Annotating Social Determinants of Health Using Active Learning, and Characterizing Determinants Using Neural Event Extraction

Kevin Lybarger[a,*], **Mari Ostendorf**[b], **Meliha Yetisgen**[a]

[a]Biomedical & Health Informatics, University of Washington, Box 358047 Seattle, WA 98109, USA

[b]Department of Electrical & Computer Engineering, University of Washington, Campus Box 352500 185, Seattle, WA 98195-2500, USA

Abstract

Social determinants of health (SDOH) affect health outcomes, and knowledge of SDOH can inform clinical decision-making. Automatically extracting SDOH information from clinical text requires data-driven information extraction models trained on annotated corpora that are heterogeneous and frequently include critical SDOH. This work presents a new corpus with SDOH annotations, a novel active learning framework, and the first extraction results on the new corpus. The Social History Annotation Corpus (SHAC) includes 4,480 social history sections with detailed annotation for 12 SDOH characterizing the status, extent, and temporal information of 18K distinct events. We introduce a novel active learning framework that selects samples for annotation using a surrogate text classification task as a proxy for a more complex event extraction task. The active learning framework successfully increases the frequency of health risk factors and improves automatic extraction of these events over undirected annotation. An event extraction model trained on SHAC achieves high extraction performance for substance use status (0.82-0.93 F1), employment status (0.81-0.86 F1), and living status type (0.81-0.93 F1) on data from three institutions.

Graphical Abstract

---

[*]Corresponding author lybarger@uw.edu (Kevin Lybarger).

1. Introduction

US life expectancy is decreasing [1], even as medical care advances. Decreasing life expectancy may be partly attributable to deteriorating social determinants of health (SDOH) [2, 3]. For example, substance abuse (including alcohol, drug, and tobacco use) is increasingly recognized as a key factor for morbidity and mortality [4–6]. More Americans are living alone, leading to increased social isolation and negative health outcomes [7]. Employment and occupation impact income, societal status, hazards encountered, and health [8]. Understanding SDOH, including behaviors influenced by these social factors, can inform clinical decision-making [9].

SDOH are characterized in the Electronic Health Record through structured data and unstructured clinical text; however, clinical text captures detailed descriptions of these determinants, beyond the representation in structured data. This text-encoded information must be automatically extracted for secondary use applications, like large-scale retrospective studies and clinical decision support systems. The automatically extracted data can augment the available structured data to create a more comprehensive patient representation in these downstream applications [10, 11].

Leveraging the social history information in clinical text requires high-quality annotated data to create machine learning-based information extraction models. This work presents a new annotated clinical corpus, referred to as Social History Annotation Corpus (SHAC). *SHAC* is comprised of 4,480 social history sections with detailed annotations for 12 critical SDOH. SHAC utilizes clinical notes from MIMIC-III [12] and an existing data set from the University of Washington (UW) and Harborview Medical Centers. It includes event-based annotations for more than 55K annotated spans and 18K distinct events across four note types.

Hand annotation of detailed SDOH information in clinical notes is costly, and many critical SDOH are infrequent. To address these budget and data sparsity limitations, the corpus development used active learning to select samples for annotation. Because extracting the event-based SDOH phenomena is a complex sequence labeling task, standard active learning

methods are not practical. This work introduces a novel active learning framework that uses a simplified surrogate task for assessing sample informativeness. Our experiments show that this method increases the diversity and richness of the annotations and improves extraction performance for a variety of event types. The largest performance gains achieved by the active learning framework are associated with infrequent, but extremely important risk factors, like drug use, homelessness, and unemployment.

With the annotated SHAC corpus, we provide a baseline neural event extractor and present the first reported extraction results on SHAC for the most frequently annotated SDOH: substance use, employment, and living status. The event extraction model identifies substance use, employment, and living status events at 0.89-0.98 F1 and characterizes the status of these determinants with 0.81-0.96 F1. The annotation guidelines and source code will be made available online[1].

2.   Related work

2.1.   SDOH Corpora

Multiple corpora with note-level SDOH annotations have been developed. For example, the i2b2 NLP Smoking Challenge introduced a publicly available corpus where tobacco use status is labeled at the note-level [13]. Gehrmann et al. [14] annotated MIMIC-III discharge summaries with note-level phenotype labels, including substance abuse and obesity. Feller et al. [15] annotated 38 different SDOH at the note-level. Annotated corpora with more detailed SDOH annotations describing status, extent, temporal information, and other characteristics also exist. For example, Wang et al. [16] introduced a corpus with detailed substance use annotations for 691 clinical notes, and Yetisgen and Vanderwende [17] created detailed annotations for 13 SDOH in a publicly available corpus of 364 notes. Both Wang et al. [16] and Yetisgen and Vanderwende [17] utilized deidentified notes from the MTSamples website[2] that were created by human transcriptionists.

To achieve high SDOH extraction performance that generalizes across clinicians, institutions, and specialties, annotated corpora must be sufficiently large and diverse. Unfortunately, existing publicly available corpora with SDOH annotations are lacking in either annotation detail, size, and/or heterogeneity. SHAC provides a relatively large corpus with high quality, detailed SDOH annotations. SHAC is heterogeneous in that it includes clinical notes from multiple institutions and note types, and in the use of active selection to encourage a richer representation of SDOH events.

2.2.   Active Learning

In annotation projects, the available unlabeled data is often significantly larger than the annotation budget. Randomly selecting samples for annotation is suboptimal from a model learning perspective, as samples vary in their usefulness, particularly when the phenomena of interest may be infrequent. Active learning identifies samples for annotation that maximize model learning [18, 19]. Samples are selected using a query function that scores

---

[1]https://github.com/uw-bionlp
[2]MTSamples website: http://www.mtsamples.com/

sample informativeness, representativeness, and/or diversity [20–22]. Informativeness describes the potential for a sample to reduce classification uncertainty. The literature varies in the usage of the terms "representativeness" and "diversity." Here, representativeness describes the degree to which a sample describes the structure of the data, and diversity characterizes the variation in the samples selected.

Active learning is well-established for classification tasks, where a single label is predicted for each sample. Multiple studies have applied active learning to text classification tasks, where a sample is a sentence or a document. Sample informativeness is derived from classification uncertainty scores, such as maximizing entropy [23] or minimizing a support vector machine margin [24, 25]. Du et al. [22] assesses diversity based on classifier posterior distributions, and Wu and Ostendorf [23] assesses diversity and representativeness based on sample similarity within the observation space.

Approaches for applying active learning to sequence tagging problems are also well-established [26–31]. Although predictions are made at the token-level, sample selection is typically performed at the sentence or document-level. Representativeness and/or diversity are often assessed by calculating sentence similarity metrics in the observation space [26–28, 30]. Sequence-level uncertainty scores are calculated by various measures, like normalized prediction sequence likelihood and minimum token-level confidence. In the clinical and biomedical domain, uncertainty scores are generated with conditional random field (CRF) models [26–30] or a neural tagger based on contextualized embeddings from ELMo and BERT [31].

Active learning is less explored in relation and event extraction tasks, where triggers (heads), arguments, and/or relations are annotated. The predictions are more complex, involving labeling and linking spans of text. Maldonado et al. [32] apply active learning to a clinical relation extraction task, selecting samples using the average entropy of all predicted phenomena as an uncertainty score. More recently, Maldonado and Harabagiu [33] explores active learning in a medical concept and relation extraction task. In lieu of a heuristic query function, an optimal selection strategy is learned from data with strong and weakly supervised labels, including 1,000 electroencephalogram (EEG) reports with automatic annotations generated by existing extraction models.

SHAC is annotated using an event-based structure, where SDOH are characterized through multiple argument types. These argument types are not equally important for secondary use applications, and the entropy of different determinant-argument combinations may differ significantly. Without sufficient annotated data to learn an optimal selection strategy, we use a simplified text classification task as a surrogate for assessing sample uncertainty, to prevent under sampling the critical phenomena. We hypothesized that the surrogate task would improve extraction performance in the more complex event extraction task and validated the hypothesis with experiments on SHAC data.

3. Materials

3.1. Data

This work utilized two clinical data sets without SDOH annotations: *MIMIC-III* and *UW Dataset*. *MIMIC-III* (referred to here as *MIMIC*) is a publicly available, deidentified health database for over 40K critical car' patients at Beth Israel Deaconess Medical Center from 2001-2012 [12]. MIMIC contains clinical notes, diagnosis codes, and other data. This work utilized 60K MIMIC discharge summaries. The *UW Dataset* is an existing clinical data set from the UW and Harborview Medical Centers generated between 2008-2019. This work utilized 83K emergency department, 22K admit, 8K progress, and 5K discharge summary notes from UW Dataset. An existing corpus with SDOH annotations created by Yetisgen and Vanderwende, *YVnotes*, was used for model training during active learning [17].

3.2. Annotation Scheme

We created detailed annotation guidelines for 12 SDOH (referred to here as *event types*), including substance use (alcohol, drug, and tobacco), physical activity, employment, insurance, living status, sexual orientation, gender identity, country of origin, race, and environmental exposure. Each *event* is a characterization of a specific SDOH instance and includes a trigger (head) and all associated arguments (attributes). These *events* capture changes to the status, extent, and temporality of SDOH in the patient timeline. Each event type is annotated across multiple dimensions. Table 1 summarizes the annotation of the most frequent SHAC event types: substance use, employment, and living status. Table A1 in the Appendix contains a summary of all annotated event types.

SDOH are annotated as events using the BRAT rapid annotation tool [34]. Figure 1 is a BRAT annotation example, describing a patient's employment and substance use. The trigger indicates the event type (e.g. *Employment* or *Tobacco*) and arguments describe the event. *Labeled* arguments, like Status, include both an annotated span and subtype label. *Span-only arguments*, like *Duration* or *History*, include an annotated span without an additional subtype.

3.3. Annotation Cycle

Social history sections, referred to here as *samples*, were extracted from MIMIC and the UW Dataset, using pattern matching to identify section headings (alphanumeric, forward slash, backslash, ampersand, or white space characters followed by a colon). SHAG includes *train, development*, and *test* sets. Samples for the train set were randomly and actively selected. Training samples were randomly selected for initial model training in active learning, then the initial model Was used in actively selecting samples to bias the training set towards diverse samples that frequently contain the phenomena of interest. All development and test samples were randomly selected to approximate the true distribution of the SDOH in the corpora used. Samples were annotated by four medical students through 12 rounds of annotation (8 randomly selected and 4 actively selected). Table A2 in the Appendix describes each round of annotation. The first two rounds were randomly sampled and double-annotated, to assess inter-annotator agreement. After the initial annotation round, the annotation guidelines were revised, and the initial annotations were updated.

3.4.   Evaluation and Annotation Scoring

We treat event annotation and extraction as a slot filling task, as this is most relevant to secondary use applications. As such, there can be multiple equivalent span annotations. Figure 2 presents the same sentence annotated by two annotators (labeled *A* and *B*), along with the populated slots. Both annotators labeled two *Drug* events: *Event 1* and *Event 2*. Event 1 describes past intravenous drug use (IVDU), and Event 2 describes current cocaine use. Event 1 is annotated identically by both annotators. However, there are differences in the annotation spans of Event 2, specifically for the *Trigger* ("cocaine" versus "cocaine use") and *Status* ("use" vs. "Recent"). From a slot perspective, the annotations for Event 2 are equivalent. Thus, scoring of automatic detection and annotator agreement is based on relaxed span match criteria, as described below. Trigger and argument performance is evaluated using precision (P), recall (R), and F1, micro averaged over the event types, argument types, and/or argument subtypes.

