MIMIC-SBDH: A Dataset for Social and Behavioral Determinants of Health

Hiba Ahsan HAHSAN@UMASS.EDU
Emmie Ohnuki EOHNUKI@UMASS.EDU
Avijit Mitra AVIJITMITRA@UMASS.EDU
College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA

Hong Yu HONG.YU@UMASSMED.EDU
College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA
Department of Computer Science, University of Massachusetts, Lowell, MA, USA
Center for Healthcare Organization & Implementation Research, Bedford, MA, USA
Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA

Abstract

Social and Behavioral Determinants of Health (SBDHs) are environmental and behavioral factors that have a profound impact on health and related outcomes. Given their importance, physicians document SBDHs of their patients in Electronic Health Records (EHRs). However, SBDHs are mostly documented in unstructured EHR notes. Determining the status of the SBDHs requires manually reviewing the notes which can be a tedious process. Therefore, there is a need to automate identifying the patients' SBDH status in EHR notes. In this work, we created MIMIC-SBDH[1], the first publicly available dataset of EHR notes annotated for patients' SBDH status. Specifically, we annotated $7,025$ discharge summary notes for the status of 7 SBDHs as well as marked SBDH-related keywords. Using this annotated data for training and evaluation, we evaluated the performance of three machine learning models (Random Forest, XGBoost, and Bio-ClinicalBERT) on the task of identifying SBDH status in EHR notes. The performance ranged from the lowest 0.69 F1 score for Drug Use to the highest 0.96 F1 score for Community-Present. In addition to standard evaluation metrics such as the F1 score, we evaluated four capabilities that a model must possess to perform well on the task using the CheckList tool (Ribeiro et al., 2020). The results revealed several shortcomings of the models. Our results highlighted the need to perform more capability-centric evaluations in addition to standard metric comparisons.

1. Introduction

Social determinants of health (SDOH) are "the conditions in which people are born, grow, live, work and age," which are "shaped by the distribution of money, power and resources" (WHO, 2008). They include factors such as socio-economic status, education, neighborhood and physical environment, employment, social support networks, and access to health care. Behavioral determinants of health include tobacco use, drug use, alcohol consumption, physical activity and diet. Together, social and behavioral determinants of health (SBDHs) are environmental and behavioral factors that impact health in significant ways.

[1] https://github.com/hiba008/MIMIC-SBDH

Several studies have shown the impact of SBDHs on health outcomes. Nijhawan et al., 2019 showed that variables derived from SBDHs such as substance use, and access to food contributed to the 30-day readmission prediction task. Takahashi et al., 2015 concluded that education level, unemployment status, and alcohol use significantly impacted the risk of hospitalization. Zheng et al., 2020 showed that substance abuse, low income, and unemployment could be used for identification of high-risk patients for suicide attempt.

Given their importance, physicians document information about their patients' SBDHs in Electronic Health Records (EHRs). Information about SBDHs is useful to both researchers and clinicians. Researchers study how SBDHs impact health outcomes whereas clinicians can identify the social and behavioral risks of their patients and consequently provide tailored care such as counselling or therapy (Haas et al., 2015) and social service (Hamilton et al., 2012). The challenge for both clinicians and researchers is that SBDHs are mostly recorded in unstructured EHR notes. A study has shown that EHR notes contain 90 times more information about SBDHs than structured EHR data (Dorr et al., 2019). However, there is no globally accepted format to consistently record SBDHs in EHR notes and manually reviewing the notes for SBDHs is a time-intensive process. Therefore, there is a need to automate identifying the patient's SBDH status in EHR notes. Although natural language processing (NLP) approaches have been developed to automatically identify the status of several SBDHs in EHR notes (Gundlapalli et al., 2013; Alzoubi et al., 2018; Feller et al., 2020), the notes and associated annotations used to develop the approaches are not publicly available.

In this work, we release MIMIC-SBDH, the first publicly available dataset of EHR notes annotated for patients' SBDH status. For this, we annotated $7,025$ discharge summaries randomly selected from the MIMIC III (Johnson et al., 2016) dataset for the following SBDHs: Community, Economics, Education, Environment, Alcohol Use, Tobacco Use, and Drug Use. In addition, we marked SBDH-related keywords to better understand the language used to discuss them. We treated the task of identifying an SBDH's status as a classification task and studied three baseline models: two tree-based approaches (Random Forest Classifier and XGBoost Classifier) and a neural network (Bio-ClinicalBERT) for each SBDH. While we evaluated the models in terms of macro-F1, the statistic alone is not enough to understand whether the models possess certain desirable capabilities or to understand where the models are failing. So with the use of the CheckList tool (Ribeiro et al., 2020), we performed behavioral testing of the models to understand the following capabilities: (1) *Negation* (2) *Attribution* (3) *Historical Phrases* and (4) *Robustness to Misspellings.*

Our contributions are three-fold: (1) We created a publicly available dataset of EHR notes with annotations for the status of 7 SBDHs (2) We evaluated three baseline machine learning models on the task of identifying an SBDH's status in EHR notes (3) We presented four capabilities that a model must possess to perform well on the task and performed behavioral testing of the models for these capabilities. The results revealed several shortcomings of the models and highlighted the need to perform more capability-centric evaluations in addition to standard metric comparisons.

Generalizable Insights about Machine Learning in the Context of Healthcare

By publicly releasing our annotated dataset, we hope to create a benchmark dataset to promote research in SBDHs. An accurate identification of the SBDHs' status can catalyze studies about how these determinants impact health outcomes. We present four capabilities that a model must possess to do well on the task of predicting an SBDH's status using EHR notes. We believe these capabilities will be helpful when designing a machine learning model for the task, regardless of which medical institution the data is sourced from. We go beyond using standard metrics to evaluate models by performing behavioral testing. The results provide us with actionable insights about a model's shortcomings and we hope that this observation will encourage researchers in incorporating more capability-centric evaluations in their experiments.

2. Related Work

Past works attempting to identify an SBDH's status in EHR notes have predominantly focused on detecting a single SBDH using various techniques. As part of the i2b2 Smoking Challenge (Uzuner et al., 2008), an EHR was classified into one of the five categories: Current smoker, Past smoker, Past or Current smoker, Non-smoker or Unknown smoker. Wicentowski and Sydes, 2008 used a rule-based models comprising of rules based on lexical and syntactic properties to solve classification the task. Carrero et al., 2006 experimented with supervised approaches as Decision Trees and Naive Bayes. Pedersen, 2006 used supervised methods such as Support Vector Machines, Decision Trees and Naive Bayes, as well as unsupervised approaches such as k-means clustering. Aramaki et al., 2006 experimented with breaking down the document into sentences and then combining sentence-based classification results to get the overall result. Jonnagaddala et al., 2015 trained a linear SVM to perform multiclass classification. They used a feature set comprising of traditional features such as unigrams and bigrams as well as topics generated using Latent Dirichlet Allocation (LDA) and Gibbs Sampling.

