Social Determinants of Health Prediction for ICD-9 Code with Reasoning Models

Sharim Khan* SHARIMK2@ILLINOIS.EDU
University of Illinois, Urbana-Champaign, USA

Paul Landes* PLANDE2@UIC.EDU
University of Illinois Chicago, Chicago, USA

Abstract

Social Determinants of Health correlate with patient outcomes but are rarely captured in structured data. Recent attention has been given to automatically extracting these markers from clinical text to supplement diagnostic systems with knowledge of patients' social circumstances. Large language models demonstrate strong performance in identifying Social Determinants of Health labels from sentences. However, prediction in large admissions or longitudinal notes is challenging given long distance dependencies. In this paper, we explore hospital admission multi-label Social Determinants of Health ICD-9 code classification on the MIMIC-III dataset using reasoning models and traditional large language models. We exploit existing ICD-9 codes for prediction on admissions, which achieved a 89% F1. Our contributions include our findings, missing SDoH codes in 139 admissions, and code to reproduce the results.

Keywords: Social determinants of health; ICD-9 V-codes; clinical notes; large language models; MIMIC-III;

Data and Code Availability This study used the publicly available, de-identified MIMIC-III v1.4 dataset, accessed under the PhysioNet Data Use Agreement. Access requires completion of credentialing and training as specified by PhysioNet. The full evaluation code and prompt templates are also available[1].

Institutional Review Board (IRB) The analyses were conducted using de-identified secondary data (MIMIC-III) and did not involve interaction with human participants. In accordance with institutional policy, this work does not constitute human subjects research and therefore did not require IRB review.

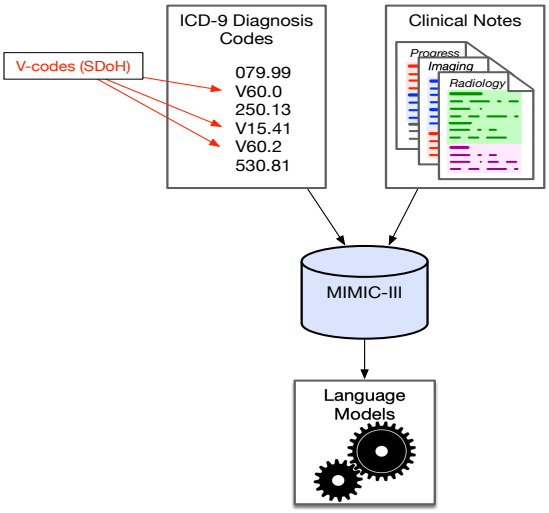

Figure 1: Free text clinical notes, and ICD-9 V-codes, are used as gold labels for prediction and to identify SDoH marked admissions.

1. Introduction

Social Determinants of Health (SDoH) are non-medical factors that account for an estimated 80–90% of modifiable contributors to health outcomes and disparities (MI Multipayer Initiatives, 2024). SDoHs may be adverse or protective. Adverse SDoHs are factors that create an additional social work or resource support need for patients (Guevara et al., 2024). Healthcare providers often document these factors in unstructured notes in Electronic Health Records, which limits their utility. The International Classification of Diseases, Ninth Revision (ICD-9) clas-

* Indicates equal contribution.
1. https://github.com/sharimk2/icd9_llm_prediction

© S. Khan & P. Landes.

Guevara Label	V-code	Z-code
Undomiciled	Homelessness (V60.0)	Z59
Unemployed	Unemployment (V62.0)	Z56

Table 1: Mapping from Guevara et al. (2024) SDoH labels to ICD-9 V-codes and ICD-10 Z-codes.

Domain	Description	Code
Housing & Resources	Homelessness	V60.0
	Inadequate material resources	V60.2
	No family caregiver	V60.4
Employment	Unemployment	V62.0
Legal	Legal problems	V62.5
Abuse History	Physical/sexual abuse	V15.41
	Emotional abuse	V15.42
Family History	Family alcoholism	V61.41

Table 2: Social Determinants of Health ICD-9 V-code taxonomy by domain.

sification system includes V-codes designed to capture social circumstances such as homelessness, unemployment, and inadequate family support. However, physicians diagnose SDoH V-codes as little as 2%, even in populations with documented social needs (Torres et al., 2017; Hatef et al., 2019).