Trigger: Triggers, T_i, are represented by a pair (event type, e_i; token indices, x_i). For *Event 2* in Figure 2, $T_{A,2} = (e_{A,2} = Drug; x_{A,2} = [8])$ and $T_{B,2} = (e_{B,2} = Drug; x_{B,2} = [8, 9])$. Triggers of the same event type, e, are aligned by minimizing the distance between span centers computed from the token indices. Trigger equivalence is defined as

$$T_i \equiv T_j \text{ if } (e_i \equiv e_j) \wedge (T_i \text{ aligned with } T_j). \tag{1}$$

Although there are two drug events in the Figure 2 example, $T_{A,2}$ aligns with $T_{B,2}$ because of the overlapping spans.

Argument: Events are aligned based on trigger equivalence, and the arguments of aligned events are compared using different criteria for *labeled arguments* and *span-only arguments*. Labeled arguments, L_i, are represented as a triple (argument type, a_i; token indices, x_i; subtype, l_i). For *Event 2* in Figure 2, $L_{A,2} = (a_{A,2} = Status; x_{A,2} = [9], l_{A,2} = current)$ and $L_{B,2} = (a_{B,2} = Status; x_{B,2} = [7], l_{B,2} = current)$. For labeled arguments, the argument type, a, and subtype, l, capture the salient information and equivalence is defined as

$$L_i \equiv L_j \text{ if } (T_i \equiv T_j) \wedge (a_i \equiv a_j) \wedge (l_i \equiv l_j). \tag{2}$$

Span-only arguments, S_i, are represented as a pair (argument type, a_i; token indices, x_i). For *Event 2* in Figure 2, $S_{A,3} = (a_{A,3} = Type; x_{A,3} = [7])$ corresponds to "cocaine." Span-only arguments are not easily mapped to a fixed set of classes, and the identified span, x, contains the most salient argument information. Span-only arguments with equivalent triggers and argument types, $(T_i \equiv T_j) \wedge (a_i \equiv a_j)$, are compared at the token-level (rather than the span-level) to allow partial matches. Partial match scoring is used as partial matches can still contain useful information.

Cohen's Kappa: We evaluate annotator agreement using Cohen's Kappa, κ, coefficient, where higher κ denotes better annotator agreement [35]. Calculating κ for the full event structure is not informative, because the probability of random agreement is close to zero.

Instead, we calculate κ for trigger annotation in the subset of sentences with zero or one trigger for a given event type in either set of annotations, which covers most of the data. We focus on this subset of sentences, because triggers for a given event type are equivalent, if the annotated sentences both include one trigger of that type. We assess annotator agreement on the full event structure using F1 scores.

3.5. Annotation Statistics

SHAC consists of 4,480 annotated social history sections (70% train, 10% development, 20% test). Table 2 presents the corpus composition by source. The SHAC training samples are 29% randomly selected and 71% actively selected. All development and test data are randomly sampled. Figure 3 presents the event type distribution. The most frequent event types are *Drug, Tobacco, Alcohol, Living status*, and *Employment*, with the remaining event types occurring infrequently.

Figure 4 presents the annotator agreement for all event types in terms of F1 score for 300 doubly annotated notes from the first two rounds of annotation. For *Alcohol, Drug, Tobacco, Employment*, and *Living status*, trigger κ is $0.94 – 0.97$. For the remaining event types, trigger κ is $0.61 – 0.90$. κ is calculated for sentences with 0-1 events for each type (99% of all sentences). The trigger agreement is very high, in terms of F1 and κ, indicating the annotators are consistently identifying and distinguishing between events. The argument agreement is also high for labeled arguments. The somewhat lower agreement for span-only arguments is primarily due to small differences in the start and end token spans (e.g. "construction worker" vs. "construction").

4. Active Learning

This section presents the active learning framework used create SHAC and describes the associated performance gains.

4.1. Methods

A portion of the SHAC training samples were selected using active learning, where a sample is a social history section. Specifically, batch-mode active learning was used to facilitate coordination with human annotators through the cyclical cyclical process shown in Figure 5.

A batch of samples, *B*, was annotated and added to the labeled pool, *L*. The surrogate classifier was trained on *L* and then generated uncertainty scores for unlabeled data *U*. Using the uncertainty scores, the query function identified the next batch of samples, *B*. This process was repeated until the annotation objective was met.

Similar to Wu and Ostendorf [23], a query score is designed to combine informativeness and diversity scores of a batch of samples, *B*. Here, the score has the form:

$$Q(B) = \sum_{i \in B} (1 - s_i)^{\alpha} u(i) \tag{3}$$

where $u(i)$ is the uncertainty entropy of sample i, s_i is the similarity score of sample i relative to B, and $(1 - s_i)$ is the diversity score. α is a weight used to balance the relative importance

of the two scores ($a > 0$). The objective is to maximize the batch score, $Q(B)$. We explored different forms for the uncertainty and similarity scores for this multi-label scenario. We implemented a greedy approach to selecting examples, as shown in Algorithm 1.

Algorithm 1:

Greedy query function

Input: unlabeled samples U, batch size N

Output: batch of samples B

$B \leftarrow \varnothing$

while $|B| < N$ **do**

$k \leftarrow \text{argmax}_{i \in U} Q(B \cup i);$

$\quad B \leftarrow \cup \{k\};$

$\quad U \leftarrow U - \{k\};$

end

Diversity: Sample diversity is assessed in the observation space using two different similarity metrics: *average similarity* and *maximum similarity*, defined as

$$s_i^a = \frac{1}{|B|} \sum_{j \in B, j \neq i} a_{j,i} \quad s_i^m = \max_{j \in B, j \neq i} a_{j,i},$$

respectively, where $a_{j,i}$ is the cosine similarity of samples j and i. The maximum similarity approach is a stricter condition that pushes the batch of samples farther apart in the observation space, especially with larger batch sizes. Similar to Lilleberg et al. [36], unsupervised vector representations of samples were learned as the TF-IDF weighted averages of pre-trained word embeddings. Word embeddings were created using the word2vec skip-gram model [37] and trained on the entirety of the MIMIC discharge summaries (not just the social history sections). Separate TF-IDF weights were calculated for MIMIC and UW Dataset samples.

Uncertainty: Active learning query functions typically assess sample informativeness (uncertainty) using the target classification task. In this work, sample uncertainty was assessed using a simplified surrogate classification task, as a proxy for the more complex event-based annotation scheme. The SHAC annotation scheme includes some arguments (e.g. *Status* for *Alcohol*) that are more predictive of negative health outcomes than others (e.g. *Type* for *Alcohol*), and the prediction uncertainty varies across event types and arguments. To ensure the query function biases selection towards the most salient arguments, each of the five most frequent event types in SHAC were represented using the single argument that is most predictive of negative health outcomes: *Alcohol-Status, Drug-Status, Tobacco-Status, Employment-Type*, and *Living status-Status*. To cover samples with multiple events of the same type (e.g. both previous and current tobacco use described), an additional class, "multiple," is added to the argument subtypes, y_l, in Table 1, $y_c = \{y_l \cup$ "*multiple*"$\}$.

The text classification model, *Surrogate Classifier* in Figure 6, was used to assess sample uncertainty. The Surrogate Classifier operates on a sample, as a single sequence of n tokens without line breaks. The input social history section is mapped to contextualized word embeddings using *Bio+Discharge Summary BERT* [38], a version of *BERT* [39] trained on clincal text from MIMIC. The BERT output feeds into a bidirectional long short-term memory (bi-LSTM) layer, the output of which feeds into event-specific output layers. Separate self-attention (Attn) output layers for each event type make sample-level predictions. Details of the Surrogate Classifier are similar to the shared and event-argument layers of the full event detec-tion system described in the next section. The Surrogate Classifier generates a set of five multi-class predictions for each sample, one for each event type.

We explored two approaches to characterizing sample uncertainty: i) the sum of the five event entropy values, similar to previous work [32, 40–42], and ii) entropy for an individual event type, iterating over all types (referred to as "loop"). As a "loop" example, *Alcohol-Status* entropy is used for sample 1, *Drug-Status* entropy is used for sample 2, and so forth, starting over with *Alcohol-Status* entropy for sample 6. The second method was motivated by the concern that summing the entropy values (referred to as "sum") could overly bias the selection process in favor of high-entropy event types, reducing the diversity of event types.

4.2. Experiments & Results

Query strategy selection: Due to limitations in the annotation budget, the query strategy was determined early in the annotation effort. We used the first 700 annotated samples, L_Q, which consists of random MIMIC samples. L_Q was partitioned into $L_Q^T \coloneqq \{620$ train samples$\}$ and $L_Q^D \coloneqq \{80$ development samples$\}$. For random sampling and each active sampling configuration, 10 runs were performed:

i. $L_{T1} \leftarrow 100$ samples from 100 samples from L_Q^T. Train model, M_1, on L_{T1}

ii. $L_{T1} \leftarrow 100$ samples from $\{L_Q^T - L_{T1}\}$ (random or active). Train model, M_2, on $\{L_{T1} \cup L_{T2}\}$

iii. Evaluate the performance of M_2 on L_Q^D

Active sampling experimentation included different uncertainty types ("loop" vs. "sum"), similarity types ("average" vs. "maximum"), and a values $\{0.1, 1, 2\}$. All active learning configurations outperform the random baseline with significance ($p < 0.05$[3]). The best configuration, uncertainty type ="sum", similarity type="maximum", and $a = 0.1$, was used in active selection. This configuration and other hyperparameters of the Surrogate Classifier were tuned on L_Q^D (for details, see Tables A3 and A4 in the Appendix).

Active learning performance: After the first round of active learning, performance of the Surrogate Classifier was evaluated to confirm the effectiveness of the active learning framework. Model training included the sets: $L_Y \coloneqq \{284$ YVnotes samples$\}$ and $L_R \coloneqq \{532$

---

[3]Significance was assessed using Welch's T-test, which is T-test variant that assumes unequal variances is the test distributions.

random MIMIC train samples}. YVnotes was used to train the Surrogate Classifier to improve its accuracy and thereby obtain a better uncertainty score. L_R was partitioned into $L_R^I \coloneqq \{288$ initial training samples} and $L_R^P \coloneqq \{244$ remaining samples, $L_R - L_R^I\}$. For the first round of active selection, an initial model, M_I, was trained on $\{L_R^I \cup L_Y\}$ and used to select 400 MIMIC samples, L_A. L_R^P was withheld when training M_I to validate the active learning approach. Hyperparameters were tuned on $L_D \coloneqq \{188$ MIMIC development samples} (parameter values in Table A4 of the Appendix).

Figure 7 presents the performance of four cases on $L_E \coloneqq \{376$ MIMIC test samples}. For *MIMIC-only initial, +random*, and *+active*, 10 runs were performed to account for variance in model initialization. For *MIMIC-only initial* and *+random*, the training sets are fixed, as all data is used each run. For *+active*, the training set varies because only a subset of L_A is randomly selected each run, so sampling variance is introduced. The error bars in Figure 7 indicate the standard deviation of the F1 scores across runs. Comparing *MIMIC-only initial* to *initial* demonstrates that including YVnotes improves performance. Adding active samples to the initial training set yields a statistically significant improvement over adding random samples ($p < 0.06$[3]), demonstrating the effectiveness of the active learning framework on the surrogate task.

The effectiveness of the active learning framework on the target event extraction task for the same conditions is presented in Figure 8, where scores are averaged across event types.[4] The details of the event extraction model are presented in Section 5. The performance achieved by adding active samples outperforms that of adding random samples for labeled argument and span-only argument extraction, with significance ($p < 0.01$[3]). The addition of actively selected notes improved extraction performance, relative to the random baseline, across most annotated phenomena. However, **the largest active learning performance gains were achieved for prominent health risk factors, including past and current drug use, current tobacco use, unemployment, homelessness, and living with others** (+0.09 F1 for *current Drug Status*, +0.14 F1 for *past Drug Status*, +0.07 F1 for *current Tobacco Status*, +0.04 F1 for unemploye. *Employment Status*, +0.06 F1 for *homeless Living Status Type*, and +0.07 F1 for *with others Living Status Type*). The difference in trigger performance is not statistically significant. This result validates the use of the simplified surrogate text classification task as a proxy for the more complex event extraction task. After validating the active learning strategy, three additional rounds of active selection were performed (see Table A2 of the Appendix for details), and the Surrogate Classifier model was retrained prior to each active round. Due to the limited number of random samples, further comparisons of active vs. random sampling are not possible.