Similar approaches have been used to detect alcohol consumption. Alzoubi et al., 2018 adopted a multi-stage process. They identified alcohol-related sentences using a rule-based approach, performed supervised classification of the sentences into Current drinker, Past drinker, Non-drinker and Unknown, and combined the classifications to a document-level classification based on a set of rules. Topaz et al., 2019 used neural embeddings such as word2vec (Mikolov et al., 2013) and phrase2vec (Wu et al., 2020) as features and Random Forest as a classifier to identify alcohol abuse in MIMIC II notes. Work has also been in detecting homelessness using EHRs. Gundlapalli et al., 2013 investigated detecting homelessness among US Veterans in clinical text using an NLP tool developed by the US Veteran Affairs.

Earlier studies to predict the status of multiple SBDHs have employed various techniques. Yetisgen and Vanderwende, 2017, Lybarger et al., 2018 looked into automatically detecting substance abuse in clinical notes using multi-task learning. Feller et al., 2020 inferred the presence of 11 SBDH categories including alcohol abuse, homelessness, and sexual orientation. They used structured as well as unstructured data and adopted a two-stage process of first identifying notes with related discussions through binary classification followed by running another set of classifiers to predict the SBDH status.

In our work, in addition to tree-based approaches, we explore the application of transfer learning by fine-tuning Bio-ClinicalBERT for the downstream task of identifying an SBDH's status. Previous work evaluate models in terms of F1 metric. We additionally evaluate models by performing behavioral testing to understand whether the models possess certain capabilities that are useful for the task.

3. MIMIC-SBDH

In this section, we describe the data and the SBDHs, the annotation guidelines and the characteristics of the annotated dataset.

3.1. EHR Data

To create our dataset, we used intensive care unit (ICU) patient notes of type 'Discharge summary' from the MIMIC III dataset. We considered discharge summaries as they contain the most comprehensive clinical information including information about a patient's social history. We extracted the social history section from each discharge summary using MedSpaCy's `clinical_sectionizer` which performs pattern-based section extraction. We specifically extracted sections with the header "social history" and randomly selected $7,025$ discharge summaries that contained this section to annotate. Two individuals annotated the sections under the supervision of a senior physician. In case of an annotation disagreement, the physician was consulted and their decision prevailed. An inter-annotator agreement (κ coefficient) of 0.898 was observed across all SBDHs on a subset of 1000 summaries.

Social History	**CP**	**CA**	**ED**	**EC**	**EN**	**AU**	**TU**	**DU**
married never smoked 2 glasses wine/day no drug history	*True*	*False*	*False*	*None*	*None*	*Present*	*Never*	*Never*
Past smoker, stopped >1year ago. He is retired.	*False*	*False*	*False*	*False*	*None*	*None*	*Past*	*None*
Pt is homeless. Denies IVDA. Smokes 2 packs a day.	*False*	*False*	*False*	*None*	*False*	*None*	*Present*	*Absent*

Table 1: Examples of annotations from the dataset. The colored text indicate SBDH-related keywords. **CP**: Community-Present, **CA**: Community-Absent, **EC**: Economics, **ED**: Education, **EN**: Environment, **AU**: Alcohol Use, **TU**: Tobacco Use, **DU**: Drug Use.

3.2. Social and Behavioral Determinants of Health

We annotated the discharge summaries based on the following SDOHs provided by the Kaiser Family Foundation (KFF) (Artiga and Hinton, 2019): (1) Community (2) Economics (3) Education (4) Environment (5) Food and (6) Healthcare. In addition to the above SDOHs, we annotated for any relevance to the following substances to capture behavioral determinants of health: (1) Alcohol Use (2) Tobacco Use and (3) Drug Use.

3.3. Annotation Guidelines

We performed two annotation tasks: (1) We labeled the discharge summaries for each SBDH based on the extracted social history section (2) We marked keywords relevant to the SBDH in the extracted section. Passages related to social support such as a family member or friend were considered relevant to the SBDH Community. Since it was possible for a patient to have active social support and to have lost social support (due to death, separation or divorce) simultaneously, we split Community further to Community-Present and Community-Absent. A discharge summary was annotated *True* for Community-Present if the discharge summary had passages related to active social support and *False* if such passages were not found in the discharge summary. A discharge summary was annotated *True* for Community-Absent if the discharge summary had passages related to the loss of social support and *False* if such passages were not found in the discharge summary. Similarly, a discharge summary was annotated as *True* for the SBDH Education if there was any passage related to the patient's education such as schooling, college or degree attained and *False* if there was no related passage.

A discharge summary was annotated as *True* for the SBDH Economics if the patient was currently employed, *False* if the patient was unemployed (including retirement) and *None* if there was no related passage. A discharge summary was annotated as *True* for the SBDH Environment if there was any indication of housing, *False* if the patient was homeless and *None* if there was no related passage.

For the behavioral factor Alcohol Use, a discharge summary was annotated as *Present* if the patient was a current consumer of alcohol, *Past* if the patient was a consumer in the past and had quit, *Never* if the patient had never consumed alcohol, *Unsure* if the discharge summary had an ambiguous passage related to alcohol consumption and *None* if there was no related text. The same rules applied to Tobacco Use and Drug Use. We did not find any passage relevant to the SBDHs Food and Healthcare in the discharge summaries and hence dropped the two categories. Table 1 shows examples of annotations from the dataset. Examples of each class for every SBDH and more annotation remarks are in Appendix A.1.

3.4. Characteristics of Dataset

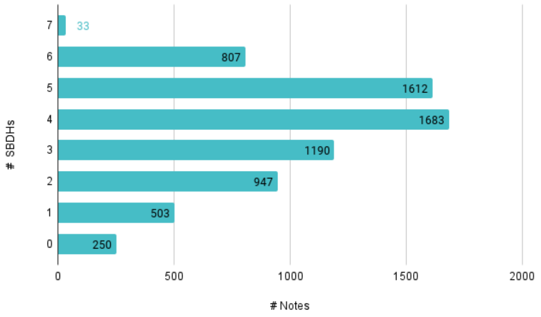


Figure 1: Frequency distribution of SBDHs per discharge summary

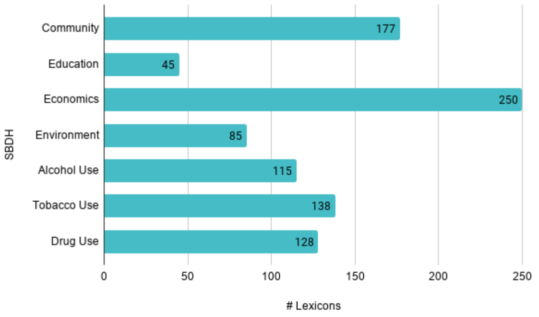


Figure 2: Frequency distribution of lexicons per SBDH

The social history sections extracted had an average length of 35.44 ± 27.69 tokens. The frequency distribution of the number of SBDHs discussed per discharge summary is shown in Figure 1. Out of the $7,025$ annotated discharge summaries, over 95% discussed at least one SBDH with Tobacco Use ($5,734$) and Alcohol Use ($5,368$) being the most frequently occurring ones and Education (201) being the least frequent.

