Question-answering system extracts information on injection drug use from clinical notes

Check for updates

Maria Mahbub [1] ✉, Ian Goethert[2], Ioana Danciu[3,4], Kathryn Knight[2], Sudarshan Srinivasan[1], Suzanne Tamang[5,6], Karine Rozenberg-Ben-Dror [7], Hugo Solares[5], Susana Martins[5], Jodie Trafton[5], Edmon Begoli [1] & Gregory D. Peterson[8]

Abstract

Background Injection drug use (IDU) can increase mortality and morbidity. Therefore, identifying IDU early and initiating harm reduction interventions can benefit individuals at risk. However, extracting IDU behaviors from patients' electronic health records (EHR) is difficult because there is no other structured data available, such as International Classification of Disease (ICD) codes, and IDU is most often documented in unstructured free-text clinical notes. Although natural language processing can efficiently extract this information from unstructured data, there are no validated tools.

Methods To address this gap in clinical information, we design a question-answering (QA) framework to extract information on IDU from clinical notes for use in clinical operations. Our framework involves two main steps: (1) generating a gold-standard QA dataset and (2) developing and testing the QA model. We use 2323 clinical notes of 1145 patients curated from the US Department of Veterans Affairs (VA) Corporate Data Warehouse to construct the gold-standard dataset for developing and evaluating the QA model. We also demonstrate the QA model's ability to extract IDU-related information from temporally out-of-distribution data.

Results Here, we show that for a strict match between gold-standard and predicted answers, the QA model achieves a 51.65% F1 score. For a relaxed match between the gold-standard and predicted answers, the QA model obtains a 78.03% F1 score, along with 85.38% Precision and 79.02% Recall scores. Moreover, the QA model demonstrates consistent performance when subjected to temporally out-of-distribution data.

Conclusions Our study introduces a QA framework designed to extract IDU information from clinical notes, aiming to enhance the accurate and efficient detection of people who inject drugs, extract relevant information, and ultimately facilitate informed patient care.

Plain language summary

There are many health risks associated with injection drug use (IDU). Identifying people who inject drugs early can reduce the likelihood of these issues arising. However, extracting information about any possible IDU from a person's electronic health records can be difficult because the information is often in text-based general clinical notes rather than provided in a particular section of the record or as numerical data. Manually extracting information from these notes is time-consuming and inefficient. We used a computational method to train computer software to be able to extract IDU details. Potentially, this approach could be used by healthcare providers to more efficiently and accurately identify people who inject drugs, and therefore provide better advice and medical care.

Injection drug use (IDU) is a critical health concern in the United States and internationally[1]. Most people begin using illicit drugs through other modes of administration, such as smoking, intranasal absorption, or oral ingestion. As dependence grows, individuals tend to prefer the intravenous (IV) route of drug administration, injecting drugs directly into the veins, as it offers stronger and more immediate effects[2]. The number of people who inject drugs increased almost fivefold from 2011 to 2018, according to estimates in[3], whereas the number of IDU-related overdoses increased eightfold from 2000 to 2018[4].

IDU can lead to complicated medical conditions such as abscesses and cutaneous infections, scarring and needle tracks, endocarditis, HIV/AIDS, Hepatitis C, overdose, and death[1,5–8]. An increase in IDU is also associated with an increase in morbidity and mortality[9–11].

Accurately identifying IDU behaviors in people who inject drugs is crucial for risk assessment and detection of patients who can benefit from harm reduction interventions to potentially prevent IDU-related morbidity and mortality[12,13]. In the literature, the study of IDU-related information extraction has been performed along with other socio-behavioral

1

determinants of health (SBDH). Considering and including SBDH such as prior incarceration, substance use (regardless of administration mode), treatment attitude, psychological distress, and interpersonal violence improve patient mortality and enhance the prediction of medication adherence, hospital readmission, and suicide attempts[14,15].

Despite the growing interest, SBDH such as IDU is not identifiable in patients' electronic health records (EHRs) through ICD codes; although not systematically assessed, it can be documented in clinical notes[16,17]. While structured data fields derived from EHRs may provide some amount of information about risky drug use behaviors and morbidities related to IDU, the clinical note is the only place it can be explicitly documented[12]. Despite being clinically meaningful and having the potential to identify patients that can benefit from harm reduction interventions, care providers often struggle to retrieve these data points from EHRs, and evidently, the exclusion of this data may result in an overall reduced quality of care[18,19].

Natural language processing (NLP) can help extract SBDH-related information from clinical notes and expand the utility of such information in patient care[20–22]. NLP is a branch of computer science that involves automated learning, understanding, and generation of natural languages, enabling the interactions between machines and human languages. Although NLP deals with a variety of tasks involving unstructured text data (e.g., event prediction[23], entity recognition[24], question-answering (QA) for information extraction[25], and relation extraction[26]), in this article, we use extractive question-answering (extractive QA) task to automatically extract information related to IDU from clinical notes in EHRs. To avoid redundancy, for the rest of the paper, we use QA in place of extractive QA. In this QA task, given a query and a clinical note, a QA model would return the relevant answer verbatim from the note as the extracted information. Thus, a QA system is tasked with learning to read and comprehend the clinical note provided a query and then extract information consisting of consecutive words from the notes relevant to the query from that note (Fig. 1).

We use QA to address the information extraction problem for the following reasons. The texts in the clinical notes are very unstructured in nature. For example, the information regarding injection drug names can be presented in the notes in multiple forms, such as—opioids: denies recent use, hx ivdu, claims last use years ago. other drugs: hx methamphetamine use, has been using daily via injecting since a relapse in December; ivdu (cocaine/methamphetamine); reports using iv meth; iv cocaine mixed with heroine use; used meth by iv drug use; or history of daily heroine use, prior ivdu. Here, ivdu refers to intravenous drug use. Given the demonstrated success of QA models in extracting information of diverse forms from clinical notes[27], we chose to focus on the QA task in NLP. Moreover, one potential implementation of this work would be to incorporate the developed model into a chatbot framework, enabling clinicians to inquire about IDU behavior in people who inject drugs at the point of care by posing questions with various syntactic structures. It would help clinicians identify people who inject drugs and pinpoint related status.

Although not specific to IDU, several studies have focused on identifying clinical concepts or information on substance use disorders (SUD) using NLP[22,28]. In these studies, various NLP techniques have been used to extract SUD-related information. The stemming algorithm has been used to identify words and phrases associated with mental illness and substance use in clinical notes[29,30]. Dependency structure has been utilized to capture

relationships between phrases and tokens in the substance use statement[28]. Word-embedding models have been employed to identify people who use alcohol and substance use status[21]. Machine reading comprehension has been applied to extract some clinical concept categories and relation categories, such as relations of medications with adverse drug events and SBDH[22]. Multi-label text classification and sequence labeling have been used to identify sentences containing labeled arguments about drug use[31]. Topic modeling and keyword matching techniques have been leveraged to extract drug use–related information[32]. Techniques such as active learning[33], multi-label classification[34–37], concept extraction, and joint extraction of entities and relations have been employed to extract information about drug use[38]. Researchers have also focused on identifying drug use information by using NLP-specific techniques to detect opioid use disorder and predict overdose[39–50]. In the literature, we came across one research study that has focused exclusively on IDU. The study has utilized rule-based algorithms, such as regular expressions (RegEx), NegEx[51], and N-grams to search for very limited IDU-related terms, with the objective of identifying people who inject drugs (PWIDs)[12]. In our study, on the other hand, we focus on extracting a broad spectrum of information on injection drug use from clinical notes. This encompasses details such as drug names, active/historical use, frequency of use, risky needle-using behavior, visible signs of IDU, last use, skin popping, harm reduction interventions, and the existence of IDU. Since evidence of IDU cannot be found in structured EHR data and therefore must be inferred from clinical notes, this study's sole focus on IDU aims to help understand how this phenomenon is represented in unstructured notes data and augment techniques that have used NLP techniques less generalizable to this population. To the best of our knowledge, to date, there has been no published attempt at developing a QA algorithm to extract IDU-related information from clinical notes.

