SBDH-Reader: an LLM-powered method for extracting social and behavioral determinants of health from medical notes

Zifan Gu, MS[1], Lesi He, MS[1], Awais Naeem[2], Pui Man Chan, MPH[1], Asim Mohamed[3], Hafsa Khalil[3], Yujia Guo, MS[1], Wenqi Shi, PhD[1], Matthew E. Dupre PhD[4,5], Guanghua Xiao, PhD[1,6,7], Eric D. Peterson, MD, MPH[8], Yang Xie, PhD[1,6,7], Ann Marie Navar, MD, PhD[8], Donghan M. Yang, PhD[1]

[1] Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, Dallas, Texas, USA
[2] School of Information, University of Texas at Austin, Austin, Texas, USA
[3] The University of Texas Southwestern Medical Center, Dallas, Texas, USA
[4] Department of Population Health Sciences, Duke University, Durham, North Carolina, USA
[5] Department of Sociology, Duke University, Durham, North Carolina, USA
[6] Department of Bioinformatics, The University of Texas Southwestern Medical Center, Dallas, Texas, USA
[7] Simmons Comprehensive Cancer Center, The University of Texas Southwestern Medical Center, Dallas, Texas, USA
[8] Department of Internal Medicine, The University of Texas Southwestern Medical Center, Dallas, TX, USA

Corresponding Author:

Donghan M. Yang, PhD

Peter O'Donnell Jr. School of Public Health
The University of Texas Southwestern Medical Center
Dallas, Texas, USA
donghan.yang@utsouthwestern.edu

ABSTRACT

Introduction: Social and behavioral determinants of health (SBDH) are increasingly recognized as essential for prognostication and informing targeted interventions. While medical notes contain rich SBDH details, these are unstructured and conventional extraction methods tend to be labor intensive, inaccurate, and/or unscalable. The emergence of large language models (LLMs) presents an opportunity to develop more effective approaches for extracting SBDH data.

Materials and Methods: We developed the SBDH-Reader, an LLM-powered method to extract structured SBDH data from full-length medical notes through prompt engineering. Six SBDH categories were queried including: employment, housing, marital relationship, and substance use including alcohol, tobacco, and drug use. The development dataset included 7,225 notes from 6,382 patients in the MIMIC-III database. The method was then independently tested on 971 notes from 437 patients at UT Southwestern Medical Center (UTSW). We evaluated SBDH-Reader's performance using precision, recall, F1, and confusion matrix.

Results: When tested on the UTSW validation set, the GPT-4o-based SBDH-Reader achieved a macro-average F1 ranging from 0.85 to 0.98 across six SBDH categories. For clinically relevant adverse attributes, F1 ranged from 0.94 (employment) to 0.99 (tobacco use). When extracting any adverse attributes across all SBDH categories, the SBDH-Reader achieved an F1 of 0.96, recall of 0.97, and precision of 0.96 in this independent validation set.

Conclusion: A general-purpose LLM can accurately extract structured SBDH data through effective prompt engineering. The SBDH-Reader has the potential to serve as a scalable and effective method for collecting real-time, patient-level SBDH data to support clinical research and care.

INTRODUCTION

Social and behavioral determinants of health (SBDH) are increasingly recognized as essential for understanding a patient's clinical presentation and as targets for interventions.[1-4] SBDH encompass a broad range of factors that describe an individual's living conditions at the personal, community, and societal levels. A growing body of research highlights the significance of integrating SBDH to enhance clinical risk assessment, outcome prediction, and clinical trial enrollment, ultimately guiding therapeutic interventions across various clinical settings.[4-11]

Despite the significance of SBDH, access to these patient-level characteristics is limited due to the lack of effective methods for high-quality and continuous data collection, particularly in real-world settings.[12,13] Previous studies have investigated collecting these data through additional patient surveys conducted during routine care.[14,15] However, to achieve scalability and generalizability, existing electronic health records (EHRs) should be leveraged. Unfortunately, when structured SBDH data fields are created in the EHR system, they tend to be poorly completed.[12,13,16] Despite this, studies demonstrate that SBDH information is often in the EHR but buried within free-text medical notes.[4,12,13,17-20]

Various natural language processing (NLP) methods have been applied to identify SBDH from clinical notes, ranging from rule-based approaches to deep learning-based models, including early language models such as BERT.[18,21-31] However, conventional NLP approaches fall short in SBDH extraction tasks due to limitations in accuracy, requirement of entity/relation-level annotations, and scalability to real-world settings. The framing of NLP tasks in previous studies is often highly crafted, constrained by training data, and impractical for large-scale EHR mining.[18,24,32,33] Common weaknesses of previous methods include rigid, often binary, coding of SBDH characteristics (e.g., predefined positive vs. negative),[25,29-31] reliance on sentence-level or paragraph-level texts due to model-imposed text length limitations, [25,30,31] and step-wise frameworks requiring separate model training for each step or task.[34]

Recently, large language models (LLMs) have demonstrated substantial potential in healthcare and biomedical applications,[35-40] particularly for extracting information from medical notes.[41-43] Unlike traditional NLP approaches, LLMs excel in text extraction tasks without the need for task-specific model development or extensive training data.[18,32,35,44] However, the application of this promising technology to SBDH data extraction remains largely unexplored. Existing studies are limited by their restriction to sentence-level input [25,29-31], lack of precise evaluation of individual SBDH categories[31,34], and a classifier-only design that requires model training or fine-tuning towards predefined, fixed SBDH coding.[25,30,31,34] Unlike medical concepts, which often fall under specialized domains, descriptions of SBDH in medical notes typically require less domain knowledge and terminology, a foundation on which advanced LLMs are well-equipped. On the other hand, categorization and coding of SBDH can vary significantly across health systems, departments, and author types. Therefore, the conventional strategy of training or fine-tuning models to fit predefined SBDH targets is challenged. Prompt engineering-based strategies may offer a more flexible, versatile, and generalizable solution.

In this study, we developed SBDH-Reader, a method for extracting SBDH from real-world clinical notes using direct prompt engineering with GPT-4o. We specifically explored whether advanced LLMs such as GPT-4o can achieve desirable and generalizable accuracy and precision in identifying a variety of SBDH attributes, without fine-tuning. Additionally, we investigated SBDH-Reader's ability to perform multi-label (SBDH categories) and multi-class (SBDH attributes) classification, extracting each SBDH instance is extracted without predefined binary coding. To enhance comprehensiveness and enable potential human verification, we also instructed the SBDH-Reader to output supporting evidence text alongside the SBDH classification.

MATERIALS AND METHODS

Datasets and Data Processing

This study involves three EHR-based datasets with clinical notes. For method development, two previously established datasets based on Medical Information Mart for Intensive Care III (MIMIC-III) database[45] were used. The first dataset, established by Guevara et al. (MIMIC-G)[29] , contains 200 notes written by physicians, nurses, and social workers for 183 patients (Table 1). To generate full-length notes as input data for SBDH-Reader, we merged the original sentence-level texts (5328 entries) and ground truths from the MIMIC-G dataset into the note level (200 entries). The second dataset, established by Ahsan et al. (MIMIC-A)[46], contains 7025 discharge summaries for 6199 patients (Table 1). Following the original annotation method, we only used "social history" section as input data.[46]

The independent validation dataset was derived from the EHRs at UT Southwestern Medical Center (UTSW). We established this dataset from a heart failure cohort of 437 patients treated at UTSW, containing 971 full-length notes of various types including progress, history and physical (H&P), consult, emergency department (ED) provider notes (Table 1). To meaningfully evaluate SBDH-Reader's performance and mitigate the intrinsic data imbalance caused by missing SBDH documentation, we applied keyword-based filtering to all available notes for this cohort in the UTSW EHRs, yielding a dataset enriched with SBDH documentation. Similarly, we extracted the "social history" section as input data.