In terms of lexical diversity of the highlighted keywords, as shown in Figure 2, Economics (250) was the most lexically diverse SBDH since the terms discussed various professions followed by Community since the terms discussed relationships. Education (45) was the least lexically diverse SBDH. Keywords across all SBDHs were primarily unigrams (token count distribution: 1.027 ± 0.18).

Tables 2 and 3 describe the class distribution among the SBDHs. SBDHs Education, Environment, Alcohol Use, Tobacco Use and Drug Use showed a significant class imbalance.

SBDH	***True***	***False***	***None***
Community-Present	4463	2562	N/A
Community-Absent	784	6241	N/A
Education	210	6815	N/A
Economics	988	1742	4295
Environment	4357	63	2605

Table 2: Class distribution for Social Determinants of Health

SBDH	***Present***	***Past***	***Never***	***Unsure***	***None***
Alcohol Use	2077	515	2444	332	1657
Tobacco Use	1006	2121	2252	355	1291
Drug Use	207	221	2136	144	4317

Table 3: Class distribution for Behavioral Determinants of Health

4. Methods

4.1. Classifiers

We treated the task of identifying an SBDH's status as a classification task. For each SBDH, we trained the following classifiers:

1. **Random Forest:** Random Forest (RF) is an ensemble of decision trees. The training data is repeatedly "bagged" to create random subsets which the trees are trained to fit to. In addition, the node splitting decision for each tree is affected by not all features but a random subset of them. Random Forest classifiers have been successfully deployed in clinical text classification tasks including recovering missing text labels in EHRs (Yang et al., 2018) and identifying patients with systemic lupus erythematosus (SLE) (Turner et al., 2017).

2. **XGBoost:** XGBoost (eXtreme Gradient Boosting) is a gradient boosting model. Gradient boosting creates a collection of weak regression trees by iteratively adding trees to optimize a differentiable loss function. XGBoost is a computationally efficient implementation that leverages optimization principles including parallelization,

distributed training and cache optimization. XGBoost has shown promising results in clinical text classification tasks such as hierarchical text classification (Stein et al., 2019) and disease classification (Garg et al., 2019).

3. **Bio-Clinical BERT:** Pre-trained deep learning models have been successfully employed in the clinical domain for several downstream NLP tasks including clinical NER and de-identification. Bio-Clinical BERT (Alsentzer et al., 2019), initialized from BioBERT (Lee et al., 2020), was fine-tuned on clinical text which makes it a contextually relevant model to explore for the task of identifying SBDH status in EHR notes.

4.2. Class Imbalance

The MIMIC-SBDH dataset has class imbalance for almost all SBDH categories particularly for Education, Environment and all behavioral factors. To avoid bias against the dominating classes, we deployed a simple upsampling strategy where we over-sampled the smaller classes so that all classes had the same number of data points in the training set. This was done independently for each SBDH.

4.3. Behavioral Testing

Previous work on predicting an SBDH's status in EHRs evaluate models in terms of metrics such as F1, precision, and recall. However, these statistics alone are not enough to understand whether the models possess certain desirable capabilities or to understand where the models are failing. Since a standard metric-based evaluation is not enough, we propose additionally evaluating a model trained for the task of predicting an SBDH's status, based on the following capabilities:

1. *Negation*: The model should be able to understand negation of the SBDH in order to currently predict its status. For the section, "Patient lost his wife recently.", predicting *True* for SBDH Community-Present is an example of error due to negation.

2. *Attribution*: Passages about social support such as family members and their habits are very common in the social history section. A model's prediction should be invariant to the presence of these passages (unless the SBDH pertains to social support like Community). A model predicting *Never* for the section, "Patient is negative for ETOH", for Alcohol Use but flipping to *Present* when the text "his son is an alcoholic" is appended to the section is an example of error due to *Attribution*.

3. *Historical Phrases*: Passages about past habits, particularly in the context of substance consumption, are a common occurrence in the social history section. The model should make predictions based on the patient's present habits instead of the past. For the section "History of alcohol abuse but no alcohol in three years.", predicting *Present* for Alcohol Use is an example of error due to presence of historical phrases.

4. *Robustness to Misspellings*: Misspellings are often present in EHR notes and a model's prediction should be invariant to these. A model flipping its prediction from *False* to

True when "Pt is homeless" is changed to "Pt is hoemless" is an example of error due to misspellings.

We use CheckList (Ribeiro et al., 2020), a tool for comprehensive behavioral testing of NLP models, to understand the above capabilities. The CheckList tool enables evaluating a capability using different test types and provides an efficient way of generating a variety of test cases. We performed Minimal Functionality Test (**MFT**) to evaluate whether a model possessed the *Negation* and *Historical Phrases* capabilities. We performed Invariance test (**INV**) to test whether a model's prediction changed in the presence of passages not pertaining to the patient (*Attribution*) or in the presence of misspellings (*Robustness to Misspellings*).

We generated 200 test cases for each capability per SBDH (we tested *Historical Phrases* for Alcohol Use, Tobacco Use and Drug Use only since this capability seemed more relevant to these factors. We did not test Education for *Negation* since we did not come across passages related to the lack of an education). The test cases were generated based on the various language styles we encountered during the annotation process. For example, "positive for tobacco", "pos for tobacco" or "he smokes 10 cigs per day" are some ways of expressing that a patient smokes cigarettes. A description of how the test cases were created is present in Appendix A.2.

5. Experiments

For each discharge summary, the input to the models was the extracted social history section. We used `scikit-learn` (Pedregosa et al., 2011) to train RF and XGBoost classifiers. Features were generated using a bag-of-words approach. The texts were lower-cased and tokenized. Non-alphanumeric characters from tokens and general stop words were removed. Term frequency-inverse document frequency (TF-IDF) features were then generated and fed to the models. We used the `Transformers` library (Wolf et al., 2020), particularly the `BertForSequenceClassification` class to fine-tune Bio-Clinical BERT for sequence classification. The texts were lower-cased and a maximum sequence length of 256 tokens was set.

We performed hyper-parameter tuning for RF and XGBoost using 5-fold cross validation. We tuned the number of learners, number of features, and minimum samples required to split for RF, and number of learners, maximum depth per learner, learning rate, and L2 regularization for XGBoost. We trained Bio-ClinicalBERT for 50 epochs using `Adam` optimizer with a linear warm-up and learning rate decay scheduler, the peak learning rate being $5e-5$. We performed 5-fold cross-validation and reported average macro-F1 along with the standard deviation across the folds.

6. Results

We first evaluated the models based on macro-F1, as shown in Table 4, and checked for statistical significance using Student's t-test with p-value 0.05. Bio-ClinicalBERT outperformed RF and XGBoost for Community-Present and Community-Absent. RF and XGBoost gave the best result for Environment. XGBoost outperformed RF and Bio-ClinicalBERT for Alcohol Use and Drug Use. We did not see a statistically significant difference in performance among the three models for Education, Economics and Tobacco Use.

