Extracting Social Determinants of Health from Pediatric Patient Notes Using Large Language Models: Novel Corpus and Methods

LREC-COLING, 2024

Authors: Yujuan Velvin Fu*, Giridhar Kaushik Ramachandran*, Nicholas J Dobbins, Namu Park, Michael Leu, Abby R. Rosenberg, Kevin Lybarger, Fei Xia, Özlem Uzuner, Meliha Yetisgen

Presenter: Yujuan Velvin Fu

Outline

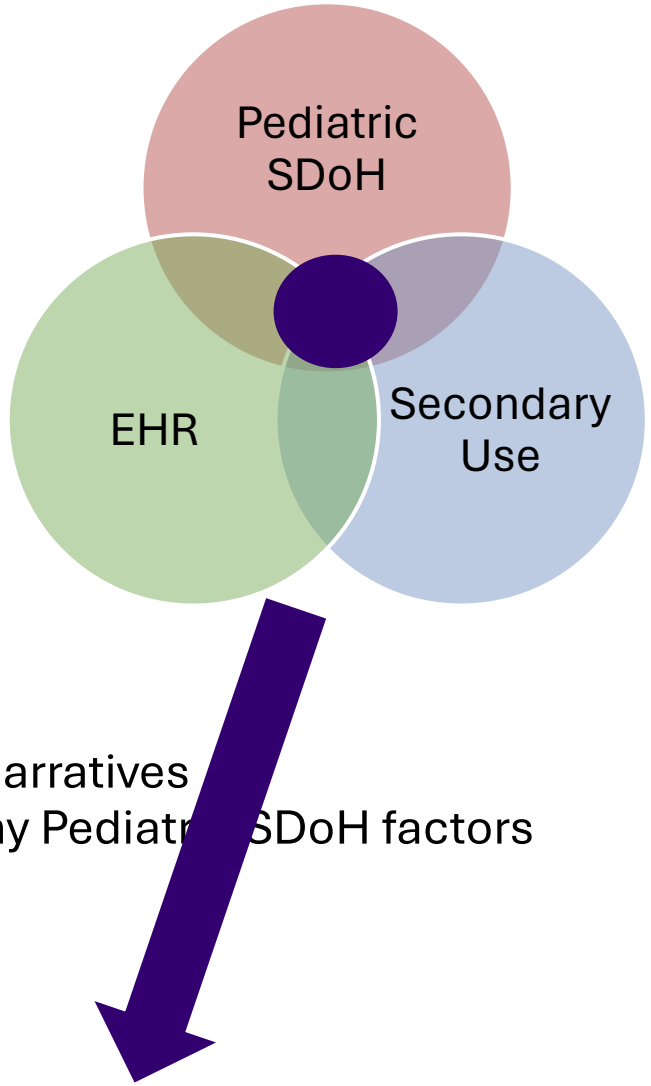
- Background
- Goal
- Related work
- Corpus development
- Information extraction (IE) approaches
- IE performance and error analysis
- Conclusion & future directions

Background

- Pediatric Social Determinants of Health (SDoH)
  - Conditions in which children born, grow, and live
  - Social, behavioral, and environmental factors
  - ➢ Knowledge of Pediatric SDoH can inform patient care
  - ➢ Long-term impact for pediatric patients

