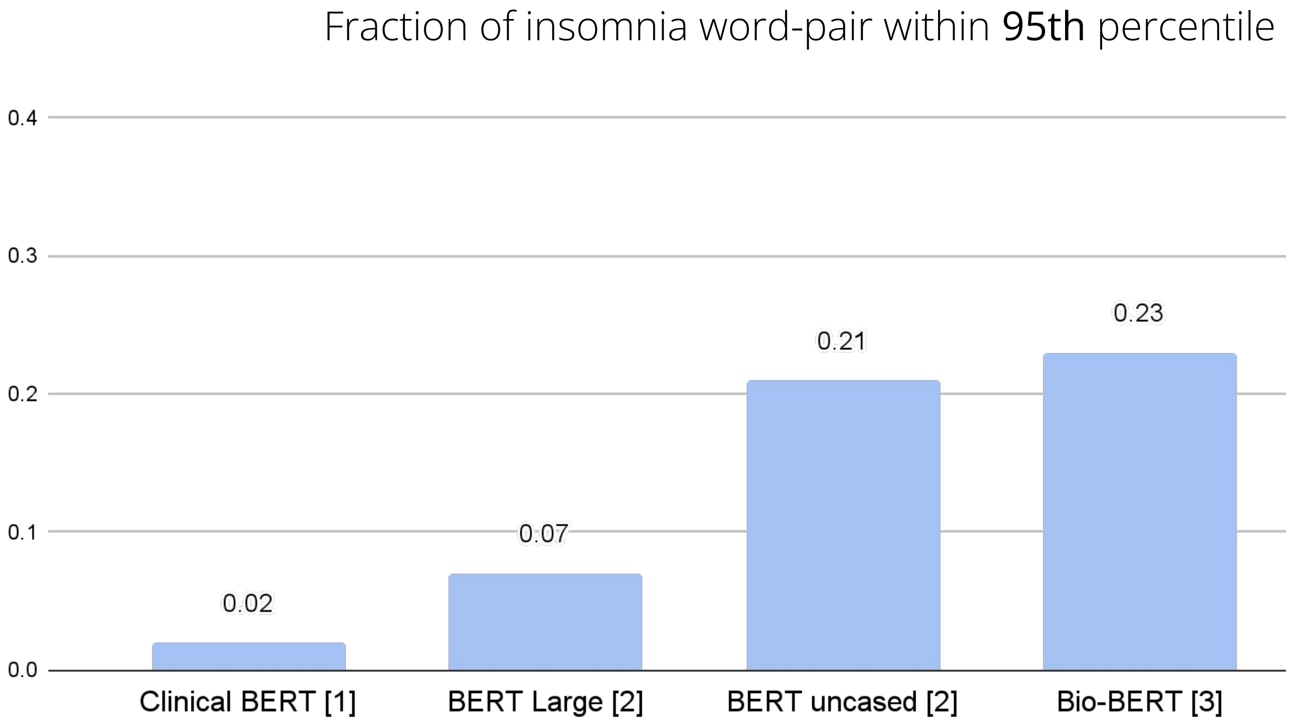
Key advantages

- **No need for complete ground truth** – handles missing relationships.
- **Encourages discovery** – does not penalize novel relationships.