We hypothesized the Surrogate Classifier uncertainty would bias the selection process to include more health risk factors (e.g. positive substance abuse, unemployment, being on disability, homelessness, etc.), which tend to be more challenging to automatically extract than less risky behavior (e.g. no sub stance use, being employed, and living with family). Active learning successfully identified samples with richer, more detailed SDOH

---

[4]For the Event Extractor, we exclude L_Y since YVnotes do not include all of the labeled phenomena of SHAC.

descriptions. Figure 9 presents the label frequency per sample (note section) for random and active samples for the entirety of SHAC. The frequency of positive substance use (*Status* ∈ {*current, past*}) is 83% higher in active samples than random samples, with the frequency of positive drug use 151% higher with active selection. Active sampling produced higher rates for all *Employment Status* labels, except *retired*. Descriptions of retirement, tend to have low entropy, because of the reliable presence of keywords like "retired" or "retirement." Regarding *Living Status*, the rate of *homeless* is 109% higher in active samples than random samples, and the rate of *with others* is 81% higher. The rate of *alone* is slightly lower in active samples, likely due to lower entropy associated with the limited vocabulary used to describe living alone (e.g. "alone" or "by herself").

5. Event Extraction

This section introduces the *Event Extractor*, which jointly predicts all the phenomena in Table 1, and presents the initial extraction results for SHAC.

5.1. Methods

The Event Extractor generates sentence and token-level predictions that are assembled into events, similar to the SHAC annotation scheme. The Event Extractor builds on our previous state-of-the-art neural multi-task extractor for substance abuse information [43]. It is a generalized version of this previous work and is shown in Figure 10.

Shared layers: Individual sentences are encoded using *Bio+Discharge Summary BERT* [44], creating an $n \times d$ matrix, where n is the sentence length in tokens and d is the BERT embedding size. BERT parameters are frozen during training (no back propagation) to limit computational cost. Similar to other work [45], only the last word piece embedding for each token is used, to simplify the downstream sequence tagging. The BERT encoding feeds into a bi-LSTM. The forward and backward outputs states of the bi-LSTM are concatenated resulting in $n \times 2u$ matrix, V, where u is the hidden size. V feeds into event type and argument-specific output layers.

Trigger: The presence of each event type is predicted using separate self-attentive binary classifiers (not present/present). Positive predictions serve as the trigger for assembling events, and the token position with the maximum attention weight serves as the trigger span. During training, event type k is considered *present*, if the sentence contains one or more events of type k. The trigger probability for event type $k \in \{1, .., m\}$ is calculated as

$$P_k^t = softmax(W_k^t(A_k^t V)^T + b_k^t) \tag{4}$$

where W_k^t, is $2 \times 2u$ weight matrix, b_k^t is a 2×1 bias vector, and A_k^t, is a $1 \times n$ vector of attention weights

$$A_k^c = softmax(Y_k^c V^T) \text{ for } k = 1, \ldots, m \tag{5}$$

and Y_k^c is another learned weight matrix. The trigger probabilities, P_k^t, are concatenated to form a $2 \times m$ matrix, P^t, for the labeled argument prediction. An event is detected if it has probability greater than 50%.

Labeled arguments: Labeled argument prediction is also treated as a text-classification task, and utilizes separate self-attentive output layers for each labeled argument. The token position with the maximum attention weight serves as the argument span. The probability of labeled argument l for event type k is calculated as

$$P_{k,l}^s = softmax(W_{k,l}^s[P^t, (A_{k,l}^s V)^T] + b_{k,l}^s) \tag{6}$$

where $W_{k,l}^s$ is a weight matrix, $A_{k,l}^s$ is a vector of attention weights, and $b_{k,l}^s$ is a bias vector. The dimension of $P_{k,l}^s$ depends on the number of possible labels for that event-argument combination. The labeled argument probabilities, $P_{k,l}^s$, are concatenated to form a 2×6 matrix, P^s, for use in span-only argument detection. Experimentation included six labeled arguments: *Status* for *Alcohol, Drug*, and *Tobacco*; *Status* for *Employment*; and *Status* and *Type* for *Living status*.

Span-only arguments: Span-only arguments are predicted using linear-chain Conditional Random Field (CRF) [46] output layers at the output of the bi-LSTM, which is a popular sequence tagging approach [47, 48]. The bi-LSTM network learns sequential word dependencies, and the CRF learns conditional dependencies between labels. A separate CRF extracts the span-only arguments for each event type (i.e. five CRF output layers), with input features V and P^s. Sequence labels are represented using the begin-inside-outside (BIO) approach. Experimentation included 20 span-only arguments: *Duration, History, Type, Amount*, and *Frequency* for *Alcohol, Drug*, and *Tobacco*; *Duration, History*, and *Type* for *Employment*; and *Duration* and *History* for *Living status*.

Training: The Event Extractor was trained on the entire SHAC train set to simultaneously extract substance abuse, living situation, and employment information. Similar to previous multitask work [49–54], the Event Extractor shares information across tasks (event types and arguments in this application). The Event Extractor hyperparameters were tuned on the development set, L_D (parameter values in Table A4 of the Appendix).

5.2. Results

Figure 11 and Table 3 present the trigger and argument performance of the Event Extractor on the MIMIC and UW Dataset test sets. As described in Section 3.4, the argument extraction performance accounts for the alignment of the event triggers (i.e. only arguments with equivalent triggers can be equivalent). Overall, performance is higher on MIMIC, even though there are more UW Dataset training samples, including more active samples. The UW Dataset portion of SHAC includes four different note types, whereas the MIMIC portion includes only one note type, which likely contributes to the lower performance on the UW Dataset.

Table 3 presents detailed results for the same Event Extractor model and data configuration as Figure 11. Trigger performance is greater than 0.89 F1 for all event types in both data sets. Labeled argument performance is similar in both data sets for *Alcohol* and *Tobacco Status*; however, there are performance differences for *Drug*, *Employment*, and *Living status* labeled arguments. In substance use *Status* prediction, the *none* label is typically less confusable and easier to predict than *past* and *current*. In the test set, the relative frequency of *none Status* labels for *Drug* events is higher in MIMIC samples (80%) than UW Dataset samples (57%), which contributes to the higher performance on MIMIC. *Living status Status* performance is lower in the UW Dataset, even though the distribution of *Status* labels is similar in both data sets. *Living status Type* performance is 0.12 F1 higher in MIMIC than the UW Dataset. In the test set, the distribution of *Living status Type* labels differs greatly between the data sets with the UW Dataset at 37% *with family*, 22% *with others*, 26% *homeless*, and 15% *alone* and MIMIC at 57% *with family*, 16% *with others*, 2% *homeless*, and 25% *alone*. For the span-only arguments, the performance is calculated at the token-level and micro averaged across the arguments for each event type. Span-only argument performance is comparable for *Alcohol*, *Tobacco*, and *Employment*. However, it is higher for *Drug* span-only arguments in MIMIC than the UW Dataset. *Living status* span-only argument performance is very low for both data sets, primarily due to sparsity in the training set (only 167 *Duration* and *History* arguments among 3,267 *Living status* events).

5.3. Limitations

Although the Event Extractor achieved high performance for most target phenomena, the extraction framework has several limitations. The Event Extractor treats trigger and labeled argument prediction as a text classification task and can only represent a single event of a given type per sentence. Figure 12a presents predicted labels for a sentence with multiple gold *Drug* events describing current marijuana use and previous cocaine use. While the *Type* predictions in this example are correct, the *Status* prediction of *past* is incorrectly associated with both marijuana and cocaine. Of the sentences with at least one event in SHAC, 6% contain multiple events of the same type. Span-only arguments for each event type are extracted using a single CRF, which cannot accommodate overlapping spans. Figure 12b presents predictions for a sentence where the gold span-only argument spans overlap. The *Amount* is correctly labeled as "about 1 pint of vodka," but there should also be a *Type* argument of "vodka." Approximately 6% of span-only arguments in events of the same type overlap in SHAC. The Event Extractor treats sentences independently. It does not incorporate context from the preceding sentences and cannot generate events that span multiple sentences. Figure 12c presents predictions for an example where past tobacco use is described in concurrent sentences. The first sentence includes a strong cue for *past Status*, "quit"; however, the *Status* in the second sentence is less clear without previous context. Fewer than 2% of SHAC events span multiple sentences.

6.   Conclusions

We present a new clinical corpus, SHAC, with detailed event-based annotations for 12 SDOH. SHAC includes approximately 4.5K social history sections from multiple institutions and note types and contains frequent descriptions of alcohol, drug, and tobacco

use, employment, and living status. Approximately 71% of the SHAC training set was selected using a novel active learning framework that utilizes a surrogate task for assessing sample uncertainty, which increased the prevalence of critical risk factors in the annotated training data, including positive substance use, unemployment, disability, and homelessness, aM increased event extraction performance, relative to using only randomly selected samples. The actively selected samples improve performance in both the surrogate task and the target event extraction task, validating the surrogate task approach. A neural multi-task model is presented for characterizing substance use, employment, and living status across multiple dimensions, including status, extent, and temporal fields. The event extractor model achieves high performance on the MIMIC and UW Dataset: 0.89-0.98 F1 in identifying distinct SDOH events, 0.82-0.93 F1 for substance use status, 0.81-0.86 F1 for employment status, and 0.81-0.93 F1 for living status type. The annotation guidelines and source code will be made available online[5].

Acknowledgements

Appendix

Table A1:

Annotation guideline summary for all event types.

Event type, e	Argument type, a	Argument subtypes, y_l	Span examples
Substance use (Alcohol, Drug, & Tobacco)	Status[*]	{none, current, past}	"denies," "smokes"
	Duration	–	"for the past 8 years"
	History	–	"seven years ago"
	Type	–	"beer," "cocaine"
	Amount	–	"2 packs," "3 drinks"
	Frequency	–	"daily," "monthly"
Employment	Status[*]	{employed, unemployed, retired, on disability, student, homemaker}	"works," "unemployed"
	Duration	–	"for five years"
	History	–	"15 years ago"
	Type	–	"nurse," "office work"
Living status	Status[*]	{current, past, future}	"lives," "lived"
	Type[*]	{alone, with family, with others, homeless}	"with husband"
	Duration	–	"for the past 6 months"
	History	–	"until a month ago"
Insurance	Status	{yes, no}	"has been off'"
Sexual orientation	Status	{current, past}	"participated in"

---

[5]https://github.com/uw-bionlp

Event type, e	Argument type, a	Argument subtypes, y_l	Span examples
	Type	{heterosexual, homosexual, bisexual}	"homosexual"
Gender identity	Status	{current, past}	"identifies as"
	Type	{cisgender, transgender}	"transgender"
Country of origin	Type	–	"England"
Race	Type	–	"African American"
	Status	{none, current, past}	"currently jogs"
	Duration	–	"for several years"
	History	–	"10 years ago"
Physical activity	Type	–	"walks"
	Amount	–	"4 miles"
	Frequency	–	"every evening"
	Status	{none, current, past}	"no history"
	Duration	–	"since 2001"
	History	–	"until a month ago"
Environmental exposure	Type	–	"asbestos"
	Amount	–	"significant"
	Frequency	–	"daily"

[*] indicates the argument is required.

Table A2:

Annotation round summary, including selection type (andom versus active) and training data used in active selection.