To solve the QA task, we use transformer-based deep learning models[52,53] that are known to be one of the most streamlined ways to solve QA tasks and achieve comparable performance in extracting targeted information from different types of biomedical documents, such as scholarly articles[52,54], clinical practice guidelines[25], electronic medical records[27], etc. Nonetheless, evidence suggests that supervised deep learning models require high-quality and large-scale annotated datasets to achieve good performance in any task[53,55,56], and the absence of such a dataset for our targeted QA task poses a critical challenge. An annotated QA dataset comprises data samples, with each sample containing a context (e.g., a clinical note), a question, and an answer extracted verbatim from the context (i.e., the extracted information). In addressing the challenge posed by the limited availability of annotated QA data for constructing an effective QA model, our study takes a two-fold approach. First, we built a high-quality gold-standard QA dataset in collaboration with a subject matter expert (SME), facilitating model training and testing. The dataset includes clinical notes as contexts and question-answer pairs specific to IDU. Then, using this meticulously curated gold-standard dataset, we dive into the primary objective of this study—develop and assess the QA system for IDU-related information extraction from clinical notes. We also perform an error analysis to identify the strengths and weaknesses of our QA system, providing valuable insights to guide future research endeavors. The QA model achieves noteworthy performance, demonstrated by the F1 score of 51.65% for a strict match between gold-standard and predicted answers, as well as

Fig. 1 | A sample clinical note featuring questions about the IDU behavior in people who inject drugs, with extracted IDU-related information color-coded in the note. pt patient, IDU injection drug use, iv intravenous.
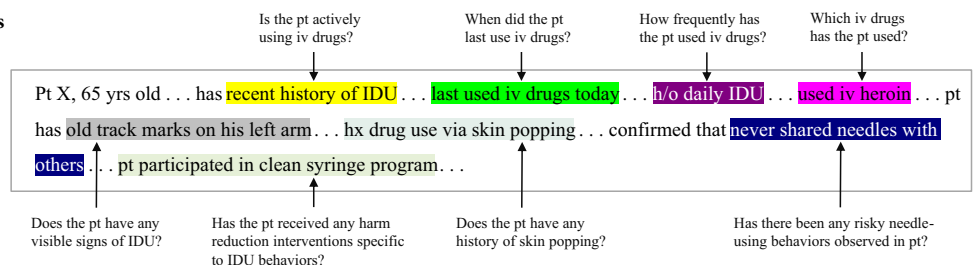
F1, Precision, and Recall scores of 78.03%, 85.38%, and 79.02%, respectively, for a relaxed match. These findings hold promising implications for the precise and efficient identification of injection drug use, enabling the extraction of relevant information from clinical notes.

Methods

In this section, we elaborate on the formulation of this study and its two components: (i) Gold-standard dataset generation and (ii) modeling (Fig. 2). Furthermore, we outline the specifications of the gold-standard dataset, the experimental setup and the metrics used to assess the performance of the QA models.

Problem formulation

We formulate the information extraction task as a QA problem in NLP in the following manner: Given a question on patients' behavior about IDU and a clinical note with IDU-related information (i.e., the context), a QA system retrieves the relevant information (i.e., the answer) from the provided note.

For example, given the question—does the patient have a history of IDU?—and the clinical note—pt X, 200 yrs old … he has a history of smoking with 50 pack years, quit 10 years ago … social ethanol user … no history of idu … remote history of marijuana use … family hx: … physical exam: … provider: name.—the QA system is expected to return the answer —no history of idu—verbatim from the note.

Gold-standard dataset generation

QA is a supervised NLP learning task and, as such, requires an annotated gold-standard dataset for model development and inference. In a QA dataset, each sample consists of the context, a question, and an answer, with the question-answer pairs serving as annotations. To generate a gold-standard dataset from clinical notes, which serve as the context, we employ a three-stage process outlined in Fig. 2: (1) question collection, (2) note enrichment, and (3) gold-standard answer extraction.

Question collection. We initialize the process of question collection for the dataset by asking SMEs about the kind of information on IDU they are interested in from the clinical notes. We then generate a set of questions based on their interest. Table 1 shows the nine categories of interest. In the rest of the paper, we use the term Query Group to imply categories of interest. Table 1 also provides sample questions and answers for each query group.

Each query group targets to extract one category of information from the notes pertaining to that group. For example, the query group—drug names—targets to extract any information about IV drug names from the notes. In our gold-standard dataset, we include multiple variations of questions for each query group. For example, for the query group—drug names, we have five different variations of questions as follows: To what IV drugs has the patient has been exposed?, Which IV drugs have the pt used?, Which intravenous drugs has the patient used?, Which injection drugs?, Which illicit drugs has the patient injected?.

We do this for the following reasons. We anticipate our system to be used as a standalone application—a more user-friendly QA tool to collect IDU evidence—and to be capable of handling different variations of questions posed by clinicians. Furthermore, we hope that different variations of questions for each query group will help increase the QA model's user-flexibility, comprehensiveness, and robustness, ultimately enhancing its

Fig. 2 | Framework for a two-part study for extracting information on IDU behavior in people who inject drugs from clinical notes. The first part consists of gold-standard dataset generation in three primary steps, and the second part consists of QA model development from implementation to inference. SMEs subject–matter experts, IDU injection drug use, QA question–answering.
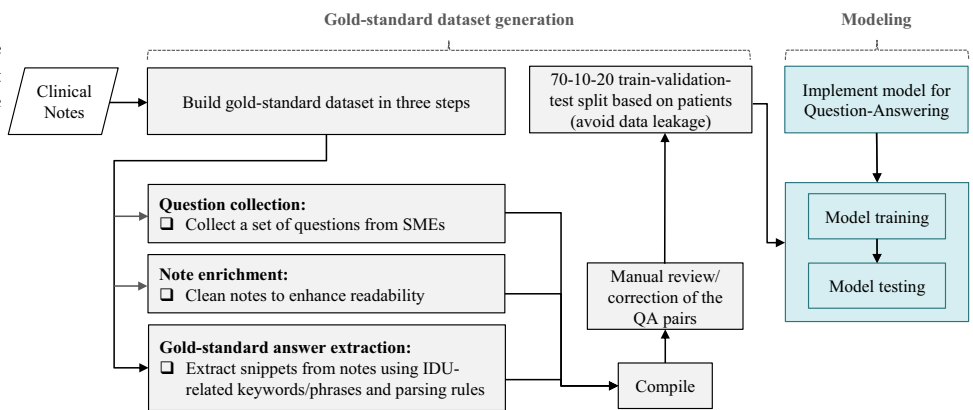
Table 1 | Types of information about IDU that are most likely to be inquired by clinicians from the clinical notes, categorized into nine query groups