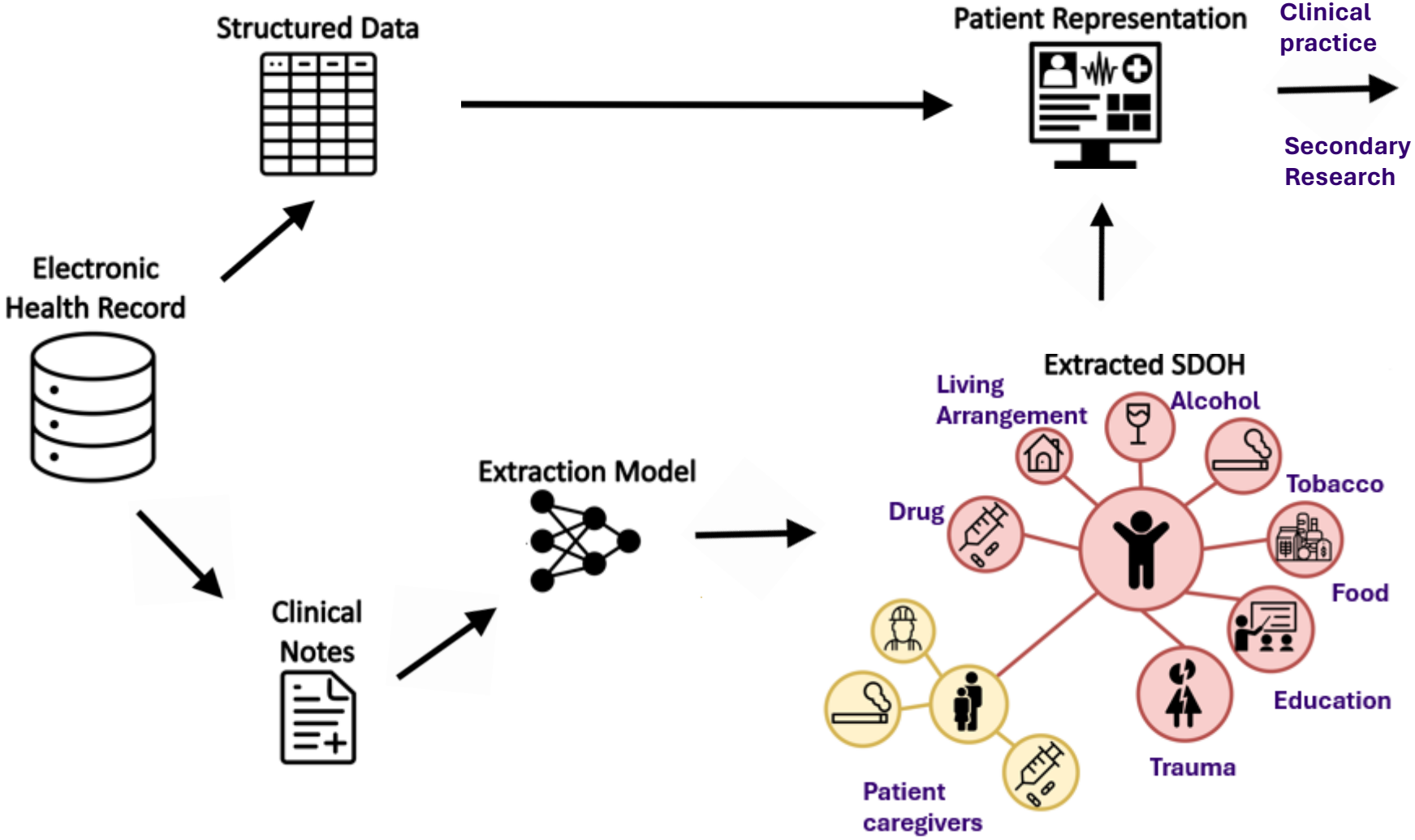
- Electronic health record (EHR)
  - Contains both structured and unstructured patient information
  - Pediatric SDoH are primarily documented in unstructured clinical narratives
  - Clinical texts contains nuanced and detailed representation of many Pediatr SDoH factors

- Healthcare data - secondary use applications
  - Real-time clinical decision-support
  - Large-scale retrospective studies



For secondary use of Pediatric SDoH from EHR, unstructured text descriptions must be mapped to a structured representation (normalization)

Motivation



Pediatric Mental & Physical Health

Related work

SDoH corpora

- Focus on a singular SDoH factor, such as substance use[1-4] ,homelessness [5-6], adverse childhood experiences from adults [6-7]
- SDoH corpora under different contexts, such as adult population (2022 n2c2 shared task) [8], sexual health [9] and hospital readmission rate [10]
- Lack comprehensive, and fine-grained SDoH corpus for pediatric patients