Round	Source	Selection	Active learning training set	Train	Dev	Test	Total
1	MIMIC	Random	–	100	–	–	100
2	MIMIC	Random	–	144	56	–	200
3	MIMIC	Random	–	288	112	–	400
4	UW Dataset	Random	–	84	140	280	504
5	MIMIC	Active	572 samples (Round 3 train + 284 YVnotes)	400	–	–	400
6	UW Dataset	Random	–	168	120	240	528
7	MIMIC	Random	–	–	20	280	300
8	UW Dataset	Random	–	112	–	–	112
9	UW Dataset	Active	1336 samples (Rounds 3-8 *train* + 284 YVnotes)	728	–	–	728
10	UW Dataset	Active	2064 samples (Rounds 3-9 *train* + 284 YVnotes)	728	–	–	728
11	MIMIC	Active	3036 samples (Rounds 1-10 *train* + 284 YVnotes)	384	–	–	384
12	MIMIC	Random	–	–	–	96	96
			TOTAL	**3136**	**448**	**896**	**4480**

Table A3:

Active learning query function tuning performance.

Uncertainty	Similarity	α	F1
loop	average	1.0	0.788[*]
loop	maximum	0.1	0.776[*]
sum	average	2.0	0.788[*]
sum	maximum	0.1	0.794[*]

[*] indicates statistical significance ($p < 0.05$) relative to a random baseline of 0.752 F1.

Table A4:

Surrogate Classifier hyperparameters

Parameter	Query function selection in Table A3	Active learning evaluation in Figure 7
batch size	20	100
learning rate	0.001	0.005
maximum gradient L2 norm	1.0	1.0
maximum length	200	200
number of epochs	500	500
LSTM hidden size	100	100
dropout, input to LSTM	0.7	0.4
dropout, output of LSTM	0.0	0.4
dropout, self-attention	0.7	0.4

Table A5:

Event Extractor hyperparameters

Parameter	Figure 8, Figure 11, and Table 3
batch size	50
learning rate	0.005
maximum gradient L2 norm	0.5
maximum length	30
number of epochs	250
LSTM hidden size	100
dropout, input to LSTM	0.6
dropout, output of LSTM	0.4
dropout, self-attention	0.4

References

[1]. Murphy SL, Xu J, Kochanek KD, Arias E, Mortality in the United States, 2017. NCHS Data Brief, no 328., National Cent. for Heal. Statistics (2018). URL: https://www.cdc.gov/nchs/data/databriefs/db328-h.pdf.

[2]. Daniel H, Bornstein SS, Kane GC, Addressing social determinants to improve patient care and promote health equity: An american college of physicians position paper, Annals of Intern. Medicine 168 (2018) 577–578. doi:10.7326/M17-2441.

[3]. Himmelstein DU, Woolhandler S, Determined action needed on social determinants, Annals of Intern. Medicine 168 (2018) 596–597. doi:10.7326/M18-0335.

[4]. Centers for Disease Control and Prevention, Annual smoking-attributable mortality, years of potential life lost, and productivity losses–united states, 1997-2001, Morbidity and Mortal. Wkly. Report 54 (2005) 625. doi:10.1001/jama.294.7.788.

[5]. World Heal. Organization, Global status report on alcohol and health 2018, World Heal. Organization (2019). URL: https://www.who.int/substance_abuse/publications/global_alcohol_report/gsr_2018/en/.

[6]. Degenhardt L, Hall W, Extent of illicit drug use and dependence, and their contribution to the global burden of disease, The Lancelot 379 (2012) 55–70. doi:10.1016/S0140-6736(11)61138-0.

[7]. Cacioppo JT, Hawkley LC, Social isolation and health, with an emphasis on underlying mechanisms, Perspectives in Biology and Medicine 46 (2003) S39–S52. doi:10.1353/pbm.2003.0063. [PubMed: 14563073]

[8]. Clougherty JE, Souza K, Cullen MR, Work and its role in shaping the social gradient in health, Annals of the New York Academy of Sciences 1186 (2010) 102–124. doi:10.1111/j.1749-6632.2009.05338.x. [PubMed: 20201870]

[9]. Blizinsky KD, Bonham VL, Leveraging the learning health care model to improve equity in the age of genomic medicine, Learn. Heal. Systems 2 (2018) e10046. doi:10.1002/lrh2.10046.

[10]. Demner-Fushman D, Chapman WW, McDonald CJ, What can natural language processing do for clinical decision support?, Journal of Biomedical Inform. 42 (2009) 760–772. doi:10.1016/j.jbi.2009.08.007.

[11]. Jensen PB, Jensen LJ, Brunak S, Mining electronic health records: towards better research applications and clinical care, Nature Rev. Genetics 13 (2012) 395. doi:10.1038/nrg3208. [PubMed: 22549152]

[12]. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016) 160035. doi:10.1038/sdata.2016.35. [PubMed: 27219127]

[13]. Uzuner O, Goldstein I, Luo Y, Kohane I, Identifying patient smoking status from medical discharge records, Journal of the American Medical Inform. Association 15 (2008) 14–24. doi:10.1197/jamia.M2408.

[14]. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, Foote J Jr, Moseley ET, Grant DW, Tyler PD, et al., Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives, Public Library of Science ONE 13 (2018) e0192360. doi:10.1371/journal.pone.0192360. [PubMed: 29447188]

[15]. Feller DJ, Zucker J, et al., Towards the inference of social and behavioral determinants of sexual health: Development of a gold-standard corpus with semi-supervised learning, in: AMIA Annual Symposium Proc, 2018, p. 422 URL: https://www.ncbi.nlm.nih.gov/pubmed/30815082.

[16]. Wang Y, Chen ES, Pakhomov S, Arsoniadis E, Carter EW, Lindemann E, Sarkar IN, Melton GB, Automated extraction of substance use information from clinical texts, in: AMIA Annual Symposium Proc, 2015, pp. 2121–30. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765598/.

[17]. Yetisgen M, Vanderwende L, Automatic identification of substance abuse from social history in clinical text, Artificial Intelligence in Medicine (2017) 171–181. doi:10.1007/978-3-319-59758-4_18.

[18]. Cohn D, Atlas L, Ladner R, Improving generalization with active learning, Machine Learn. 15 (1994) 201–221. doi:10.1007/BF00993277.

[19]. Cohn DA, Ghahramani Z, Jordan MI, Active learning with statistical models, Journal of Artificial Intelligence Res. 4 (1996) 129–145. doi:10.1613/jair.295.

[20]. Shen D, Zhang J, Su J, Zhou G, Tan C-L, Multi-criteria-based active learning for named entity recognition, in: Association for Computational Linguistics, 2004, p. 589–596. doi:10.3115/1218955.1219030.

[21]. Yang Y, Ma Z, Nie F, Chang X, Hauptmann AG, Multi-class active learning by uncertainty sampling with diversity maximization, International Journal of Computer Vision 113 (2015) 113–127. doi:10.1007/s11263-014-0781-x.

[22]. Du B, Wang Z, Zhang L, Zhang L, Liu W, Shen J, Tao D, Exploring representativeness and informativeness for active learning, IEEE Transactions on Cybernetics 47 (2017) 14–26. doi:10.1109/TCYB.2015.2496974. [PubMed: 26595936]

[23]. Wu W, Ostendorf M, Graph-based query strategies for active learning, IEEE/ACM Transactions on Audio, Speech, and Lang. Processing 21 (2013) 260–269. doi:10.1109/TASL.2012.2219525.

[24]. Tong S, Koller D, Support vector machine active learning with applications to text classification, Journal Machine Learning Res. 2 (2002) 45–66.

[25]. Park S, Lee W, Moon I-C, Efficient extraction of domain specific sentiment lexicon with active learning, Pattern Recognit. Letters 56 (2015) 38 – 44. doi:10.1016/j.patrec.2015.01.004.

[26]. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H, A study of active learning methods for named entity recognition in clinical text, Journal of Biomedical Inform. 58 (2015) 11–18. doi:10.1016/j.jbi.2015.09.010.

[27]. Chen Y, Lask TA, Mei Q, Chen Q, Moon S, Wang J, Nguyen K, Dawodu T, Cohen T, Denny JC, et al., An active learning-enabled annotation system for clinical named entity recognition, BMC Medical Inform. and Decision Mak. 17 (2017) 82. doi:10.1186/s12911-017-0466-9.

[28]. Kholghi M, De Vine L, Sitbon L, Zuccon G, Nguyen A, Clinical information extraction using small data: An active learning approach based on sequence representations and word embeddings, Journal of the Association for Information Science and Technology 68 (2017) 2543–2556. doi:10.1002/asi.23936.

[29]. Li M, Scaiano M, El Emam K, Malin BA, Efficient active learning for electronic medical record de-identification, AMIA Summits on Translational Science Proc. 2019 (2019) 462 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568071/.

[30]. Gao J, Chen J, Zhang S, He X, Lin S, Recognizing biomedical named entities by integrating domain contextual relevance measurement and active learning, in: IEEE Information Technology, Networking, Electronic and Automation Control Conference, 2019, pp. 1495–1499. doi:10.1109/ITNEC.2019.8728991.

[31]. Shelmanov A, Liventsev V, Kireev D, Khromov N, Panchenko A, Fedulova I, Dylov DV, Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts, in: IEEE International Conference on Bioinform. and Biomedicine, 2019, pp. 482–489. doi:10.1109/BIBM47256.2019.8983157.

[32]. Maldonado R, Goodwin TR, Harabagiu SM, Active deep learning-based annotation of electroencephalography reports for cohort identification, AMIA Summits on Translational Science Proc. 2017 (2017) 229 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543351/.

[33]. Maldonado R, Harabagiu SM, Active deep learning for the identification of concepts and relations in electroencephalography reports, Journal of Biomedical Inform. 98 (2019) 103265. doi:10.1016/j.jbi.2019.103265.

[34]. Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J, BRAT: a web-based tool for NLP-assisted text annotation, in: Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107. URL: https://www.aclweb.org/anthology/E12-2021.

[35]. Cohen J, A coefficient of agreement for nominal scales, Educational and psychological meas. 20 (1960) 37–46. doi:10.1177/001316446002000104.

[36]. Lilleberg J, Zhu Y, Zhang Y, Support vector machines and word2vec for text classification with semantic features, in: International Conference on Cognitive Inform. & Cognitive Computing, 2015, pp. 136–140. doi:10.1109/ICCI-CC.2015.7259377.

[37]. Mikolov T, Chen K, Corrado G, Dean J, Efficient estimation of word representations in vector space, in: International Conference on Learn. Representations, 2013, pp. 1–12. URL: https://arxiv.org/abs/1301.3781.

[38]. Alsentzer E, Murphy J, Boag W, Weng W-H, Jin D, Naumann T, McDermott M, Publicly available clinical BERT embeddings, in: Clinical Natural Language Processing Workshop, 2019, pp. 72–78. doi:10.18653/v1/W19-1909.

[39]. Devlin J, Chang M-W, Lee K, Toutanova K, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[40]. Yang B, Sun J-T, Wang T, Chen Z, Effective multi-label active learning for text classification, in: International Conference on Knowl. Discov. and Data Min, 2009, p. 917–926. doi:10.1145/1557019.1557119.

[41]. Wu J, Sheng VS, Zhang J, Zhao P, Cui Z, Multi-label active learning for image classification, in: IEEE International Conference on Image Processing, 2014, pp. 5227–5231. doi:10.1109/ICIP.2014.7026058.

[42]. Reyes O, Ventura S, Evolutionary strategy to perform batch-mode active learning on multilabel data, ACM Transactions on Intelligent Systems and Technology 9 (2018). doi:10.1145/3161606.

[43]. Lybarger K, Yetisgen M, Ostendorf M, Using neural multi-task learning to extract substance abuse information from clinical notes, in: AMIA Annual Symposium Proc, 2018, p. 1395 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371261/.

[44]. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinform. (2019). doi:10.1093/bioinformatics/btz682.