Query Groups	Sample Question → Sample Answer
Drug names	Which iv drugs has the patient used? → hx of iv heroin abuse, cocaine, and bnz
Visible signs of IDU	Does the patient have any needle track marks? → track marks noted over bilateral upper extremities
Risky needle-using behavior	Has the patient ever shared needles? → hx of ivdu, has shared needles in the past few weeks
Active/historical use	Is the patient actively using iv drugs? → h/o active iv drug use
Frequency of use	How frequently has the patient used iv drug? → history of ivdu (reports daily use of heroin)
Last use	When did the patient last use iv drugs? → daily use of iv heroin with last use 4 days prior to admission
Skin popping	Does the patient have any history of skin popping? → diffuse scarring from skin popping on lower extremities
Harm reduction interventions	Has the patient been counseled on safe injection techniques? → discussed the importance of using clean needles with patient should he continue to inject drugs
Existence of IDU	Does the patient have any history of IDU? → no ivdu or h/o sharing needles

IDU injection drug use, ivdu intravenous drug use, bnz benzodiazepines, h/o history of.

performance in real-world applications, as follows: (i) Users may pose questions in different ways based on their preferences or understanding. A QA model trained with diverse question variations is more adaptable and capable of accommodating the linguistic diversity inherent in user queries. (ii) Including variations of questions during training helps the QA model become more robust by exposing it to diverse ways the same question can be asked, preparing the model to handle real-world scenarios where questions may be phrased differently but still seek the same information. (iii) Variations of questions during training enable the QA model to generalize its understanding. Instead of memorizing specific phrasings, the model learns the underlying patterns and associations between questions and answers, improving its ability to respond accurately to novel queries.

We use abbreviations, synonyms, and syntactical variations to introduce variations in the questions for each query group, as follows: (i) abbreviations: Is the patient actively using intravenous drugs? → Is the pt actively using intravenous drugs?, Is the patient actively using intravenous drugs? → Is the patient actively using iv drugs?. (ii) synonyms: Does the pt have a history of using intravenous drugs? → Does the pt have a history of using injection drugs?, Does the pt have a history of IDU? → Does the pt have a history of IVDU?. (iii) syntactical variations: Which iv drugs has the patient used? → To which iv drugs has the patient been exposed?, Does the pt have a history of IVDU? → Has the pt ever used IV drugs?

It should be noted that when identifying abbreviations and synonyms to be used in questions, we only choose terms and variants that clinicians commonly use. Examples of these terms and variants include patient and pt, intravenous and iv, history and hx, and IVDU and IDU. And, to ensure that we were able to accurately capture the nuances of possible language usage in the questions with regard to syntactical variations, we sought the guidance of SMEs.

Note enrichment. The contexts in the gold-standard dataset are clinical notes that contain some IDU-related information. As such, we select a cohort of patients whose notes have a higher chance of containing IDU-related information, such as patients who have been diagnosed with Hepatitis C. To guarantee that the clinical notes include information relevant to IDU and narrow down the notes accordingly, we use a list of keywords/phrases that are indicative of IDU (refer to Table 2) and has been developed by SMEs. SMEs followed an iterative approach to create this list. They began by compiling a list of common terms related to IDU, which they then refined by reviewing the associated snippet. They removed terms that caused excessive noise, such as—slamming and drug paraphernalia—and added terms like—skin popping—to enhance granularity. The experts received extensive training to sort and/or define the snippet categories, and they validated the terms to ensure their accuracy.

For our study, we assumed that the presence of any of these IDU-related keywords indicates the presence of relevant information pertaining to IDU in the note. Hence, we discard the notes that do not contain any of the words/phrases provided in Table 2, suggesting the possible non-existence of any IDU-related information in that note. As shown in Table 2,

this list can be categorized into the following groups: IV drug names, visible signs of IDU, risky needle-using behavior, skin popping, harm reduction interventions, and generic IDU terms.

To enhance the readability of clinical notes and make them more suitable for automated processing, we conduct rigorous manual exploration of the final set of notes, identifying some common patterns that can help clean them using RegEx. It is important to note that to preserve crucial information in the clinical notes, we perform minimal data cleaning, as follows: (i) Remove newlines following within-sentence punctuation marks, such as commas, semicolons, or colons. For instance, removing the newline (\n) highlighted in the sentence—Veteran reported using iv meth,\n iv cocaine, and etoh. (ii) Remove newlines appearing before punctuation marks, such as period, comma, or semicolon. For example, removing the newline (\n) highlighted in the sentence—Veteran reported using iv meth, iv cocaine, and etoh\n. (iii) Remove newlines positioned between words within the same sentence. For example, removing the newline (\n) highlighted in the sentence—Veteran reported\n using iv meth, iv cocaine, and etoh. (iv) Consolidate multiple consecutive occurrences of newlines, white spaces, or punctuations into single instances. For example, replacing multiple periods with a single period in the sentence—Veteran reported using iv meth, iv cocaine, and etoh............. We perform these steps to clean all the notes used for training, validation, and testing.

Gold-standard answer extraction. The next step in our dataset generation process is to extract gold-standard answers (i.e., information related to IDU) from the clinical notes. Clinical notes are inherently lengthy, and manually extracting the gold-standard answers from them requires a substantial amount of time, rendering the process unfeasible. Therefore, we devise a pre-annotation strategy involving an automated step-by-step answer extraction process that integrates rule-based NLP techniques. The primary objective of this phase is to substantially reduce the manual annotation/review effort. Nevertheless, to ensure the utmost quality of the gold-standard dataset, the outputs from this pre-annotation phase, along with the associated questions, underwent subsequent manual review and correction by a subject-matter expert with a Ph.D. in Psychology and an extensive background in substance use disorder, counseling, and treatment. Our pre-annotation strategy is based on three assumptions:

Assumption 1. Our QA task only tackles information extraction (i.e., answering questions) from one single place (a sentence) in the note at a time.

Assumption 2. The inquired information can be found in a single sentence in the note. This assumption stems from our rigorous manual exploration of the notes during the note enrichment step, where we find RegEx patterns. Our observation indicates that, in most instances, a single sentence per question suffices to capture the relevant answer. Nonetheless, we acknowledge that this straightforward sentence selection process may not always be optimal. Unstructured clinical notes often deviate from grammatical rules. Additionally, information presentation in these notes may

Table 2 | A list of IDU keywords/phrases provided by SMEs

Keyword Groups	Keywords/phrases
IV drug names	iv/intravenous/inject(s/ed) heroin/meth/cocaine/crack, speedball
Visible signs of IDU	track marks, skin popping
Risky needle-using behavior	sharing/shared/dirty needle
Skin popping	skin popping
Harm reductioninterventions	community/clean/safe syringe service/program, ssp, ris4e, counseled on safe(r) injection, safe injection technique
Generic IDU terms	ivdu, idu, ivda, iv/intravenous/injection drug use/abuse, inject/injected drug, drug(s) by injection, iv/intravenous drug injector/injection, illicit iv/intravenous drug, iv/intravenous drug paraphernalia, suspect injecting, pwid

IDU injection drug use, *ivdu* intravenous drug use, *ssp* syringe services programs, *ivda* intravenous drug abuse, *ris4e* resists infection by sterile syringe safe sex and education, *PWID* people who inject drugs.

vary, adopting styles such as questionnaires or bulleted lists. As a result, a single sentence in the traditional sense occasionally leads to either a larger text segment or a fragmented part of a single piece of information. These instances lead to the inclusion of irrelevant or incomplete information in the answers, and we address and rectify these issues during our manual review phase.

Assumption 3. If the note contains IDU-related information in multiple locations, each is considered a separate answer string. Furthermore, multiple answer strings from the same note are expected to contain different kinds of information that should be answered by different questions. For example, in the note snippet—pt has a history of smoking with 50 pack years, quit 10 years ago … social ethanol user … has h/o ivdu … remote history of marijuana use … last used iv meth 2 years ago …—there are two locations where IDU-related information can be found—has h/o ivdu and last used iv drugs 2 years ago. In such cases, we consider them as separate answers that are retrieved when asked the following questions: Does the pt have a history of IDU? and When did the pt last use IV drugs?.