IE methods for SDoH

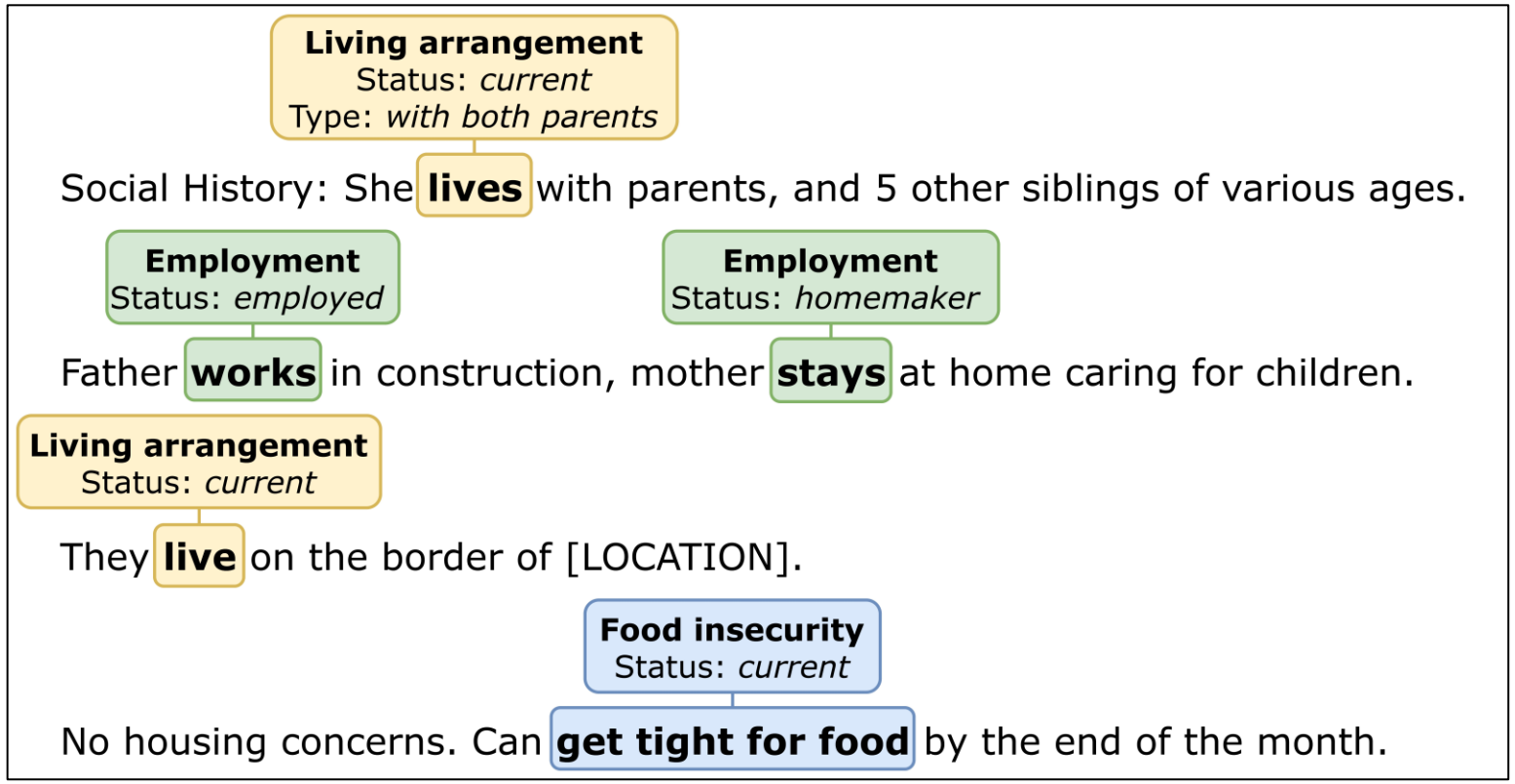
- Rule-, machine-learning- and BERT-based models [11-18]
- GPT-4 in-context learning for clinical IE[19-20]
- Limited exploration of generative large language models (LLMs) with different learning strategies, such as fine-tuning and prompt engineering.

Pediatric Social History Annotation Corpus (PedSHAC)

Pediatric Population
- Patients under 18 years old
- 10-year period
- 198k distinct notes
- 36k distinct patients
- University of Washington (UW)

PedSHAC annotated data
- **1,260** social history sections
- **10** Pediatric SDoH events

Pediatric SDoH events
- Trigger span
- Labeled arguments: normalization



Example social history section from EHR with Pediatric SDoH events

PedSHAC

Dataset

- Train/valid/test: 894, 121, and 245 social history sections

SDoH event evaluation

- Trigger: span overlap (relaxed)
- Labeled arguments: label-only (normalization)
- Same as 2022 n2c2 SDoH challenge

Inter-annotator agreement (IAA)

- Double annotation on the test set + 96% validation set.