[45]. Kitaev N, Cao S, Klein D, Multilingual constituency parsing with self-attention and pretraining, in: Association for Computational Linguistics, 2019, p. 3499–3505. doi:10.18653/v1/P19-1340.

[46]. Lafferty J, McCallum A, Pereira FC, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: International Conference on Machine Learn, 2001, pp. 282–289. URL: https://repository.upenn.edu/cis_papers/159/.

[47]. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C, Neural architectures for named entity recognition, in: North American Chapter of the Association for Computational Linguistics, 2016, pp. 260–270. doi:10.18653/v1/N16-1030.

[48]. Luan Y, Ostendorf M, Hajishirzi H, Scientific information extraction with semi-supervised neural tagging, in: Conference of the Empir. Methods in Natural Lang. Processing, 2017, pp. 2641–2651. doi:10.18653/v1/D17-1279.

[49]. Collobert R, Weston J, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: International Conference on Machine Learn, 2008, pp. 160–167. doi:10.1145/1390156.1390177.

[50]. Luo Y, Cheng Y, Uzuner O, Szolovits P, Starren J, Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes, Journal of the American Medical Inform. Association 25 (2017) 93–98. doi:10.1093/jamia/ocx090.

[51]. Jaques N, Taylor S, Nosakhare E, Sano A, Picard R, Multi-task learning for predicting health, stress, and happiness, in: NIPS Work. on Machine Learn. For Heal., 2016, pp. 1–5. URL: https://affect.media.mit.edu/pdfs/16.Jaques-Taylor-et-al-PredictingHealthStressHappiness.pdf.

[52]. Liu B, Lane I, Attention-based recurrent neural network models for joint intent detection and slot filling, in: INTERSPEECH, 2016, pp. 685–689. doi:10.21437/Interspeech.2016-1352.

[53]. Luan Y, He L, Ostendorf M, Hajishirzi H, Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction, in: Conference of the Empir. Method in Natural Lang. Processing, 2018, pp. 3219–3232. doi:10.18653/v1/D18-1360.

[54]. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A, Multitask learning and benchmarking with clinical time series data, Scientific Data 6 (2019) 96. doi:10.1038/s41597-019-0103-9. [PubMed: 31209213]

Highlights

We present a corpus with annotations for 12 social determinants of health (SDOH).

Annotations benefit from active learning using a surrogate classification task.

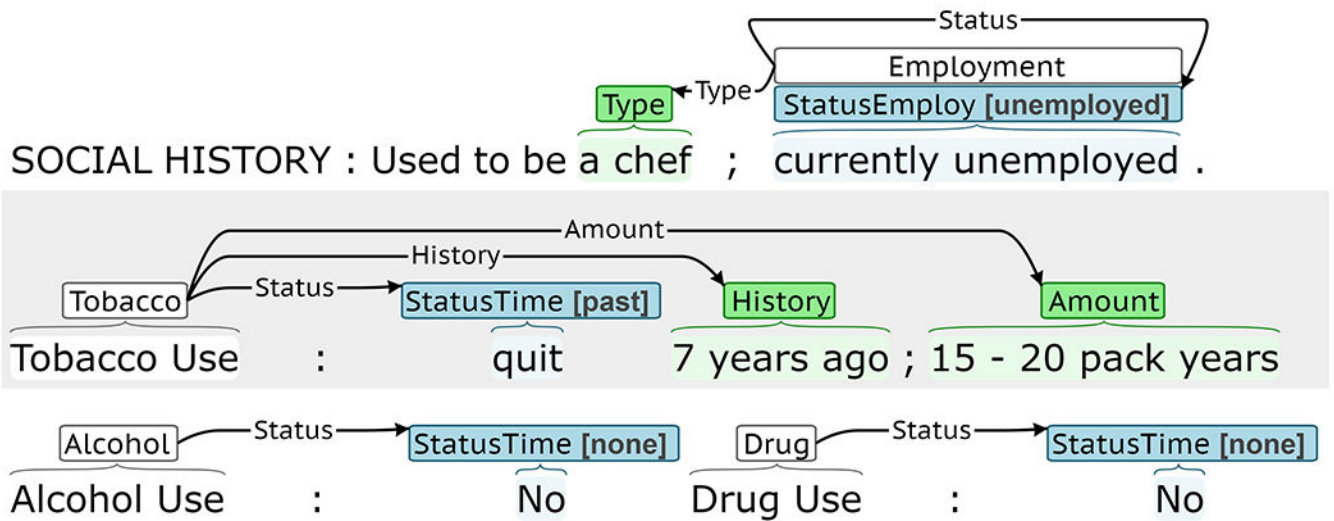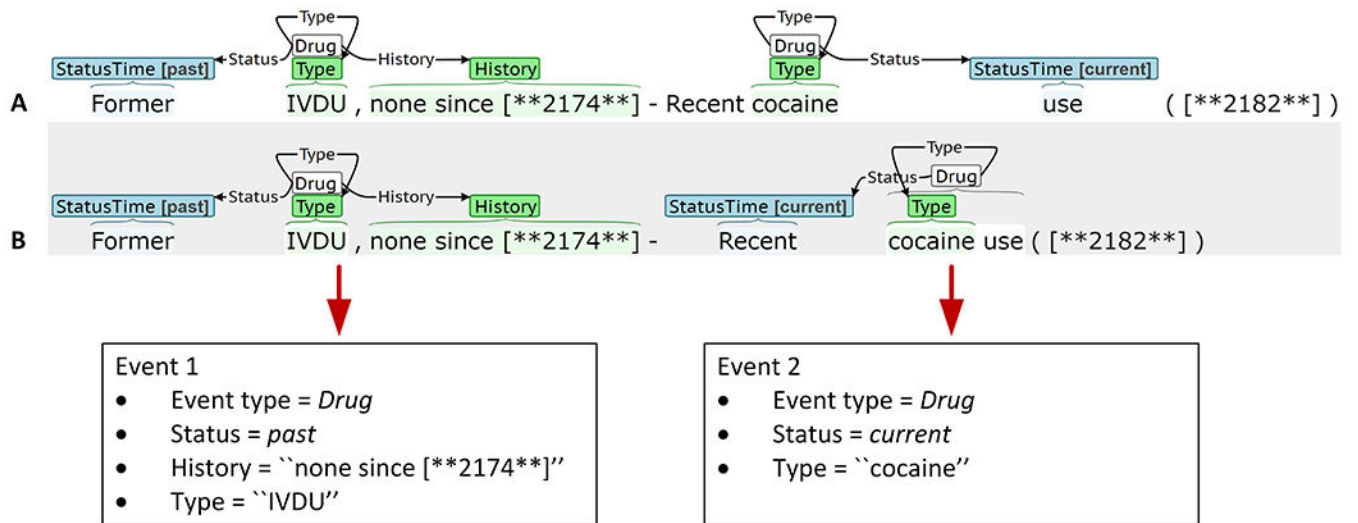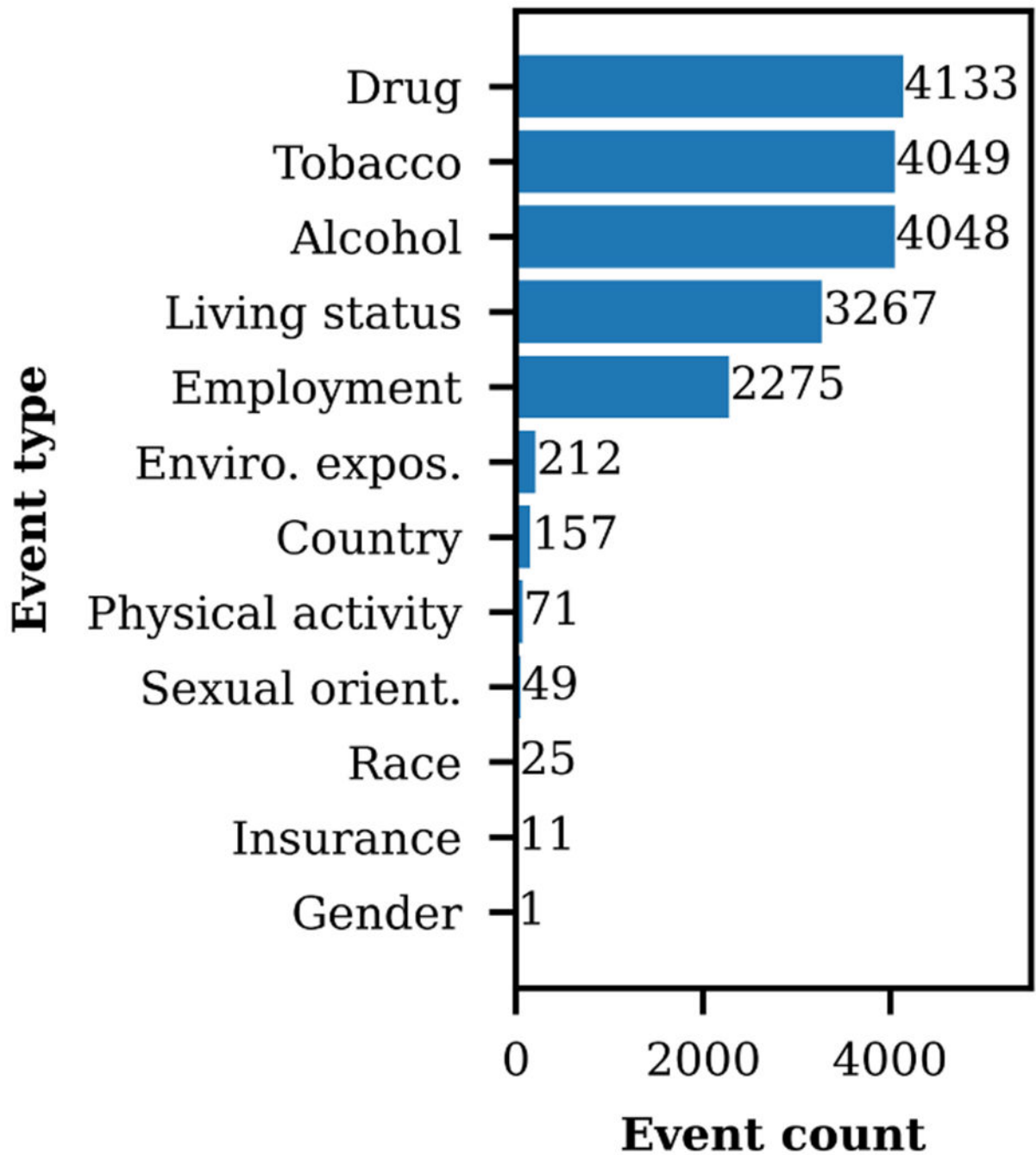A neural multi-task event extractor achieves high performance extracting SDOH.