Given clinical notes, we extract the automated gold-standard answers using rule-based NLP techniques as follows:

Step 1: Tokenize the sentences in the notes. Here, we define a sentence in the traditional sense, ending with a period. Therefore, for the sentence tokenization, we use periods to indicate the end-of-sentence.

Step 2: Identify sentences that contain any of the IDU keywords from Table 2 using regular expression string matching and discard the rest.

Step 3: At this point, the sentences containing the IDU keywords can be ideally considered gold-standard answers (i.e., extracted information relevant to IDU). Nonetheless, our primary aim is to extract IDU-related information from the notes, but we also want the extracted information to be as precise as possible containing lesser nonessential information. A full-sentence answer is most likely to include nonessential information, which can be further reduced by using parsing rules. Parsing rules refer to NLP techniques that can identify specific patterns of text within a string that represent the concepts of interest while ignoring the remaining text. An example of removing nonessential information from the answer can be transforming the sentence—social history: pt lives with family in [location], quit smoking 10 y ago, occ etoh, ... hx methamphetamine use, has been using daily via injecting since a relapse in December.—into the phrase—hx methamphetamine use, has been using daily via injecting since a relapse in December.

To create the parsing rules in this study, we randomly sample a set of sentences and focus on identifying specific phrases that occur together before or after the IDU keywords and modify or provide information that is crucial to the IDU-related history of the patient (refer to Table 1). These phrases can be adjacent to or distant from the keywords. For example, pt lives with family, denies ivdu—versus—pt lives with family, denies any tobacco, etoh or ivdu In this example, the phrase—denies—provides crucial information on the IDU behavior of the patient.

In Supplementary Table 1, we provide a detailed list of these phrases along with the targeted pattern type, parsing rules, and examples of how they help reduce the nonessential information from the answers. The parsing rules mainly focus on identifying patterns stating negative IDU mentions, temporal information, opioid use disorder specific to IDU, and status of track marks.

Although these parsing rules can extract the correct concise gold-standard answers from the clinical notes in numerous cases, manual review reveals instances where the rules failed to accurately identify these answers. This discrepancy was primarily attributed to the unstructured nature of information within the notes.

Question-to-answer mapping. Finally, to generate the labels (question-answer pairs) of our gold-standard dataset, we create mappings between the questions from Section Question collection and the gold-standard answers from Section Gold-standard answer extraction. We achieve this by considering the query groups in Table 1. For each query group, we identify a group of words in the gold-standard answers that are most likely to provide the information inquired by that query group. To compile this group of words, we engage in meticulous manual exploration, reading sentences containing IDU keywords. Depending on the kind of information we are interested in (reflected by the query groups), these words can be either the keywords in Table 2 or the words (Supplementary Table 1) that co-occur with the keywords and can help convey the information inquired by the user. For example, co-occurring words—daily and last—describe the frequency of use and the last use of IDU, respectively.

Thus, for each query group, we decide on a group of words that are most likely to help convey the inquired information and map the answers that contain these words to the questions in that query group (Table 1). The resulting compilation is presented in the—Words in Gold-standard Answers Most Likely to Provide Inquired Information—column of Supplementary Table 2. It is important to note, however, that this list is not exhaustive and represents only what we observe during our exploration, not an all-encompassing collection of potential phrases indicating the inquired information. Considering this, in our manual review phase, we manually correct annotations that are overlooked or mislabeled by these rules.

In Supplementary Table 2, we present the mappings between the query groups and the words in gold-standard answers that are most likely to provide inquired information. We also demonstrate sample answers for each mapping. Note that the answers in one query group and the answers in a different query may not be mutually exclusive. This is because if we find words in an answer that belong to multiple query groups, then that answer is mapped to all questions from these query groups. For example, the first sample answer from Supplementary Table 2—recent ivdu with meth and heroin—contains the words—recent and heroin/meth—from query groups active/historical use and drug names, respectively. Hence, this answer will be mapped to all questions in these two query groups.

The well-known ConText rules[57] in the literature use a similar rule-based approach to identify the negation or temporality of a condition. They used a specific set of words tailored to the types of notes used in their study. On the contrary, although the words utilized in our study share some commonalities, they exhibit notable differences from those employed in the ConText algorithm. This distinction arises from variations in the notes used in our experiments and the specific information we target to extract from the notes. Our study exclusively focuses on injection drug use. In contrast, the error analysis of ConText indicates its unsatisfactory performance in identifying temporality related to chronic conditions and risk factors, i.e., alcohol, and drugs in clinical notes. Additionally, while ConText explicitly identifies historical versus recent conditions, our question-answering system concentrates on extracting any temporal information regarding injection drug use, leaving the determination of whether the status is recent or historical to clinicians.

Regarding the query group—last use, it is crucial to note that a patient may have multiple note entries, each with its own last use. Given our study's emphasis on extracting information from one clinical note at a time, the definition of—last use—is confined to—last use per note.

After generating the labels (i.e., question-answer pairs), we manually review the whole dataset in collaboration with a subject-matter expert to ensure that our gold-standard dataset is of high quality and accuracy.

Modeling with question–answering system
In the next step of our study, we develop the QA model for extracting IDU-related information using the gold-standard QA dataset from Section Gold-standard dataset generation. We use Bidirectional Encoder Representations from Transformers (BERT)[53]-based deep learning QA models where the feature extractor is a trainable pre-trained BERT-based language model, and the QA task layer is a single-layer feed-forward neural network.

We experiment with four state-of-the-art pre-trained language models —BERT[53], BioBERT[52], BlueBERT[58], and ClinicalBERT[59]—as trainable feature extractors and develop four QA models.

Provided a sequence of tokens (words or pieces of words) in a question and a clinical note, the QA model returns the start and end token of the answer span. Any text between the start and end tokens included is then considered as the answer (i.e., the extracted information). Together with the question and the note, the maximum allowable number of input tokens in these BERT-based QA models is 512. To handle samples with longer clinical notes, we follow a widely known technique in QA modeling—sliding window with a document stride[53].

Below we provide a brief description of this technique: Given an input question consisting of 20 tokens, the remaining allowable number of input tokens for the note is limited to 492 (which is 512 minus the 20 tokens in the question). If the note exceeds this limit, we employ a sliding window technique to split it into multiple chunks using a document stride of 128 tokens. The document stride determines the starting token of each subsequent chunk. After this preprocessing step, each chunk prepended with the original question tokens is considered a separate data sample.

Dataset statistics

We use clinical notes sourced from the VA Corporate Data Warehouse (CDW) to construct the gold-standard dataset. The clinical notes in the VA CDW are fully identified. The selected notes correspond to the period of January 2022 and belong to patients with the Hepatitis C diagnosis. The identification of Hepatitis C-positive patients is performed using ICD-10 codes. We select the cohort of patients with Hepatitis C as their clinical notes are more likely to include information related to IDU. As explained in Section Note enrichment, we narrow down the clinical notes using a list of keywords/phrases indicative of IDU (refer to Table 2).

To reduce computational overhead during training and because unusually large notes (determined by the outliers in the distribution of note lengths) may contain templated nonessential information that is not relevant to any specific patient, we remove some outlier notes based on the interquartile range of the note lengths. We later show in Section Error analysis that note length does not affect the performance of the model at the time of inference.

We also analyze the types of notes included in this study. Our analysis reveals that there are 411 different types of notes. Supplementary Fig. 1 displays the 20 most frequently encountered note types in this study. Notably, internal medicine notes and primary care notes emerge as the two most prevalent types. We also find that addendum notes rank third in frequency. Addendum notes serve as supplements to notes of other types.