Event	Trigger & Arg.	Trigger examples & Argument subtypes	# labels			IAA F1
			Train	Validation	Test	
Adoption	Trigger	"adopted", ...	27	4	9	
Education Access	Trigger	"5th grade" , "junior year", ...	227	35	74	
	Status	(yes,no)	227	35	74	
Employment	Trigger	"Employment: ... ", "works", ...	390	45	117	
	Status	(employed, unemployed, retired, on disability, student, homemaker)	390	45	117	
Food Insecurity	Trigger	"food stamps", "food insecurity", ...	37	5	8	
	Status	(current, past, none)	37	5	8	
Living Arrangement	Trigger	"lives", "foster care", ...	676	101	195	
	Status	(current, past, future)	676	101	195	
	Type*	(with both parents, with single parent, with other relatives, with foster family, with strangers)	566	86	160	
	Residence*	(home, institution, homeless)	136	22	38	
Mental Health	Trigger	"depression", "self-harm", ...	45	11	15	
	Status	(current, past, none)	45	11	15	
	Experiencer	(patient, parent/caregiver)	45	11	15	
Substance Use - Alcohol / Drug / Tobacco	Trigger	"meth", "alcohol", "smokes",...	265	38	78	
	Status	(current, past, none)	265	38	78	
	Experiencer	(patient, parent/caregiver)	265	38	78	
Trauma	Trigger	"mentally abusive", "bullying", ...	132	23	33	
	Status	(yes, no)	132	23	33	
	Type	(divorce / separation, loss, psychological, physical, domestic violence, sexual)	132	23	33	

PedSHAC

Dataset

- Train/valid/test: 894, 121, and 245 social history sections

SDoH event evaluation

- Trigger: span overlap (relaxed)
- Labeled arguments: label-only (normalization)
- Same as 2022 n2c2 SDoH challenge

Inter-annotator agreement (IAA)

- Double annotation on the test set + 96% validation set.

Event	Trigger & Arg.	Trigger examples & Argument subtypes	# labels			IAA
			Train	Validation	Test	F1
Adoption	Trigger	"adopted", ...	27	4	9	100.0
Education Access	Trigger	"5th grade" , "junior year", ...	227	35	74	80.0
	Status	(yes,no)	227	35	74	80.0
Employment	Trigger	"Employment: ... ", "works", ...	390	45	117	81.1
	Status	(employed, unemployed, retired, on disability, student, homemaker)	390	45	117	77.8
Food Insecurity	Trigger	"food stamps", "food insecurity", ...	37	5	8	40.0
	Status	(current, past, none)	37	5	8	40.0
Living Arrangement	Trigger	"lives", "foster care", ...	676	101	195	90.4
	Status	(current, past, future)	676	101	195	88.5
	Type*	(with both parents, with single parent, with other relatives, with foster family, with strangers)	566	86	160	88.4
	Residence*	(home, institution, homeless)	136	22	38	38.1
Mental Health	Trigger	"depression", "self-harm", ...	45	11	15	66.7
	Status	(current, past, none)	45	11	15	53.3
	Experiencer	(patient, parent/caregiver)	45	11	15	66.7
Substance Use - Alcohol / Drug / Tobacco	Trigger	"meth", "alcohol", "smokes",...	265	38	78	86.4
	Status	(current, past, none)	265	38	78	85.7
	Experiencer	(patient, parent/caregiver)	265	38	78	73.2
Trauma	Trigger	"mentally abusive", "bullying", ...	132	23	33	88.9
	Status	(yes, no)	132	23	33	88.9
	Type	(divorce / separation, loss, psychological, physical, domestic violence, sexual)	132	23	33	84.6