This underutilization often stems from documenting social risk factors within clinical narratives, which require substantial time and expertise to diagnose as billing codes (Torres et al., 2017). Reasoning-capable language models present an opportunity to predict SDoH by using coding guidelines in prompts and training on V-codes (a subset of ICD-9 codes used by insurance companies for billing) as labels already diagnosed by physicians in the MIMIC-III dataset as shown in Figure 1.

Existing large language models (LLMs) studies on SDoH classify text into broad social risk categories such as housing or social support, as shown by Guevara et al. (2024). This approach does not operate at the admission level nor predict specific billing codes. In contrast, our work directly predicts fine-grained ICD-9 V-codes from the entire set of clinical documents for an admission, bridging the gap between unstructured SDoH documentation and structured diagnostic coding.

The semantics of the clinical notes with the diagnosed ICD-9 V-codes map to the descriptions of the labels from the annotation guide provided by Guevara et al. (2024) as shown in Table 1. Undomiciled and unemployed are the only adverse SDoH categories with overlap between this study and the work done by Guevara et al. (2024). This study evaluates the feasibility and performance of reasoning language models to predict V-codes from clinical documentation across open-source and closed frontier models.

2. Related Work

Initial approaches to SDoH identification employed rule-based systems and traditional machine learning methods to detect SDoH from clinical text (Chen et al., 2020). However, these methods demonstrated limited generalizability due to institution-specific documentation patterns. Subsequent advances in neural architectures, including convolutional models (Mullenbach et al., 2018) and transformer architectures such as ClinicalBERT (Alsentzer et al., 2019), improved automated medical coding performance on diagnosis and procedure codes.

Researchers have explored large language models for SDoH mention classification in clinical notes. Guevara et al. (2024) achieved 70% macro-F1 for adverse SDoH classification at the sentence-level. Their work identified 93.8% of patients with adverse SDoH compared to 2.0% captured by ICD-10 codes. Their taxonomy of labels maps to broad ICD-10 Z-code categories.

Hybrid architectures combining language models with traditional neural networks have demonstrated improvements in both accuracy and computational efficiency for SDoH extraction (Landes et al., 2025).

3. Dataset

Unstructured free text clinical notes from the Medical Information Mart for Intensive Care (MIMIC-III v1.4) was used as input to several LLMs. MIMIC-III is a publicly available dataset containing de-identified health records from over 58,000 hospital admissions at Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). We predicted ICD-9 V-codes[2] identified as SDoH by Torres et al.

2. MIMIC-III predates the 2015 ICD-10 transition.

Model	Parameters	EM	wP	wR	wF1	mP	mR	mF1	MP	MR	MF1
Frontier models											
GPT-3.5-turbo	Unknown	0.144	0.718	0.338	0.452	0.529	0.338	0.413	0.208	0.162	0.167
GPT-4o-mini	Unknown	0.691	0.864	**0.870**	0.855	0.845	**0.870**	0.857	0.719	**0.642**	0.645
GPT-5-mini	Unknown	**0.722**	**0.939**	0.847	**0.882**	**0.939**	0.847	**0.891**	**0.797**	0.608	**0.665**
Open-source models											
Llama-3.1	8B	0.103	0.527	0.765	0.573	0.412	0.765	0.536	0.383	0.521	0.351
Llama-3.3	70B	0.355	0.765	**0.850**	0.740	0.598	**0.850**	0.702	0.678	**0.663**	0.573
Qwen3	32B	**0.655**	**0.918**	0.770	**0.829**	0.925	0.770	0.840	0.769	0.577	0.644
GPT-OSS	20B	0.652	0.896	0.775	0.824	**0.952**	0.775	**0.854**	**0.795**	0.580	**0.652**

Table 3: Model results metrics on SDoH code prediction. Metrics include (**E**)xact (**M**)atch, (**P**)recision, (**R**)ecall, and F1 scores reported as (**w**)eighted, (**m**)icro, and (**M**)acro averages. Model parameters shown in billions (B). **Bolded** measures indicate the highest performing for each subgroup.

