Annotation and Information Extraction of Social Determinants of Health from Social Worker Notes of Pediatric Transplantation

Harjith Pradeep[1], Luis Sanchez[1], Zhe He[2], and Xiaoyu Wang[3]
1.*College of Arts and Sciences, Florida State University*
2.*School of Information, Florida State University*
3.*Department of Statistics, Florida State University*

Introduction

Social determinants such as socioeconomic status, healthcare access, education, and family support shape pediatric transplant outcomes. Children in underserved communities often face delays, limited care, and added complications due to financial and racial disparities.

Besides medical factors, a child's home environment and caregiver well-being also impact recovery. Analysis of de-identified social worker notes from UF Health Shands Children's Hospital will be conducted to identify key social factors. In addition an annotation framework will be developed to train a model capable of identifying key SDoH factors linked to poorer outcomes. This tool will help providers predict risks and make more informed decisions to support pediatric transplant patients and their families.

Method

- **Phase 1: Annotation Framework**
  - Collaborated with social work experts and a UF annotator.
  - Developed guidelines with two levels:
    - **Level 1:** Identify key SDoH trigger words.
    - **Level 2:** Add contextual details.
- **Phase 2: Model Training**
  - Trained transformer models (Roberta-Large & BioBERT) to extract SDoH trigger words from EHRs.
- **Phase 3: Evaluation**
  - Measured performance (precision, recall, F-1 score) using strict and relaxed matching.

See Figure 1 for the overall process.

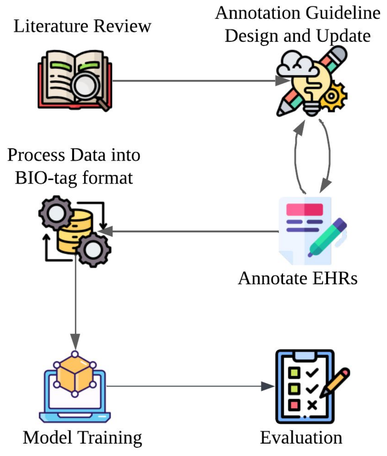

Fig.1 Study Pipeline

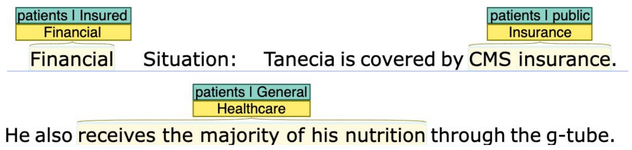

Fig.2 Annotation Samples

Category	Strict			Sentence		
	precision	recall	f-1	precision	recall	f-1
BioBert	0.79	0.78	0.78	0.81	0.79	0.8
RoberTa	0.79	0.72	0.76	0.84	0.73	0.78

Table.1 Performance from each model

Category	Sentence-level		
	precision	recall	f-1
Adherence	0.85	0.86	0.83
Alcohol	0.4	0.4	0.4
Concern	0.7	0.52	0.55
Drug	0	0	0
Education	0.67	0.64	0.62
Employment	0.8	0.84	0.82
Financial	0.8	0.85	0.82
Healthcare	0.58	0.63	0.59
Insurance	0.94	0.94	0.94
Literacy	0.07	0.2	0.1
Living	0.79	0.82	0.8
MentalHealth	0.97	0.92	0.94
Recommendation	0.89	0.89	0.88
Smoke	0.98	1	0.99
Social	0.79	0.75	0.77
SubstanceUse	0.93	0.94	0.93
Transportation	0.68	0.73	0.67
Trauma	0.72	0.77	0.73
overall	0.79	0.78	0.78

Table.2 Performance from each category

Results

- **Data Annotation:**
  - Annotated 19 EHRs with 1250 sentences.
  - Identified SDoH factors: financial concerns (131), living situation (120), employment (94), healthcare access (92), social support (72), insurance (70), mental health (50), trauma (57), and treatment adherence (29).
  - Figure 1 shows an example with two annotation levels for financial, insurance, and healthcare tags.
- **Model Performance:**
  - **BioBERT:** F1-score of 0.782 (strict) and 0.798 (sentence-level).
  - **RoBERTa:** F1-score of 0.755 (strict) and 0.781 (sentence-level),
- **Category Breakdown (Table 2):**
  - High scores: Insurance (0.94), Mental Health (0.94), Recommendation (0.88), Smoking (0.99).
  - Lower scores: Concern (0.55), Education (0.62), Alcohol (0.64), suggesting areas for improvement.

Conclusion

We developed a detailed guideline to annotate social determinants of health (SDoH), adding extra details like who is affected and specific attributes. Unlike previous methods that used a simpler, single-layer approach with a T5 model, our framework covers 17 categories with more depth.

Using transformer models, we extracted trigger words and categories. BioBERT and RoBERTa performed similarly, with RoBERTa doing better when evaluating larger text spans. Although our performance didn't fully meet expectations—likely due to the complex criteria—we plan to refine our models and annotation strategies in future work.

Reference

Lybarger, K., Yetisgen, M., & Uzuner, Ö. (2023). The 2022 n2c2/UW shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association, 30(8), 1367–1378.*