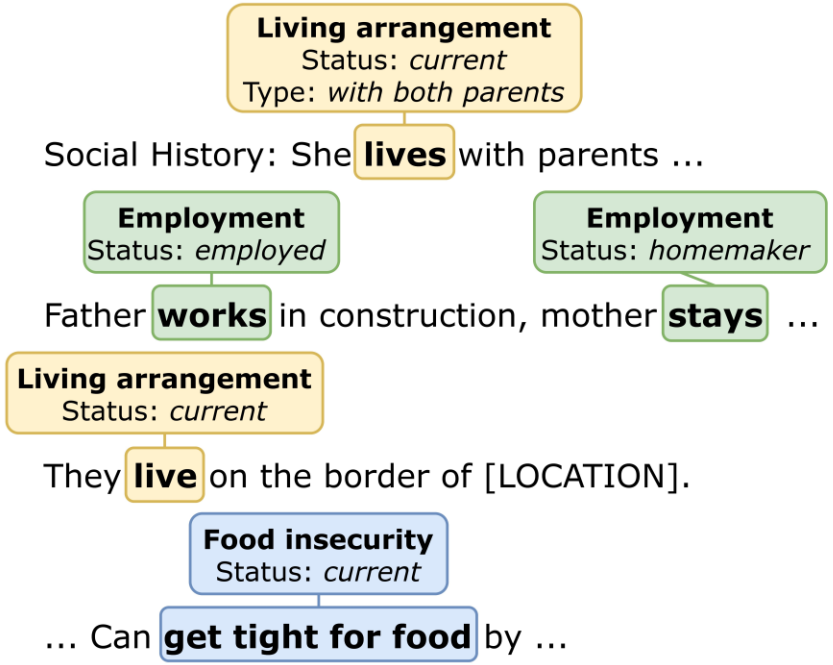
Pediatric SDoH Information Extraction (IE)

	Encoder-only LM	Generative LM
Fine tuning	**mSpERT** (Lybarger et al.,2023) • Multi-label span classification • Relation prediction omitted	**Flan-T5** **Prompting strategies** Single-step event extraction (Event) Two-step question-answering (2sQA)

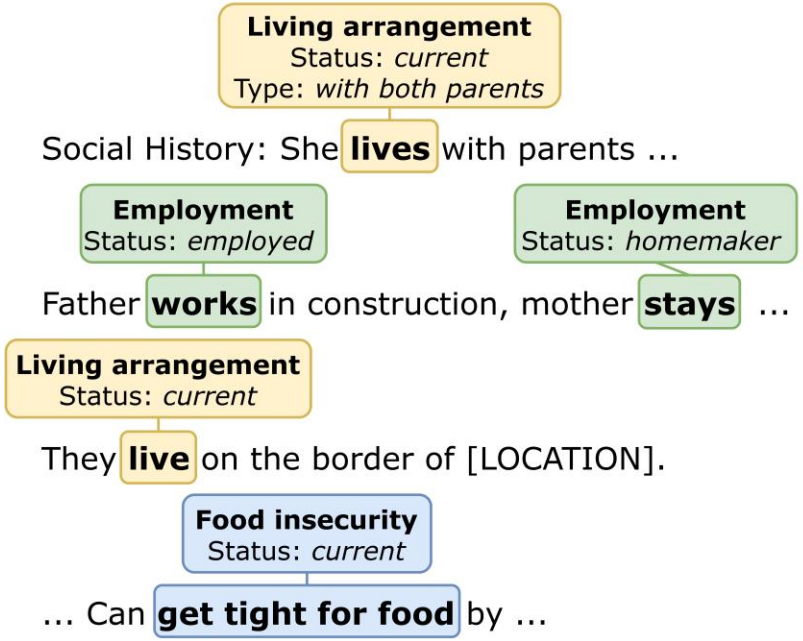
In-context learning		**GPT-4**
		• In-context learning
		  • Guideline summary
		  • Few-shot examples
		• HIPAA-compliant Azure instance

Prompting Strategy – *Event*

- Adapted from Romanowski et al., 2023[21]



Prompting Strategy – *2sQA*



IE performance

Event	Trigger & Argument	# Gold	F1						
			mSpERT	Flan-T5 -Event	Flan-T5 -2sQA	GPT-Event	GPT -2sQA	GPT -2sQA +guideline	GPT -2sQA +guideline +few-shot
Micro Avg.	Trigger	529	0.80	0.80	0.81	0.70	0.71	0.80†	**0.82**†
	Arguments	844	0.75	0.76	**0.78***	0.62	0.60	0.70†	0.72^{i}

Prompting

- No significant difference between *Event* and *QA*

Significance

- pairwise non-parametric bootstrap test ($p<0.05$)
- * indicates > mSpERT
- † in-context learning > GPT-QA