Supplementary Table 3 shows the statistics of our gold-standard dataset. Our cohort consists of 1145 patients with a total of 2323 notes that have an average length of 1013 words. Words are identified based on whitespace. In addition, we examine the distribution of the query groups outlined in Table 1 and Supplementary Table 2 within the gold-standard dataset. This analysis is illustrated by the pie chart depicted in Supplementary Fig. 2. The dataset is dominated by QA pairs related to active/historical use, as demonstrated. Following closely behind are QA pairs about the existence of IDU and drug names, whereas the least frequent QA pairs in the dataset are those pertaining to skin-popping and harm reduction interventions.

Ethics: This project was conducted as a national quality improvement effort to improve care for Veterans with substance use being treated in the Veterans Health Administration (VHA). Models were designed to be implemented into VHA decision support systems and are not expected to be generalizable or valid for application outside of notes from the VHA Computerized Patient Record System (CPRS). As such, this work is considered non-research by VHA (as per ProgramGuide-1200-21-VHA-Operations-Activities.pdf (va.gov)). However, Oak Ridge National Laboratory (ORNL) required additional oversight of this VHA clinical quality improvement project as local standard practice for all uses of patient medical record data within their institution, with approval of the project by the Oak Ridge National Laboratory IRB. The need for the veterans whose medical records were used in the study to give informed consent for the study was waived by the ORNL IRB.

Experimental setup

For experimentation, we divide our gold-standard dataset into train, validation, and test sets using a 70-10-20 split based on patients to avoid any data leakage. To implement the QA models, we use PyTorch[60]. We use the pre-trained language models from the huggingface API[61].

Based on the statistics of our gold-standard dataset, we choose 512 as the maximum sequence length, 20 as the query length, and 100 as the answer length. After reviewing the hyperparameters utilized in various QA tasks as outlined in[25,52,53,58,62–67], we set the document stride to 128 and opted for a batch size of 32, a learning rate of $3e^{-5}$, and a training epoch count of 5 for the training configurations. We performed all experiments using a single GPU on a Linux virtual machine with two GRID V100-32C GPUs.

Metrics

To assess the performance of QA models in extracting IDU-related information, we utilize strict matching criteria to compute the F1 score[68]. It involves verifying if the prediction precisely matches the gold-standard answer character by character, resulting in a strict F1 score per sample that can be either 1 or 0. Additionally, we use relaxed matching criteria to measure the F1, precision, and recall scores[68]. A relaxed match determines if there is any overlap between the prediction and the gold-standard answer. The recall or sensitivity score per sample reveals the proportion of words in the gold-standard answer that is identified correctly in the predicted answer. Precision or positive predictive value (PPV) score per sample informs us about the proportion of words in the predicted answer that are actually correct. In the context of QA problem, when calculating these metrics, true positive refers to the count of tokens that both the predicted answer and the gold-standard answer share, false positive represents the number of tokens found solely in the predicted answer, and false negative indicates the number of tokens only in the gold-standard answer and not in the predicted one. The relaxed F1, precision, or recall scores per sample can range from 0 to 1. Following[55], we report the macro-averaged F1 score, accompanied by macro-averaged precision and recall scores on the test sets.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results and discussion

In this section, we report and discuss the findings from the experiments with QA models. Furthermore, we conduct a comprehensive error analysis to demonstrate the capabilities and limitations of the QA models in extracting information related to IDU from clinical notes.

Results on gold-standard test set

This section focuses on examining the experimental outcomes of the QA models and demonstrates their performance on the test set of our gold-standard dataset. As shown in Table 3, ClinicalBERT outperforms other BERT-based QA models. A strict F1 score of 52% for ClinicalBERT implies that the QA model can extract IDU-related information 52% of the time with a strict match to the gold-standard answers. A relaxed recall score of 79% on the test set suggests that overall there is a substantial degree of word overlap between the predicted answers and gold-standard answers. We further analyze the recall score in Section Analysis of recall score. On the other hand, a relaxed precision score of 85% in the test set indicates that a higher percentage of terms retrieved as answers by the QA model are included in the gold-standard answers. A relaxed F1 score of 78% indicates that the ClinicalBERT model can extract a high percentage of correct information while achieving high precision in those extracted answers.

Temporal out-of-distribution testing

The writing style of clinical notes may change over time because of changes in clinicians, health care facilities, patients, etc.[69]. Given the purpose of our QA model, it is imperative to examine whether the performance of our QA models is retained over time. Therefore, we perform additional testing of the

Table 3 | Performance scores of QA models on the test set

Model	Strict match	Relaxed match		
	F1	F1	Precision	Recall
BERT	48.10	75.88	81.57	78.82
BioBERT	46.49	74.99	81.09	76.74
BlueBERT	43.03	71.21	79.14	73.77
ClinicalBERT	**51.65**	**78.03**	**85.38**	**79.02**

*The bold font highlights the best scores.

Table 4 | Performance scores of the QA models on the additional test datasets built using Cohort-Short and Cohort-Long

Model	Test dataset (Cohort-Short)			
	Strict match	Relaxed match		
	F1	F1	Precision	Recall
BERT	47.10	74.14	77.32	76.97
BioBERT	46.95	74.25	77.36	**79.12**
BlueBERT	41.76	71.71	76.47	76.18
ClinicalBERT	**55.31**	**76.12**	**81.34**	76.63
Model	Test dataset (Cohort-Long)			
	Strict match	Relaxed match		
	F1	F1	Precision	Recall
BERT	48.38	74.38	79.57	78.99
BioBERT	50.63	73.89	**84.44**	74.15
BlueBERT	51.53	73.46	81.17	78.52
ClinicalBERT	**53.42**	**75.16**	81.32	**80.20**

*The bold font highlights the best scores.

models from Table 3 on unseen data. We examine the QA models' short-term and long-term information extraction capabilities by testing on clinical notes from two additional cohorts. For testing the short-term capability, we randomly select 100 patients and use their notes from February 2022. Similarly, for testing the longer-term capability, we randomly select 100 patients and use their notes from November 2022. Due to the limitations in our data availability at the time of this study, we were unable to include clinical notes beyond November 2022 for testing the longer-term information extraction capability of the QA models. In future endeavors, we aim to assess the performance of QA models on more recent notes as part of ongoing research.

To avoid data leakage, we use patients and their notes that did not appear in the gold-standard dataset generated by using notes from January 2022. We use the method described in Section Gold-standard dataset generation for building the test datasets using these notes. Similar to the gold-standard dataset, we manually review these test datasets in collaboration with a subject-matter expert. For the rest of the paper, we use the terms Cohort-Short and Cohort-Long to represent temporally out-of-distribution notes in February and November, respectively. Supplementary Table 4 shows the statistics of the test datasets built using Cohort-Short and Cohort-Long. We also show the distribution of query groups in these test datasets in Supplementary Fig. 3. As shown, the distribution of the query groups is similar for the additional test sets and our original gold-standard dataset (refer to Supplementary Fig. 2).

Table 4 shows the performance of the QA models. As shown, for both test datasets, the ClinicalBERT model performs with overall high scores, reflecting its competence in extracting information over time.

Error analysis

In this section, we provide a comprehensive analysis of the strengths and weaknesses of our best-performing model, which is the ClinicalBERT QA model, in extracting IDU-related information. We perform a fivefold analysis as follows: Examine the (i) confidence intervals of the performance scores, the effect of (ii) note length, (iii) question length, and (iv) gold-standard answer length on the performance of the QA model, and (v) the performance of the QA model for each query group. Furthermore, by analyzing the recall scores, we showcase the proficiency of the QA model in retrieving IDU-related information. For our error analysis, we consider all three of our test sets—the test set in our gold-standard dataset and the test datasets from Cohort-Short and Cohort-Long.