Data from MIMIC-III were de-identified at the source.[45] Data from UTSW were de-identified in house. The LLM used in this study, GPT-4o, was operated on UTSW's private Microsoft Azure OpenAI service. Per contractual agreements with Microsoft, data processed on this service will remain inaccessible to OpenAI or other customers and will not be used to improve OpenAI models or any Microsoft or 3rd party products or services. Use of MIMIC-III data in this setting is in compliance with the PhysioNet Credentialed Health Data Use Agreement 1.5.0[47] and the policy of "Responsible use of MIMIC data with online services like GPT"[48]. This study protocol was approved by UTSW's institutional review board (Protocol #STU-2024-0087).

Task Definition and Data Annotation

In this study, the task of extracting SBDH from a given note was defined as a multi-label, multi-class classification problem, where each label corresponds to an SBDH category. This study

targets six SBDH categories: employment, housing, marital relationship, and substance use including alcohol, tobacco, and drug use. An SBDH terminology was constructed by integrating the terminologies used in the two MIMIC-based studies [29,46] and the UTSW Epic system, in consultation with a medical sociologist (MED) and two physicians (AMN, EDP) (Supplementary Table 1). This terminology defines the permissible attributes (classes) in each category. Notably, the terminologies defined in the MIMIC-G and MIMIC-A studies resulted in ground truth annotations with varying granularities and standards, as commonly seen in the field[25,29-31,34,46]. To ensure consistency across different standards, we tasked the SBDH-Reader with classifying only at the granular attribute level defined in Supplementary Table 1, without allowing the LLM to determine these as positive or negative attributes. Then, to enable evaluation against the original attributes used in the MIMIC-G and MIMIC-A studies[29,46], we applied a rule-based mapping of the SBDH-Reader-identified attributes to three post hoc classes—adverse attributes, non-adverse attributes, or no information (Supplementary Table 1).

All human annotations in this study followed the same two-step approach: first, annotating the granular attributes, and then applying rule-based mapping to the three evaluative classes. To generate ground truths for the UTSW datasets, two medical students (AM, HK) annotated all notes under the supervision of two data scientists (LH, DMY) and one physician (AMN) using a local instance of the INCEpTION platform[49]. Initially, the two annotators practiced on the first 100 notes in the UTSW datasets, discussed the results with the supervisors, and repeated the process. Once a good agreement was achieved (Cohen's Kappa $\geq$ 0.85, Supplementary Table 2), they proceeded with independent annotation of the entire dataset. Any remaining discrepancies were resolved by LH and DMY to generate the final ground truths.

To ensure consistent annotation standards across all three datasets, after the annotation of the UTSW datasets, three data scientists (ZG, LH, DMY) conducted a thorough review of the original annotations provided with the two MIMIC datasets following the UTSW annotation guidelines. Mislabelings were corrected to generate the final ground truths for the MIMIC-G and MIMIC-A used in this study.

Model and Prompt Engineering

We used GPT-4o (version 2024-05-13) in this study. Through prompting, the LLM was instructed to process an entire note entry and summarize and classify all relevant SBDH descriptions into pre-defined categories and attributes based on the established terminology (Figure 1). Each note was processed in a single-round chat completion, with all involved SBDH categories handled in one prompting.

The SBDH-Reader prompt consists of two sections: instruction and input (Figure 1; Supplementary Info 1). The input section corresponds to a single note entry. The instruction section follows a modular design by sequentially listing each involved SBDH category along with category-specific definitions of permissible attributes and detailed descriptions. For example, the instruction under Employment status defines six permissible attributes: employed, unemployed, underemployed, retired, student, and no information (Supplementary Table 1). Definition and specification

surrounding each attribute is provided along with the prompt. For instance, under Alcohol Use we specified that any mention of polysubstance abuse implies alcohol use unless otherwise stated. To ensure structured outputs, we prompted the LLM to generate responses in JSON format, where the first key represents the identified attribute for each category, and the second key provides supporting evidence extracted from the text. The JSON format was further enforced by specifying the *response_format* parameter within the Azure OpenAI API. Last, a global finisher defines general instructions applicable across all categories. For example, we explicitly instructed the LLM to prioritize the current SBDH status at the time of the note, and not focus on any speculative statements about future changes (Supplementary Info 1).

We developed the prompt based on the MIMIC-G and MIMIC-A datasets. To iteratively refine the prompt, we incorporate more precise specifications for each attribute by sampling the training dataset and summarizing common patterns and characteristics. In independent testing with the UTSW dataset, no prompt changes were allowed.

Evaluation

The metrices we used to evaluate SBDH-Reader's performance include precision (positive predictive value), recall (sensitivity), F1, and confusion matrix. Considering the high imbalance across various attributes, including "no information", in all three datasets, we present the evaluations on the level of three combined classes: adverse attributes, non-adverse attributes, or no information. For precision, recall, and F1, we first report macro-average across these three classes for each SBDH category to show general performance. Then, we report these metrics on the adverse attributes alone to show its performance in this clinically relevant class. Median and 95% confidence interval were estimated by bootstrapping with 500 iterations, sampling the entire dataset with replacement. The housing category from MIMIC-G was not evaluated due to low SBDH documentation rate (99% missing, Table 1).

RESULTS

In this study, we included 7,225 clinical notes from 6,382 patients for method development and 971 notes from 437 patients for independent validation (Table 1). The combined patient cohort from the two MIMIC-III datasets was predominantly female (56.3%), white (72.5%), and non-Hispanic (86.1%). Similarly, the UTSW cohort was predominantly female (51.5%), white (55.1%), and non-Hispanic (79.6%). The UTSW cohort had a higher proportion of Black (33.0%) and Hispanic (14.0%) populations compared to the two MIMIC-III cohorts (Table 1). For all SBDH categories except housing, the UTSW datasets had a higher documentation rate. Additional patient characteristics are summarized in Table 1.

The distribution of human-annotated SBDH ground truths across all three datasets are summarized in Supplementary Figure 1. Incorporation of different SBDH terminologies in these datasets highlight the need for a generic and granular terminology to guide prompt design. When annotating the ground truths for the UTSW testing dataset, high inter-annotator agreement (Cohen's Kappa ≥ 0.85; > 98% agreement) was achieved during the initial practice phase (Supplementary Table 2). In total, the original ground truths from 29 notes in the MIMIC-G dataset

and 31 notes in the MIMIC-A dataset were corrected by our team (see examples in Supplementary Table 3).