(2017), eight of which were found in the MIMIC-III dataset and listed in Table 2.

Evaluation Dataset The evaluation dataset comprised 139 MIMIC-III admissions selected to capture varied SDoH documentation patterns. Selection criteria included admissions with existing V-code assignments as well as admissions without SDoH codes. We sampled 357 admissions with at least 500 characters from discharge summaries, nursing notes, and social work evaluation note categories.

4. Experimental Setup

We identified admissions with notes containing the V-codes listed in Table 2 or SDoHs not diagnosed by physicians. We labeled these admissions with the additional V-codes for related SDoHs, which we refer to this as "amended codes". For example, "He is *unemployed* and receives SSI secondary to anxiety, problems with depth perception," appears in an admission only labeled with `homelessness` (V60.0). The *unemployed* token is a clear indication of a `unemployment` (V62.0), but is missing from the admission, and therefore, added as an amended code. We evaluate the models on the amended codes to avoid penalizing for missing undiagnosed SDoHs. The procured dataset is available via `PyHealth` (Yang et al., 2023)

Prompting Few-shot prompting (Brown et al., 2020) is used to predict SDoH labels from clinical documentation. Our prompts' role is a certified medical coder, specify labeling criteria, and include coding example (both positive and negative). Our prompts are given in Appendix A.

Evaluation We evaluated seven LLMs spanning frontier closed-source, open-source, reasoning-capable, and traditional architectures on multi-label classification of V-codes that represent adverse SDoH

Code	Precision (%)	Recall (%)	F1 (%)	Count
Homelessness (V60.0)	100.0	94.8	97.4	97
Inadequate material resources (V60.2)	100.0	70.8	82.8	8
No family caregiver (V60.4)	95.8	52.8	67.5	12
Unemployment (V62.0)	95.9	84.5	89.9	28
Legal problems (V62.5)	56.1	45.2	50.0	14
Physical abuse (V15.41)	72.3	73.8	73.0	14
Emotional abuse (V15.42)	100.0	25.0	40.0	4
Family alcoholism (V61.41)	96.8	100.0	98.4	30

Table 4: By code performance of GPT-5-mini on the evaluation dataset and count of admissions containing the code.

labels. We processed each clinical note through the model independently across admission, then aggregated predictions at the admission level.

5. Results

The results of the models we tested are in Table 3. GPT-5-mini achieved the highest overall performance with 72.2% exact match and 89.1% micro-F1, which demonstrates frontier models' capability in extracting V-codes in clinical documentation. Open-source models demonstrated competitive capabilities, with Qwen3-32B and GPT-OSS-20B achieving exact match scores of 65.5% and 65.2%, respectively. This performance suggests that effective SDoH prediction does not require proprietary systems.

Traditional language models showed significantly lower performance, with Llama-3.3-70B and GPT-3.5 achieving exact match scores of 35.5% and 14.4%, respectively. The performance difference between traditional and reasoning-capable models indicates that multi-label SDoH classification presents challenges that benefit from advanced reasoning capabilities, particularly when processing lengthy clinical narratives with distributed evidence.

Reasoning models leverage the chain-of-thought, which is a series of intermediate reasoning steps the model performs before producing an output (Wei et al., 2023). These models perform better given the length of the prompt, and complexity of conjunctions (i.e. inclusion and exclusion clauses), and negative examples. For example, our prompt contains hierarchical code exclusion rules (such as identifying `homelessness` in an example, but instructing the exclusion of `inadequate material resources`).

Label-Specific Performance GPT-5-mini's performance varied across SDoH categories (Table 4). The model achieved high performance for `homelessness` and `family alcoholism`, where documentation typically follows consistent terminology. Moderate performance was observed for `unemployment` and `inadequate material resources`. Performance decreased for labels requiring inference across scattered evidence, including `no family caregiver`, `legal problems`, and `emotional abuse`. While this pattern suggests labels with explicit documentation are more reliably predicted, the conservative prompting strategy for codes with low support may have limited the model's ability to capture more nuanced cases, and further prompt optimization could improve performance on complex inference tasks.