Confidence intervals of performance scores. We calculate the confidence intervals (CI) for strict F1 score and relaxed F1, precision, and recall scores achieved by the best-performing QA model to represent how good these estimates are and thus quantify their uncertainty. Smaller confidence intervals demonstrated in Table 5 indicate that our estimates are precise with a high level (95%) of confidence.

Effect of note length. Clinical notes have varying lengths—they can be as short as 30 words up to as long as 5747 words, based on the statistics of our test datasets. Therefore, we want the QA model to perform consistently well for all lengths of clinical notes. To identify the effect of note length on the QA model's performance, we calculate the length of the

contexts (i.e., notes) in the three test sets and bin them into four quartiles based on their ascending lengths. Supplementary Data 1 and the x-axis in Fig. 3a show the length range of these bins, whereas the green bars with the right y-axis show the sample count for each bin. We find that note length does not have any notable effect on the model's performance scores, demonstrated in Supplementary Data 1 and on the left y-axis of Fig. 3a.

Effect of question length. We also examine the effect of question length on the performance of the QA model. For this analysis, we adopt the same binning approach as the one on note length. Figure 3b and Supplementary Data 1 show that, similar to note length, question length also has no effect on the model's performance scores.

Effect of gold-standard answer length. In our test sets, we have varying lengths for the gold-standard answers (i.e., extracted information). For successful implementation, it is essential for the QA model to be able to extract different lengths of information from the clinical notes. Using the binning approach described earlier for the analysis on note length, we find that the QA model struggles to extract longer gold-standard answers with a strict match—demonstrated by the strict F1 score in Fig. 3c and Supplementary Data 1. Nevertheless, higher relaxed metric scores demonstrated by the QA model indicate its capability to identify the location of the correct answers. To improve the QA model's proficiency in extracting longer answers with a strict match, additional research is required.

Performance for query groups. Based on the information we are interested in extracting from the clinical notes, we create nine query groups, as shown in Table 1. Supplementary Data 1 and the green bars along with the right y-axis in Fig. 3d show the sample count (in log scale) for the query groups in the test sets. The query group active/historical use dominates the datasets, followed by the query group existence of IDU and drug names. Interestingly, we find that the model performs the best on the query group visible signs of IDU and existence of IDU. Presumably, the query group visible signs of IDU has an overall higher performance despite having the third lowest sample count in the test sets and fifth lowest sample count in the gold-standard dataset because the information queried by this query group usually has some consistent terms in them such as track marks or needle track marks along with some other limited relevant information, for example, fresh track marks on his forearms. We hypothesize that the information extracted by this query

Table 5 | Performance scores (with 95% confidence intervals) of the best-performing ClinicalBERT QA model on the test set in the gold-standard dataset and test datasets from cohort-short and cohort-long

Model	Strict match	Relaxed match		
	F1 (95% CI)	F1 (95% CI)	Precision (95% CI)	Recall (95% CI)
Gold-standard	51.65 (49.92–53.39)	78.03 (76.99–79.08)	85.38 (84.40–86.37)	79.02 (77.89–80.15)
Cohort-Short	55.31 (53.10–57.10)	76.12 (74.75–77.48)	81.34 (80.01–82.67)	76.63 (75.19–78.06)
Cohort-Long	53.42 (50.45–56.35)	75.16 (73.29–77.03)	81.32 (79.50–83.14)	80.20 (78.28–82.12)
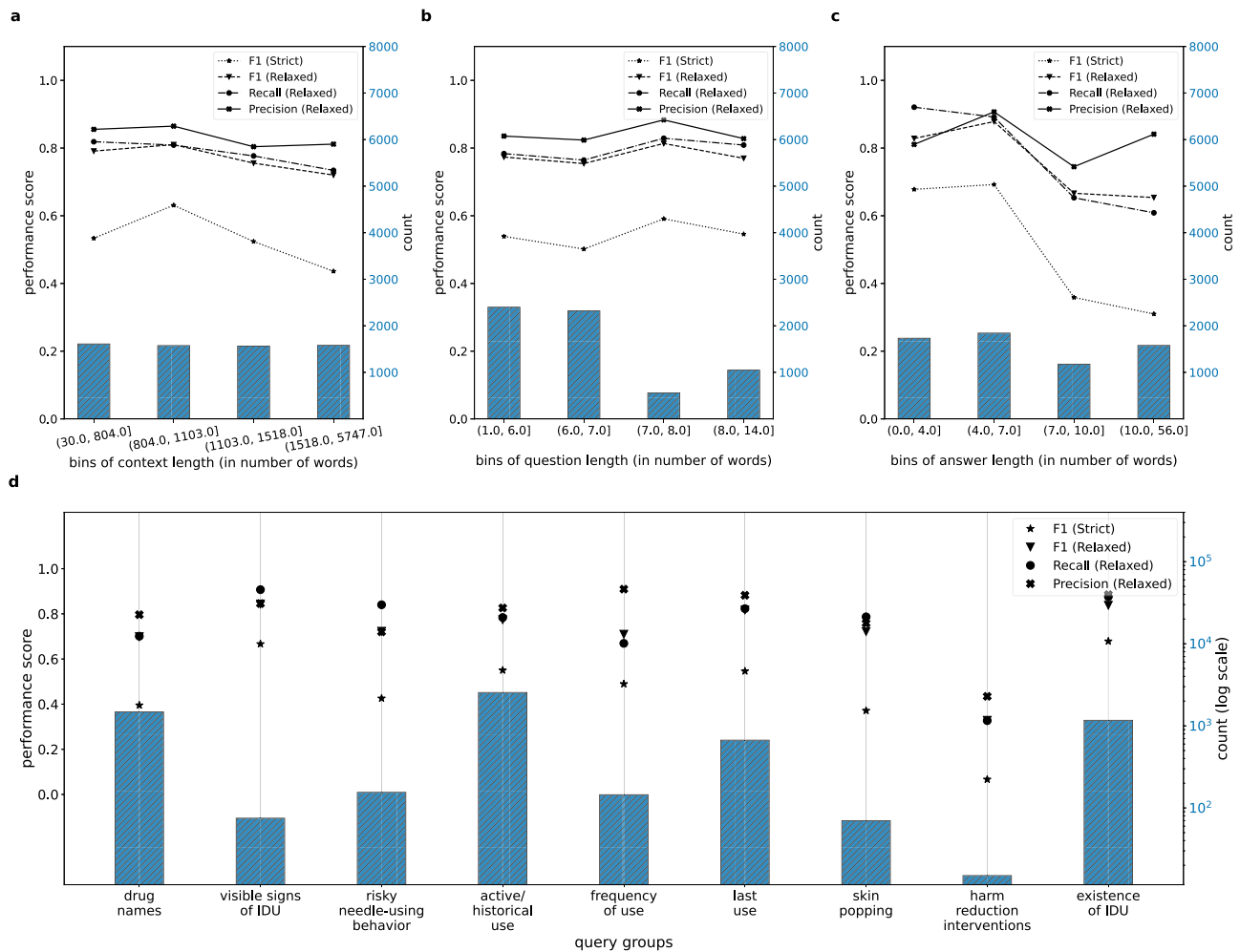
Fig. 3 | Error analysis of the QA model. Blue bars refer to the sample count. Effect of context (note) length (**a**), question length (**b**), and gold-standard answer length (**c**) on QA model's performance. **d** Performance of QA model for each query group. IDU injection drug use.

group may be easier for the QA model to comprehend. However, further evaluation of the QA model is necessary to corroborate this hypothesis. Figure 3d also shows that the QA model struggles the most with the group harm reduction interventions. It may happen because harm reduction interventions have the least number of samples in the gold-standard dataset, possibly causing difficulty for the model to learn from training samples. It also has the least number of samples in the test sets to obtain a comprehensive overview of the model's performance.