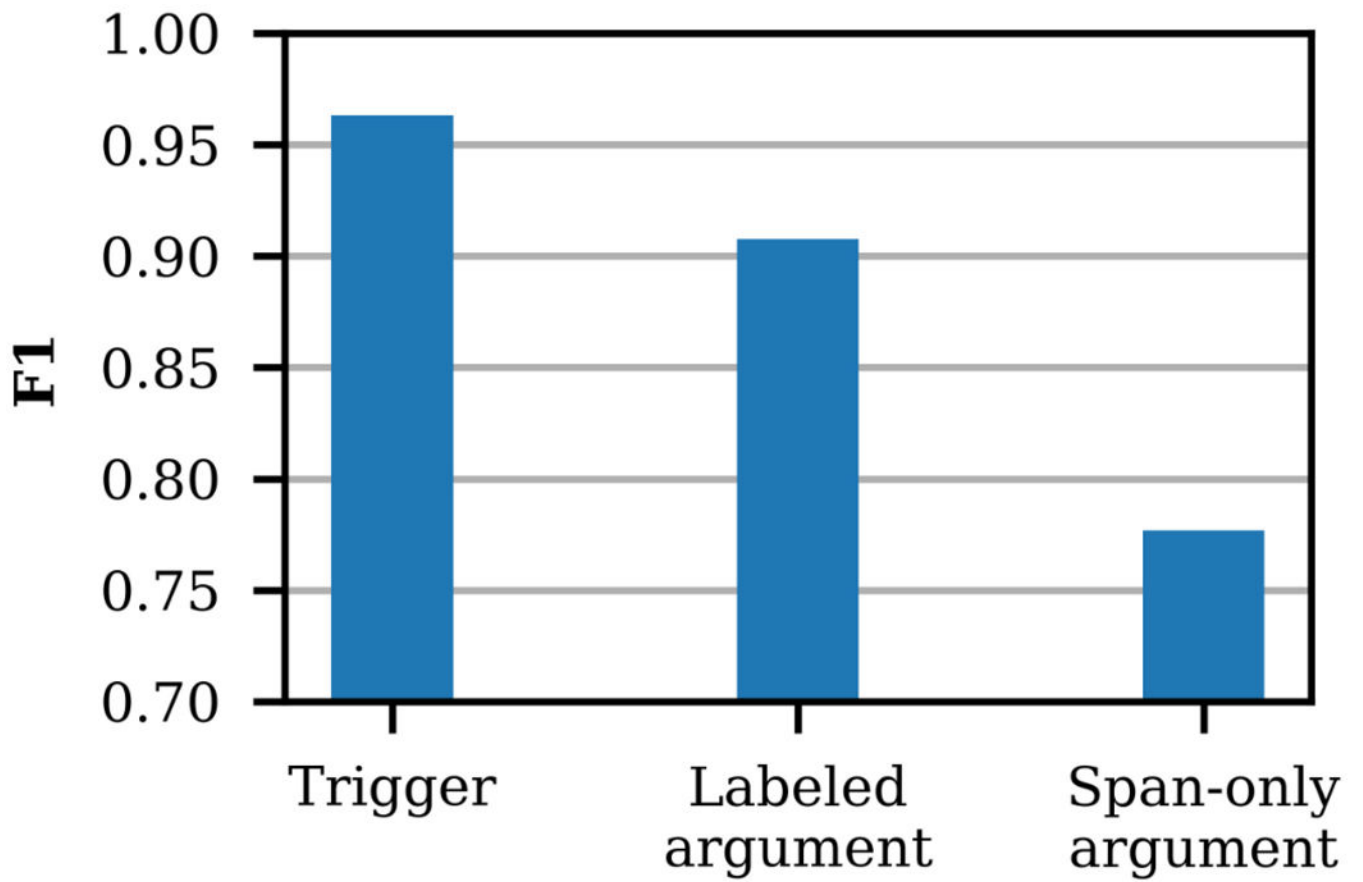**Figure 1:**
BRAT annotation example

Figure 2:
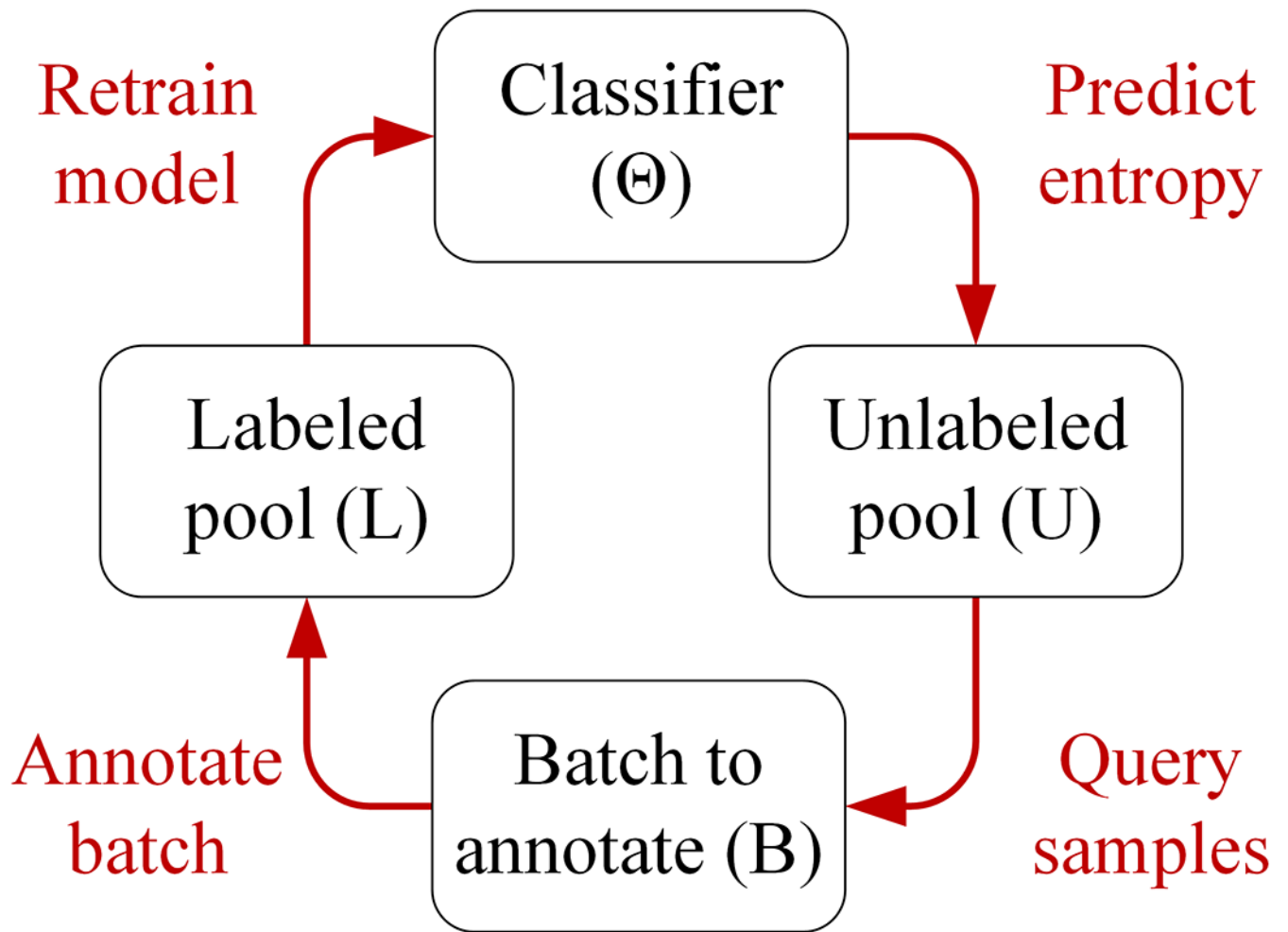Annotation examples describing event extraction as a slot filling task
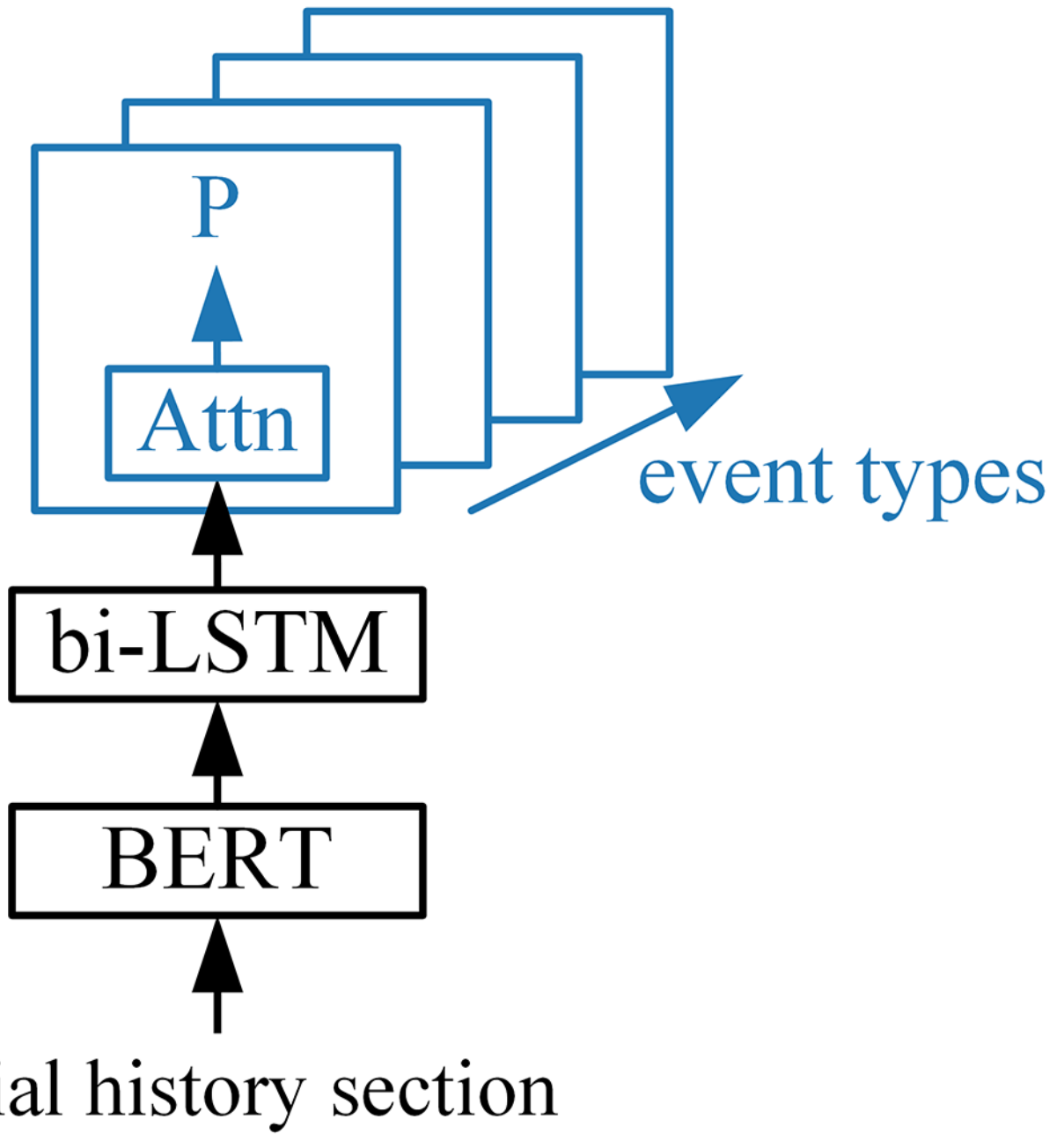
Figure 3:
Event type distribution

Figure 4:
Annotator agreement for 300 doubly annotated MIMIC samples

Retrain model

Classifier (Θ)

Predict entropy

Labeled pool (L)

Unlabeled pool (U)

Annotate batch

Batch to annotate (B)