Through iterative prompt development using MIMIC-G and MIMIC-A, SBDH-Reader achieved high performance in extracting all SBDH categories from these two datasets (Table 2). The macro-average F1 across attributes ranged from 0.83 (MIMIC-A, employment) to 0.98 (MIMIC-G, employment). When focusing on adverse attributes, F1 ranged from 0.79 (MIMIC-A, drug use) to 1.00 (MIMIC-G, employment). For the potential clinical application of capturing adverse SBDH factors, SBDH-Reader demonstrated high sensitivity, with recall $\geq$ 0.87 across all categories. To extract any adverse attributes across all categories, SBDH-Reader achieved an F1 of 0.92 (MIMIC-G) and 0.89 (MIMIC-A), with recall of 0.97 (MIMIC-G) and 0.93 (MIMIC-A). Overall, lower performance was observed in MIMIC-A, with employment (macro-average F1: 0.83) and drug use (0.86) being the lowest-performing categories.

In the independent validation with the UTSW dataset, SBDH-Reader delivered an overall higher performance across all categories when compared with the development datasets (Table 2). The macro-average F1 across attributes ranged from 0.85 (housing) to 0.98 (alcohol use). When focusing on adverse attributes, F1 ranged from 0.94 (employment) to 0.99 (tobacco use). For capturing adverse SBDH factors, SBDH-Reader demonstrated high sensitivity, with recall $\geq$ 0.92 across all categories, while maintaining a high precision $\geq$ 0.90. Performance in identifying adverse substance use (F1: 0.97–0.99) was slightly better than that extracting adverse employment, housing, and marital relationship (F1: 0.94–0.95). To extract any adverse attributes across all categories, SBDH-Reader achieved an F1 of 0.96, recall of 0.97, and precision of 0.96. Class-wise performance details are shown by confusion matrices (Figure 2). High rate of missing documentation was found for housing and drug use. SBDH-Reader demonstrated consistent performance when detecting adverse and non-adverse attributes. In one exception, SBDH-Reader rendered a high proportion (115 out of 810) of missing housing documentation as non-adverse attributes. An additional test of repeatability of SBDH-Reader's performance shows minimal variation across five independent rounds of LLM prompting.

Discussion

The SBDH-Reader is designed to extract and classify detailed SBDH information from real-world clinical notes through direct LLM prompting. Previous NLP-based approaches for SBDH identification have been constrained by problem framing (e.g., fixed entity or relation definitions),[25,29-31] limitations on input length due to model constraints,[25,30,31] and the high labor costs associated with defining and annotating ground truths for model training.[24,32] The emergence of the latest generation of LLMs offers an opportunity to redefine the conventional NLP approach to SBDH extraction. Given that SBDH documentation in medical notes largely aligns with general domain knowledge and common language patterns—unlike specialized medical concepts—vertically training or fine-tuning a model for this task might not be necessary. SBDH-Reader serves as a proof of concept, demonstrating that effective SBDH extraction can be achieved through direct LLM prompting.

One key advantage of direct LLM prompting is the ability to quickly adapt to new SBDH terminologies in a granular and flexible manner during inference. Traditionally, SBDH ontology and terminology lack widely accepted, standardized frameworks, and training models on locally defined SBDH terminologies can limit the accuracy and applicability of the results.[4,17,18] In SBDH-Reader, the active SBDH terminology is invoked solely within the modular prompt instructions rather than embedded in the foundational model itself, allowing category- or attribute-level adjustments depending on future application scenario. This study demonstrates that SBDH-Reader can generate granular attribute-level classifications without imposing predefined positive or negative labels, unlike some existing studies. Such dichotomized distinctions often depend on the specific clinical context and should not be enforced during model training, but rather applied post hoc based on the intended use case, as shown by the evaluation protocol in this study. Additionally, SBDH-Reader's prompt content can be flexibly modified—without altering its core structure—and re-validated to accommodate ad hoc definitions of SBDH terminology. New SBDH categories can also be incorporated into the prompt to expand its domain coverage as needed.

In this study, we conducted the evaluation using three post hoc combined attribute groups—adverse attributes, non-adverse attributes, and no information—for two main reasons. First, clinical applications of SBDH data often prioritize identifying adverse SBDH factors without requiring granular details. Second, the two MIMIC-III datasets lack a consistent SBDH terminology, leading to ground truth labels with varying levels of granularity.[29,46] We included the "no information" group in our evaluation alongside the other two SBDH-containing groups because accurately identifying missing SBDH documentation can serve as a quality-of-care measure, highlighting gaps in documentation across systems, care teams, or patient populations. We reported macro-average metrics instead of micro-average metrics, as seen in some studies, because most SBDH categories are dominated by missing data (Table 1), leading to a highly imbalanced dataset. Reporting only micro-averages, including the "no information" group, could misrepresent performance by obscuring performance in the adverse and non-adverse groups. Additionally, we specifically reported SBDH-Reader's performance in detecting adverse attributes due to their relevance in many clinical applications.

Overall, we observed strong performance across both the prompt development and testing datasets, as measured by F1, precision, and recall. Due to fundamental differences in task framing and in-house corrections to several ground truths in the MIMIC-G and MIMIC-A datasets (Supplementary Table 3), a direct comparison to previous works on this subject was not feasible. However, with an overall F1 > 0.80 and most categories exceeding 0.90, SBDH-Reader demonstrates high performance relative to other reports. Interestingly, SBDH-Reader's performance in the UTSW testing dataset surpassed that in the MIMIC-III development datasets. This improvement may be attributed to the more structured SBDH documentation embedded within UTSW's Epic system. Additionally, based on the extracted evidence text, we found that the supporting text in the UTSW dataset was generally shorter than in the MIMIC-III datasets. This suggests that SBDH documentation in the UTSW cohort was inherently more concise, potentially due to differences in patient populations, healthcare systems, and note types between the two datasets.

Limitations and Future Directions

The primary objective of this study is to establish a proof-of-concept method for extracting SBDH data through LLM prompting. Our results demonstrated minimal variation when re-prompting GPT-4o (Supplementary Table 4), indicating its stability. Future evaluations can be conducted as newer versions of GPTs or other advanced LLMs become available in regulatory-compliant environments. Since SBDH-Reader does not rely on model training or fine-tuning, validating it for new use cases or with different LLMs remains straightforward and efficient. In addition, SBDH-Reader is API based, enabling easier integration into existing care systems and workflows compared with hosting an LLM locally.

The current version of SBDH-Reader can process long medical texts, up to the prompt length limit of the underlying LLM. This provides a significant advantage over earlier methods that focused on sentence- or paragraph-level texts.[25,30,31] For example, GPT-4o can handle up to 128k tokens, which equates to roughly 170k English words. This is more than sufficient for typical medical notes, particularly for capturing the SBDH sections. However, future iterations could extend the scope to incorporate longer sequences of notes for even more comprehensive SBDH information collection. Combining notes written across different care settings and by various authors over a short period could yield more accurate and holistic data about a patient's SBDH. Additionally, a patient-level longitudinal SBDH knowledge base could be established to store long-range SBDH contexts. Such a database could act as a "memory bank," making it available for retrieval-augmented generation in more advanced LLM methods.

To thoroughly validate SBDH-Reader for real-world applications, testing across multiple health systems, different care settings, and diverse patient populations is necessary. During large-scale deployment, the overall cost burden may become a concern depending on data volume and system constraints. Further investigations into which note types and authors yield more accurate and abundant SBDH documentation could help optimize processing priorities and reduce costs. Since most medical notes only dedicate a few sections to SBDH information, a method for effectively trimming full-length notes to focus on these sections could improve efficiency. This trimming process could be rule-based, assuming the use of structured documentation templates by a care team, or LLM-based, where prompts are designed to filter out the relevant SBDH sections or paragraphs.