EPFL AMLD

Evaluation of word embedding

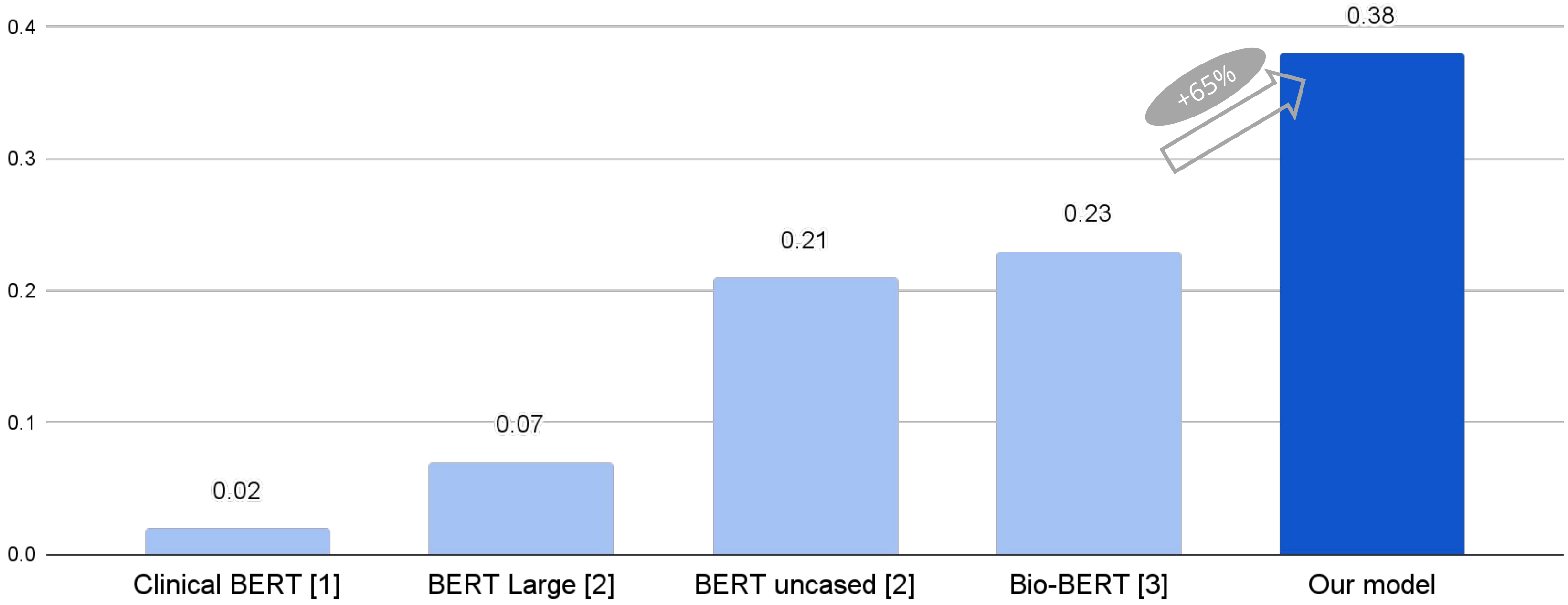
Language models performance



[1] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234-1240.
[2] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[3] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T. and McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Better than language models

Fraction of insomnia word-pair within **95th** percentile



[1] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234-1240.
[2] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[3] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T. and McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Why is a simple embedding better?

- These results were obtained pre-LLM
- However, off-the-shelf LLM have struggled to obtain relevant ontologies

Table 3. Zero-shot results across 11 LLMs and finetuned Flan-T5-Large and Flan-T5-XL LLMs results reported for ontology learning Task A i.e. term typing in MAP@1, and as F1-score for Task B i.e. type taxonomy discovery, and Task C i.e. type non-taxonomic relation extraction. The results are in percentages.

		Zero-Shot Testing											Finetuned	
Task	**Dataset**	**BERT-Large**	**PubMedBERT**	**BART-Large**	**Flan-T5-Large**	**Flan-T5-XL**	**BLOOM-1b7**	**BLOOM-3b**	**GPT-3**	**GPT-3.5**	**LLaMA-7B**	**GPT-4**	**Flan-T5-Large***	**Flan-T5-XL***
	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	**91.7**	81.4	90.1	76.9	**86.3**
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	**43.3**	16.9	18.4
A	*NCI*	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	**16.1**	31.9	**32.8**
	SNOMEDCT_US	21.1	28.5	19.8	24.3	31.6	37.0	**37.7**	24.4	25.0	13.8	27.8	33.4	**43.4**
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	**29.8**	25.7	23.9	4.9	23.7	38.4	**51.8**
	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	**67.8**	33.5	55.4	62.5	59.1
B	*UMLS*	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	**78.1**	53.4	**79.3**
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	**74.4**	33.8	74.3	91.7	**91.7**
C	*UMLS*	40.1	42.7	42.4	46.0	**49.5**	43.1	42.7	38.8	37.5	20.3	41.3	49.1	**53.1**

- In the case of clinical notes, specialized terminology was not captured well by older models
- These limitations can probably be overcome by reasoning models

[3] Babaei Giglou, H., D'Souza, J., & Auer, S. (2023, October). LLMs4OL: Large language models for ontology learning. In International Semantic Web Conference (pp. 408-427)

Measures of data quality

Claims data with clinical experts' input

Which diagnosis codes are relevant to day-time impairment?