Metric	Amended	Non-Amended
Exact Match	72.2	48.9
Macro-F1	66.5	32.2
Micro-F1	89.1	67.0

Table 5: GPT-5-mini multi-label classification performance against the amended codes versus non-amended diagnoses codes

Evaluation Against Original Coding We evaluated model performance on the amended codes and the non-amended codes (Table 5). Evaluated on the amended codes, GPT-5-mini achieved 72.2% exact match and 89.1% micro-F1. Performance against non-amended codes showed 48.9% exact match and 67.0% micro-F1. The difference in performance reflects instances where clinical notes contained evidence for additional SDoH labels beyond those originally diagnosed.

6. Discussion

The evaluation metrics support that LLMs can predict adverse SDoH V-codes from clinical notes with high accuracy (89.1% micro-F1). Furthermore, we identified clinical notes that contained adverse SDoH semantics beyond those originally diagnosed. The performance difference between the prediction of amended codes and the non-amended codes suggests non-agreement in the diagnosis of V-codes among physicians. There is potential for these multi-label classification models to bridge the gap between unstructured documentation of SDoHs and missed diagnoses.

The model performed best on labels with standard terminology and struggled with recall when documentation varied, such as `history of emotional abuse`. However, further experimentation is needed improve precision-recall tradeoff. Open-source models demonstrated competitive performance, suggesting effective V-code prediction may be achievable without requiring commercial models.

7. Conclusion

Reasoning LLMs can effectively predict adverse SDoH labels from clinical documentation through few-shot prompting approaches. The competitive performance of open-source models compared to commercial models demonstrates that effective SDoH prediction does not require proprietary systems, making these capabilities accessible to resource-constrained healthcare settings. Our identification of admissions with documented but uncoded SDoH factors highlights a gap between clinical documentation and coding practices. As healthcare systems increasingly recognize SDoH as critical quality indicators, classification models can improve both clinical care and population health management.

8. Limitations

The use of MIMIC-III admissions potentially limits generalizability to other institutions with varied documentation practices. The study examined eight ICD-9 V-codes, a subset of available SDoH codes. The evaluation dataset was developed by a single annotator without an inter-rater reliability assessment.

The notes used in our experiments are limited to those exported in the MIMIC-III dataset, which include only discharge summaries and those from the ICU department Landes et al. (2022). Notes during the hospital stay for patients that are moved to other departments during their recovery are omitted and likely have an impact on model performance.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*, 2019. URL https://arxiv.org/abs/1904.03323.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

Min Chen, Xuemei Tan, and Rema Padman. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *Journal of the American Medical Informatics Association*, 27(11): 1764–1773, November 2020. doi: 10.1093/jamia/ocaa143.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin Kann, Shalini Moningi, Jack Qian, Madeleine Goldstein, Susan Harper, Hugo J. W. L. Aerts, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine*, 7(1):—, 2024. doi: 10.1038/s41746-023-00970-0.

Elham Hatef, Masoud Rouhizadeh, Iddrisu Tia, Elyse Lasser, Felicia Hill-Briggs, Jill Marsteller, and Hadi Kharrazi. 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 Medical Informatics*, 7(3):e13802, 2019. doi: 10.2196/13802.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3: 160035, 2016. doi: 10.1038/sdata.2016.35.

Paul Landes, Kunal Patel, Sean S. Huang, Adam Webb, Barbara Di Eugenio, and Cornelia Caragea. A New Public Corpus for Clinical Section Identification: MedSecId. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3709–3721. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.326.

Paul Landes, Jimeng Sun, and Adam Cross. Integration of large language models and traditional deep

learning for social determinants of health prediction. *arXiv preprint arXiv:2505.04655*, 2025. URL https://arxiv.org/abs/2505.04655.

MI Multipayer Initiatives. Social determinants of health: Discussion paper 101 for health care. Technical report, MI Multipayer Initiatives, 2024. URL https://mimultipayerinitiatives.org/wp-content/uploads/2024/07/sdoh-discussion-paper-101-for-health-care.pdf. Accessed: 2025-08-25.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT 2018*, pages 1101–1111, New Orleans, LA, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100/.

