# Research and Applications

# The 2022 n2c2/UW shared task on extracting social determinants of health

**Kevin Lybarger** [ID]**[1], Meliha Yetisgen[2], and Özlem Uzuner** [ID]**[1]**

[1]Department of Information Sciences and Technology, George Mason University, Fairfax, Virginia, USA [2]Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, Washington, USA

Corresponding Author: Kevin Lybarger, PhD, Department of Information Sciences and Technology, George Mason University, 4400 University Dr. MSN 1G8, Fairfax, VA 22030, USA; klybarge@gmu.edu

## ABSTRACT

**Objective:** The n2c2/UW SDOH Challenge explores the extraction of social determinant of health (SDOH) information from clinical notes. The objectives include the advancement of natural language processing (NLP) information extraction techniques for SDOH and clinical information more broadly. This article presents the shared task, data, participating teams, performance results, and considerations for future work.

**Materials and Methods:** The task used the Social History Annotated Corpus (SHAC), which consists of clinical text with detailed event-based annotations for SDOH events, such as alcohol, drug, tobacco, employment, and living situation. Each SDOH event is characterized through attributes related to status, extent, and temporality. The task includes 3 subtasks related to information extraction (Subtask A), generalizability (Subtask B), and learning transfer (Subtask C). In addressing this task, participants utilized a range of techniques, including rules, knowledge bases, n-grams, word embeddings, and pretrained language models (LM).

**Results:** A total of 15 teams participated, and the top teams utilized pretrained deep learning LM. The top team across all subtasks used a sequence-to-sequence approach achieving 0.901 F1 for Subtask A, 0.774 F1 Subtask B, and 0.889 F1 for Subtask C.

**Conclusions:** Similar to many NLP tasks and domains, pretrained LM yielded the best performance, including generalizability and learning transfer. An error analysis indicates extraction performance varies by SDOH, with lower performance achieved for conditions, like substance use and homelessness, which increase health risks (risk factors) and higher performance achieved for conditions, like substance abstinence and living with family, which reduce health risks (protective factors).

**GRAPHICAL ABSTRACT**
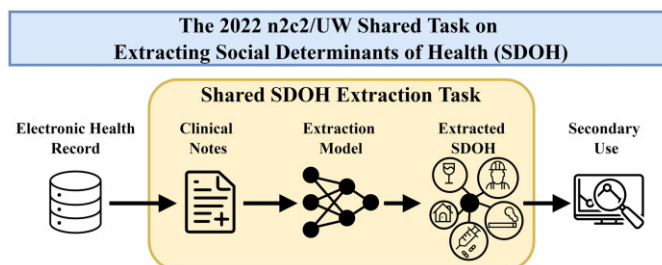
## BACKGROUND AND SIGNIFICANCE

Social determinants of health (SDOH) are the conditions in which people live that affect quality-of-life and health, including social, economic, and behavioral factors.[1] SDOH include *protective factors*, like social support, which reduce health risks and *risk factors*, like housing instability, which increase health risks.[2] SDOH are increasingly recognized for their impact on health outcomes and may contribute to decreased life expectancy.[3–6] As examples, substance abuse,[7–9] living alone,[10,11] housing instability,[12] and unemployment[13,14] are health risk factors. Knowledge of SDOH is important to clinical decision-making, improving health outcomes, and advancing health equity.[6,15]

The electronic health record captures SDOH information through structured data and unstructured clinical notes. However, clinical notes contain more detailed and nuanced descriptions of many SDOH than are available through structured sources. Utilizing textual information from clinical notes in large-scale studies, clinical decision-support systems, and other secondary use applications, requires automatic extraction of key characteristics using natural language processing (NLP) information extraction techniques. Automatically extracted structured representations of SDOH can augment available structured SDOH data to produce more comprehensive patient representations.[16–19] Developing high-performing information extraction models requires annotated data for supervised learning, and extraction performance is influenced by corpus size, heterogeneity, and annotation uniformity.

## OBJECTIVE

This article summarizes the National NLP Clinical Challenges (n2c2) extraction task, *Track 2: Extracting Social Determinants of Health* (n2c2/UW SDOH Challenge). The n2c2/UW SDOH Challenge utilized a novel corpus of annotated clinical text to evaluate the performance of a wide range of SDOH information extraction approaches from rules to deep learning language models (LM). The findings provide insight regarding extraction architecture development and learning strategies, generalizability to new domains, and remaining SDOH extraction challenges.

## RELATED WORK

SDOH are increasingly being recognized for their impact on health, and the body SDOH extraction research is also increasing.[20] Clinical corpora have been annotated for a range of SDOH, including substance use, employment, living situation, environmental factors, physical activity, sexual factors, transportation, education, and language.[21–31] Annotated SDOH datasets typically assign labels at the note-level or use a more granular relation-style structure.

Many annotated SDOH datasets have sentence or note-level labels, approaching extraction as text classification.[22–28] The i2b2 NLP Smoking Challenge introduced a corpus of 502 notes with tobacco use status labels.[22] Gehrmann et al.[23] annotated 1610 notes with phenotype labels, including substance abuse and obesity. Feller et al.[24] created a dataset with 3883 notes annotated for sexual health factors, substance use, and housing. Chapman et al.[25] annotated 621 notes for housing instability. Yu et al.[26,27] created a corpus of 500 notes annotated with 15 SDOH concepts (marital status, education,

occupation, etc.). Han et al.[28] annotated 3504 sentences for 13 SDOH including social environment, support networks, other factors.

While most SDOH corpora use coarser note-level labels, some SDOH corpora utilize a more granular annotation scheme where spans and links between span are identified. Wang et al.[29] annotated 691 notes using a relation-based scheme for substance use. Yetisgen et al.[30] annotated 364 notes using an event-based scheme for 13 SDOH (substance use, living situation, etc). Reeves et al.[31] annotated 8 SDOH concepts (living situation, language, etc.) with assertion values (present vs absent) in a corpus of 160 notes.

SDOH information extraction techniques include rules, supervised learning, and unsupervised learning.[20,32] Rules-based approaches include curated lexicons, regular expressions, and term expansion.[19,20,25,31–33] Supervised extraction approaches include discrete input representations, like n-grams, term frequency-inverse document frequency (TF-IDF), part-of-speech tags, and medical concepts. Discrete classification architectures include Support Vector Machines, random forest, logistic regression, maximum entropy, and conditional random fields.[20,28–30] Recent supervised extraction approaches utilize deep learning architectures, such as convolutional neural networks, recurrent neural networks, and transformers.[20,21,23,26–28]

## MATERIALS AND METHODS

### Data

The n2c2/UW SDOH Challenge used SHAC for model training and evaluation.[21] SHAC includes social history sections from clinical notes annotated for SDOH using an event-based scheme.[21] SHAC was annotated using the BRAT, which is available online (https://brat.nlplab.org/).[34] Figure 1 presents an annotation example. Each event includes exactly one trigger (in white) and one or more arguments that characterize the event. There are 2 argument categories: *span-only* (in green) and *labeled* (in blue). The trigger anchors the event and indicates the event type (eg, *Employment*). Span-only arguments include an annotated span and argument type (eg, *Duration*). *Labeled arguments* include an annotated span, argument type (eg, *Status Time*), and argument subtype (eg, *past*). Argument roles connect triggers and arguments and can be interpreted as binary connectors, because there is one valid argument role type per argument type. For example, all *Status Time* arguments connect to triggers through a *Status* argument role. Table 1 summarizes the annotated phenomena.