Search for these codes in claims data

Clinical Expert

A123
B234
C567
D123
E234
F567
G567
H567
I123
...

EPFL AMLD

Claims with AI and clinical input

Apply AI on Physician Notes

xxxxxxxxxxxxxxxxxxxxxxxxx**stamina**
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxx**lively**xxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxx**afraid**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxx**agnosia**xxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Share with Clinical Experts

Which diagnosis codes are relevant to day-time impairment?

Clinical Expert

A123	Z567
B234	D567
C567	C567
D123	W123
E234	Q567
F567	Q567
G567	O567
H567	P123
I123	…
…	…
	…

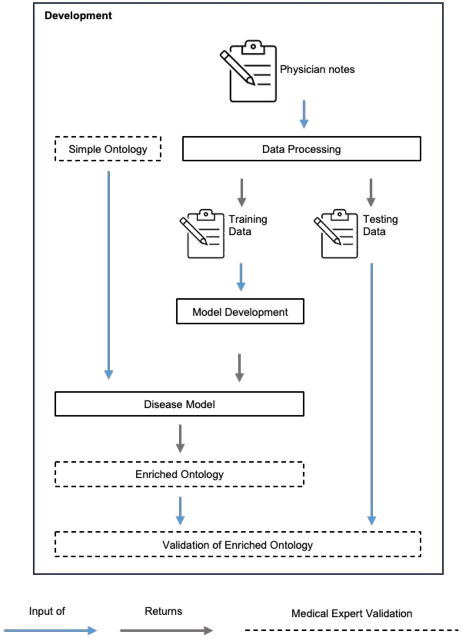
Search for these codes in claims data

More extensive selection of codes

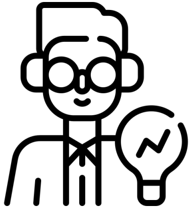
EPFL AMLD

Notes with AI and clinical input

AI generated ontology validated by Clinical Expert



Validate with Clinical Experts

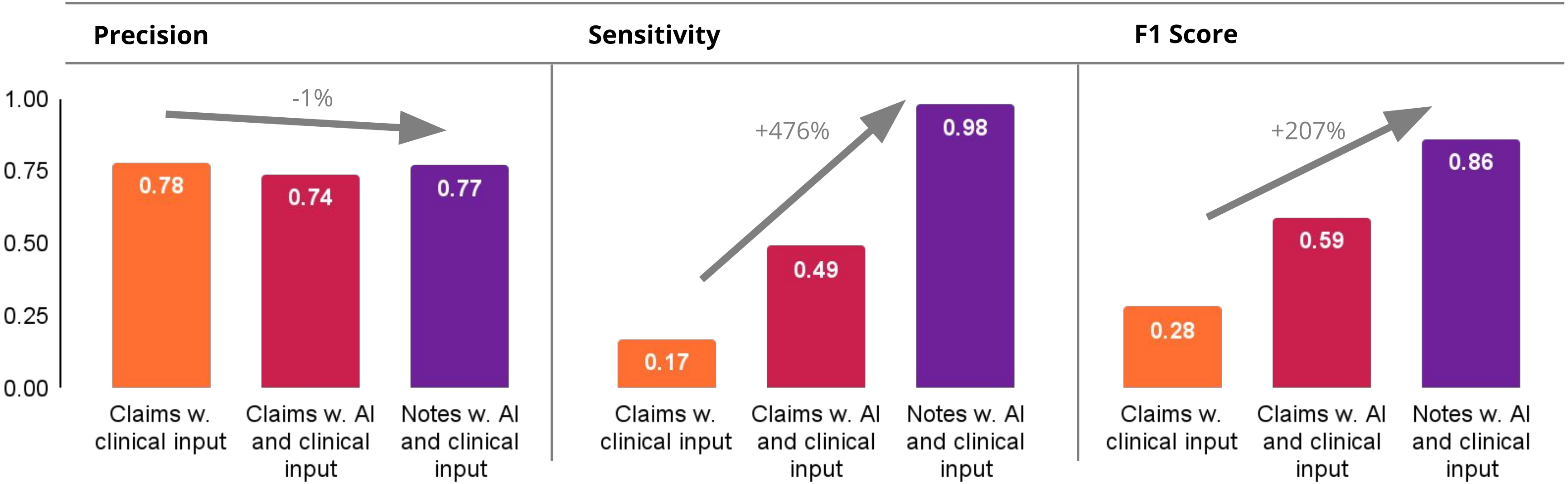