Conclusion

We developed SBDH-Reader, an LLM-powered method for extracting patient-level, categorized, and granular SBDH information from real-world medical notes. Through effective prompt engineering, the SBDH-Reader demonstrated high performance in extracting and classifying SBDH data across six key categories: employment, housing, marital relationship, and substance use, including alcohol, tobacco, and drug use. While additional validation across diverse health systems, care settings, and patient populations is needed, the SBDH-Reader shows potential as an effective tool for collecting real-world SBDH data to support clinical research and patient care.

Acknowledgements

This study was supported in part by the National Institutes of Health under award numbers R01GM140012 (GX), R01DE030656 (GX), R01GM115473 (GX), U01CA249245 (GX), U01AI169298 (YX), R35GM136375 (YX), the Cancer Prevention and Research Institute of Texas under award numbers RP180805 (YX), RP240521 (YX), RP230330 (GX), and the Texas Health Resources Clinical Scholars Program (DMY).

Competing Interests

The authors have no conflict of interest to disclose.

Data Availability

The MIMIC-III data are available to credentialed users at https://physionet.org. The corrected ground truths generated from this study for the two published MIMIC-III datasets will be made available on PhysioNet. The UTSW dataset is not publicly available due to data and privacy protection policies at UTSW.

Table 1. Cohort characteristics. P-value was based on chi-square test for patient-level characteristics between each MIMIC-III dataset and the UTSW dataset.

	MIMIC-G	**MIMIC-A**	**UTSW**
Role	Development	Development	Testing
Patient-level Characteristics			
Number of patients	183	6199	437
Sex	P < 0.05	P < 0.05	
Male	101 (55.2%)	2687 (43.3%)	212 (48.5%)
Female	82 (44.8%)	3512 (56.7%)	225 (51.5%)
Race	P < 0.05	P < 0.05	
White	131 (71.6%)	4496 (72.5%)	241 (55.1%)
Black	16 (8.7%)	540 (8.7%)	144 (33.0%)
Other	9 (4.9%)	301 (4.9%)	10 (2.3%)
Unknown	27 (14.8%)	862 (13.9%)	42 (9.6%)
Ethnicity	P < 0.05	P < 0.05	
Non-Hispanic	156 (85.2%)	5337 (86.1%)	348 (79.6%)
Hispanic	12 (6.6%)	214 (3.4%)	61(14.0%)
Unknown	15 (8.2%)	648 (10.5%)	28 (6.4%)
Note-level Characteristics			
Number of notes	200	7025	971
Number of notes by type (n)	Nursing (65) Physician (70) Social work (65)	Discharge summary (7025)	Consults (372) H&P (333) Progress (230) ED Provider (36)
Number of notes with SBDH information			
Employment	36 (18.0%)	2730 (38.9%)	382 (39.3%)
Housing	2 (1.0%)	4429 (63.0%)	161 (16.6%)
Marital relationship	64 (32.0%)	N.A.	830 (85.5%)
Alcohol use	N.A.	5036 (71.7%)	899 (92.6%)
Tobacco use	N.A.	5379 (76.6%)	910 (93.7%)
Drug use	N.A.	2564 (36.5%)	884 (91.0%)

N.A.: The ground truths for the SBDH category was not provided in the published dataset.

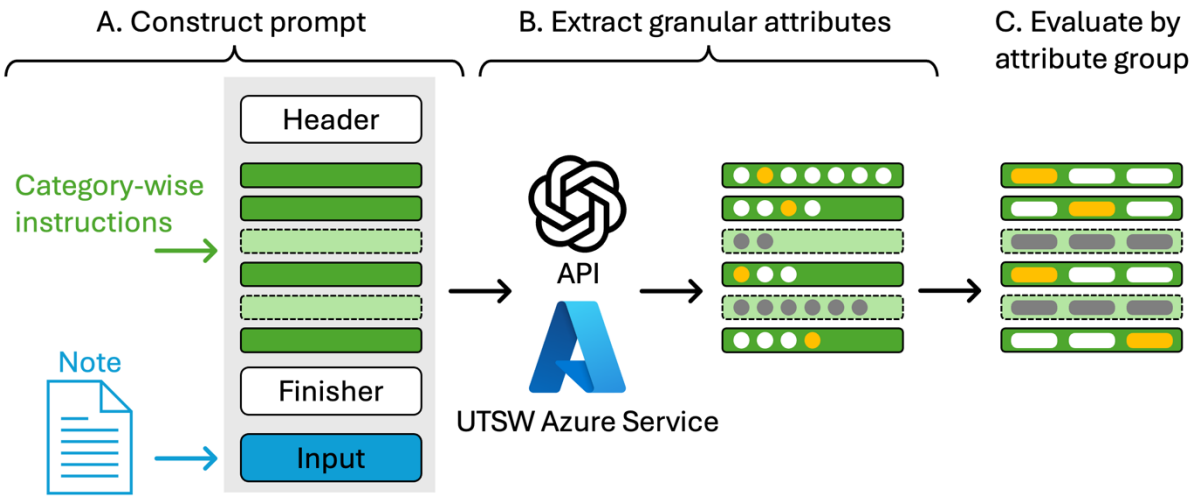


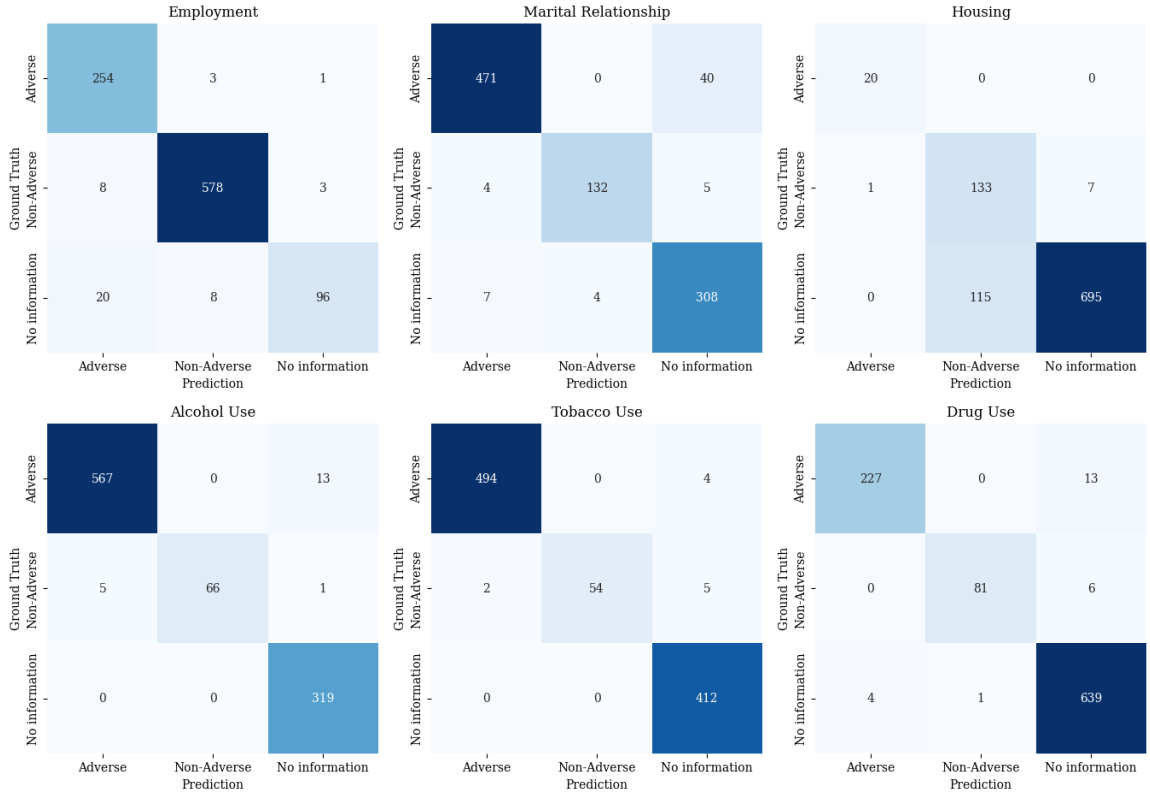
Figure 1. Overview of SBDH-Reader. The three key components of developing and validating SBDH-Reader are: (A) prompt construction, (B) result generation, and (C) evaluation. (A) The LLM prompt for SBDH-Reader is structured into a global header, category-wise instructions (green boxes), a global finisher, and input note (blue box). Each dark green box with a solid edge represents instructions for an included SBDH category, while each light green box with a dashed-line edge represents instructions available in SBDH-Reader but skipped during a given LLM query based on specific operational conditions (e.g., data, task, etc.). (B) The constructed prompt is sent to the UTSW private Azure OpenAI service via API to classify each included SBDH category into granular attributes (yellow circles). All circles represent permissible granular attributes for the LLM to target, while gray circles indicate skipped generations due to excluded SBDH categories (light green boxes). (C) For evaluation, granular attributes are mapped onto three attribute groups—adverse attributes, non-adverse attributes, and no information—using a rule-based mapping (Supplementary Table 1).