Table 5 shows results of behavioral testing. All models have a high failure rate in understanding *Negation* for Community-Present (RF:40.5%, XGBoost: 27.2%, Bio-ClinicalBERT: 44.5%) but there is an improvement in the same for Community-Absent (RF:11.0%, XGBoost: 14.5%, Bio-ClinicalBERT: 13.3%). RF and XGBoost perform poorly in understanding *Attribution* in all SBDHs. Bio-ClinicalBERT too has a high failure rate for SBDHs Education (70.9%), Alcohol Use (43.0%) and Tobacco Use (17.5%) for *Attribution.* In terms of substance use, all models have a high failure rate in understanding *Historical Phrases.* All models show sensitivity to *Misspellings* across all SBDHs but with Community being relatively more robust.

SBDH	**RF**	**XGBoost**	**Bio-ClinicalBERT**
Community-Present	0.9325 ± 0.0085	0.9314 ± 0.0034	**0.9578 ± 0.0027***
Community-Absent	0.8686 ± 0.0199	0.8631 ± 0.0180	**0.9068 ± 0.0206***
Education	0.8249 ± 0.0323	0.8242 ± 0.0231	**0.8386 ± 0.0424**
Economics	0.8886 ± 0.0143	0.8879 ± 0.0154	**0.8964 ± 0.0073**
Environment	**0.9069 ± 0.0284***	0.8881 ± 0.0340	0.7969 ± 0.0703
Alcohol Use	0.7071 ± 0.0173	**0.7348 ± 0.0148***	0.7049 ± 0.022
Tobacco Use	0.7368 ± 0.0176	0.7399 ± 0.0195	**0.7450 ± 0.0222**
Drug Use	0.6697 ± 0.0075	**0.6942 ± 0.0195***	0.5812 ± 0.0501

*p-value< 0.05

Table 4: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for all SBDHs.

7. Discussion

7.1. Impact of Class Imbalance and Oversampling

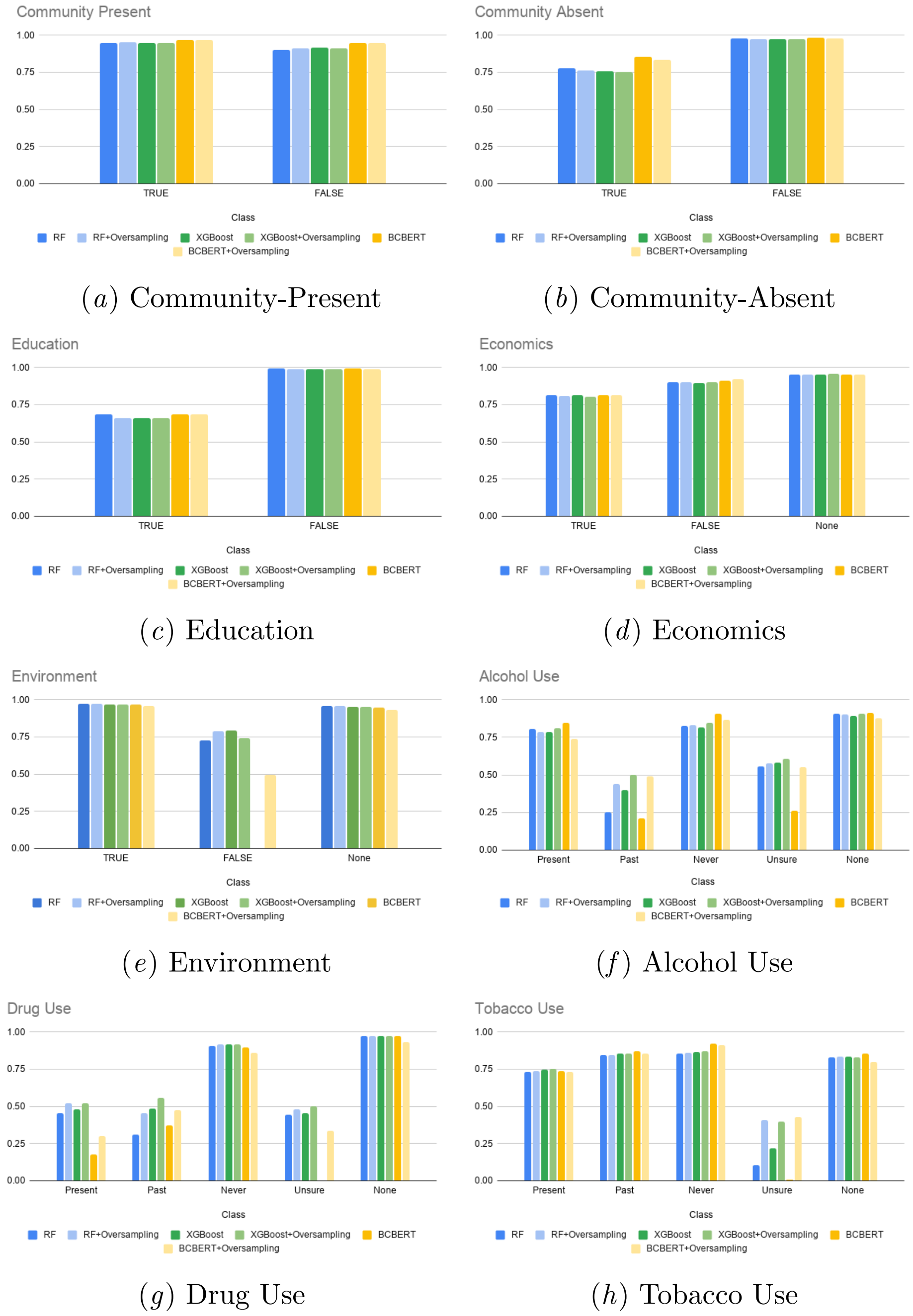
Bio-ClinicalBERT was sensitive to extreme class imbalance compared to RF and XGBoost. We inferred this from its relatively poor performance on SBDH Environment, which has a *True:False:None* class ratio of $\sim 83 : 1 : 53$. In addition, the model performed worse than RF and XGBoost for SBDHs Alcohol Use and Drug Use, both of which have more than one rare class (*Past* and *Unsure* for Alcohol Use, and *Present, Past and Unsure* for Drug Use).

Table 6 contains macro-F1 scores of the models when oversampling was not performed. Figure 3 and Appendix A.3 show F1 scores per class for all SBDHs with and without oversampling. We checked for a statistically significant improvement in macro-F1 when oversampling is performed using Student's t-test with p-value of 0.05. Alcohol Use, Tobacco Use and Drug Use showed an improvement across all models. Performance improvement in the rare classes varied across the models. In case of Alcohol Use, all models performed better on class *Past* when oversampling was performed. Bio-ClinicalBERT improved in terms of class *Unsure* as well. In case of Tobacco Use, while RF and XGBoost saw an improvement in performance on class *Past*, Bio-ClinicalBERT improved in terms of class *Unsure*. For Drug Use, all models performed better on class *Unsure* when oversampling was performed.