IE performance

Event	Trigger & Argument	# Gold	F1						
			mSpERT	Flan-T5 -Event	Flan-T5 -2sQA	GPT-Event	GPT -2sQA	GPT -2sQA +guideline	GPT -2sQA +guideline +few-shot
Micro Avg.	Trigger	529	0.80	0.80	0.81	0.70	0.71	0.80†	**0.82**†
	Arguments	844	0.75	0.76	**0.78***	0.62	0.60	0.70†	0.72†

Prompting

- No significant difference between *Event* and *QA*

Significance

- pairwise non-parametric bootstrap test ($p<0.05$)
- * indicates > mSpERT
- † in-context learning > GPT-QA

IE performance

Event	Trigger & Argument	# Gold	F1						
			mSpERT	Flan-T5 -Event	Flan-T5 -2sQA	GPT-Event	GPT -2sQA	GPT -2sQA +guideline	GPT -2sQA +guideline +few-shot
Micro Avg.	Trigger	529	0.80	0.80	0.81	0.70	0.71	0.80†	**0.82**†
	Arguments	844	0.75	0.76	**0.78***	0.62	0.60	0.70†	0.72†

Prompting

- No significant difference between *Event* and *QA*

Significance

- pairwise non-parametric bootstrap test ($p<0.05$)
- * indicates > mSpERT
- † in-context learning > GPT-QA

IE performance

Event	Trigger & Argument	# Gold	F1						
			mSpERT	**Flan-T5 -Event**	**Flan-T5 -2sQA**	**GPT-Event**	**GPT -2sQA**	**GPT -2sQA +guideline**	**GPT -2sQA +guideline +few-shot**
Micro Avg.	Trigger	529	0.80	0.80	0.81	0.70	0.71	0.80†	**0.82†**
	Arguments	844	0.75	0.76	**0.78***	0.62	0.60	0.70†	0.72†

Prompting

- No significant difference between *Event* and *QA*

Significance

- pairwise non-parametric bootstrap test ($p<0.05$)
- * indicates > mSpERT
- † in-context learning > GPT-QA

IE performance

Event	Trigger & Argument	# Gold	F1						
			mSpERT	**Flan-T5 -Event**	**Flan-T5 -2sQA**	**GPT-Event**	**GPT -2sQA**	**GPT -2sQA +guideline**	**GPT -2sQA +guideline +few-shot**
Micro Avg.	Trigger	529	0.80	0.80	0.81	0.70	0.71	0.80†	**0.82**†
	Arguments	844	0.75	0.76	**0.78***	0.62	0.60	0.70†	0.72†

Prompting

- No significant difference between *Event* and *QA*

Significance

- pairwise non-parametric bootstrap test ($p<0.05$)
- * indicates > mSpERT
- † in-context learning > GPT-QA

Error analysis

Frequent vs. Infrequent event types

Fine-tuned models

- High precision
- Low recall: poor generalization

In-context learning

- Low precision: especially for living arrangements, with false positives such as
  - ❑ “Dad </name>, Mom </name> ”
- High recall: great generalization for infrequent subtypes.

Event	Trigger & Argument	# Gold	F1 mSpERT	F1 Flan-T5 -2sQA	F1 GPT -2sQA +guideline +few-shot
Adoption	Trigger	9	**0.84**	**0.84**	0.55
Education Access	Trigger	74	0.78	0.84	**0.86**†
	Status	74	0.78	0.84	0.85†
Employment	Trigger	117	0.75	0.81	**0.89***†
	Status	117	0.71	0.74	**0.81***†
Food Insecurity	Trigger	8	**0.93**	**0.93**	0.88
	Status	8	**0.93**	**0.93**	0.88
Living Arrangement	Trigger	195	0.85	0.85	0.84
	Status	195	0.83	**0.84**	0.78
	Type	160	0.83	**0.89***	0.78
	Residence	38	0.64	0.62	0.29
Mental Health	Trigger	15	0.38	0.36	0.52
	Status	15	0.29	0.35	**0.44**
	Experiencer	15	0.10	0.17	**0.44***
Substance Use	Trigger	78	**0.86***	0.82	0.80†
	Status	78	0.81	**0.82**	0.77†
	Experiencer	78	0.75	**0.81**	0.80†
Trauma	Trigger	33	0.62	0.53	**0.70**
	Status	33	0.52	0.54	**0.63**
	Type	33	0.55	0.54	**0.67**
Micro Avg.	Trigger	529	0.80	0.81	**0.82**†
	Arguments	844	0.75	**0.78***	0.72†

Error analysis

Frequent vs. Infrequent event types

Fine-tuned models

- High precision
- Low recall: poor generalization

In-context learning

- Low precision: especially for living arrangements, with false positives such as
  - ❑ “Dad </name>, Mom </name> ”
- High recall: great generalization for infrequent subtypes.