Table 2. Performance of SBDH-Reader on development and validation datasets. Note-level results are presented as median followed by (lower–upper) bounds of the 95% confidence interval, estimated by bootstrapping with 500 iterations, sampling the entire dataset with replacement.

	Macro average			Adverse attributes only		
Category	F1	Precision	Recall	F1	Precision	Recall
MIMIC-G						
Employment	0.98 (0.94–1.00)	0.96 (0.91–1.00)	0.99 (0.99–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
Marital Relationship	0.90 (0.83–0.95)	0.89 (0.82–0.95)	0.92 (0.85–0.96)	0.84 (0.69–0.95)	0.75 (0.56–0.94)	0.94 (0.79–1.00)
Overall*	0.94 (0.90–0.96)	0.93 (0.89–0.96)	0.95 (0.92–0.98)	0.92 (0.84–0.98)	0.86 (0.78–0.97)	0.97 (0.89–1.00)
MIMIC-A						
Employment	0.83 (0.83–0.84)	0.82 (0.81–0.83)	0.86 (0.85–0.87)	0.84 (0.83–0.85)	0.80 (0.78–0.82)	0.88 (0.86–0.89)
Housing	0.91 (0.89–0.92)	0.90 (0.88–0.93)	0.92 (0.90–0.93)	0.88 (0.83–0.92)	0.83 (0.76–0.89)	0.93 (0.88–0.98)
Alcohol Use	0.91 (0.91–0.92)	0.93 (0.92–0.94)	0.90 (0.90–0.91)	0.93 (0.92–0.93)	0.89 (0.87–0.90)	0.98 (0.97–0.98)
Tobacco Use	0.94 (0.93–0.94)	0.95 (0.95–0.96)	0.93 (0.92–0.93)	0.97 (0.96–0.97)	0.95 (0.94–0.95)	0.99 (0.98–0.99)
Drug Use	0.86 (0.85–0.87)	0.83 (0.82–0.85)	0.90 (0.89–0.91)	0.79 (0.75–0.81)	0.71 (0.67–0.76)	0.87 (0.84–0.90)
Overall*	0.90 (0.90–0.91)	0.90 (0.89–0.91)	0.91 (0.90–0.91)	0.89 (0.88–0.90)	0.86 (0.84–0.88)	0.93 (0.92–0.94)
UTSW						
Employment	0.93 (0.91–0.95)	0.95 (0.93–0.96)	0.91 (0.89–0.94)	0.94 (0.92–0.96)	0.90 (0.86–0.93)	0.98 (0.97–1.00)
Housing	0.85 (0.82–0.88)	0.81 (0.76–0.85)	0.94 (0.92–0.95)	0.95 (0.86–1.00)	0.91 (0.76–1.00)	1.00 (1.00–1.00)
Marital Relationship	0.94 (0.92–0.95)	0.94 (0.92–0.95)	0.94 (0.92–0.96)	0.95 (0.93–0.96)	0.98 (0.96–0.99)	0.92 (0.90–0.94)
Alcohol Use	0.98 (0.96–0.99)	0.98 (0.97–0.99)	0.97 (0.95–0.99)	0.98 (0.96–0.99)	0.99 (0.99–1.00)	0.97 (0.96–0.99)
Tobacco Use	0.97 (0.96–0.99)	0.99 (0.99–1.00)	0.96 (0.93–0.98)	0.99 (0.99–1.00)	1.00 (0.99–1.00)	0.99 (0.98–1.00)
Drug Use	0.97 (0.95–0.98)	0.98 (0.97–0.99)	0.96 (0.94–0.98)	0.97 (0.95–0.98)	0.98 (0.97–1.00)	0.95 (0.92–0.98)
Overall	0.94 (0.93–0.95)	0.94 (0.93–0.95)	0.95 (0.94–0.95)	0.96 (0.95–0.97)	0.96 (0.93–0.98)	0.97 (0.96–0.98)

* The overall metrices report macro-average across all SBDH categories.

Figure 2. Confusion matrices for each SBDH category in UTSW validation dataset.



Supplementary Info 1. Example prompt and output. Prompt structure is outlined (Figure 1). For privacy protection, the example outputs for various SBDH categories were extracted from different patients.

Example Prompt

<global header>

```
   This prompt consists of two sections: "Instruction", and "Input" which contains a clinical
note from EHR.
   Your task is to follow the "Instruction" and extract from the "Input" info about Social and
Behavioral Determinants of Health (SBDH).
```

<category-wise instructions>

```
   Section 1: Instruction
   From the double bracketed <<text>> at the end, for each of the SBDH Categories listed below,
determine the SBDH Attribute value that accurately describes the status of the patient at the
time of <<text>>.
   In each section below, each SBDH Category is shown in double quote and permissible SBDH
Attributes are shown in square brackets:

    "EMPLOYMENT": ['EMPLOYMENT_Employed', 'EMPLOYMENT_Unemployed', 'EMPLOYMENT_Underemployed',
'EMPLOYMENT_Retired', 'EMPLOYMENT_Student', 'EMPLOYMENT_Unknown'];

    "HOUSING": ['HOUSING_Issue', 'HOUSING_NoIssue', 'HOUSING_Unknown'];

    "MARITAL": ['MARITAL_Married', 'MARITAL_Partnered', 'MARITAL_Widowed', 'MARITAL_Divorced',
'MARITAL_OtherSingleness', 'MARITAL_Other', 'MARITAL_Unknown'];

    "ALCOHOL": ['ALCOHOL_Current', 'ALCOHOL_Past', 'ALCOHOL_Never', 'ALCOHOL_Unknown'];

    "TOBACCO": ['TOBACCO_Current', 'TOBACCO_Past', 'TOBACCO_Never', 'TOBACCO_Unknown'];

    "DRUG": ['DRUG_Current', 'DRUG_Past', 'DRUG_Never', 'DRUG_Unknown'];

Further explanations/specifications of each Attribute is provided below:

    The permissible EMPLOYMENT attributes are: ['EMPLOYMENT_Employed', 'EMPLOYMENT_Unemployed',
'EMPLOYMENT_Underemployed', 'EMPLOYMENT_Retired', 'EMPLOYMENT_Student', 'EMPLOYMENT_Unknown'].
```