Out of RF, XGBoost and Bio-ClinicalBERT, Bio-ClinicalBERT benefited the most from oversampling, with improvements in Environment, Alcohol Use, Tobacco Use and Drug

SBDH	**Test**	**Failure Rate**			**Example Test Cases, *Label***
		RF	XGB	BCB	
Community (Present)	N	40.5	27.2	44.5	Husband died recently. *False*
	M	1.9	3.8	1.9	Patient ilves with her grandson. *True*
Community (Absent)	N	11.0	14.5	13.3	Patient lost her child in an accident. *True*
	M	2.5	1.7	2.5	Patient is separtaed from his wife. *True*
Education	A	83.7	73.5	70.9	Daughter is a medical student. *False*
	M	6.5	7.7	13.1	Son is a honors high school studnet. *False*
Economics	N	11.6	15.2	8.0	Former ICU nurse. *False*
	A	76.5	71.5	9.0	He is not working. Daughter-in-law works as chemist. *False*
	M	11.2	11.2	12.2	He is a retierd dentist. *False*
Environment	N	24.7	15.6	1.3	Intermittently lives with friends but currently homeless. *False*
	A	31.2	11.7	1.3	Homeless last 3 months. His family lives nearby. *False*
	M	9.1	10.4	7.8	Homeelss, searching for apt. *False*
Alcohol Use	N	28.0	11.1	27.1	Negative for alcohol use. *Never*
	A	48.0	53.5	43.0	Negative for ETOH. Son has history of alcohol abuse. *Never*
	HP	68.3	68.3	77.8	Was sober for a few months, but started drinking again. *Present*
	M	20.0	22.0	18.0	He dneies any history of alcohol intake. *Never*
Tobacco Use	N	10.0	11.7	0.6	Cigs: none ETOH: none. *Never*
	A	27.5	39.0	17.5	Denied smoking, EtOH or drugs. Wife is a smoker. *Never*
	HP	73.9	49.7	46.4	Denies recent tobacco (history of abuse but none in years) *Past*
	M	8.5	19.0	12.4	Unclear when he started msoking again. *Never*
Drug Use	N	0.6	0.6	4.3	Negative for smoking and drug. *Never*
	A	87.0	88.0	6.0	Denied smoking, EtOH or drugs. Wife uses marijuana. *Never*
	HP	58.8	68.6	98.7	Had stopped opiates. Started again 1 month ago. *Present*
	M	9.2	17.0	3.3	Hx of IVDU abuse ubt quit. *Past*

Table 5: Results of behavioral testing using CheckList. The example test cases are those in which at least one model failed. Model abbreviations - RF: Random Forest, XGB: XGBoost, BCB: Bio-ClinicalBERT. Test abbreviations - N: Negation, A: Attribution, HP: Historical Phrases, M: (Robustness to) Misspellings



(a) Community-Present

(b) Community-Absent

(c) Education

(d) Economics

(e) Environment

(f) Alcohol Use

(g) Drug Use

(h) Tobacco Use

Figure 3: F1 scores of RF, XGBoost and Bio-ClinicalBERT for all SBDHs with and without oversampling at class level

SBDH	RF	XGBoost	Bio-ClinicalBERT
Community-Present	0.9245 ± 0.0073	0.9329 ± 0.0067	0.9568 ± 0.0039
Community-Absent	0.8775 ± 0.022	0.8657 ± 0.0173	0.9186 ± 0.019
Education	0.8399 ± 0.0322	0.8266 ± 0.0211	0.8201 ± 0.0214
Economics	0.8904 ± 0.0191	0.8876 ± 0.0171	0.8941 ± 0.0157
Environment	0.8861 ± 0.0384	0.8881 ± 0.034	0.6386 ± 0.0034
Alcohol Use	0.6694 ± 0.022	0.6954 ± 0.0187	0.6275 ± 0.0345
Tobacco Use	0.6726 ± 0.0198	0.6438 ± 0.015	0.6787 ± 0.0232
Drug Use	0.6187 ± 0.0302	0.6606 ± 0.0094	0.4828 ± 0.0407

Table 6: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT without oversampling for all SBDHs.

Use. Oversampling helped improve Bio-ClinicalBERT's performance significantly on rare classes. For instance, the model's performance on class *Unsure* improved from 0.0 ± 0.0 to 0.3355 ± 0.0714 and 0.4272 ± 0.0883 for Drug Use and Tobacco Use respectively. Similarly, the model's performance on class *False* improved from 0.0 ± 0.0 to 0.4960 ± 0.2014 for SBDH Environment.

7.2. Behavioral Testing

Behavioral testing helped evaluate specific capabilities and understand the shortcomings of the models. While Bio-ClinicalBERT evaluates well in terms macro-F1(0.9578 ± 0.0027) for Community-Present, it lacks the capability of understanding *Negation* well (failure rate of 44.5%). Bio-ClinicalBERT possesses the *Negation* capability in better measure for Community-Absent (failure rate of 13.3%). We looked at the top important features of the RF and XGBoost and observed that terms such as 'divorced', 'widowed', 'passed', 'died' were among the most important features for Community-Absent. This explains why the models did well in terms of *Negation* since these terms were primarily used in the discharge summaries to express the absence of community. We did not see these terms in the top important features in case of Community-Present. The important features consisted primarily of relationships such as 'daughter', 'son' and 'husband' which explained why the models had a higher failure rate for Negation in Community-Present. Our future work intends to look at this in greater detail.

RF and XGBoost perform poorly in understanding *Attribution* in all SBDHs. This can be reasoned by the bag-of-words based features used to train the models. The models are more attuned to detecting the presence of SBDH-related keywords but not understanding who the keywords pertain to. Bio-ClinicalBERT has a particularly high *Attribution* failure rate for SBDH Education (70.9%). This can be attributed to general lack of education-related passages in the dataset and hence a lack of diverse set of related contexts that would lead to better learning.

Understanding *Historical Phrases* is crucial in correctly classifying substance use status. We observed that all models performed poorly with respect to this capability. This could be because of insufficient data with historical passages as is seen from the class distribution. The failure rate is lower for Tobacco Use compared to Alcohol Use and Drug Use. Tobacco

Use has relatively more samples for class *Past* and that could help in acquiring a better understanding of related phrases.

Limitations Our dataset was created using data from a single institution which may limit a model's ability to generalize and perform well on other datasets when trained solely on ours. Given the size of the dataset, the same k-fold validation was used to perform both hyperparameter tuning and model selection for RF and XGBoost classifiers which may lead to optimistic results. We could not perform hyperparameter tuning for Bio-ClinicalBERT models for the same reason.

8. Conclusion

In this work, we created a novel dataset MIMIC-SBDH, containing $7,025$ discharge summaries annotated for the status of seven SBDHs. We studied three baselines: Random Forest, XGBoost and Bio-ClinicalBERT for the task of predicting an SBDH's status from an EHR. We highlighted a set of capabilities a model must possess to perform well on the task and performed behavioral testing using CheckList for these capabilities. While behavioral testing helps understand where the models are failing, the question regarding why the models are failing is a work in itself and needs more investigation. Future research will involve first identifying the reasons behind the lack of certain capabilities and then proposing approaches to address these shortcomings.