Query samples

Figure 5:
Active learning annotation cycle

P

Attn

event types
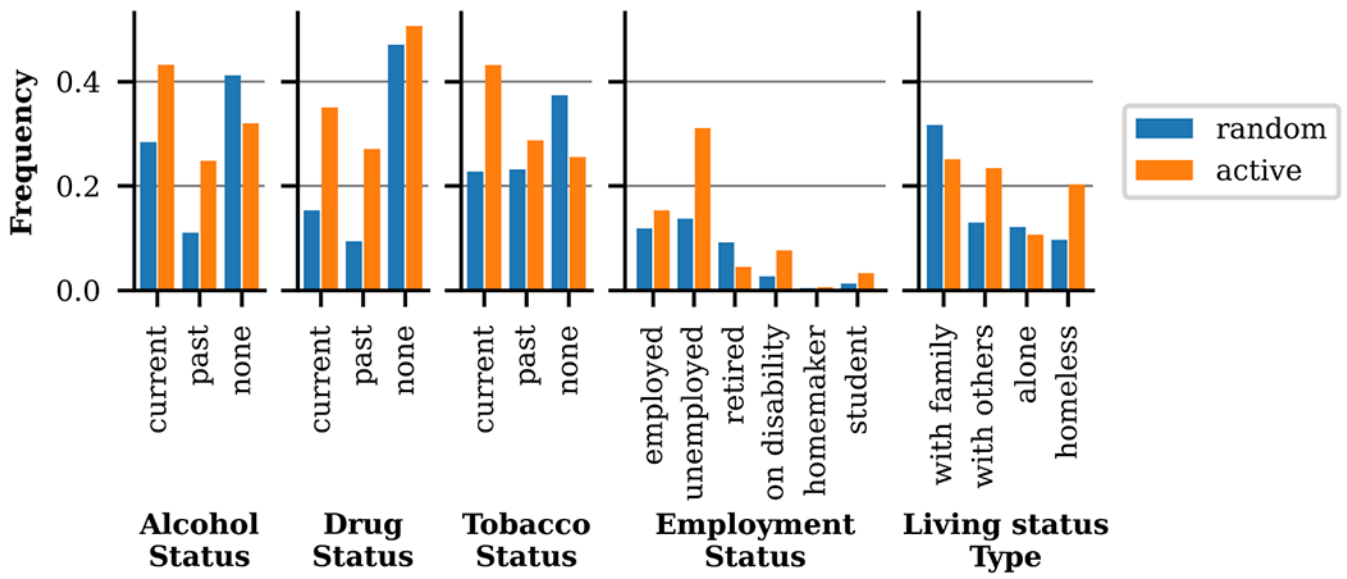
bi-LSTM

BERT

social history section

Figure 6:
Surrogate Classifier used to assess sample uncertainty in active learning
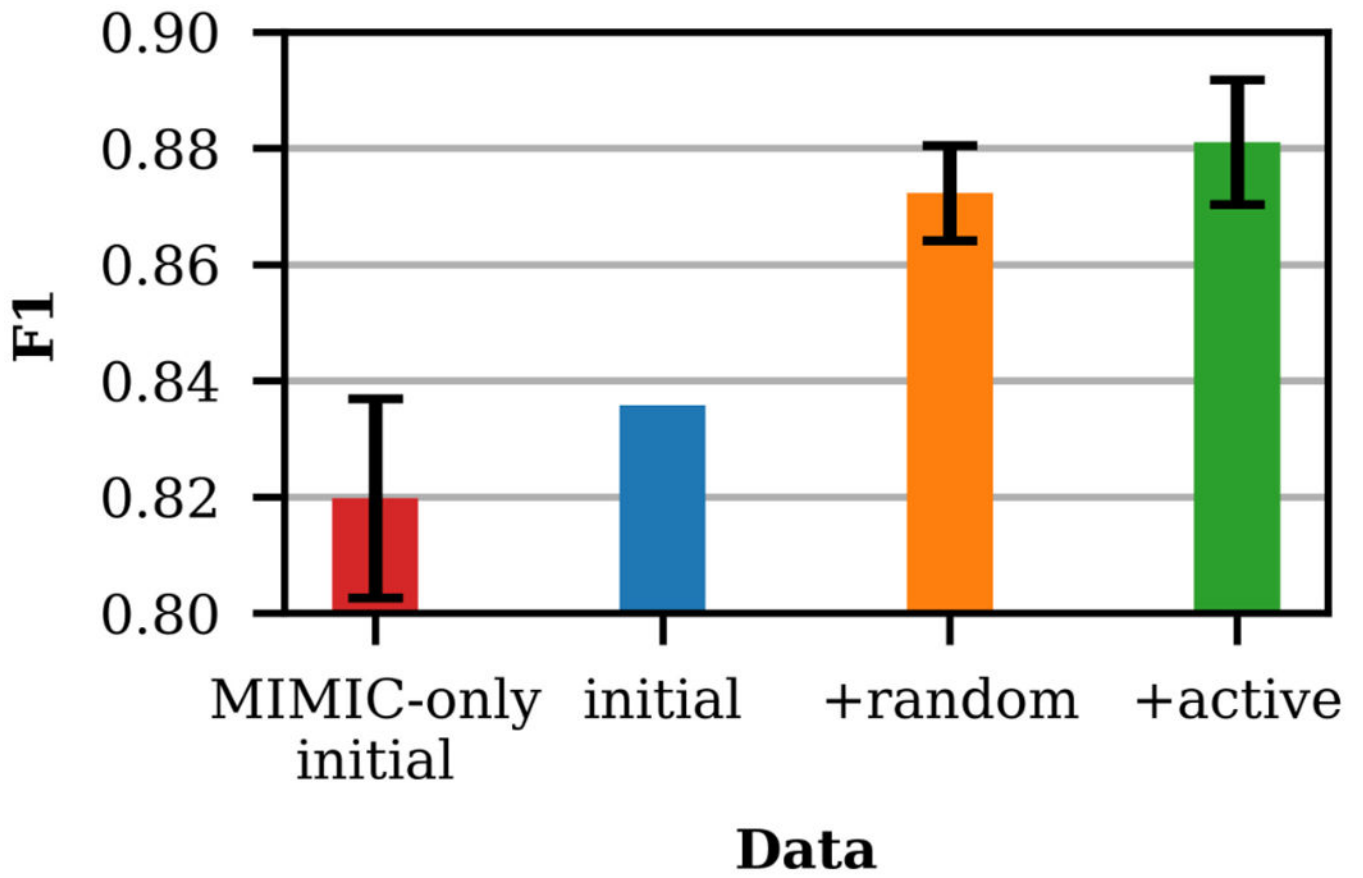
Figure 7:
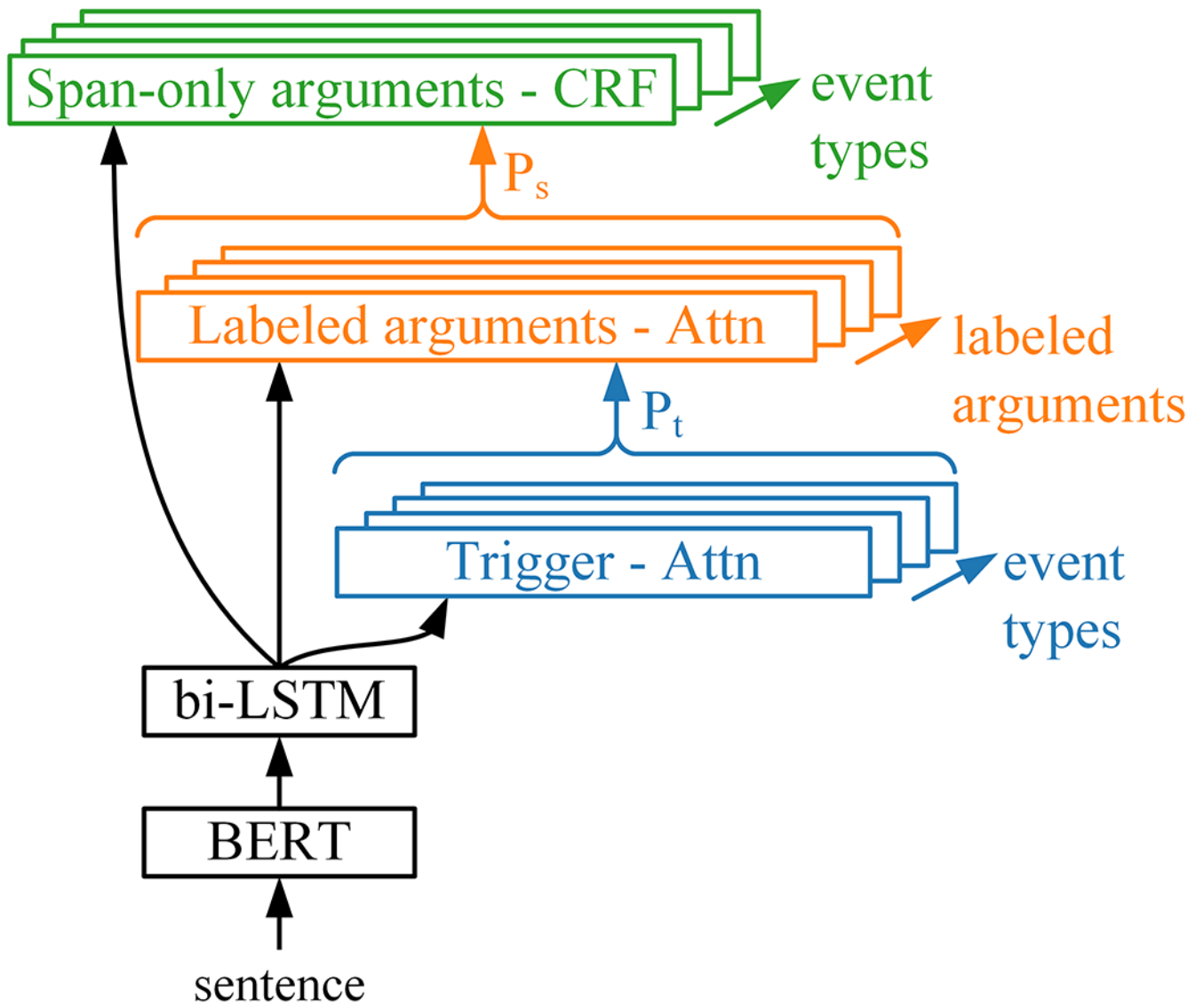Surrogate Classifier performance with random and active samples, evaluated on MIMIC test samples.
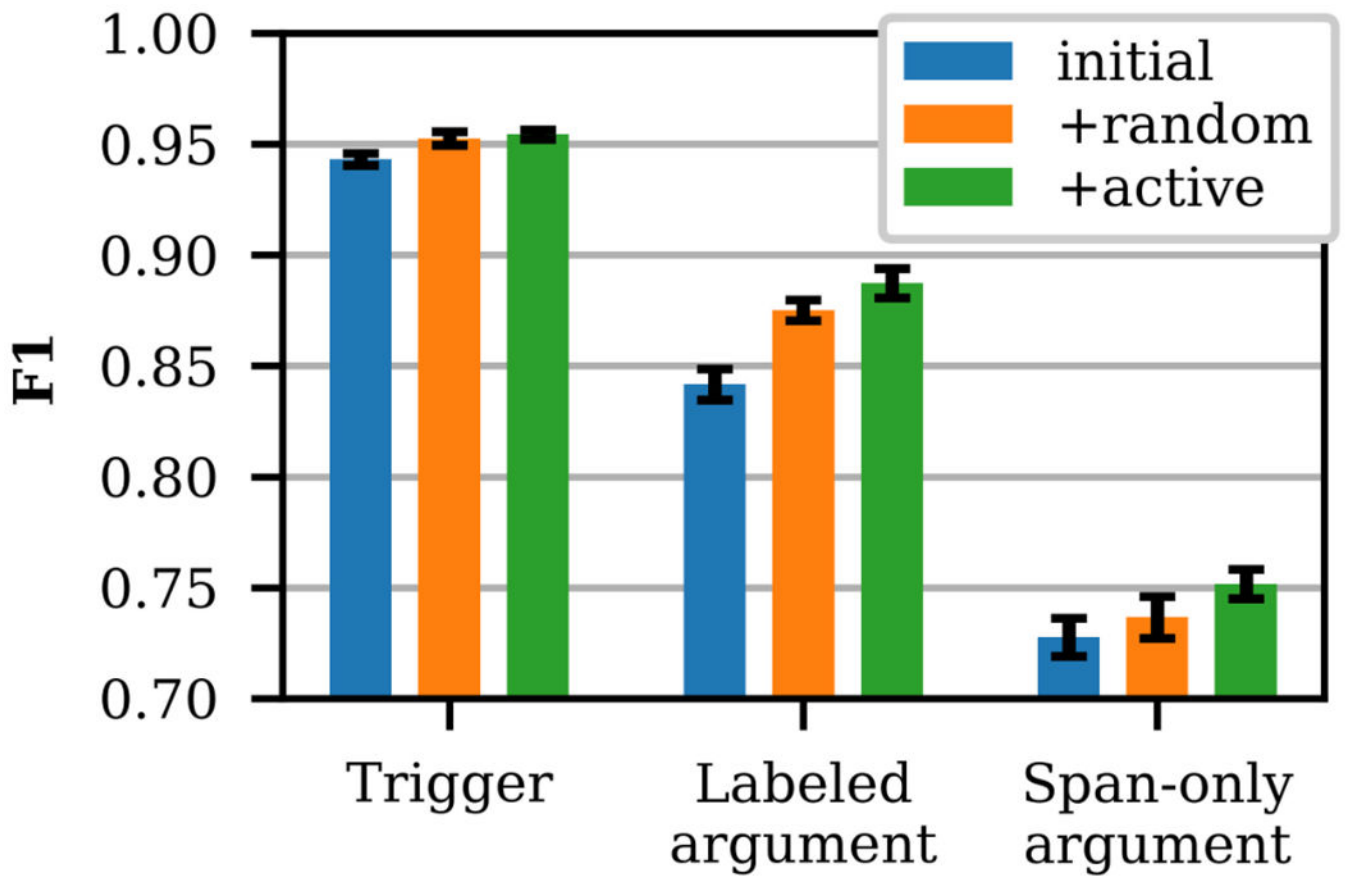
Figure 8:
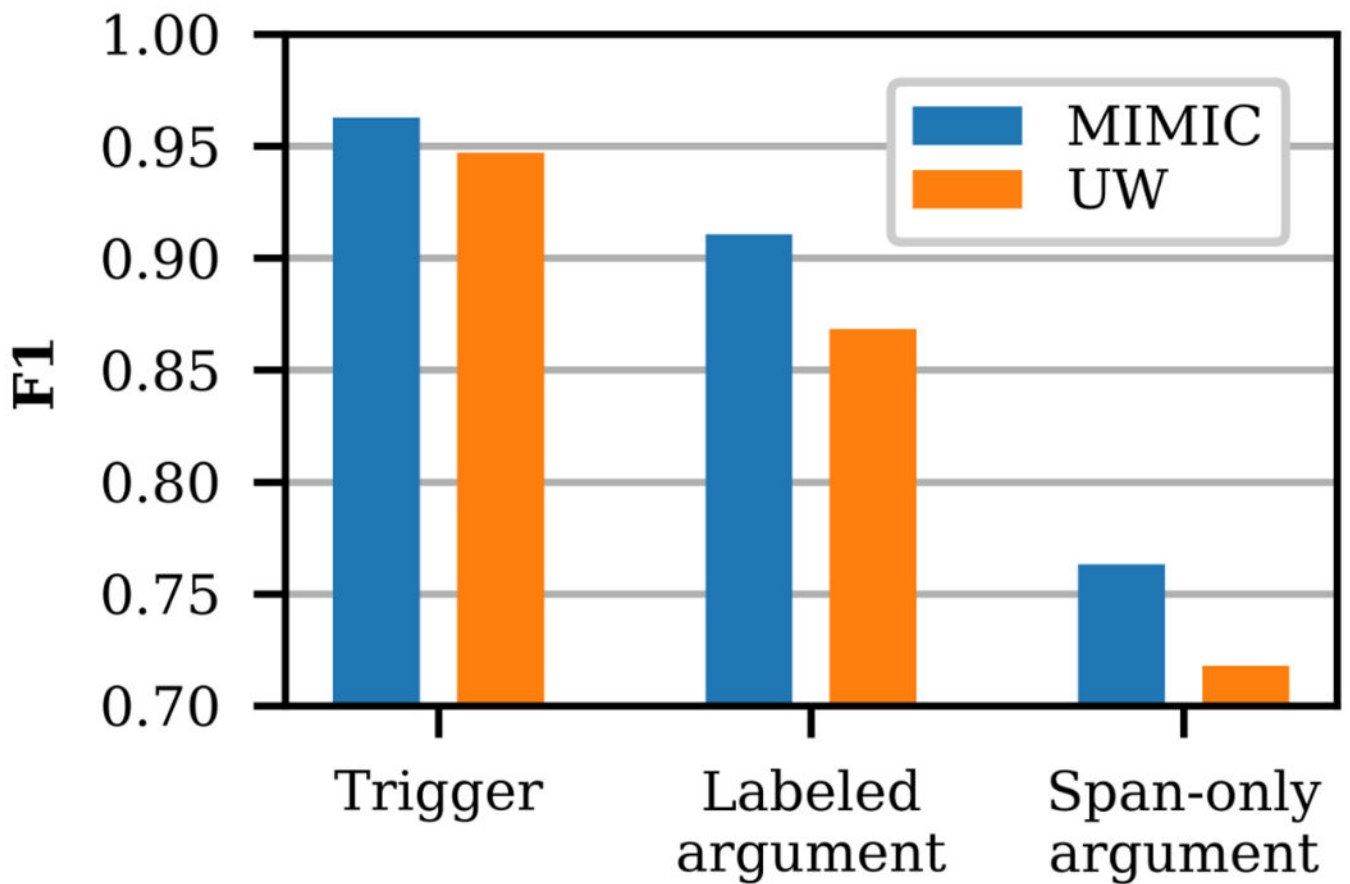Event Extractor performance with random and active samples, evaluated on MIMIC test samples.