SHAC includes 4405 social history sections from MIMIC-III and UW. MIMIC-III is a publicly available, deidentified health database for critical care patients at Beth Israel Deaconess Medical Center from 2001 to 2012.[35] SHAC samples were selected from 60K MIMIC-III discharge summaries. The UW dataset includes 83K emergency department, 22K admit, 8K progress, and 5K discharge summary notes from the UW and Harborview Medical Centers generated between 2008 and 2019. We refer to these social history sections as *notes*, even though each social history section is only part of the original note. Table 2 summarizes the SHAC train, development, and test partitions. The development ($\mathcal{D}_{dev}$) and test ($\mathcal{D}_{test}$) partitions were randomly sampled. The train partition ($\mathcal{D}_{train}$) was 71% actively selected, and the UW train partition ($\mathcal{D}_{train}^{uw}$) was 79% actively selected using a novel active learning framework. The active learning framework selected batches for annotation based on diversity and informativeness. To assess diversity, samples were mapped
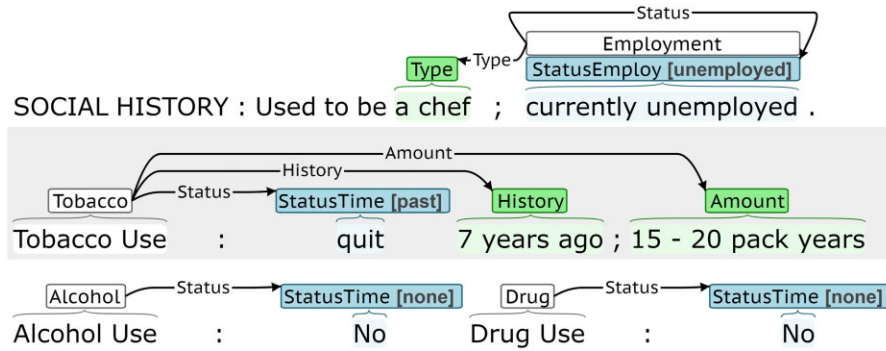
**Figure 1.** BRAT annotation example.

**Table 1.** Annotation guideline summary

| Event type | Argument type | Argument subtypes | Span examples |
|---|---|---|---|
| Alcohol, Drug, & Tobacco | Status Time* | {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 Employ* | {employed, unemployed, retired, on disability, student, homemaker} | "works," "unemployed" |
| | Duration | — | "for five years" |
| | History | — | "15 years ago" |
| | Type | — | "nurse," "office work" |
| Living Status | Status Time* | {current, past, future} | "lives," "lived" |
| | Type Living* | {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.** SHAC note counts by source

| Source | Train | Dev | Test | Total |
|---|---|---|---|---|
| MIMIC-III | 1316 ($\mathcal{D}_{train}^{mimic}$) | 188 ($\mathcal{D}_{dev}^{mimic}$) | 373 ($\mathcal{D}_{test}^{mimic}$) | 1877 ($\mathcal{D}^{mimic}$) |
| UW | 1751 ($\mathcal{D}_{train}^{uw}$) | 259 ($\mathcal{D}_{dev}^{uw}$) | 518 ($\mathcal{D}_{test}^{uw}$) | 2528 ($\mathcal{D}^{uw}$) |
| Total | 3067 ($\mathcal{D}_{train}$) | 447 ($\mathcal{D}_{dev}$) | 891 ($\mathcal{D}_{test}$) | 4405 ($\mathcal{D}$) |

into a vector space using TF-IDF weighted averages of pre-trained word embeddings. To assess informativeness, simplified note-level text classification tasks for substance use, employment, and living status were derived from the event annotations and used as a surrogate for event extraction. Sample informativeness was assessed as the entropy of note-level predictions for: *Status Time* for *Alcohol*, *Drug*, and *Tobacco*; *Type Employ* for *Employment*; and *Status Living* for *Living Status*. These labels capture normalized representations of social protective and risk factors. Active learning increased sample diversity and the prevalence of risk factors, like housing instability and polysubstance use. It improved extraction performance, with the largest performance gains associated with important risk factors, such as drug use, homelessness, and unemployment.[21]

To prepare it for release, SHAC ($\mathcal{D}$) was reviewed, and annotations were adjusted to improve uniformity. The UW partition ($D^{uw}$) was deidentified using an in-house deidentification model[36] and through manual review of each note by 3 annotators. Most protected health information (PHI) was replaced with special tokens related to age, contact information, dates, identifiers, locations, names, and professions (eg, "[DATE]"). However, the meaning of some annotations relies on spans identified as PHI. To preserve annotation meaning, some PHI spans were replaced with randomly selected surrogates from curated lists. For example, named homeless shelters (eg, "Union Gospel Mission") are important to the *Type Living* label *homeless*, and named shelters were replaced with random selections from a list of regional shelters. Surrogates were also manually created to reduce specificity. For example, an *Employment Type* span, like "UPS driver," may be replaced with "delivery driver."

### Subtasks
The challenge includes 3 subtasks:

- *Subtask A (Extraction)* focuses on in-domain extraction, where the training and evaluation data are from the same domain. The training data include the MIMIC-III train and development partitions ($\mathcal{D}_{train}^{mimic}$, $\mathcal{D}_{dev}^{mimic}$), and the evaluation data include the MIMIC-III test partition ($\mathcal{D}_{test}^{mimic}$).
- *Subtask B (Generalizability)* explores generalizability to an unseen domain, where the training and evaluation data are from
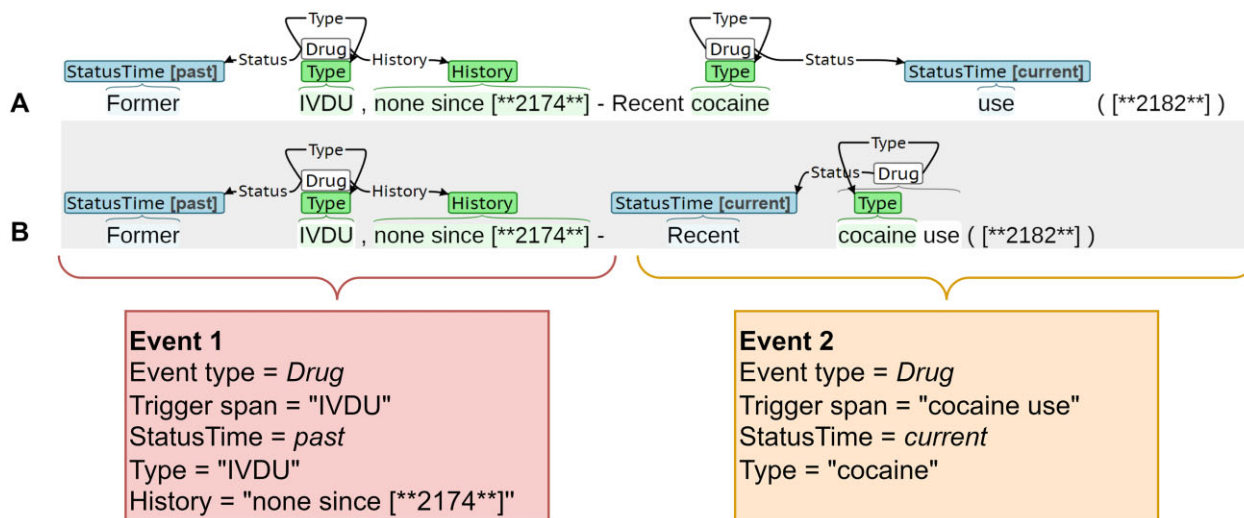
**Figure 2.** Annotation examples, *A* and *B*, describing event extraction as a slot filling task.

different domains. The training data are the same as Subtask A, and the evaluation data are the UW train and development partitions ($\mathcal{D}^{uw}_{train}$, $\mathcal{D}^{uw}_{dev}$).