Jacqueline M. Torres, John Lawlor, Jeffrey D. Colvin, Marion R. Sills, Jessica L. Bettenhausen, Amber Davidson, Gretchen J. Cutler, Matt Hall, and Laura M. Gottlieb. Icd social codes: An underutilized resource for tracking social needs. *Medical Care*, 55(9):810–816, 2017. doi: 10.1097/MLR.0000000000000764.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. 2023. URL https://arxiv.org/abs/2201.11903.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin Danek, and Jimeng Sun. PyHealth: A deep learning toolkit for healthcare predictive modeling. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2023*, 2023. URL https://github.com/sunlabuiuc/PyHealth.

Appendix A. Final Prompt for SDoH Coding

```
# SDOH Medical Coding Task

You are an expert medical coder. Your job is to find
Social Determinants of Health (SDOH) codes in clinical notes.

## CRITICAL PREDICTION RULES:
1. IF YOU FIND EVIDENCE FOR A CODE -> YOU MUST PREDICT THAT CODE
2. BE CONSERVATIVE: Only predict codes with STRONG evidence
3. Evidence detection = Automatic code prediction

## CODE HIERARCHY - ABSOLUTELY CRITICAL:
HOMELESS (V600) COMPLETELY EXCLUDES V602:
- [X] If you see homelessness -> ONLY predict V600, NEVER V602
- [X] "Homeless and can't afford X" -> STILL only V600
- [OK] V602 is EXCLUSIVELY for housed people with explicit financial hardship

V602 ULTRA-STRICT REQUIREMENTS:
- [OK] Must have exact quotes: "cannot afford medications", "no money for food"
- [X] NEVER infer from: homelessness, unemployment, substance use, mental health
- [X] NEVER predict from: "poor", "disadvantaged", social circumstances

## Available ICD-9 Codes:

Housing & Resources:
- V600: Homeless (living on streets, shelters, hotels, motels, temporary housing)
- V602: Cannot afford basic needs. ULTRA-STRICT RULES:
  - [OK] ONLY if EXPLICIT financial statements:
        "cannot afford medications", "unable to pay for treatment",
        "no money for food", "financial hardship preventing care"
  - [X] NEVER predict from social circumstances: substance abuse, unemployment,
        mental health, divorce
  - [X] NEVER predict from discharge inventory: "no money/wallet returned"
  - [X] NEVER predict for insurance/benefits mentions: "on disability", "has
      Medicaid"
- V604: No family/caregiver available. CRITICAL RULES:
  - [X] NOT just "lives alone" or "elderly"
  - [OK] ONLY if: "no family contact" AND "no support" AND "no one to help"
  - [X] NOT for: "lives alone but daughter visits", "independent"

Employment & Legal:
- V620: Unemployed (current employment status) -- EXPLICIT ONLY
  - [OK] Predict only if one of these exact employment-status phrases appears:
    unemploy(ed|ment), jobless, no job, out of work, without work,
    not working, between jobs, on unemployment, receiving unemployment,
    laid off, fired, terminated, lost job
  - [X] Do NOT infer from "used to work", "formerly employed", etc.
  - [X] Contradictions (block V620 if nearby): "employed", "works at", "working",
      etc.
  - [X] Non-employment uses of "work": "work of breathing", "social work", etc.
  - [X] Exclusions: retired, student, stay-at-home parent, medical leave