<attribute-wise specifications are omitted in this example>

```
    The permissible HOUSING attributes are: ['HOUSING_Issue', 'HOUSING_NoIssue',
'HOUSING_Unknown'].
```

<attribute-wise specifications are omitted in this example>

```
    The permissible MARITAL attributes are: ['MARITAL_Married',
'MARITAL_Partnered','MARITAL_Widowed','MARITAL_Divorced','MARITAL_OtherSingleness','MARITAL_Other
','MARITAL_Unknown',].
```

<attribute-wise specifications are omitted in this example>

```
    The permissible ALCOHOL attributes are: ['ALCOHOL_Current', 'ALCOHOL_Past', 'ALCOHOL_Never',
'ALCOHOL_Unknown'].
```

<attribute-wise specifications are omitted in this example>

```
    The permissible TOBACCO attributes are: ['TOBACCO_Current', 'TOBACCO_Past', 'TOBACCO_Never',
'TOBACCO_Unknown'].
```

<attribute-wise specifications are omitted in this example>

The permissible DRUG attributes are: ['DRUG_Current', 'DRUG_Past', 'DRUG_Never', 'DRUG_Unknown'].

<attribute-wise specifications are omitted in this example>

<global finisher>

You must create key-value pairs in JSON formats:
<CATEGORY>: <estimated Attribute from the list of permissible Attribute values>,
<CATEGORY>_evidence: < evidence extracted from <<text>> supporting the Attribute you choose >.

Even if the patient expresses concerns/predictions over an upcoming change in their SBDH Attribute status, you should still assign Attribute values only based on patient's status maintained at the time of <<text>>.
For example "She works full time at a company, but fears that she will lose her current job" is considered “EMPLOYMENT_Employed”.
Focus on the status of the patient only, and NOT the status of their family members or other personnel involved in the <<text>>.
Do not infer the impact of <<text>> onto the SBDH Attributes, focus on the factual information only.

Section 2: Input
<<Clinical note entry>>

<input note text>

Example Prompt


```
{
  "EMPLOYMENT": "EMPLOYMENT_Employed",
  "EMPLOYMENT_evidence": "he has got a steady job at a [redacted]",
  "HOUSING": "HOUSING_Issue",
  "HOUSING_evidence": "Patient is homeless",
  "MARITAL": "MARITAL_Unknown",
  "MARITAL_evidence": "No information provided",
  "ALCOHOL": " ALCOHOL_Current",
  "ALCOHOL_evidence": "She occasionally has wine.",
  "TOBACCO": "TOBACCO_Past",
  "TOBACCO_evidence": "Absence of current tobacco use;",
  "DRUG": "DRUG_Unknown",
  "DRUG_evidence": " No mention of drug usage in the text."
}
```

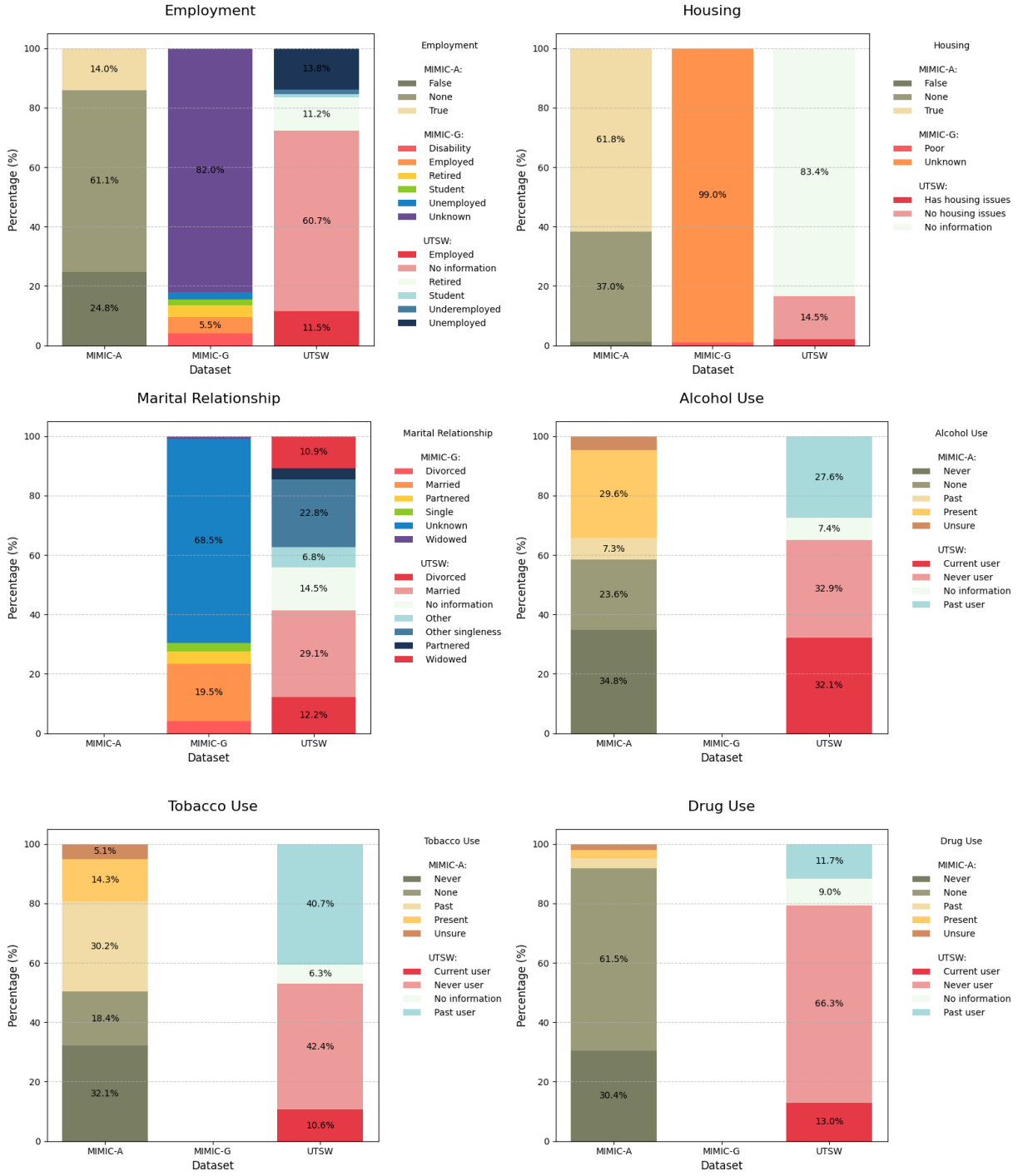
Supplementary Table 1. SBDH terminology. Each category also allows a universal "no information" attribute when no explicit evidence of other attributes is identified.