9. Acknowledgement

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award number R01MH125027. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Hadeel Alzoubi, Naeem Ramzan, Raid Alzubi, and Ehsan Mesbahi. An automated system for identifying alcohol use status from clinical text. In *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pages 41–46. IEEE, 2018.

Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Patient status classification by using rule based sentence extraction and bm25 knn-based classifier. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

Samantha Artiga and Elizabeth Hinton. Beyond health care: the role of social determinants in promoting health and health equity. *Health*, 20(10):1–13, 2019.

F Carrero, JG Hidalgo, E Puertas, M Maña, and J Mata. Quick prototyping of high performance text classifiers. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

David Dorr, Cosmin A Bejan, Christie Pizzimenti, Sumeet Singh, Matt Storer, and Ana Quinones. Identifying patients with significant problems related to social determinants of health with natural language processing. *Studies in health technology and informatics*, 264:1456–1457, 2019.

Daniel J Feller, Oliver J Bear Don't Walk, Jason Zucker, Michael T Yin, Peter Gordon, Noémie Elhadad, et al. Detecting social and behavioral determinants of health with structured and free-text clinical data. *Applied clinical informatics*, 11(01):172–181, 2020.

Ravi Garg, Elissa Oh, Andrew Naidech, Konrad Kording, and Shyam Prabhakaran. Automating ischemic stroke subtype classification using machine learning and natural language processing. *Journal of Stroke and Cerebrovascular Diseases*, 28(7):2045–2051, 2019.

Adi V Gundlapalli, Marjorie E Carter, Miland Palmer, Thomas Ginter, Andrew Redd, Steven Pickard, Shuying Shen, Brett South, Guy Divita, Scott Duvall, et al. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among us veterans. In *AMIA Annual Symposium Proceedings*, volume 2013, page 537. American Medical Informatics Association, 2013.

Jennifer S Haas, Jeffrey A Linder, Elyse R Park, Irina Gonzalez, Nancy A Rigotti, Elissa V Klinger, Emily Z Kontos, Alan M Zaslavsky, Phyllis Brawarsky, Lucas X Marinacci, et al. Proactive tobacco cessation outreach to smokers of low socioeconomic status: a randomized clinical trial. *JAMA internal medicine*, 175(2):218–226, 2015.

Alison B Hamilton, Ines Poza, Vivian Hines, and Donna L Washington. Barriers to psychosocial services among homeless women veterans. *Journal of Social Work Practice in the Addictions*, 12(1):52–68, 2012.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, 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(1):1–9, 2016.

Jitendra Jonnagaddala, Hong-Jie Dai, Pradeep Ray, and Siaw-Teng Liaw. A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. In *Proceedings of BioNLP 15*, pages 147–151, 2015.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Kevin Lybarger, Meliha Yetisgen, and Mari Ostendorf. Using neural multi-task learning to extract substance abuse information from clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1395. American Medical Informatics Association, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Ank E Nijhawan, Lisa R Metsch, Song Zhang, Daniel J Feaster, Lauren Gooden, Mamta K Jain, Robrina Walker, Shannon Huffaker, Michael J Mugavero, Petra Jacobs, et al. Clinical and sociobehavioral prediction model of 30-day hospital readmissions among people with hiv and substance use disorder: beyond electronic health record data. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 80(3):330–341, 2019.

Ted Pedersen. Determining smoker status using supervised and unsupervised learning with lexical features. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*. Citeseer, 2006.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

Roger Alan Stein, Patricia A Jaques, and Joao Francisco Valiati. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471:216–232, 2019.

Paul Y Takahashi, Euijung Ryu, Janet E Olson, Erin M Winkler, Matthew A Hathcock, Ruchi Gupta, Jeff A Sloan, Jyotishman Pathak, Suzette J Bielinski, and James R Cerhan. Health behaviors and quality of life predictors for risk of hospitalization in an electronic health record-linked biobank. *International journal of general medicine*, 8:247, 2015.

Maxim Topaz, Ludmila Murga, Ofrit Bar-Bachar, Kenrick Cato, and Sarah Collins. Extracting alcohol and substance abuse status from clinical notes: The added value of nursing data. *Studies in health technology and informatics*, 264:1056–1060, 2019.

Clayton A Turner, Alexander D Jacobs, Cassios K Marques, James C Oates, Diane L Kamen, Paul E Anderson, and Jihad S Obeid. Word2vec inversion and traditional text classifiers for phenotyping lupus. *BMC medical informatics and decision making*, 17(1): 1–11, 2017.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.

WHO. *Closing the gap in a generation: health equity through action on the social determinants of health: final report of the commission on social determinants of health*. World Health Organization, 2008.

Richard Wicentowski and Matthew R Sydes. Using implicit information to identify smoking status in smoke-blind medical discharge summaries. *Journal of the American Medical Informatics Association*, 15(1):29–31, 2008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yongliang Wu, Shuliang Zhao, and Wenbin Li. Phrase2vec: Phrase embedding based on parsing. *Information Sciences*, 517:100–127, 2020. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.12.031. URL https://www.sciencedirect.com/science/article/pii/S0020025519311429.

Bokai Yang, Guangzhe Dai, Yujie Yang, Darong Tang, Qi Li, Denan Lin, Jing Zheng, and Yunpeng Cai. Automatic text classification for label imputation of medical diagnosis notes based on random forest. In *International Conference on Health Information Science*, pages 87–97. Springer, 2018.

Meliha Yetisgen and Lucy Vanderwende. Automatic identification of substance abuse from social history in clinical text. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 171–181. Springer, 2017.

Le Zheng, Oliver Wang, Shiying Hao, Chengyin Ye, Modi Liu, Minjie Xia, Alex N Sabo, Liliana Markovic, Frank Stearns, Laura Kanov, et al. Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. *Translational psychiatry*, 10(1):1–10, 2020.

Appendix A.

A.1. Annotation

SBDH	**Class**	**Discussion and Examples**
Community-Present	*True*	Presence of social support. Examples: “...Lives with his **wife**.”
	False	No passages about presence of social support. Examples: “No tobacco, ethanol or drug use.”
Community-Absent	*True*	Lack/loss of social support. Examples: “**Widowed** with three children...”.
	False	No passages about loss/lack of social support. Examples: “Former smoker, no EtOH.”
Education	*True*	Passages about patient’s education. Examples: “...,finished **law school**...”