- V625: Legal problems. ACTIVE LEGAL PROCESSES ONLY:
  - [OK] Criminal: arrest, jail, prison, parole, probation, charges, court case
  - [OK] Civil: restraining order, custody case, litigation
```

```
  - [OK] Guardianship: legal capacity hearing, court-appointed guardian
  - [OK] Child welfare: CPS/DCF removal, court-ordered custody
  - [X] Do NOT code for history of crime, substance use, or homelessness
        without legal process

Health History:
- V1541: Physical/sexual abuse (violence BY another person). CRITICAL:
  - [OK] Physical violence, sexual abuse, assault, domestic violence, rape
  - [X] NOT accidents, falls, or when patient is aggressor
- V1542: Pure emotional/psychological abuse. MUTUALLY EXCLUSIVE WITH V1541:
  - [OK] ONLY if NO physical/sexual abuse mentioned AND explicit emotional abuse:
        witnessed violence, verbal abuse, psychological abuse,
        emotional manipulation, intimidation
  - [X] NEVER predict if physical/sexual abuse mentioned
  - [X] NEVER predict both V1541 AND V1542 for same patient
  - [X] Depression/anxiety/suicidality alone = NO CODE

Family History:
- V6141: Family alcoholism
  - [OK] Kinship + alcohol terms (father alcoholic, mother ETOH abuse, etc.)
  - [X] Negations: "denies family history of alcoholism"
  - [X] NEVER for patient's own history

## ENHANCED NEGATIVE EXAMPLES:
V602 FALSE POSITIVES TO AVOID:
- [X] "Homeless patient" -> V600 ONLY
- [X] "Lives in shelter, gets food stamps" -> V600 ONLY
- [X] "Homeless, on disability" -> V600 ONLY
- [X] "No permanent address, has Medicaid" -> V600 ONLY
- [X] "Homeless and can't afford medications" -> V600 ONLY
- [X] "Lives in poverty" -> NO CODE
- [X] "Financial strain from divorce" -> NO CODE

V604 FALSE POSITIVES TO AVOID:
- [X] "82 year old lives alone" -> NO CODE
- [X] "Lives by herself" -> NO CODE
- [X] "Widowed, lives alone, son calls daily" -> NO CODE

V1542 FALSE POSITIVES TO AVOID:
- [X] "History of physical and sexual abuse" -> V1541 ONLY
- [X] "PTSD from rape at age 7" -> V1541 ONLY
- [X] "Childhood sexual abuse by uncle" -> V1541 ONLY
- [X] "History of domestic abuse" -> V1541 ONLY
- [X] "Depression and anxiety" -> NO CODE
- [X] "Suicide attempts" -> NO CODE
- [X] "History of abuse" (unspecified) -> NO CODE
- [X] "Recent argument with partner" -> NO CODE

V1542 TRUE POSITIVES TO CAPTURE:
- [OK] "Witnessed violence as a child" -> V1542
- [OK] "Emotionally abusive relationship for 14 years" -> V1542
- [OK] "Verbal abuse from controlling partner" -> V1542
- [OK] "Jealous and controlling behavior" -> V1542

## CONFIDENCE RULES:
```

```
HIGH CONFIDENCE (Predict):
- Direct statement
- Multiple evidence pieces
- Explicit matching phrase

LOW CONFIDENCE (Don't Predict):
- Ambiguous language
- Single weak indicator
- Contradictory evidence

## Key Rules:
1. Precision over Recall: Better to miss a code than falsely predict
2. Evidence-Driven: Strong evidence required for prediction
3. Multiple codes allowed: But each needs independent evidence
4. Conservative approach: When in doubt, don't predict

## Output Format:
Return applicable codes separated by commas, or "None".

Examples:
V600, V625
None
```