Figure 9:
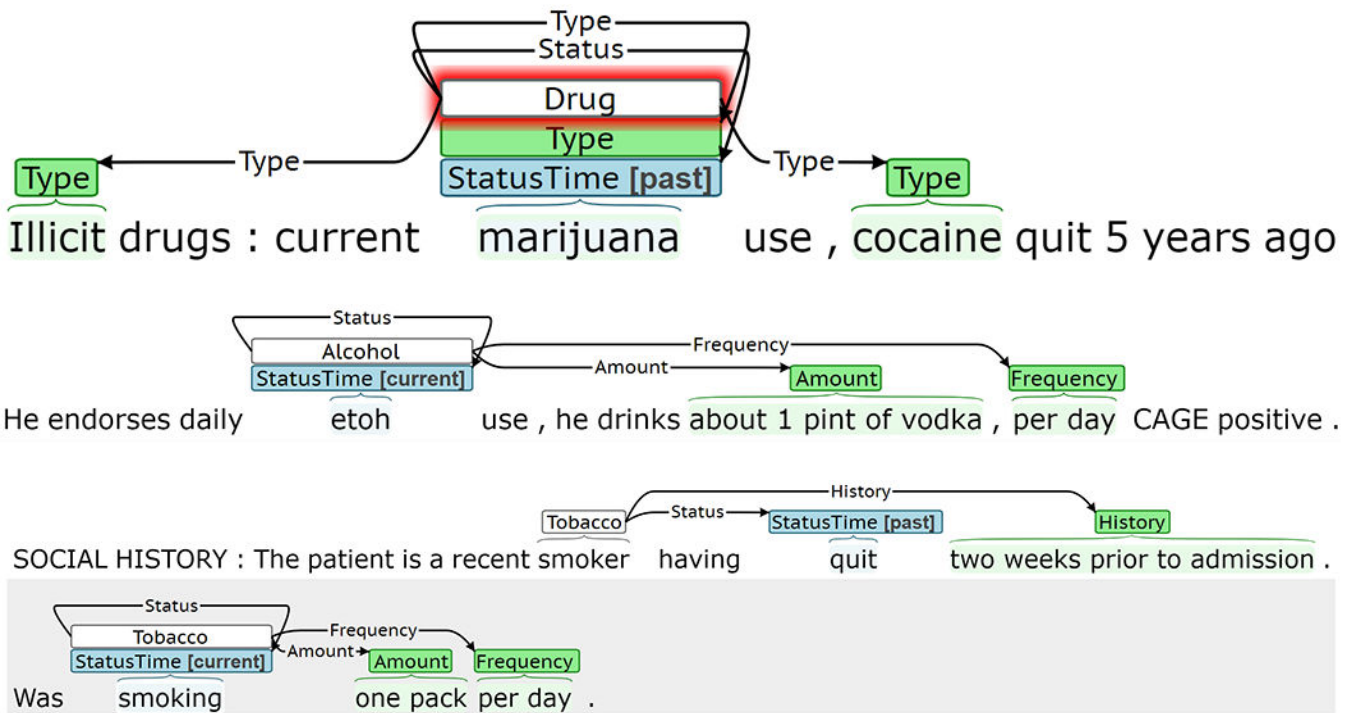Label frequency per social history section, comparing random and active sampling

Figure 10:
Event Extractor model

Figure 11:
Event Extractor average trigger and argument performance, comparing the MIMIC and UW Dataset test

Figure 12:
Error analysis examples

Table 1:

Annotation guideline summary for the most frequent event types.

Event type, e	Argument type, a	Argument subtypes, y_l	Span examples
Substance use (Alcohol, Drug, & Tobacco)	Status[*]	{none, current, past}	"denies," "smokes"
	Duration	–	"for the past 8 years"
	History	–	"seven years ago"
	Type	–	"beer," "cocaine"
	Amount	–	"2 packs," "3 drinks"
	Frequency	–	"daily," "monthly"
Employment	Status[*]	{employed, unemployed, retired, on disability, student, homemaker}	"works," "unemployed"
	Duration	–	"for five years"
	History	–	"15 years ago"
	Type	–	"nurse," "office work"
Living status	Status[*]	{current, past, future}	"lives," "lived"
	Type[*]	{alone, with family, with others, homeless}	"with husband," "alone"
	Duration	–	"for the past 6 months"
	History	–	"until a month ago"

[*] indicates the argument is required.

Table 2:

Corpus composition by source

Source	Train	Dev	Test
MIMIC	1,316	188	376
UW Dataset	1,820	260	520
TOTAL	3,136	448	896

Table 3:

Event Extractor trigger and argument role performance trained on the entire SHAC train set, evaluated on the MIMIC and UW Dataset test sets.

Field	Event type	Argument	MIMIC				UW			
			#	P	R	F1	#	P	R	F1
Trigger	Alcohol	–	314	0.99	0.96	0.97	404	0.97	0.99	0.98
	Drug	–	194	0.96	0.95	0.96	481	0.97	0.92	0.94
	Tobacco	–	324	0.98	0.95	0.97	432	0.97	0.97	0.97
	Employment	–	169	0.93	0.96	0.94	148	0.86	0.91	0.89
	Living status	–	244	0.96	0.97	0.97	343	0.93	0.88	0.90
Labeled argument	Alcohol	Status	314	0.92	0.89	0.90	404	0.92	0.94	0.93
	Drug	Status	194	0.91	0.89	0.90	481	0.85	0.80	0.82
	Tobacco	Status	324	0.91	0.89	0.90	432	0.91	0.90	0.90
	Employment	Status	169	0.84	0.88	0.86	148	0.79	0.83	0.81
	Living status	Status	244	0.96	0.95	0.96	343	0.92	0.86	0.89
		Type	244	0.93	0.93	0.93	343	0.85	0.78	0.81
Span-only argument	Alcohol	Amount, Duration, Frequency, History, Type	396	0.70	0.74	0.72	420	0.67	0.80	0.73
	Drug		219	0.67	0.75	0.71	583	0.62	0.63	0.62
	Tobacco		799	0.81	0.83	0.82	880	0.78	0.81	0.79
	Employment	Duration, History, Type	441	0.80	0.74	0.77	261	0.77	0.77	0.77
	Living status	Duration, History	21	0.21	0.57	0.31	57	0.19	0.26	0.22