Analysis of recall score. In this part of the discussion, we analyze the recall scores of the QA model to shed light on its overall capability to extract gold-standard answers. In cases where the strict F1 score for the predicted answer is 0, the recall score can demonstrate the overlap between the gold-standard and predicted answers. For the test set in our gold-standard dataset, our QA model achieved a strict F1 score of

approximately 52%. For the remaining 48%, we examine the recall scores by binning them into 12 intervals (shown in Table 6). We also perform similar analyses for cohort-short and cohort-long. As indicated in Table 6, 14% of the predictions for the gold-standard test set, although lacking a strict match, exhibit a 100 Similarly, for Cohort-Short and Cohort-Long, respectively, 7% and 15% of the predicted answers have a 100% overlap with the gold-standard answers while not having a strict match. One potential issue while considering 100% overlap without a strict match is the predicted answer being the entire context. To address this concern, we compare the ratio of the predicted answers (that do not have a strict F1 score of 1) to the contexts with the ratio of the gold-standard answers to the contexts. Figure 4 and Supplementary Data 2 show that the distribution of the percentage ratios of the predicted answers to the contexts is similar to that of the gold-standard answers to the contexts.

Table 6 | Analysis of recall scores for cases where the predicted answers do not have a strict match with the gold-standard answer

Recall (%)	Sample count (%)		
	Gold-standard	Cohort-Short	Cohort-Long
(99–100]	13.76	6.65	15.32
(90–99]	0.65	0	0.45
(80–90]	2.23	0.6	0.54
(70–80]	2.81	2.62	2.43
(60–70]	1.33	3.43	2.7
(50–60]	2.38	1.76	3.87
(40–50]	5.87	3.58	2.61
(30–40]	6	16.07	1.08
(20–30]	3.37	2.87	7.39
(10–20]	4.3	1.91	4.59
(0–10]	2.07	1.56	2.61
0	3.56	3.63	2.97
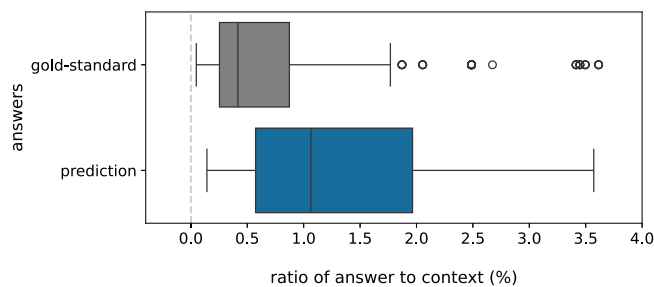
Fig. 4 | Distributions of the ratios of the predicted answers to the contexts and the ratios of the gold-standard answers to the contexts across 747 QA samples. This analysis specifically focuses on cases where the predicted answers exhibit 100% overlap with gold-standard answers without adhering to strict matching criteria. The ratios here are presented in the form of percentages.

Examples of predicted answers. We demonstrate the capability of the QA model by showing some randomly selected examples of the predicted answers along with the questions and gold-standard answers in Supplementary Table 5.

Analysis of model's capability to identify whether a note contains IDU-related information or not. Our study focuses on extracting IDU-related information from clinical notes, but ideally, we also want our QA model to identify whether the note contains IDU-related information or not. As such, as an additional analysis, we examine the QA model's ability to identify clinical notes that do not contain any mention of IDU keywords (Table 2) and, as such, are assumed to have no information about IDU. We hypothesize that given a clinical note with no mentions of IDU, the QA model should return an empty string because it could not find the information it was asked to retrieve.

To test this, we use patients from the test set in the gold-standard dataset. Recall that in our context processing step in Section Note enrichment, we remove notes that do not contain any IDU keywords. For this analysis, we incorporate 443 notes from 226 patients with no mentions of IDU keywords. We ensure that the notes only belong to the patients in the test set.

To annotate these notes, we use the query group—existence of IDU—as questions and empty strings as answers. For example, given a note with no mentions of IDU and the question—Has the pt ever injected drugs?, the QA model should return an empty string.

To measure the performance, we consider only the strict F1 score. Thus, if the predicted answer matches with the empty string, we consider that a success (strict F1 score = 1) and otherwise a failure (strict F1 score = 0). We find that our QA model can identify approximately 88% of the clinical notes that do not contain any IDU-related information. Interestingly, we find that for 10% of the mispredicted answers, the model returned the string—empty. Additionally, we observe that the model returned the string with a single period, constituting the second most frequently mispredicted answer, accounting for 0.5% of the predictions. Therefore, we can say that while our QA model can extract IDU-related information from clinical notes, it also has the potential to identify the notes that do not contain any.

Study limitations

This study has some limitations. First, the QA model was trained and tested on a dataset that had already undergone a fair amount of NLP pre-processing. Therefore, the model's performance may be limited when generalized to raw, source clinical notes. Further evaluation is needed to prove otherwise. Second, in many cases, we have noticed the use of the terms—patient denied, or veteran tells me—for IDU-related information in the clinical notes. The QA model's capabilities are limited to the text from which it can extract the pertinent information. Therefore, the QA model must be implemented with supervision in the clinical setting. Third, our list of IDU keywords/phrases provided by SMEs to filter notes for generating gold-standard datasets is not exhaustive. Notably, drug names such as fentanyl or xylazine are absent from the list. Further assessment is required to measure the QA model's capability to extract information related to these substances. Fourth, the datasets used in this study have been manually reviewed by one reviewer. Including a second reviewer in the manual review process may ensure more diverse perspectives, reducing the likelihood of individual biases or errors.

Conclusion

Detection of injection drug use (IDU) behavior among patients is crucial for informed patient care. In this paper, we tackle the challenging task of IDU-related information extraction from clinical notes. We build a QA system that takes in a clinical note and an end-user query on IDU and returns the information on IDU extracted from the note. We hope to potentially integrate the QA model from this study into a user-friendly chatbot framework, enabling clinicians to inquire about information related to nine categories, as identified in this study, with a view to collecting IDU evidence through an interactive platform. We evaluate our QA system on a gold-standard dataset built using clinical notes from VA CDW and a combination of manual exploration, rule-based NLP techniques, and subject–matter expert validation. We also perform an additional evaluation to examine the capability of our QA model to extract information from temporally out-of-distribution notes. We then investigate the strengths and limitations of the QA model and identify potential avenues for future research by performing rigorous error analysis.

We have identified the following next steps for this research: (i) Examine the QA model's capability to extract information from temporally out-of-distribution clinical notes by testing the model on a more recent set of clinical notes. (ii) Examine/enhance the QA model's capability to handle raw clinical notes without the data-cleaning steps. (iii) Examine/enhance the QA model's capability to extract information on illicit injection drugs that are not covered in this study, for example, xylazine. (iv) The extractive QA problem may benefit from the named entity recognition (NER) task[70,71]. Subsequent research could explore the integration of NER into the QA task for further investigation. (v) Expand the applications of QA tasks to extract other types of information from clinical notes, such as information related to alcohol use disorder and substance use disorder.

This method can support the accurate and efficient identification of people who inject drugs and relevant information extraction using their clinical notes to facilitate harm-reduction interventions and care.