Clinical Expert

Extract from notes

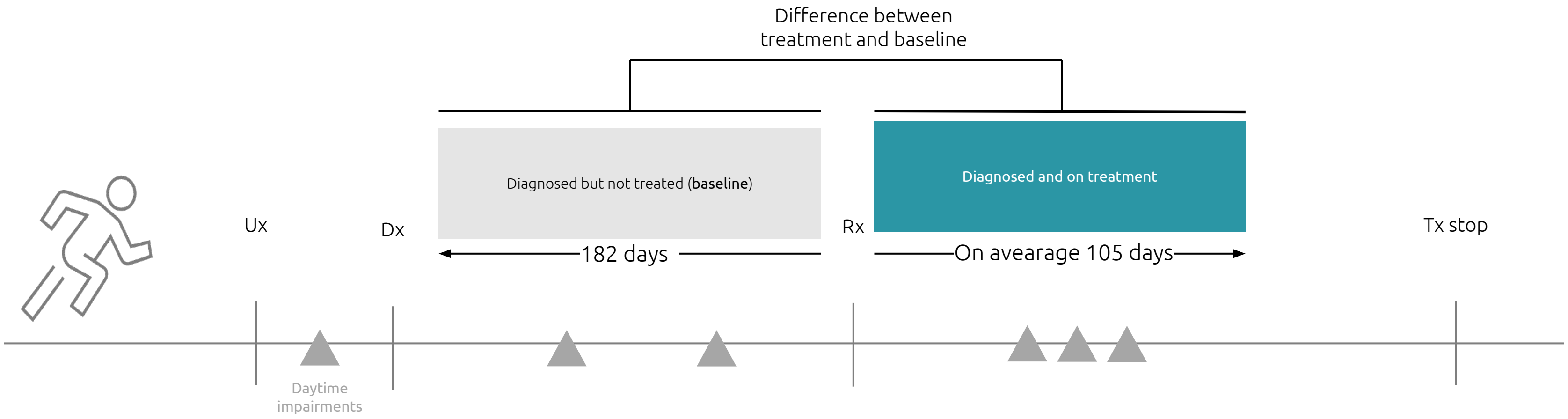
Apply AI on Physician Notes

The AI-derived ontology **significantly increases data quality** of secondary data sources



Real World Evidence implications

A within study design was chosen to control for static, non-time varying variables. The results **do not** allow for **statement of causality** between **treatments** and **daytime impairments**, only **association**



Benzo treated patients are associated with an increased rate of day-time impairments

Percentage increase of events per 100 patients per year [statistically sig. numbers in gray]
(within subject - untreated vs. Benzo treated insomnia period, n= 1045)

Daytime impairment events ontology

Daytime impairment
19%

Mood domain
26%

Alert/Co. domain
7%

Sleep domain
21%

EPFL AMLD

Daytime impairment events ontology

Percentage increase of events per 100 patients per year

[statistically sig. numbers in gray]