- *Subtask C (Learning Transfer)* investigates learning transfer, where the training data include in-domain and out-domain data. The training data are the MIMIC-III and UW train and development partitions ($\mathcal{D}^{mimic}_{train}$, $\mathcal{D}^{mimic}_{dev}$, $\mathcal{D}^{uw}_{train}$, $\mathcal{D}^{uw}_{dev}$), and the evaluation data are the UW test partition ($\mathcal{D}^{uw}_{test}$).

### Shared task structure

The n2c2/UW SDOH Challenge was conducted in 2022. To access SHAC, participants were credentialed through PhysioNet for MIMIC-III[35] and submitted data use agreements for the UW partition. For Subtasks A and B, training data were provided on February 14, unlabeled evaluation data were provided on June 6, and system predictions were submitted by participants on June 7. For Subtask C, training data were released on June 8, unlabeled evaluation data were released on June 9, and system predictions were submitted by participants on June 10. Teams were allowed to submit 3 sets of predictions for each subtask, and the highest performing submission for each subtask was used to determine team rankings.

### Scoring

Extraction requires identification of trigger and argument spans, resolution of trigger-argument connections, and prediction of argument subtypes. The evaluation criteria interprets the extraction task as a slot filling task, as this is most relevant to secondary use applications. Figure 2 presents the same sentence with 2 sets of annotations, *A* and *B*, along with the populated slots. Both annotations identify 2 *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 *A* and *B*. However, there are differences in the annotated spans of Event 2, specifically for the trigger ("cocaine" vs "cocaine use") and *Status Time* ("use" vs "Recent"). From a slot perspective, the annotations for Event 2 could be considered equivalent. The evaluation uses relaxed criteria for triggers and labeled arguments that reflect the clinical meaning of the extraction. The criteria and justification are presented below.

*Trigger:* A trigger is defined by an event type (eg, *Drug*) and multiword span (eg, phrase "cocaine use"). In SHAC, the primary function of

the trigger is to anchor the event and aggregate related arguments. The text of the trigger span typically does not meaningfully contribute to the event meaning. Consequently, trigger equivalence is defined using *any overlap* criteria where 2 triggers are equivalent if: (1) the event types match and (2) the spans overlap by at least one character. For *Event 2* in Figure 2, the trigger in *A* has the event type *Drug* and span "cocaine," and the trigger in *B* has the event type *Drug* and span "cocaine use." These triggers are equivalent under this *any overlap* criteria.

*Arguments:* Events are aligned based on trigger equivalence, and the arguments of aligned events are compared using different criteria for *span-only* and *labeled arguments*.

*Span-only arguments:* A span-only argument is defined by an argument type (eg, *Type*), span (eg, phrase "cocaine"), and connection to a trigger. Span-only argument equivalence uses *exact match* criteria, where 2 span-only arguments are equivalent if: (1) the connected triggers are equivalent, (2) the argument types match, and (3) the spans match exactly.

*Labeled arguments:* A labeled argument is defined by an argument type (eg, *Status Time*), argument subtype (eg, *current*), argument span (eg, phrase "Recent"), and connection to a trigger. The argument subtypes normalize the span text and capture a majority of the span semantics. There was often ambiguity in the labeled argument span annotation. To focus the evaluation on the most salient information (subtype labels) and address the ambiguity in argument span annotation, labeled argument equivalence is defined using a *span agnostic* approach, where 2 labeled arguments are equivalent if: (1) the connected triggers are equivalent, (2) the argument types match, and (3) the argument subtypes match.

Performance is evaluated using precision (P), recall (R), and F1, micro averaged over the event types, argument types, and argument subtypes. Submissions were compared using an overall F1 score calculated by summing the true positives, false-negatives, and false-positives across all annotated phenomena. The scoring routine is available (https://github.com/Lybarger/brat_scoring). Significance testing was performed on the overall F1 scores using a paired bootstrap test with 10 000 repetitions, where samples were selected at the note-level.

### Systems

This section summarizes the methodologies of the participating teams. Table 3 summarizes the best performing system for each

**Table 3.** Summary of top performing system for each team

| Team (in alphabetical order) | Team abbrv. | Language rep. | Architecture | External data | Subtasks |
|---|---|---|---|---|---|
| Children's Hospital of Philadelphia* | CHOP | LM$^c$ | ST+TC | — | A, B, C |
| IBM* | IBM | LM$^c$ | unknown | — | C |
| Kaiser Permanente Southern CA* | KP | LM$^{c,b}$ | ST+TC | — | A, B |
| Medical Univ. of South Carolina* | MUSC | n-grams | KB$^u$, rules | — | A, B, C |
| Microsoft* | MS | LM$^t$ | seq2seq | $\mathcal{U}, \mathcal{L}$ | A, B, C |
| Philips Research North America* | PR | LM$^t$ | seq2seq | — | A, B, C |
| University Medical Center Utrecht* | UMCU | LM$^i$, WE | ST+TC | $\mathcal{U}$ | A, B, C |
| University of Florida* | UFL | LM$^i$ | ST+TC | $\mathcal{U}$ | A, B, C |
| University of Massachusetts | UMass | unknown | unknown | — | A, C |
| University of Michigan* | UM | n-grams | ST+TC, rules | — | A, B |
| University of New South Wales | UNSW | n-grams, WE | ST+TC, rules | — | A, B |
| University of Pittsburgh | Pitt | n-grams | rules | — | A |
| University of Texas at San Antonio* | UTSA | LM$^r$, WE | ST+TC | — | A, B, C |
| University of Utah* | UU | LM$^c$ | ST+TC | — | A, C |
| Verily Life Sciences | Verily | LM$^c$ | ST+TC | — | A |

*Indicates the team submitted an abstract.
$^c$Indicates ClinicalBERT.[37]
$^b$Indicates BioBERT.[38]
$^t$Indicates T5.[39]
$^i$Indicates an in-house BERT variant pretrained on institutional data.
$^u$Indicates the Unified Medical Language System (UMLS).[40]
$^r$Indicates RoBERTa.[41]

team, including language representation, architecture, and external data. Language representation includes: (1) *n-grams*—discrete word representations, (2) *pretrained word embeddings (WE)*—vector representation of words, like word2vec,[42] and (3) *pretrained language models (LM)*—deep learning LM, like BERT[43] and T5.[39] Architecture includes: (1) *rules*—hand-crafted rules, (2) *knowledge based (KB)*—dictionaries and ontologies, (3) *sequence tagging + text classification (ST+TC)*—combination of sequence tagging and text classification layers, and (4) *sequence-to-sequence (seq2seq)*—text generation models, like T5, transform unstructured input text into a structured representation. External data includes: (1) *unlabeled ($\mathcal{U}$)*—unsupervised learning with unlabeled data and (2) *labeled ($\mathcal{L}$)*—supervised learning with labeled data other than SHAC. Many teams used pretrained WE and LM; however, $\mathcal{U}$ is only assigned for additional pretraining, beyond publicly available models. For example, ClinicalBERT[37] would not be assigned $\mathcal{U}$; however, additional pretraining of ClinicalBERT would be assigned $\mathcal{U}$.

Below is a summary of the teams that achieved first and second place in each subtask. The top teams were determined based on the performance in Table 4 and are presented below alphabetically.