Category	Attribute	Class (evaluation only)	Definition
Employment	Employed	Non-adverse	Engaged in paid work, typically full-time or part-time position.
	Unemployed	Adverse	Without paid work but actively seeking job opportunities or employment.
	Underemployed	Adverse	Employed below skill level, earning insufficient income or part-time.
	Retired	Adverse	No longer working, typically after career completion, receiving retirement benefits.
	Student	Non-adverse	Actively enrolled in academic programs, not yet in full-time employment.
Housing	No housing issues	Non-adverse	In a non-adverse housing status, or implies a non-adverse housing status, such as "lives with spouse".
	Has housing issues	Adverse	In an adverse housing status, such as homeless, undomiciled for various reasons, or financial strains for housing.
Marital Relationship	Married	Non-adverse	Legally married to a spouse.
	Partnered	Non-adverse	In a long-term, committed relationship, often cohabiting, but not legally married.
	Widowed	Adverse	A spouse or partner has passed away.
	Divorced	Adverse	Legally dissolved a marriage.
	Other singleness	Adverse	Not currently in a marital or partnered relationship
	Other	Adverse	Otherwise specified
Alcohol Use	Current user	Adverse	Patient is a current user.
	Past user	Adverse	Patient was a user in the past but is not a current user.
	Never user	Non-adverse	Patient had never been a user.
Tobacco Use	Current user	Adverse	Patient is a current user.
	Past user	Adverse	Patient was a user in the past but is not a current user.
	Never user	Non-adverse	Patient had never been a user.
Drug Use	Current user	Adverse	Patient is a current user.
	Past user	Adverse	Patient was a user in the past but is not a current user.
	Never user	Non-adverse	Patient had never been a user.

Supplementary Table 2. Inter-annotator agreement for the first 100 notes in UTSW dataset.

Category	Cohen's Kappa	Percent Agreement
Employment	0.97	98%
Housing	0.85	98%
Marital Relationship	0.97	98%
Alcohol Use	0.97	98%
Tobacco Use	0.98	99%
Drug Use	0.98	99%

Supplementary Figure 1. Summary of SBDH information for three datasets based on human-annotated ground truths. Ground-truth SBDH attributes follow different terminology in the two MIMIC-III datasets.



Supplementary Table 3. Examples of corrected ground truths in the MIMIC-G and MIMIC-A datasets.

Category	Note ID	Description of the GT Error	Original GT	Corrected GT
Employment	674757	Note indicates that the patient is a manager at company [redacted]	Unemployed	Employed
Housing	41036	Note describes that the patient lives in homeless shelter.	No Issue	Issue
Marital Relationship	611542	This note comes from a social worker who documented information provided by the patient's brother and husband. The note mentions that while the brother's wife is deceased, the patient's husband has been actively discussing her condition with the physician and staff.	Widowed	Married

GT: Ground truth.

Supplementary Table 4. SBDH-Reader's performance variation across five independent rounds of LLM prompting. The median and standard deviation are estimated from five independent promptings using GPT-4o.

	Macro average			Adverse attributes only		
Category	F1	Precision	Recall	F1	Precision	Recall
MIMIC-G						
Employment	0.98 ± 0.00	0.96 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Marital Relationship	0.90 ± 0.01	0.89 ± 0.00	0.91 ± 0.02	0.83 ± 0.01	0.75 ± 0.00	0.94 ± 0.03
MIMIC-A						
Employment	0.83 ± 0.00	0.81 ± 0.01	0.86 ± 0.00	0.84 ± 0.00	0.80 ± 0.00	0.87 ± 0.00
Housing	0.90 ± 0.00	0.89 ± 0.01	0.91 ± 0.00	0.85 ± 0.01	0.79 ± 0.02	0.92 ± 0.00
Alcohol use	0.91 ± 0.00	0.93 ± 0.00	0.90 ± 0.00	0.93 ± 0.00	0.89 ± 0.00	0.98 ± 0.00
Tobacco use	0.94 ± 0.00	0.95 ± 0.01	0.93 ± 0.00	0.97 ± 0.00	0.95 ± 0.00	0.99 ± 0.00
Drug use	0.88 ± 0.00	0.86 ± 0.00	0.91 ± 0.00	0.82 ± 0.01	0.77 ± 0.00	0.87 ± 0.00
UTSW						
Employment	0.93 ±0.00	0.95 ±0.00	0.91 ±0.00	0.94 ±0.00	0.90 ±0.00	0.98 ±0.00
Housing	0.85 ±0.00	0.81 ±0.01	0.94 ±0.00	0.95 ±0.01	0.91 ±0.02	1.00 ±0.00
Marital Relationship	0.94 ±0.00	0.94 ±0.00	0.94 ±0.00	0.95 ±0.00	0.98 ±0.00	0.92 ±0.00
Alcohol use	0.97 ±0.01	0.98 ±0.00	0.97 ±0.01	0.98 ±0.00	0.99 ±0.00	0.97 ±0.01
Tobacco use	0.97 ±0.00	0.99 ±0.00	0.96 ±0.00	0.99 ±0.00	1.00 ±0.00	0.99 ±0.00
Drug use	0.97 ±0.00	0.98 ±0.00	0.96 ±0.00	0.97 ±0.01	0.98 ±0.00	0.95 ±0.01

References

1. Marmot M, Friel S, Bell R, Houweling TA, Taylor S, Commission on Social Determinants of H. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet*. Nov 8 2008;372(9650):1661-9. doi:10.1016/S0140-6736(08)61690-6
2. Adler NE, Glymour MM, Fielding J. Addressing Social Determinants of Health and Health Inequalities. *JAMA*. Oct 25 2016;316(16):1641-1642. doi:10.1001/jama.2016.14058
3. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychol Rev*. 2015;9(3):323-44. doi:10.1080/17437199.2014.941722
4. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *J Am Med Inform Assoc*. Nov 1 2020;27(11):1764-1773. doi:10.1093/jamia/ocaa143
5. Lu MLR, Davila CD, Shah M, et al. Marital status and living condition as predictors of mortality and readmissions among African Americans with heart failure. *International Journal of Cardiology*. Nov 1 2016;222:313-318. doi:10.1016/j.ijcard.2016.07.185
6. Sterling MR, Ringel JB, Pinheiro LC, et al. Social Determinants of Health and 90-Day Mortality After Hospitalization for Heart Failure in the REGARDS Study. *J Am Heart Assoc*. May 5 2020;9(9):e014836. doi:10.1161/JAHA.119.014836
7. Segar MW, Hall JL, Jhund PS, et al. Machine Learning-Based Models Incorporating Social Determinants of Health vs Traditional Models for Predicting In-Hospital Mortality in Patients With Heart Failure. *JAMA Cardiol*. Aug 1 2022;7(8):844-854. doi:10.1001/jamacardio.2022.1900
8. Hiatt RA, Breen N. The social determinants of cancer - A challenge for transdisciplinary science. *American Journal of Preventive Medicine*. Aug 2008;35(2):S141-S150. doi:10.1016/j.amepre.2008.05.006
9. Adkins-Jackson PB, George KM, Besser LM, et al. The structural and social determinants of Alzheimer's disease related dementias. *Alzheimers Dement*. Jul 2023;19(7):3171-3185. doi:10.1002/alz.13027
10. Sekar RR, Herrel LA, Stensland KD. Social Determinants of Health and the Availability of Cancer Clinical Trials in the United States. *JAMA Netw Open*. May 1 2024;7(5):e2410162. doi:10.1001/jamanetworkopen.2024.10162
11. Rae S, Shaya S, Taylor E, et al. Social determinants of health inequalities in early phase clinical trials in Northern England. *Br J Cancer*. Sep 2024;131(4):685-691. doi:10.1038/s41416-024-02765-w
12. Hatef E, Rouhizadeh M, Tia I, et al. 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*. Aug 2 2019;7(3):e13802. doi:10.2196/13802
13. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of Social Determinants of Health ICD-10 Z-Codes Among Hospitalized Patients in the United States, 2016-2017. *Med Care*. Dec 2020;58(12):1037-1043. doi:10.1097/MLR.0000000000001418