	Treated with Trazadone	Treated with Z Drug	Treated with Benzo
Daytime impairment	9 %	16 %	19 %
Mood domain	13 %	22 %	26 %
Alert/Co. domain	4 %	16 %	7 %
Sleep domain	7 %	12 %	21 %

EPFL AMLD

Conclusion

Conclusion

- We have demonstrated that we can learn a representation of a disease in an **unsupervised** manner that is reflective on how a condition is describe in **real-world** setting
- By doing so we have significantly increased the data quality and thus the reliability of resulting studies
- Extracting clinically relevant endpoints form textual data remains a difficult challenge. We believe the success of this project is rooted in …
  - setting realistic expectations on the program’s outcome
  - clear focus on one indication
  - strong cross-functional collaboration between AI experts and clinical specialists,
  - rigorous validation of the system’s reliability
- While traditional LLMs are not good at creating meaningful ontologies, more recent reasoning systems offer better capabilities in this space.

Acknowledgments

Experts involved and contributions to the Study

VISIUM	Development of AI, data pipelines	Renato Durrer, Tobias Ochsner Moritz Freidank, Thibault Viglino Matteo Togninalli
ETH	Overall methodological validation	Prof. Karsten Borgwardt, ETH
IBM	Validation environment setup	Vlad Zamfirescu
Medical College of Georgia, Augusta University 1828	Clinical validation	M.D. William Vaughn McCall
The University of Arizona	Clinical validation	M.D. Michael Grandner
idorsia	Clinical	Andrea Beyer
	Medical Affairs	Antonio Olivieri
	Value and Access	Paulien Meijer
	Communication	Andrew Jones, Agnes Lei
	GIS	Oliver Cotto, Thomas Straehl

Evaluation metrics

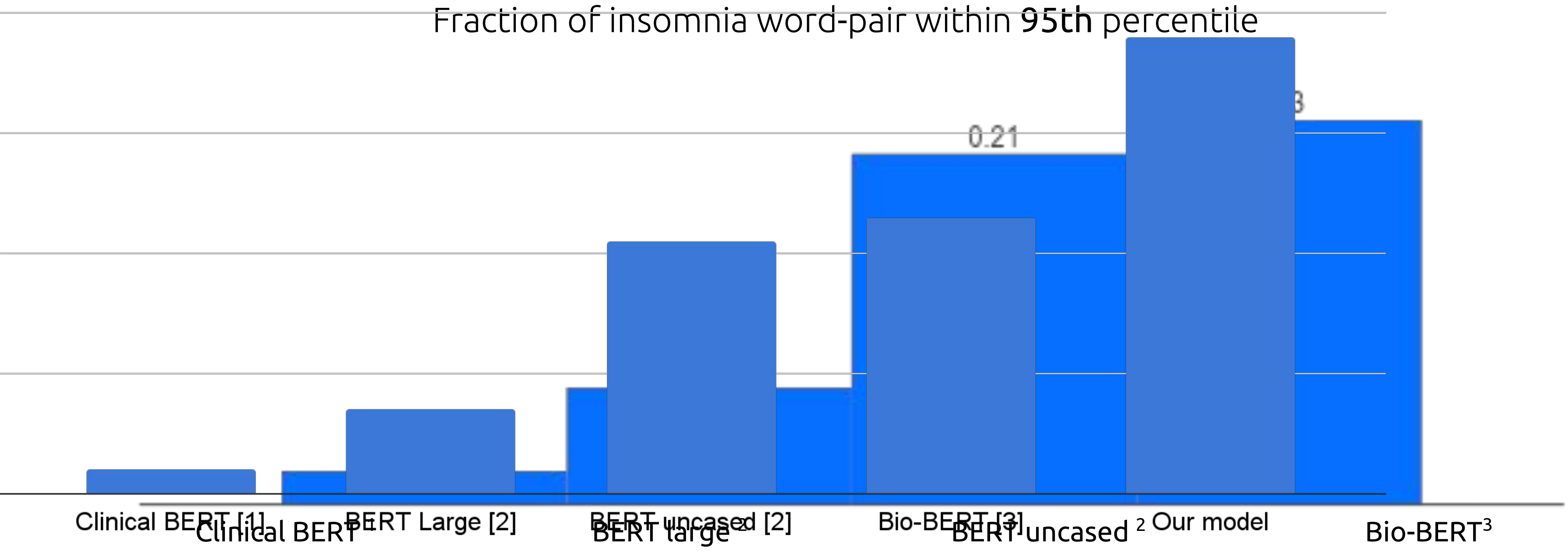
Sensitivity
Measures what fraction of people that have a condition are recognized as such.