- *Children's Hospital of Philadelphia (CHOP)* designed a BERT-based pipeline consisting of trigger identification and argument resolution. Triggers were extracted as a sequence tagging task using BERT with a linear prediction layer at the output hidden states. For argument extraction, a unique input representation was created for each identified trigger, where the segment IDs were 1 for the target trigger and 0 elsewhere. Arguments were identified using multiple linear layers applied to the BERT hidden states, to allow multi-label token predictions. Encoding the target trigger using the segment IDs allowed BERT to focus the argument prediction on a single event. Argument subtype labels were resolved by creating separate tags for each subtype (eg, "LivingStatus=alone").

- *Kaiser Permanente Southern CA (KP)* assumed there is at most one event per event type per sentence, similar to the original SHAC paper. Trigger and labeled argument prediction was performed as a binary text classification task (eg, no alcohol vs has alcohol; no current alcohol use vs current alcohol use) using BERT with a linear layer applied to the pooled state. Trigger and span-only arguments spans were extracted using BERT with a linear layer applied to the hidden states, with separate models for each. Within a sentence, the extracted trigger and arguments for a given event type were assumed to be part of the same event. The trigger-argument linking was implicit in the assumption that there is at most one event per SDOH per sentence.

- *Microsoft (MS)* developed a seq2seq approach with T5-large as the pretrained encoder-decoder. T5 was further pretrained on MIMIC-III notes. In fine-tuning, the input was note text, and the output was a structured text representation of SDOH. Additional negative samples were incorporated using MIMIC-III text that does not include SDOH, and additional positive samples were created from the annotated data by excerpting shorter samples ($\approx$9 new samples per note). In addition to the challenge data, MS used an in-house dataset with SDOH annotations; however, an MS ablation study indicated this in-house data had a marginal impact on performance. A constraint solver post-processed the structured text output to generate the text offsets for the final span predictions.

- *University of Florida (UFL)* designed a multi-step approach: (1) span extraction, (2) relation prediction, and (3) argument subtype classification. Trigger and argument types were divided into 5 groups, where overlapping spans are minimized within each group. The trigger and argument spans for each group were then extracted using a separate BERT model with a linear layer at the hidden states. Relation prediction was a binary text classification task using BERT, where the input included special tokens demarking target trigger and argument

**Table 4**. Performance of top performing teams

| Team | P | R | F1 | Significance | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MS | UFL | KP | CHOP | PR | UTSA | UMCU | Verily |
| MS | 0.909 | 0.893 | 0.901 | — | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| UFL | 0.878 | 0.908 | 0.893 | — | — | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ |
| KP | 0.884 | 0.884 | 0.884 | — | — | — | ✗ | ✓ | ✓ | ✓ | ✓ |
| CHOP | 0.870 | 0.889 | 0.879 | — | — | — | — | ✓ | ✓ | ✓ | ✓ |
| PR | 0.862 | 0.842 | 0.852 | — | — | — | — | — | ✗ | ✓ | ✓ |
| UTSA | 0.847 | 0.827 | 0.837 | — | — | — | — | — | — | ✓ | ✓ |
| UMCU | 0.864 | 0.653 | 0.744 | — | — | — | — | — | — | — | ✓ |
| Verily | 0.797 | 0.529 | 0.636 | — | — | — | — | — | — | — | — |

(a) Subtask A

| Team | P | R | F1 | Significance | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MS | KP | CHOP | UFL | UTSA | PR | — | — |
| MS | 0.811 | 0.740 | 0.774 | — | ✗ | ✓ | ✓ | ✓ | ✓ | — | — |
| KP | 0.790 | 0.748 | 0.768 | — | — | ✗ | ✓ | ✓ | ✓ | — | — |
| CHOP | 0.761 | 0.770 | 0.766 | — | — | — | ✓ | ✓ | ✓ | — | — |
| UFL | 0.761 | 0.713 | 0.736 | — | — | — | — | ✓ | ✓ | — | — |
| UTSA | 0.737 | 0.665 | 0.699 | — | — | — | — | — | ✗ | — | — |
| PR | 0.739 | 0.655 | 0.695 | — | — | — | — | — | — | — | — |

(b) Subtask B

| Team | P | R | F1 | Significance | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MS | CHOP | UTSA | RP | UFL | UMCU | IBM | — |
| MS | 0.891 | 0.887 | 0.889 | — | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | — |
| CHOP | 0.874 | 0.888 | 0.881 | — | — | ✓ | ✓ | ✓ | ✓ | ✓ | — |
| UTSA | 0.880 | 0.841 | 0.860 | — | — | — | ✓ | ✓ | ✓ | ✓ | — |
| PR | 0.885 | 0.780 | 0.829 | — | — | — | — | ✓ | ✓ | ✓ | — |
| UFL | 0.923 | 0.683 | 0.785 | — | — | — | — | — | ✓ | ✓ | — |
| UMCU | 0.886 | 0.553 | 0.681 | — | — | — | — | — | — | ✓ | — |
| IBM | 0.538 | 0.788 | 0.640 | — | — | — | — | — | — | — | — |

(c) Subtask C

Significance evaluated for F1 score at $P < .05$. ✓ indicates the systems *were* statistically different, and ✗ indicates the system *were not* statistically different.

spans. Argument subtype classification was similar to relation prediction, except the targets were the subtype labels. UFL used an internal BERT variant, GatorTron, which was trained on in-house and public data.

## RESULTS

Table 4 presents the overall performance for Subtasks A, B, and C, including significance testing. Table 4 includes the results from teams with F1 greater than the mean across all teams. Since the original SHAC publication, the UW SHAC partition was deidentified, and the entirety of SHAC was reviewed to improve annotation consistency. Additionally, the scoring routine used in the challenge differs from the original SHAC publication. Consequently, the results in Table 4 should be considered the current state-of-the-art for SHAC. MS achieved the highest overall F1 across all 3 subtasks; however, the MS performance was not statistically different from the second team in each subtask: UFL for Subtask A, KP for Subtask B, and CHOP for Subtask C. Performance was substantively lower for Subtask B than Subtasks A and C. No in-domain training data was used in Subtask B.

Additionally, the Subtask A and Subtask C evaluation data were randomly selected; however, the evaluation data for Subtask B included actively selected samples that intentionally biased the data toward risk factors.

## Error analysis

We explored the performance by event type, argument subtype, and event frequency per note. The error analysis presents the micro-averaged performance of the top 3 teams in each subtask (Subtask A—MS, UFL, and KP; Subtask B—MS, KP, and CHOP; and Subtask C—MS, CHOP, and UTSA). Subtask B performance is presented separately for $\mathcal{D}_{train}^{uw}$ and $\mathcal{D}_{dev}^{uw}$, because of the differing sampling strategies (active vs random).