	False	No passages about patient’s education. Examples: “Former smoker, no EtOH. Lives with his wife.”
Economics	*True*	Patient is employed. Examples: “...,a **technician** at...”
	False	Not currenttly employed (including retirement). Examples: “**Retired** teacher...” “Formerly worked in...not working currently...”
	None	No passages about patient’s employment. Examples: “Former smoker, no EtOH. Lives with his wife.
Environment	*True*	Indication of housing. Examples: “...,**lives** at **home** with mother...”
	False	Lack of housing. Examples: “**Homeless**...”
	None	No passages about patient’s housing. Examples: “Former smoker, no EtOH”.
Alcohol Use	*Present*	Patient is a current consumer of alcohol. Examples: “...,rare **ETOH** use...” “glass of **wine** everyday.”
	Past	Past consumer of alcohol. Examples: “...,no **alcohol** use since...” ”quit **alcohol** [de-ID] weeks ago...”
	Never	Has never consumed **alcohol**. Examples: “...,**Alcohol**: none...” “...denies **alcohol**..”
	Unsure	Ambiguous passages about patient’s consumption. Examples: “No **ETOH** abuse.” “...questionable history of **ETOH** use...”
	None	No passages about patient’s alcohol consumption. Examples: “Retired. Lives with his wife.”
Tobacco Use	*Present*	Patient is a current consumer of tobacco. Examples: “...,she is still **smoking**...” “**smokes** a **pack** per day...”
	Past	Past consumer of **tobacco**. Examples: “...,quit **smoking** 20 years ago...” “former **smoker**...”
	Never	Has never consumed tobacco.

		Examples: “...,no **tobacco** use...” “...denies **tobacco**..”
	Unsure	Ambiguous passages about patient’s consumption. Examples: “...[de-ID] **tobacco**...” “...denied **tobacco** abuse...”
	None	No passages about patient’s tobacco consumption. Examples: “Retired. Lives with his wife.”
Drug Use	*Present*	Patient is a current consumer of a drug. Examples: “...,positive for **cocaine**...” “**intravenous drug** abuse...”
	Past	Past consumer and does not consume any drug anymore. Examples: “...,minimal **marijuana** years ago...” “remote **cocaine** use in the past...”
	Never	Has never consumed a drug. Examples: “...,no **illicit drug** use...” “...denies **recreation drug** use..”
	Unsure	Ambiguous passages about patient’s drug consumption. Examples: “...no **cocaine** abuse...” “...no **IV drug** abuse...”
	None	No passages about patient’s drug consumption. Examples: “Retired. Lives with his wife.”

Table 7: Examples of each class per SBDH from the dataset

1. For Environment, we annotated discharge summaries with passages such as “lives with wife” as *True* and marked the token **lives** as a related keyword. Discharge summaries stating that the patient lived in locations such as long-term care facilities, rehab were also annotated as *True*.

2. For Alcohol Use, Tobacco Use and Drug Use, a discharge summary was annotated as *Unsure* if it had an ambiguous passage related to the substance. For instance, “Pt has a history of alcohol abuse” was annotated as *Unsure* since the discharge summary did not mention what the current status was. There is a possibility that the patient does not abuse alcohol anymore but consumes it in a controlled manner. If the de-identification hampered inferring the consumption status, we annotated such discharge summaries as *Unsure* as well. Example: “ [de-identified token]tobacco”.

A.2. Test Cases for Behavioral Testing

We generated our test cases using the language styles we encountered during the annotation process. We made use of *templates* to generalize and diversify test cases. We created sets of text to embed within the templates. Some sets that were used across the SBDHs were:

1. $OTHER$: This contained a set of sentences that did not pertain to the SBDH being tested. For example, when testing Alcohol Use, $OTHER$ comprised of sentences such as "He is a retired postal worker", "Patient is widowed" and "Married with four kids". The set $OTHER$ varies with each SBDH. Using this set ensured that the test cases did not just always contain passages relevant to the SBDH being tested. Since there is a possibility that a discharge summary may actually just be discussing the SBDH being tested, we kept an empty string in the set as well.

2. $RELATION$: This set contained relationships such as son, daughter, grandchild etc.

3. $SPOUSE$: This set contained spouse-related terms such as husband, wife, partner and spouse.

4. $NUMBER$: This set contained positive integers.

5. $TIME$: This set contained terms indicating duration such as weeks, months and years.

1. Community-Present

   (a) Negation: Test cases were generated using templates such as "Patient's $\{RELATION\}$ died recently. $\{OTHER\}$.", "Patient is separated from his $\{SPOUSE\}$. $\{OTHER\}$". We also added non-*template* based test cases such as "Patient is widowed", "Pt is divorced".

   (b) Misspelling: We used CheckList's *Perturb.add_typo* functionality to introduce misspellings in sentences.

2. Community-Absent: We used the same approach as that used for Community-Present. In case of *Negation*, the model was expected to predict True instead of False.

3. Education: We created education-related sets such as $EDUCATION = \{$high school, medical,...$\}$, $FIELD = \{$art, medical,...$\}$.

   (a) Attribution: Test cases were created using templates such as "Patient's $\{RELATION\}$ is a $\{EDUCATION\}$ student", "$\{RELATION\}$ attends $\{FIELD\}$ school".

   (b) *Perturb.add_typo* functionality to introduce misspellings in sentences.

4. Economics: We created a set $PROFESSION$ that contained professions, set $PAST$ that contained terms indicating past employment and set $VERB$ containing employment-related verb.

   (a) Negation: Test cases were created using templates that indicated absence of current employment such as "$\{PAST\}$ $\{PROFESSION\}$. $\{OTHER\}$", "Patient is not currently $\{VERB\}$", where $PROFESSION = \{$doctor,high school teacher...$\}$, $PAST = \{$former, retired,...$\}$, $VERB = \{$working,employed...$\}$.

   (b) Attribution: Test cases that indicated the patient was employed and the relationship wasn't and vice versa were generated using templates such as “$\{PAST\}$ $\{PROFESSION\}$. $\{RELATION\}$ works as $\{PROFESSION\}$”, “$\{RELATION\}$ is not currently $\{VERB\}$. Patient is a $\{PROFESSION\}$.

   (c) *Perturb.add_typo* functionality to introduce misspellings in sentences.

5. Environment

   (a) Negation: Test cases that indicated patient's lack of housing were created using templates such as “Patient has been homeless for the last $\{NUMBER\}$ months. $\{OTHER\}$.”, “Patient intermittently lives with $\{RELATION\}$ but is homeless. $OTHER$.”

   (b) Attribution: Test cases indicating that patient lacked housing and the relationship didn't were created using templates such as “Patient has been homeless for the last $\{NUMBER\}$ months. His $\{RELATION\}$ lives nearby. $\{OTHER\}$.”, “Patient intermittently lives with $\{RELATION\}$ but is homeless. $\{OTHER\}$. His $\{RELATION\}$ lives in an apartment close by.”

   (c) *Perturb.add_typo* functionality to introduce misspellings in sentences.

6. Alcohol Use, Tobacco Use, Drug Use: We created a set $SUBSTANCE$ containing relevant keywords that we came across during annotation for each substance such {alcohol, ETOH, wine, brandy,...} for alcohol, {Tobacco, cigarettes, cigars, cigs, ppd...} for tobacco, {drug, marijuana, cocaine, intraveneous drug, IVDU, Special K,...} for drugs as well as verbs $VERB$ such as {drink, consume,...} for alcohol, {smoke, consume} for tobacco and {consume} for drug.