Figure 2: Final optimized prompt used for admission-level SDoH ICD-9 V-code prediction. The prompt encodes hierarchy rules, precision-first criteria, negative examples, and explicit output formatting. This full text is provided to ensure reproducibility.

Appendix B. Annotation Guide for SDoH ICD-9 V-codes

B.1. V60.0: Homelessness

Definition: Patient has no permanent residence or lives in transient arrangements such as streets, shelters, or motels.
Approximate Synonyms: Homeless, currently homeless.
Applies To: Hobos, tramps, vagabonds, social migrants, transients.
Inclusion Criteria: Explicit statements such as "Patient is homeless," "Lives in shelter," or "No permanent address."
Exclusion Criteria: Financial hardship only (V60.2), statements like "lives alone" or "poor" without loss of housing.
Examples:

- Positive: "Lives in a homeless shelter after eviction."
- Negative: "Struggling to pay rent but living with family."

B.2. V60.2: Inadequate Material Resources

Definition: Patient lacks financial means for necessities such as food, medication, or housing.
Approximate Synonyms: Financial problem.
Applies To: Economic problem, poverty not otherwise stated.
Inclusion Criteria: Explicit financial statements (e.g., "Cannot afford medications," "No money for food").

Exclusion Criteria: Any homelessness (V60.0), implicit hardship without direct evidence.
Examples:

- Positive: "Financial hardship preventing purchase of insulin."
- Negative: "Homeless and can't afford meds." (V60.0 only)

B.3. V60.4: No other household member able to render care

Definition: Indicates no family or caregiver available to provide care or support.
Applies To: Family member too ill or unsuited to render care, partner temporarily away, or person away from usual home.
Inclusion Criteria: Explicit lack of family support (e.g., "No family contact and no one to help with care.").
Exclusion Criteria: "Lives alone," "independent," or "widowed" without need for care.
Examples:

- Positive: "No family or friends available to assist after surgery."
- Negative: "Lives alone, independent in activities of daily life."

B.4. V62.0: Unemployment

Definition: Patient currently lacks employment.
Excludes: Main problem due to economic inadequacy (V60.2).
Inclusion Criteria: Direct mention of unemployment or job loss ("unemployed," "laid off," "no job").
Exclusion Criteria: Retired, student, stay-at-home parent, or on medical leave; "used to work."
Examples:

- Positive: "Unemployed, seeking work."
- Negative: "Previously worked as nurse."

B.5. V62.5: Legal Problems

Definition: Current involvement with legal or judicial processes (criminal, civil, custody, or guardianship).
Approximate Synonyms: Arrested, charged with crime, imprisonment, litigation, parole, bail, probation, court case pending.
Applies To: Imprisonment, legal investigation, litigation, prosecution.
Inclusion Criteria: Active legal process (arrest, trial, custody case, guardianship hearing).
Exclusion Criteria: History of crime or incarceration without current legal process.
Examples:

- Positive: "Currently on probation after DUI arrest."
- Negative: "History of incarceration in 2010."

B.6. V15.41: History of Physical or Sexual Abuse

Definition: Past exposure to physical or sexual violence inflicted by another person.
Approximate Synonyms: History of physical abuse, sexual abuse, child abuse, domestic violence, rape.
Applies To: Rape.
Inclusion Criteria: Mentions such as "History of sexual abuse" or "Physically abused by spouse."
Exclusion Criteria: Accidents, self-injury, patient as aggressor, emotional abuse only (V15.42).
Examples:

- Positive: "Reports childhood sexual abuse by uncle."
- Negative: "Verbal abuse from partner." (V15.42)

B.7. V15.42: History of Emotional Abuse or Neglect

Definition: History of psychological, emotional, or neglectful abuse without physical or sexual component.
Approximate Synonyms: Emotional abuse, neglect, psychological abuse, verbal abuse, child neglect.
Applies To: Neglect.
Inclusion Criteria: "Emotionally abusive relationship," "childhood neglect," "verbal abuse."
Exclusion Criteria: Physical or sexual abuse (V15.41), depression/anxiety/PTSD alone.
Examples:

- Positive: "Endured years of emotional abuse from spouse."
- Negative: "PTSD from physical assault." (V15.41)

B.8. V61.41: Family History of Alcoholism

Definition: Indicates alcoholism or alcohol dependence in the patient's family.
Approximate Synonyms: Family history of alcohol dependence, FHX of alcoholism, adult child of alcoholic.
Inclusion Criteria: Mentions such as "Father alcoholic," "Mother with ETOH abuse," or "Family history of alcoholism."
Exclusion Criteria: Patient's own alcohol use; negations ("Denies family history of alcoholism").
Examples:

- Positive: "Mother and uncle both heavy drinkers."
- Negative: "Patient with long history of alcohol use disorder."

General Annotation Rules

1. Precision over recall: only label with explicit evidence.
2. Mutual exclusions: V60.0 supersedes V60.2; V15.41 and V15.42 are mutually exclusive.
3. Each code requires independent evidence.
4. When in doubt, do not assign a code.

Figure 3: Annotation guide used to label SDoH circumstances in clinical text for ICD-9 V-code prediction. Each section defines the code, domain, inclusion and exclusion criteria, and examples. Annotators must use precision-first judgment and follow hierarchy rules as outlined.