Figure 3 presents the extraction performance by subtask and includes the total number of gold events ($n$) and average number of gold events per note ($+$). Subtask A focused on in-domain extraction, and Subtask B focused on generalizability. The performance drops from Subtask A to Subtask B ($\mathcal{D}_{dev}^{uw}$) at $-0.04$ ΔF1 for triggers, $-0.07$ ΔF1 for labeled arguments, and $-0.07$ ΔF1 span-only arguments (*All* category in figure). The largest performance drop is associated with *Living Status* at $-0.14$ ΔF1 for triggers and $-0.17$ ΔF1 for labeled arguments. Subtask C explored learning transfer, where both in-
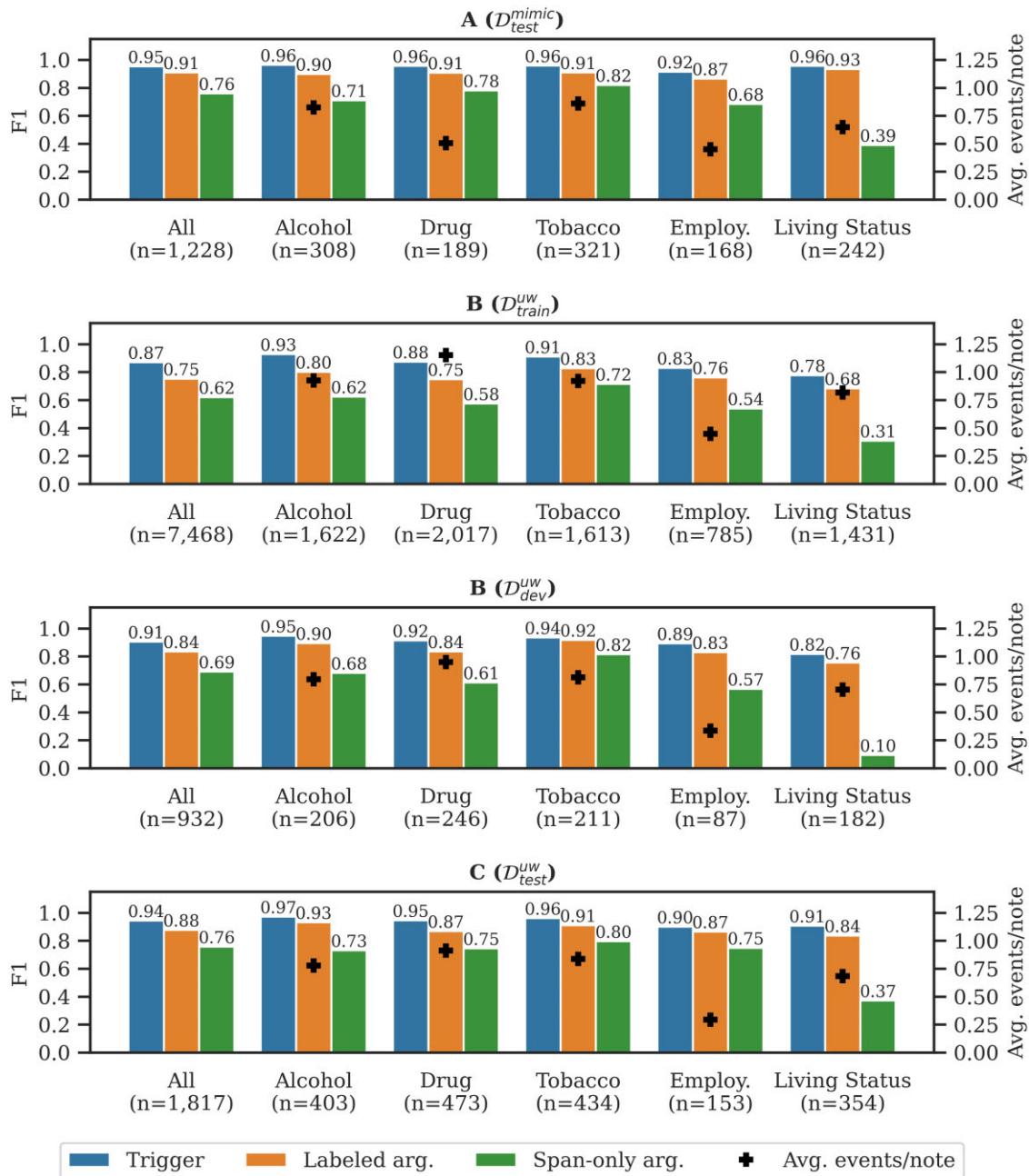
**Figure 3.** Performance for top performing teams in Subtasks A, B, and C. The left-hand *y*-axis is the micro-averaged F1 for triggers, labeled arguments, and span-only arguments (vertical bars). The right-hand *y*-axis is the average number of gold events per note (+).

domain and out-domain data are available. Comparing Subtask B ($\mathcal{D}_{dev}^{uw}$) and Subtask C, incorporating in-domain data increased performance, such that performance for Subtasks A and C are similar. The performance difference between $\mathcal{D}_{train}^{uw}$ and $\mathcal{D}_{dev}^{uw}$ in Subtask B demonstrates the actively selected samples are more challenging extraction targets.

SHAC annotations capture normalized SDOH factors through the arguments: *Status Time* for substance use, *Status Employ* for *Employment*, and *Type Living* for *Living Status*. Figure 4 presents the performance for these factors and the average number of gold events per note (+). The *student* and *homemaker* labels for *Status Employ* are omitted, due to low frequency. For substance use, the *none* label has the highest performance where descriptions tend to

be relatively concise and have low linguistic variability (eg, "Tobacco use: none" or "Denies EtOH"). Performance is lower for *current* and *past* subtype labels, which are associated with more linguistic diversity and have higher label confusability. Performance is lower for *Drug* than *Alcohol* and *Tobacco* because of the more heterogeneous descriptions of drug use and types (eg, "cocaine remote," "smokes MJ," and "injects heroin"). Regarding *Employment*, the *retired* label has the highest performance, due to the consistent use of the keyword "retired." The *employed* and *unemployed* labels have similar performance that is incrementally less than *retired*. The *on disability* label has the lowest performance, which is partially attributable to annotation inconsistency and classification
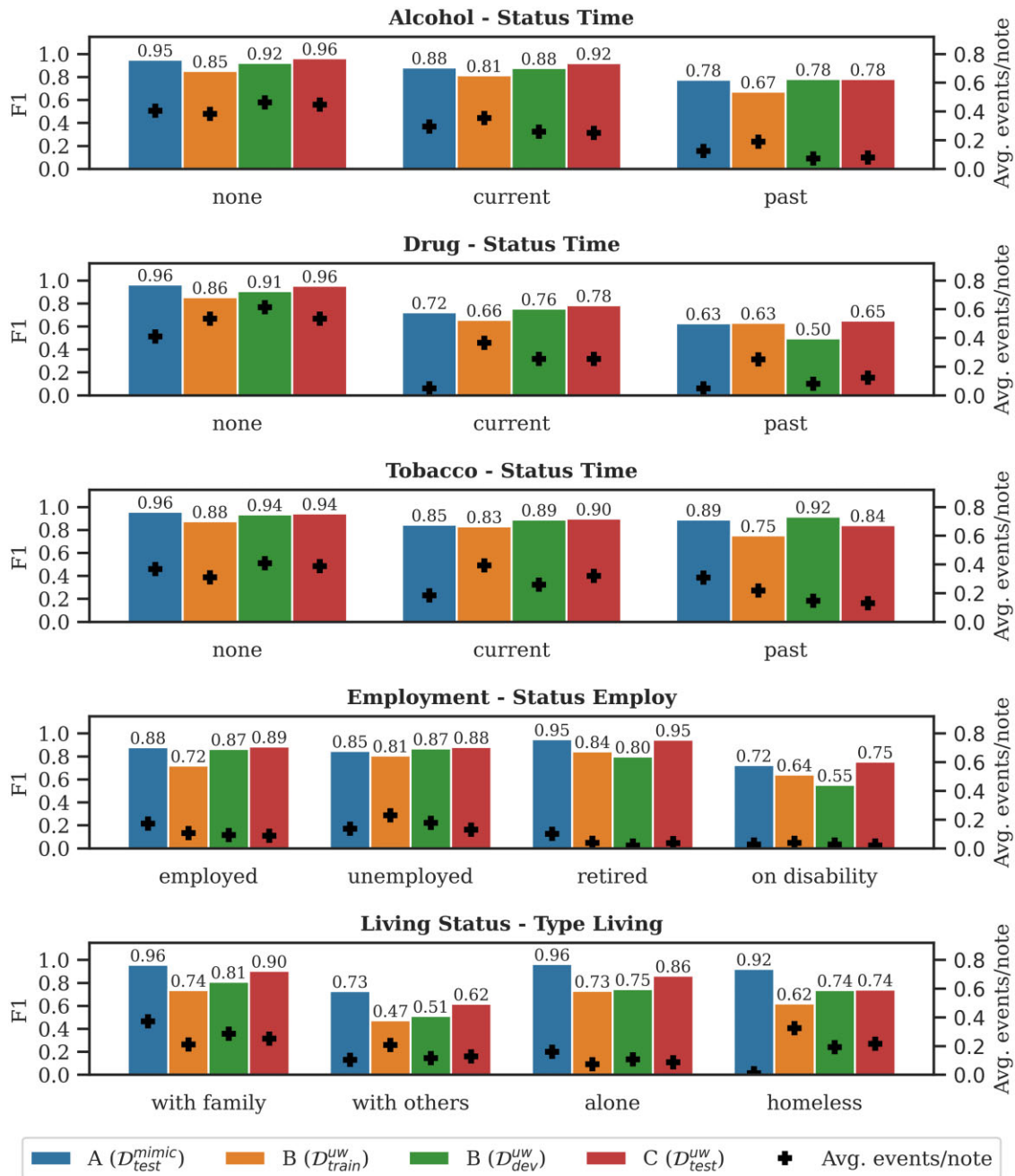
**Figure 4.** Performance breakdown by argument subtype label for Subtasks A, B, and C. The left-hand *y*-axis is the micro-averaged F1 for the argument subtype labels (vertical bars). The right-hand *y*-axis is the average number of gold events per note (+).