Event	Trigger & Argument	# Gold	F1 mSpERT	F1 Flan-T5 -2sQA	F1 GPT -2sQA +guideline +few-shot
Adoption	Trigger	9	**0.84**	**0.84**	0.55
Education Access	Trigger	74	0.78	0.84	**0.86**†
	Status	74	0.78	0.84	0.85†
Employment	Trigger	117	0.75	0.81	**0.89***†
	Status	117	0.71	0.74	**0.81***†
Food Insecurity	Trigger	8	**0.93**	**0.93**	0.88
	Status	8	**0.93**	**0.93**	0.88
Living Arrangement	Trigger	195	0.85	0.85	0.84
	Status	195	0.83	**0.84**	0.78
	Type	160	0.83	0.**89***	0.78
	Residence	38	0.64	0.62	0.29
Mental Health	Trigger	15	0.38	0.36	0.52
	Status	15	0.29	0.35	**0.44**
	Experiencer	15	0.10	0.17	**0.44***
Substance Use	Trigger	78	**0.86***	0.82	0.80†
	Status	78	0.81	**0.82**	0.77†
	Experiencer	78	0.75	**0.81**	0.80†
Trauma	Trigger	33	0.62	0.53	**0.70**
	Status	33	0.52	0.54	**0.63**
	Type	33	0.55	0.54	**0.67**
Micro Avg.	Trigger	529	0.80	0.81	**0.82**†
	Arguments	844	0.75	**0.78***	0.72†

Error analysis

Challenges for both models

- Distinguishing past and current events
  - ❑ “Lived with grandmom. Now dad.”
- Implicit reasoning
  - ❑ “Father has him 3 days a week. Live with Mom in other time.”

Event	Trigger & Argument	# Gold	F1		
			mSpERT	Flan-T5 -2sQA	GPT -2sQA +guideline +few-shot
Adoption	Trigger	9	**0.84**	**0.84**	0.55
Education Access	Trigger	74	0.78	0.84	**0.86**†
	Status	74	0.78	0.84	0.85†
Employment	Trigger	117	0.75	0.81	**0.89***†
	Status	117	0.71	0.74	**0.81***†
Food Insecurity	Trigger	8	**0.93**	**0.93**	0.88
	Status	8	**0.93**	**0.93**	0.88
Living Arrangement	Trigger	195	0.85	0.85	0.84
	Status	195	0.83	**0.84**	0.78
	Type	160	0.83	0.**89***	0.78
	Residence	38	0.64	0.62	0.29
Mental Health	Trigger	15	0.38	0.36	0.52
	Status	15	0.29	0.35	**0.44**
	Experiencer	15	0.10	0.17	**0.44***
Substance Use	Trigger	78	**0.86***	0.82	0.80†
	Status	78	0.81	**0.82**	0.77†
	Experiencer	78	0.75	**0.81**	0.80†
Trauma	Trigger	33	0.62	0.53	**0.70**
	Status	33	0.52	0.54	**0.63**
	Type	33	0.55	0.54	**0.67**
Micro Avg.	Trigger	529	0.80	0.81	**0.82**†
	Arguments	844	0.75	**0.78***	0.72†

Conclusion

Our **contributions** include

- A novel corpus, PedSHAC, annotated for fine-grained 10 SDoH factors from 1,260 social history sections from real pediatric clinical notes.
- Exploring IE across multiple dimensions, including
  - ❑ pre-trained transformer architectures: encoder-only (BERT), encoder-decoder (Flan-T5), decoder-only (GPT-4)
  - ❑ learning strategies: fine-tuning and in-context methods
  - ❑ prompting approaches: one-step text-to-event and two-step QA.
- Demonstrating that detailed SDoH representations can be extracted from pediatric narratives with performance comparable to human annotators

Future directions would include:

- Effective data selection strategies to save annotation costs: such as active learning in the annotation
- GPT-4 prompt-tuning: involvement of medical experts, automatic prompt generation, self-verification to improve the response quality

References

1. Yan Wang, Elizabeth S Chen, Serguei Pakhomov, Elliot Arsoniadis, Elizabeth W Carter, Elizabeth Lindemann, Indra Neil Sarkar, and Genevieve B Melton. 2015. Automated extraction of substance use information from clinical texts. In *AMIA Annu Symp Proc*, volume 2015, page 2121. AMIA.
2. Meliha Yetisgen and Lucy Vanderwende. 2017. Automatic identification of substance abuse from social history in clinical text. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, pages 171–181. Springer
3. David S Carrell, David Cronkite, Roy E Palmer, Kathleen Saunders, David E Gross, Elizabeth T Masters, Timothy R Hylan, and Michael Von Korff. 2015. Using natural language processing to identify problem usage of prescription opioids. *Int. J. Med. Inform.*, 84(12):1057–1064.
4. Raid Alzubi, Hadeel Alzoubi, Stamos Katsigiannis, Daune West, and Naeem Ramzan. 2022. Automated detection of substance-use status and related information from clinical text. *Sensors*, 22(24):9609
5. Adi V Gundlapalli, Marjorie E Carter, Miland Palmer, Thomas Ginter, Andrew Redd, Steven Pickard, Shuying Shen, Brett South, Guy Divita, Scott Duvall, et al. 2013. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among us veterans. In *AMIA Annu Symp Proc*, volume 2013, page 537. AMIA.
6. Cosmin A Bejan, John Angiolillo, Douglas Conway, Robertson Nash, et al. 2018. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*, 25(1):61–71.
7. Jinge Wu, Rowena Smith, and Honghan Wu. 2022b. Ontology-driven self-supervision for adverse childhood experiences identification using social media datasets.
8. Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023b. The 2022 n2c2/UW shared task on extracting social determinants of health. *J Am Med Inform Assoc*, 30(8):1367–1378.
9. Daniel J Feller, Jason Zucker, Bharat Srikishan, Roxana Martinez, Henry Evans, Michael T Yin, Peter Gordon, Noémie Elhadad, et al. 2018. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. In *AMIA Annu Symp Proc*, volume 2018, page 422. AMIA
10. Amol S Navathe, Feiran Zhong, Victor J Lei, Frank Y Chang, Margarita Sordo, Maxim Topaz, Shamkant B Navathe, Roberto A Rocha, and Li Zhou. 2018. Hospital readmission and social risk factors identified from physician notes. *Health Serv. Res.*, 53(2):1110–1136.
11. Braja G Patra, Mohit M Sharma, Veer Vekaria, et al. 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*, 28(12):2716–2727.
12. Iham Hatef, Masoud Rouhizadeh, Iddrisu Tia, et al. 2019. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med Inform*, 7(3):e13802.
13. Cheryl Clark, Kathleen Good, Lesley Jezierny, Melissa Macpherson, Brian Wilson, and Urszula Chajewska. 2008. Identifying smokers with a medical extraction system. *J Am Med Inform Assoc*, 15(1):36–39.
14. Yan Wang, Elizabeth S Chen, Serguei Pakhomov, Elliot Arsoniadis, Elizabeth W Carter, Elizabeth Lindemann, Indra Neil Sarkar, and Genevieve B Melton. 2015. Automated extraction of substance use information from clinical texts. In *AMIA Annu Symp Proc*, volume 2015, page 2121. AMIA.

References

15. Cosmin A Bejan, John Angiolillo, Douglas Conway, Robertson Nash, et al. 2018. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc*, 25(1):61–71
16. Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS One*, 13(2).
17. Anusha Bompelli, Yanshan Wang, Ruyuan Wan, et al. 2021. Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: A scoping review. *Health Data Sci*, 2021.
18. Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. 2023. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*, 30(8):1389–1397.
19. Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on EMNLP*, pages 1998–2022, Abu Dhabi, United Arab Emirates. ACL
20. Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*
21. Brian Romanowski, Asma Ben Abacha, and Yadan Fan. 2023. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. *J Am Med Inform Assoc*, page ocad071.

Thank you!

QR code for the manuscript

QR code for the GitHub

Dataset to be released, after the IRB approval from our home institution, and the de-identification step.