Precision
Measures what fraction of people that are recognized to have a condition actually have it.

F1 Score
F1 is the harmonic mean between Sensitivity and precision.

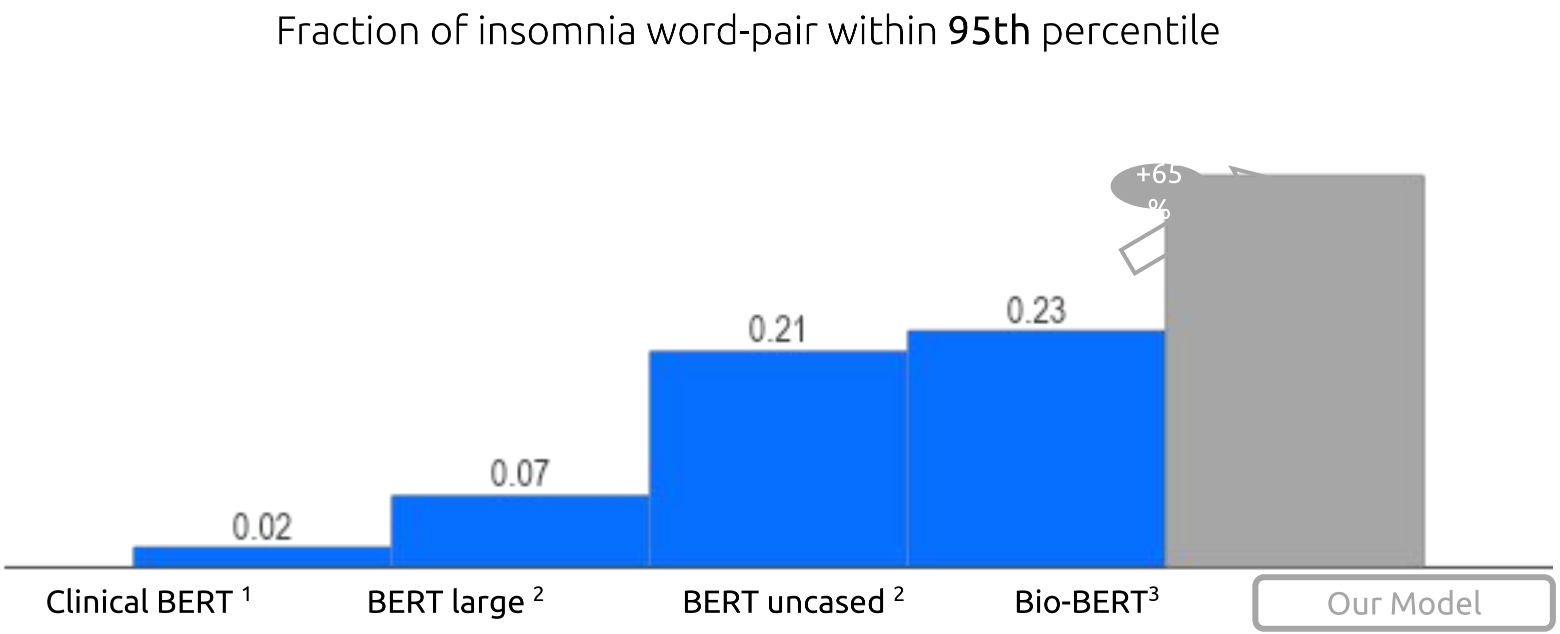
A high F1 score corresponds to good **data quality**

Language models performance



[1] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234-1240.
[2] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[3] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T. and McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