confusability associated with disambiguating the presence of disability and receiving disability benefits (eg, "She does not work because of her disability" vs "On SSI disability").

Figure 5 presents the performance for Subtask B ($\mathcal{D}_{train}^{uw}$) by the number of gold events per note for a given event type (referred to as *event density*): 1 (1), 2 (2), and 3 or more (3+) events per note. Figure 5 also includes the total number of gold events (+). Across event types, performance is lower for multi-event notes (density > 1) than single-event notes (density = 1). Multi-event notes tend to have more detailed and nuanced options of the SDOH. All aspects of the extraction task, including span identification, span linking, and

argument subtype resolution, are more challenging in multi-event notes. The Supplementary Appendix includes similar figures for Subtask A, Subtask B ($\mathcal{D}_{dev}^{uw}$), and Subtask C. Subtask B ($\mathcal{D}_{train}^{uw}$) is presented here because it has the highest event density.

Formatting, linguistic, and content differences between the MIMIC and UW data resulted in lower performance in Subtask B (train on MIMIC, evaluate on UW), relative to Subtask A (train and evaluate on MIMIC). The inclusion of UW data improved performance in Subtask C (train on MIMIC and UW, evaluate on UW), largely mitigating the performance degradation. We manually reviewed the predictions from the top 3 teams in each subtask (same
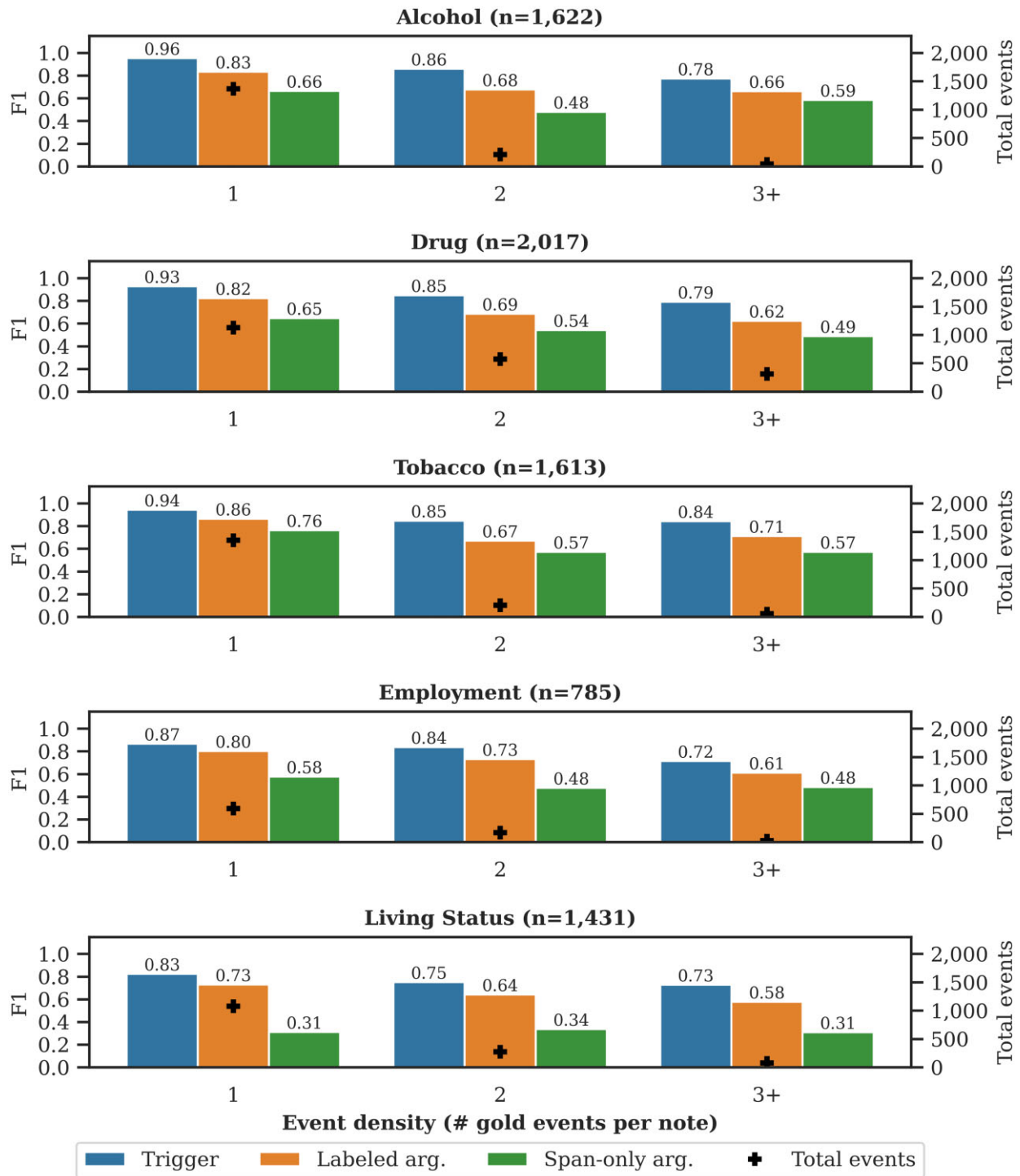
**Figure 5.** Performance breakdown by event density for Subtask B ($\mathcal{D}_{train}^{uw}$). The left-hand *y*-axis is the micro-averaged F1 for triggers, labeled arguments, and span-only arguments (vertical bars). The right-hand *y*-axis is the total number of gold events (+).

data as Figures 3–5) focusing on the notes with lower performance, to explore domain mismatch errors and challenging classification targets.