   (a) Negation: Test cases were generated using templates such as “Patient does not $\{VERB\}$ $\{KEYWORD\}$.”, “Negative for $\{KEYWORD\}$”, “There is no history of $\{VERB\}$ intake”.

   (b) Attribution: Test cases in which the patient's substance status differed from the relationship's status were created using templates such as “Patient $\{VERB\}$ $\{KEYWORD\}$. $\{RELATION\}$ does not $\{VERB\}$ $\{KEYWORD\}$. $\{OTHER\}$”, “Patient has a history of $\{KEYWORD\}$ abuse but quit. $\{RELATION\}$ does not $\{VERB\}$ $\{KEYWORD\}$. $\{OTHER\}$.”

   (c) Historical Phrases: Test cases in which the patient's historical substance status was discussed along with his present consumption status were created using templates such as “Patient consumed $\{KEYWORD\}$ in the past but quit $\{NUMBER\}$ $\{TIME\}$ ago.”, “Patient quit $\{VERB\}$ $\{KEYWORD\}$ but started again $\{NUMBER\}$ $\{TIME\}$ ago.”.

   (d) *Perturb.add_typo* functionality to introduce misspellings in sentences.

A.3. Results

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
True	0.9513 ± 0.0058	0.9489 ± 0.0016	0.9688 ± 0.0011	0.9464 ± 0.0052	0.9504 ± 0.004	0.9681 ± 0.0019
False	0.9137 ± 0.0118	0.9138 ± 0.0066	0.9468 ± 0.0046	0.9026 ± 0.0104	0.9153 ± 0.01	0.9456 ± 0.0061

Table 8: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Community-Present.

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
True	0.7634 ± 0.0357	0.754 ± 0.033	0.8343 ± 0.0374	0.7779 ± 0.0406	0.7569 ± 0.0327	0.8536 ± 0.0337
False	0.9738 ± 0.0045	0.9722 ± 0.0036	0.9792 ± 0.0041	0.977 ± 0.004	0.9746 ± 0.0026	0.9834 ± 0.0045

Table 9: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Community-Absent.

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
True	0.6601 ± 0.0657	0.658 ± 0.0476	0.6861 ± 0.0841	0.6878 ± 0.0657	0.662 ± 0.0406	0.6486 ± 0.0434
False	0.9898 ± 0.0019	0.9905 ± 0.0021	0.991 ± 0.0025	0.9919 ± 0.0019	0.9913 ± 0.0034	0.9915 ± 0.0028

Table 10: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Education.

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
True	0.8081 ± 0.0245	0.8064 ± 0.0286	0.8166 ± 0.0124	0.8124 ± 0.0343	0.8117 ± 0.0312	0.8134 ± 0.0307
False	0.9036 ± 0.0155	0.9006 ± 0.0181	0.9208 ± 0.0142	0.9032 ± 0.0192	0.8971 ± 0.0177	0.9134 ± 0.0171
None	0.9541 ± 0.0068	0.9568 ± 0.0045	0.9519 ± 0.0066	0.9556 ± 0.0069	0.9541 ± 0.0055	0.9555 ± 0.0044

Table 11: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Economics.

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
True	0.9734 ± 0.0026	0.969 ± 0.0036	0.9597 ± 0.0037	0.9732 ± 0.0034	0.9708 ± 0.0027	0.969 ± 0.0031
False	0.7883 ± 0.0831	0.7421 ± 0.0943	0.496 ± 0.2014	0.7258 ± 0.1112	0.7948 ± 0.0774	0.0 ± 0.0
None	0.959 ± 0.0042	0.9531 ± 0.005	0.9351 ± 0.0091	0.9592 ± 0.0042	0.955 ± 0.004	0.9468 ± 0.0078

Table 12: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Environment.

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
Present	0.7868 ± 0.0087	0.8114 ± 0.0105	0.7371 ± 0.023	0.8065 ± 0.0101	0.7879 ± 0.0098	0.8486 ± 0.0143
Past	0.441 ± 0.0239	0.5015 ± 0.035	0.4882 ± 0.0544	0.2489 ± 0.0278	0.3992 ± 0.0703	0.2092 ± 0.1255
Never	0.8313 ± 0.0197	0.848 ± 0.0132	0.8682 ± 0.0283	0.8287 ± 0.0159	0.8186 ± 0.0161	0.9094 ± 0.0088
Unsure	0.5745 ± 0.0841	0.6068 ± 0.0675	0.5525 ± 0.0558	0.5572 ± 0.0844	0.5802 ± 0.0651	0.258 ± 0.1765
None	0.9017 ± 0.0231	0.9062 ± 0.0252	0.8785 ± 0.0345	0.9059 ± 0.0243	0.891 ± 0.0179	0.9124 ± 0.0337

Table 13: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Alcohol Use.

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
Present	0.7379 ± 0.0339	0.7521 ± 0.0266	0.7317 ± 0.0318	0.7318 ± 0.0196	0.7487 ± 0.0296	0.7381 ± 0.0717
Past	0.8436 ± 0.0187	0.8548 ± 0.0121	0.8554 ± 0.0052	0.8454 ± 0.0182	0.8564 ± 0.0089	0.8729 ± 0.0265
Never	0.8616 ± 0.0163	0.8698 ± 0.0137	0.9122 ± 0.0114	0.8532 ± 0.0166	0.8655 ± 0.0152	0.9203 ± 0.0055
Unsure	0.4057 ± 0.1147	0.3952 ± 0.1017	0.4272 ± 0.0883	0.1014 ± 0.0657	0.2181 ± 0.0521	0.008 ± 0.016
None	0.835 ± 0.0322	0.8278 ± 0.0366	0.7985 ± 0.04	0.8311 ± 0.0234	0.8339 ± 0.025	0.8541 ± 0.0212

Table 14: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Tobacco Use.

	With Oversampling			**Without Oversampling**		
	RF	XGBoost	Bio-ClinicalBERT	RF	XGBoost	Bio-ClinicalBERT
Present	0.5221 ± 0.0428	0.5194 ± 0.0796	0.2999 ± 0.102	0.4538 ± 0.0823	0.4776 ± 0.0688	0.1735 ± 0.1204
Past	0.4547 ± 0.0151	0.5582 ± 0.0462	0.4749 ± 0.0636	0.3119 ± 0.0711	0.4821 ± 0.0312	0.3698 ± 0.1631
Never	0.9194 ± 0.013	0.9188 ± 0.0126	0.8606 ± 0.0625	0.908 ± 0.0138	0.9172 ± 0.0094	0.8989 ± 0.0186
Unsure	0.4793 ± 0.0301	0.5011 ± 0.0706	0.3355 ± 0.0714	0.4437 ± 0.0732	0.4535 ± 0.0496	0.0 ± 0.0
None	0.9732 ± 0.0049	0.9736 ± 0.0061	0.935 ± 0.0077	0.9761 ± 0.0055	0.9723 ± 0.007	0.9719 ± 0.004

Table 15: Macro-F1 scores of RF, XGBoost and Bio-ClinicalBERT for Drug Use.