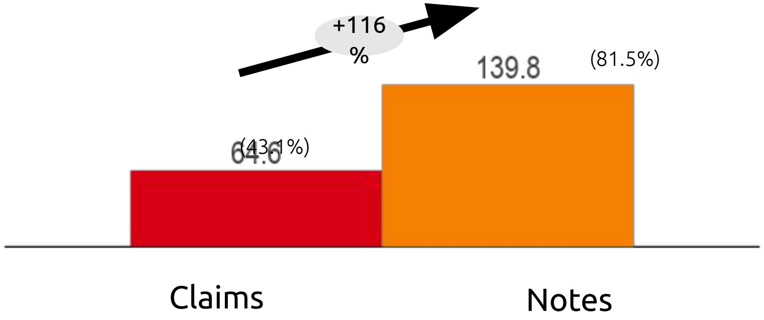
Better than State-of-the-art language models



[1] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234-1240.
[2] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[3] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T. and McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

The prevalence of daytime impairments has been strongly underestimated in claims data

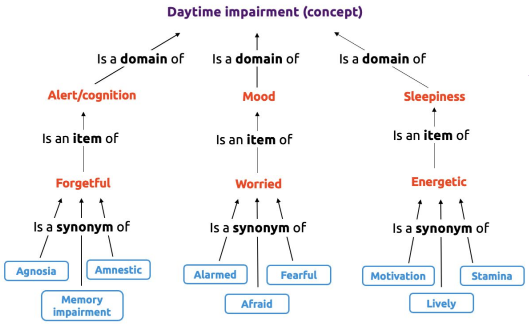
The number of daytime impairment in **100 patient years** has been calculated for a fixed insomnia population



Notes with AI and clinical input

AI generated ontology validated by Clinical Expert

Apply AI on Physician Notes

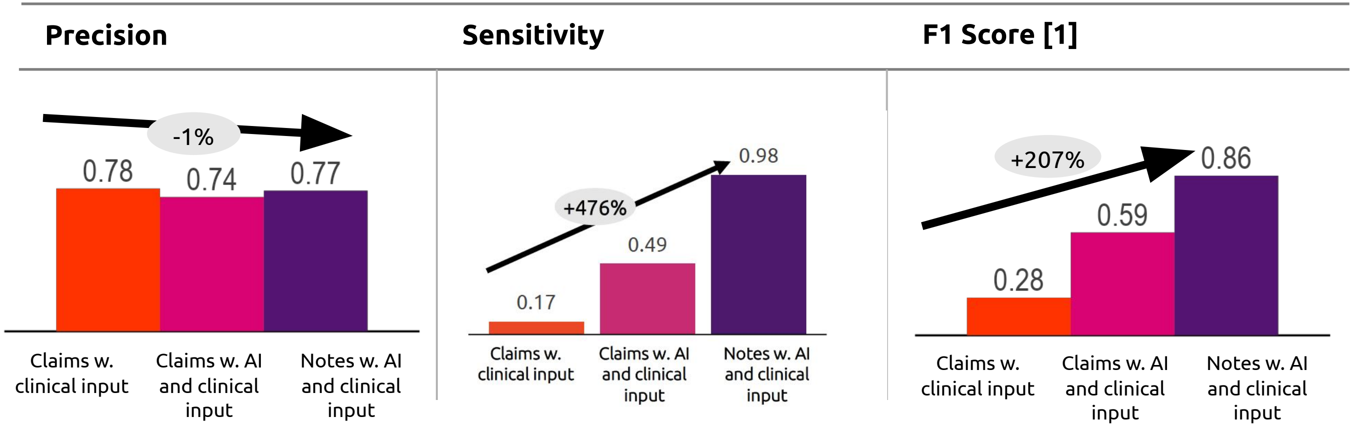


Validate with Clinical Experts

Clinical Expert

Extract from notes

The AI-derived ontology significantly increase data quality of secondary data sources



[1] Van Rijsbergen, C. J. (1979) Information Retrieval. London: Butterworths