The UW data include templated substance use, which differs in format from MIMIC. In Subtask B, unpopulated templates, like "Tobacco use: _ Alcohol: _ Drug Use: _," resulted in false positives. There are

differences in the substance use vocabulary that resulted in false negatives in Subtask B. For example, the UW data include descriptions of "toxic habits" and uses "nicotine" as a synonym for tobacco.

The UW data contain living situation descriptions that lack sufficient information to resolve the *Type Living* label, a necessary condition for a *Living Status* event. In Subtask B, there were frequent
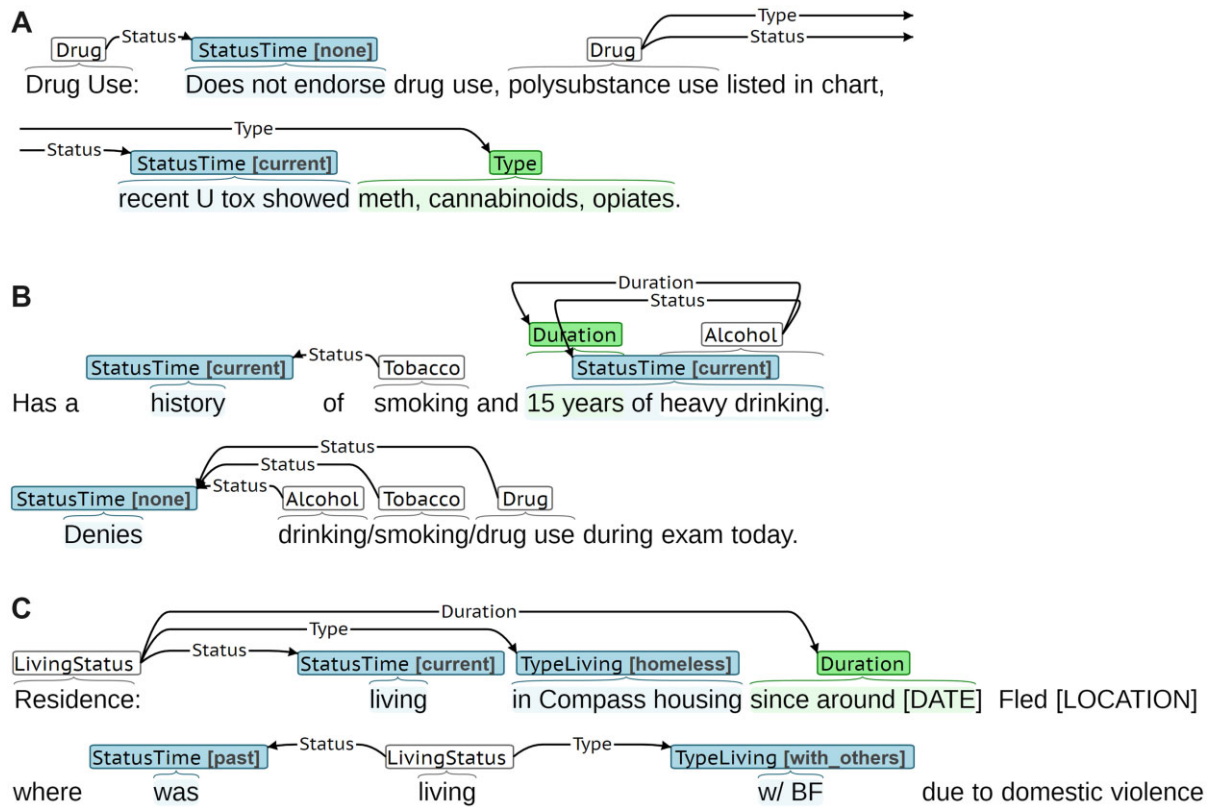
**Figure 6.** Error analysis examples from Subtask C ($\mathcal{D}_{test}^{uw}$). (A) Multiple gold *Drug* events in a sentence with conflicting *Status Time* labels. (B) Multiple gold *Tobacco* and *Alcohol* events in note with conflicting *Status Time* labels. (C) Multiple gold *Living Status* events in a sentence with conflicting *Status Time* and *Type Living* labels.

false positives associated with living situation descriptions, like "Lives in a private residence," which are not present in MIMIC. The MIMIC data include homogeneous descriptions of housing instability (eg, "homeless"), whereas the UW data include more heterogeneous descriptions (eg, "undomiciled" or "living on street") and references to regional named shelters (eg, "Union Gospel Mission"). In Subtask B, this difference resulted in false negatives and misclassified *Type Living*. The UW data also include residences that are uncommon in the MIMIC data, like skilled nursing facilities or adult family homes, contributing to false negatives in Subtask B.

*Employment* performance in Subtask B was negatively impacted by references to the employment of family members, confusability between having a disability and receiving disability benefits, and linguistically diversity. The linguistic diversity was especially challenging when a common trigger phase, like "works," is omitted, as in "Cleans carpets for a living."

More nuanced SDOH descriptions, where determinants are represented through multiple events, were more challenging classification targets. Figure 6 presents gold examples from Subtask C that were challenging targets. Figure 6A presents a sentence describing drug use, including the denial of use and positive toxicology. Extraction requires creating separate events for the drug denial and positive toxicology report, including correctly identifying the appropriate triggers among several commonly annotated trigger spans ("Drug Use," "drug use," and "polysubstance use"). Figure 6B describes smoking and alcohol use history in the first sentence and refutes substance use in the second sentence. SHAC

annotators annotated events using intra-sentence information, where possible, resulting in the conflicting *Status Time* labels. Extraction requires identifying multiple triggers and resolving the inconsistent *Status Time* labels. Figure 6C presents a *Living Status* example where separate events are needed to capture previous and current living situations and there are multiple commonly annotated trigger spans ("Residence" and "living"). Here, homelessness must be inferred from "Compass housing," a regional shelter. The nuanced descriptions of SDOH in Figure 6 are also more challenging to annotate consistently, contributing to the extraction challenge.

## DISCUSSION

The top-performing submissions utilized pretrained LM. A majority of the submissions, including the top-performers, utilized a combination of sequence tagging and text classification. However, the best performing system across subtasks was Microsoft's seq2seq approach. The success of Microsoft's seq2seq approach may be related to the classifier architecture (seq2seq vs sequence tagging and text classification), LM architecture (T5 vs BERT), data augmentation (additional shorter training samples), LM pretraining, and/or use of additional supervised data. Systems that utilized data-driven neural approaches performed better than systems based on rules, knowledge sources, and n-grams.

The top-performing teams demonstrated good domain generalizability to the UW data in Subtask B ($\mathcal{D}_{dev}^{uw}$) for alcohol, drug, tobacco, and employment, with lower generalizability for living status.

Incorporating in-domain data in Subtask C increased performance, resulting in similar performance in both domains (MIMIC for Subtask A and UW for Subtask C). The error analysis indicates SDOH risk factors tend to be extracted with lower performance than protective factors: (1) performance for current and past substance use is lower than substance abstinence, (2) performance for current and past drug use is lower than for alcohol and tobacco, (3) performance for being on disability is lower than other employment categories, and (4) performance for homelessness is lower than living with family or (stably) alone. The error analysis also indicates extraction performance decreases as the density of SDOH events increases. Qualitatively, notes with more SDOH events tend to express more severe social needs (eg, polysubstance use, long history of tobacco use, and more frequent living situation transitions). The analysis of SDOH factors (Figure 4) and event density (Figure 5) suggests notes describing higher risk SDOH are more challenging extraction targets.