14. Dupre ME, Nelson A, Lynch SM, et al. Socioeconomic, Psychosocial and Behavioral Characteristics of Patients Hospitalized With Cardiovascular Disease. *Am J Med Sci*. Dec 2017;354(6):565-572. doi:10.1016/j.amjms.2017.07.011
15. Dupre ME, Nelson A, Lynch SM, et al. Identifying Nonclinical Factors Associated With 30-Day Readmission in Patients with Cardiovascular Disease: Protocol for an Observational Study. *JMIR Res Protoc*. Jun 15 2017;6(6):e118. doi:10.2196/resprot.7434
16. Guo Y, Chen Z, Xu K, et al. International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. *Medicine (Baltimore)*. Dec 24 2020;99(52):e23818. doi:10.1097/MD.0000000000023818
17. Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform*. Nov 2017;107:101-106. doi:10.1016/j.ijmedinf.2017.09.008
18. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*. Nov 25 2021;28(12):2716-2727. doi:10.1093/jamia/ocab170
19. Mullangi S, Aviki EM, Hershman DL. Reexamining Social Determinants of Health Data Collection in the COVID-19 Era. *JAMA Oncol*. Dec 1 2022;8(12):1736-1738. doi:10.1001/jamaoncol.2022.4543
20. Ong JCL, Seng BJJ, Law JZF, et al. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: Current state and future directions. *Cell Rep Med*. Jan 16 2024;5(1):101356. doi:10.1016/j.xcrm.2023.101356
21. Bompelli A, Wang Y, Wan R, et al. Social and Behavioral Determinants of Health in the Era of Artificial Intelligence with Electronic Health Records: A Scoping Review. *Health Data Sci*. 2021;2021:9759016. doi:10.34133/2021/9759016
22. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying Patients with Significant Problems Related to Social Determinants of Health with Natural Language Processing. *Stud Health Technol Inform*. Aug 21 2019;264:1456-1457. doi:10.3233/SHTI190482
23. Reeves RM, Christensen L, Brown JR, et al. Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. *J Biomed Inform*. Aug 2021;120:103851. doi:10.1016/j.jbi.2021.103851
24. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J Biomed Inform*. Jan 2021;113:103631. doi:10.1016/j.jbi.2020.103631
25. Han S, Zhang RF, Shi L, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J Biomed Inform*. Mar 2022;127:103984. doi:10.1016/j.jbi.2021.103984
26. Raza S, Dolatabadi E, Ondrusek N, Rosella L, Schwartz B. Discovering social determinants of health from case reports using natural language processing: algorithmic development and validation. *BMC Digital Health*. 2023/09/11 2023;1(1):35. doi:10.1186/s44247-023-00035-y

27. Wu W, Holkeboer KJ, Kolawole TO, Carbone L, Mahmoudi E. Natural language processing to identify social determinants of health in Alzheimer's disease and related dementia from electronic health records. *Health Serv Res*. Dec 2023;58(6):1292-1302. doi:10.1111/1475-6773.14210
28. Roy S, Morrell S, Zhao L, Homayouni R. Large-scale identification of social and behavioral determinants of health from clinical notes: comparison of Latent Semantic Indexing and Generative Pretrained Transformer (GPT) models. *BMC Med Inform Decis Mak*. Oct 10 2024;24(1):296. doi:10.1186/s12911-024-02705-x
29. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. Jan 11 2024;7(1):6. doi:10.1038/s41746-023-00970-0
30. Gabriel RA, Litake O, Simpson S, Burton BN, Waterman RS, Macias AA. On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proc Natl Acad Sci U S A*. Sep 24 2024;121(39):e2320716121. doi:10.1073/pnas.2320716121
31. Keloth VK, Selek S, Chen Q, et al. Large Language Models for Social Determinants of Health Information Extraction from Clinical Notes - A Generalizable Approach across Institutions. *medRxiv*. May 22 2024;doi:10.1101/2024.05.21.24307726
32. Lybarger K, Yetisgen M, Uzuner O. The 2022 n2c2/UW shared task on extracting social determinants of health. *J Am Med Inform Assoc*. Jul 19 2023;30(8):1367-1378. doi:10.1093/jamia/ocad012
33. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. *Appl Clin Inform*. Jan 2020;11(1):172-181. doi:10.1055/s-0040-1702214
34. Yu Z, Peng C, Yang X, et al. Identifying social determinants of health from clinical narratives: A study of performance, documentation ratio, and potential bias. *J Biomed Inform*. May 2024;153:104642. doi:10.1016/j.jbi.2024.104642
35. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. Aug 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8
36. Will ChatGPT transform healthcare? *Nat Med*. Mar 2023;29(3):505-506. doi:10.1038/s41591-023-02289-5
37. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. Mar 2023;5(3):e107-e108. doi:10.1016/S2589-7500(23)00021-3
38. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. *JAMA*. Sep 5 2023;330(9):866-869. doi:10.1001/jama.2023.14217
39. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. Jun 1 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838
40. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*. Apr 2023;90:104512. doi:10.1016/j.ebiom.2023.104512
41. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. Dec 26 2022;5(1):194. doi:10.1038/s41746-022-00742-2

42. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ Digit Med*. May 1 2024;7(1):106. doi:10.1038/s41746-024-01079-8
43. Nezafati K, Wang L, Rong R, et al. Assessing disease severity in cutaneous lupus patients using natural language processing: Preliminary data from a cohort study. *J Am Acad Dermatol*. Nov 28 2024;doi:10.1016/j.jaad.2024.10.105
44. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open*. Jul 2021;4(3):ooaa069. doi:10.1093/jamiaopen/ooaa069
45. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. May 24 2016;3:160035. doi:10.1038/sdata.2016.35
46. Ahsan H, Ohnuki E, Mitra A, Yu H. MIMIC-SBDH: A Dataset for Social and Behavioral Determinants of Health. *Proc Mach Learn Res*. Aug 2021;149:391-413.
47. PhysioNet. PhysioNet Credentialed Health Data License 1.5.0. Accessed 2/19/2025, https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150/
48. PhysioNet. Responsible use of MIMIC data with online services like GPT. Accessed 1/19/2025, 2025. https://physionet.org/news/post/gpt-responsible-use
49. Klie JC, Bugert M, Boullosa B, Eckart de Castilho R, Gurevych I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. presented at: In Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics; 2018; Santa Fe, NM.