## CONCLUSIONS

SDOH impact individual and public health and contribute to morbidity and mortality. The n2c2/UW SDOH Challenge explores the automatic extraction of substance use, employment, and living status from clinical text using SHAC.[21] The top-performing teams in this extraction challenge used pretrained LM, with most using a combination of sequence tagging and text classification. A novel seq2seq approach achieved the best performance across all subtasks at 0.901 F1 for Subtask A, 0.774 F1 Subtask B, and 0.889 F1 for Subtask C. SDOH risk factors tend to be more challenging extraction targets than protective factors. Future work related to SDOH extraction should consider this potential for bias, especially for patients with higher-risk SDOH.

## FUNDING

## AUTHOR CONTRIBUTIONS

All authors contributed to the organization of the shared task and the n2c2 workshop where the shared task results were presented. KL performed the data analysis and drafted the initial manuscript. All authors contributed to the interpretation of the data, manuscript revisions, and intellectual value to the article.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## ACKNOWLEDGMENTS

## CONFLICT OF INTEREST STATEMENT

None declared.

## DATA AVAILABILITY

The SHAC dataset used in the shared task will be made available through the University of Washington.

## REFERENCES

1. Social Determinants of Health. 2021. https://www.cdc.gov/socialdeterminants/index.htm. Accessed November 1, 2022.
2. Alderwick H, Gottlieb LM. Meanings and misunderstandings: a social determinants of health lexicon for health care systems. *Milbank Q* 2019; 97 (2): 407–19.
3. Friedman NL, Banegas MP. Toward addressing social determinants of health: a health care system strategy. *Perm J* 2018; 22: 18–95.
4. Daniel H, Bornstein SS, Kane GC, *et al.*; Health and Public Policy Committee of the American College of Physicians. Addressing social determinants to improve patient care and promote health equity: an American College of Physicians Position Paper. *Ann Intern Med* 2018; 168 (8): 577–8.
5. Himmelstein DU, Woolhandler S. Determined action needed on social determinants. *Ann Intern Med* 2018; 168 (8): 596–7.
6. Singh GK, Daus GP, Allender M, *et al.* Social determinants of health in the United States: addressing major health inequality trends for the nation, 1935-2016. *Int J MCH AIDS* 2017; 6 (2): 139–64.
7. Centers for Disease Control and Prevention. Annual smoking-attributable mortality, years of potential life lost, and productivity losses–United States, 1997-2001. *MMWR Morb Mortal Wkly Rep* 2005; 54 (25): 625.
8. Degenhardt L, Hall W. Extent of illicit drug use and dependence, and their contribution to the global burden of disease. *Lancet* 2012; 379 (9810): 55–70.
9. World Heal Organization. *Global Status Report on Alcohol and Health 2018*. World Health Organization; 2019. https://www.who.int/publications/i/item/9789241565639. Accessed November 1, 2022.
10. Cacioppo JT, Hawkley LC. Social isolation and health, with an emphasis on underlying mechanisms. *Perspect Biol Med* 2003; 46 (3 Suppl): S39–52.
11. Hawkley LC, Capitanio JP. Perceived social isolation, evolutionary fitness and health outcomes: a lifespan approach. *Philos Trans R Soc Lond B Biol Sci* 2015; 370 (1669): 20140114.
12. Oppenheimer SC, Nurius PS, Green S. Homelessness history impacts on health outcomes and economic and risk behavior intermediaries: new insights from population data. *Fam Soc* 2016; 97 (3): 230–42.
13. Clougherty JE, Souza K, Cullen MR. Work and its role in shaping the social gradient in health. *Ann N Y Acad Sci* 2010; 1186: 102–24.
14. Dooley D, Fielding J, Levi L. Health and unemployment. *Annu Rev Public Health* 1996; 17 (1): 449–65.
15. Blizinsky KD, Bonham VL. Leveraging the learning health care model to improve equity in the age of genomic medicine. *Learn Health Syst* 2018; 2 (1): e10046.
16. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009; 42 (5): 760–72.
17. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012; 13 (6): 395–405.
18. Navathe AS, Zhong F, Lei VJ, *et al.* Hospital readmission and social risk factors identified from physician notes. *Health Serv Res* 2018; 53 (2): 1110–36.
19. Hatef E, Rouhizadeh M, Tia I, *et al.* Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med Inform* 2019; 7 (3): e13802.

20. Patra BG, Sharma MM, Vekaria V, *et al*. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc* 2021; 28 (12): 2716–27.

21. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J Biomed Inform* 2021; 113: 103631.

22. Uzuner Ö, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008; 15 (1): 14–24.

23. Gehrmann S, Dernoncourt F, Li Y, *et al*. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One* 2018; 13 (2): 1–19.

24. Feller DJ, Zucker J, Srikishan B, *et al*. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. *AMIA Annu Symp Proc* 2018; 2018: 422–29.

25. Chapman AB, Jones A, Kelley AT, *et al*. ReHouSED: A novel measurement of Veteran housing stability using natural language processing. *J Biomed Inform* 2021; 122: 103903.

26. Yu Z, Yang X, Dang C, *et al*. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. *AMIA Annu Symp Proc* 2021; 2021: 1225–33.

27. Yu Z, Yang X, Guo Y, *et al*. Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives. *Front Public Health* 2022; 10: 778463.

28. Han S, Zhang RF, Shi L, *et al*. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J Biomed Inform* 2022; 127: 103984.

29. Wang Y, Chen ES, Pakhomov S, *et al*. Automated extraction of substance use information from clinical texts. *AMIA Annu Symp Proc* 2015; 2015: 2121–30.

30. Yetisgen M, Vanderwende L. Automatic identification of substance abuse from social history in clinical text. *Artif Intell Med* 2017; 10259: 171–81.

31. Reeves RM, Christensen L, Brown JR, *et al*. Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. *J Biomed Inform* 2021; 120: 103851.

32. Bompelli A, Wang Y, Wan R, *et al*. Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: a scoping review. *Health Data Sci* 2021; 2021: 1–19.

33. Lowery B, D'Acunto S, Crowe RP, *et al*. Using natural language processing to examine social determinants of health in prehospital pediatric encounters and associations with EMS transport decisions. *Prehosp Emerg Care* 2022; 2022: 1–8.

34. Stenetorp P, Pyysalo S, Topić G, *et al*. BRAT: a web-based tool for NLP-assisted text annotation. In: *Conference of the European Chapter of the Association for Computational Linguistics*; 2012; 2012: 102–7; Avignon, France.

35. Johnson AE, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3: 160035.

36. Lee K, Dobbins NJ, McInnes B, *et al*. Transferability of neural network clinical deidentification systems. *J Am Med Inform Assoc* 2021; 28 (12): 2661–9.

37. Alsentzer E, Murphy J, Boag W, *et al*. Publicly available clinical BERT embeddings. In: *Assoc Comput Linguist, Clinical Natural Language Processing Workshop*. 2019: 72–8; Minneapolis, MN. doi:10.18653/v1/W19-1909.

38. Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.

39. Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020; 21 (140): 1–67.

40. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–70.

41. Zhuang L, Wayne L, Ya S, *et al*. A robustly optimized BERT pre-training approach with post-training. In: *Chinese National Conference on Computational Linguistics*. 2021: 1218–27; Huhhot, China. https://aclanthology.org/2021.ccl-1.108.

42. Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. In: *International Conference on Learned Representations*. 2013; Scottsdale, AZ. https://arxiv.org/abs/1301.3781.

43. Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. In: *N Am Chapter Assoc Comput Linguist*. 2019: 4171–86; Minneapolis, MN. doi:10.18653/v1/N19-1423.