Data availability

The dataset developed for this study is not accessible to the public under requirements of the Health Insurance Portability and Accountability Act of 1996 and related privacy and security concerns. The underlying electronic health record data can only be used for improving treatment for patients receiving services from the Veterans Health Administration (VHA). Those interested in accessing VHA EHR data extracts curated for this quality improvement project to replicate and validate findings may contact the corresponding author regarding access via VHA collaboration. Source Data for Figs. 3 and 4 are provided in Supplementary Data files 1 and 2, respectively.

Code availability

The code used to develop the QA models is a modified version of the publicly available huggingface example for the question-answering task, which can be found here: https://github.com/huggingface/transformers/blob/master/examples/legacy/question-answering/run_squad.py. The modified code is stored in a GitHub[72].

References

1. Goel, N., Munshi, L. B. & Thyagarajan, B. Intravenous drug abuse by patients inside the hospital: a cause for sustained bacteremia. *Case Rep. Infect. Dis.* **2016**, 1738742 (2016).
2. O'Brien, C. P. Drug addiction and drug abuse. *Goodman Gilman's Pharmacol. Basis Therap.* **11**, 607–627 (2006).
3. Bradley, H. et al. Estimated number of people who inject drugs in the United States. *Clin. Infect. Dis.* **76**, 96–102 (2023).
4. Hall, E. W. et al. Estimated number of injection-involved drug overdose deaths, United States, 2000–2018. *Drug Alcohol Depend.* **234**, 109428 (2022).
5. Cornford, C. & Close, H. The physical health of people who inject drugs: complexities, challenges, and continuity. *Br. J. Gen. Pract.* **66**, 286-287 (2016).
6. Marks, L. R., Nolan, N. S., Liang, S. Y., Durkin, M. J. & Weimer, M. B. Infectious complications of injection drug use. *Med. Clin.* **106**, 187–200 (2022).
7. Powell, D., Alpert, A. & Pacula, R. L. A transitioning epidemic: how the opioid crisis is driving the rise in hepatitis c. *Health Aff.* **38**, 287–294 (2019).
8. Strathdee, S. A. et al. Preventing HIV outbreaks among people who inject drugs in the United States: plus ça change, plus ça même chose. *AIDS* **34**, 1997 (2020).
9. Wurcel, A. G. et al. Increasing infectious endocarditis admissions among young people who inject drugs. In *Open forum infectious diseases*, vol. 3 (Oxford University Press, 2016).
10. Sredl, M., Fleischauer, A. T., Moore, Z., Rosen, D. L. & Schranz, A. J. Not just endocarditis: hospitalizations for selected invasive infections among persons with opioid and stimulant use diagnoses-North Carolina, 2010–2018. *J. Infect. Dis.* **222**, S458–S464 (2020).
11. See, I. et al. National public health burden estimates of endocarditis and skin and soft-tissue infections related to injection drug use: a review. *J. Infect. Dis.* **222**, S429–S436 (2020).
12. Goodman-Meza, D. et al. Natural language processing and machine learning to identify people who inject drugs in electronic health records. In *Open Forum Infectious Diseases*, vol. 9, ofac471 (Oxford University Press US, 2022).
13. Edwards, A. E. & Collins Jr, C. B. Exploring the influence of social determinants on HIV risk behaviors and the potential application of structural interventions to prevent HIV in women. *J. Health Disparities Res. Pract.* **7**, 141 (2014).
14. Nijhawan, A. E. et al. Clinical and sociobehavioral prediction model of 30-day hospital readmissions among people with HIV and substance use disorder: beyond electronic health record data. *J. Acquired Immune Defic. Syndr. (1999)* **80**, 330 (2019).
15. Chen, M., Tan, X. & Padman, R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *J. Am. Med. Inform. Assoc.* **27**, 1764–1773 (2020).
16. Patra, B. G. et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J. Am. Med. Inform. Assoc.* **28**, 2716–2727 (2021).
17. Feller, D. J. et al. Detecting social and behavioral determinants of health with structured and free-text clinical data. *Appl. Clin. Inform.* **11**, 172–181 (2020).
18. Gottlieb, L. M., Tirozzi, K. J., Manchanda, R., Burns, A. R. & Sandel, M. T. Moving electronic medical records upstream: incorporating social determinants of health. *Am. J. Prevent. Med.* **48**, 215–218 (2015).
19. Weir, C. R. et al. A qualitative evaluation of the crucial attributes of contextual information necessary in ehr design to support patient-centered medical home care. *BMC Med. Inform. Decis. Mak.* **15**, 1–8 (2015).
20. Hayes, C. J. et al. Using data science to improve outcomes for persons with opioid use disorder. *Subst Abus.* **43**, 956–963 (2022).
21. Topaz, M., Murga, L., Bar-Bachar, O., Cato, K. & Collins, S. Extracting alcohol and substance abuse status from clinical notes: The added value of nursing data. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, 1056–1060 (IOS Press, 2019).
22. Peng, C. et al. Clinical concept and relation extraction using prompt-based machine reading comprehension. *J. Am. Med. Inform. Assoc.* **30**, 1486–1493 (2023).
23. Mahbub, M. et al. Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PloS ONE* **17**, e0262182 (2022).
24. Li, J., Sun, A., Han, J. & Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **34**, 50–70 (2020).
25. Mahbub, M. et al. cpgqa: A benchmark dataset for machine reading comprehension tasks on clinical practice guidelines and a case study using transfer learning. *IEEE Access* **11**, 3691–3705 (2023).
26. Eberts, M. & Ulges, A. Span-based joint entity and relation extraction with transformer pre-training. In *24th European Conference on Artificial Intelligence (ECAI 2020)* (Santiago de Compostela, Spain, 2020).
27. Pampari, A., Raghavan, P., Liang, J. & Peng, J. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2357–2368 (Association for Computational Linguistics, Brussels, Belgium, 2018). https://aclanthology.org/D18-1258.
28. Wang, Y. et al. Automated extraction of substance use information from clinical texts. In *AMIA Annual Symposium Proceedings*, vol. 2015, 2121 (American Medical Informatics Association, 2015).
29. Ridgway, J. P. et al. Natural language processing of clinical notes to identify mental illness and substance use among people living with HIV: retrospective cohort study. *JMIR Med. Inform.* **9**, e23456 (2021).
30. Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. Natural language processing: an introduction. *J. Am. Med. Inform. Assoc.* **18**, 544–551 (2011).
31. Torii, M. et al. Task formulation for extracting social determinants of health from clinical narratives. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.11386 (2023).
32. Feller, D. J., Zucker, J., Yin, M. T., Gordon, P. & Elhadad, N. Using clinical notes and natural language processing for automated HIV risk assessment. *J. Acquired Immune Defic. Syndr. (1999)* **77**, 160 (2018).
33. Lybarger, K., Ostendorf, M. & Yetisgen, M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J. Biomed. Inform.* **113**, 103631 (2021).

34. Han, S. et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J. Biomed. Inform.* **127**, 103984 (2022).

35. Yu, Z., Yang, X., Guo, Y., Bian, J. & Wu, Y. Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives. *Front. Public Health* **10**, 778463 (2022).

36. Feller, D. J., 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 Proceedings*, vol. 2018, 422 (American Medical Informatics Association, 2018).

37. Ahsan, H., Ohnuki, E., Mitra, A. & You, H. Mimic-sbdh: A dataset for social and behavioral determinants of health. In *Machine Learning for Healthcare Conference*, 391–413 (PMLR, 2021).

38. Lybarger, K. et al. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *J. Am. Med. Inform. Assoc.* **30**, 1389–1397 (2023).

39. Carrell, D. S. et al. Using natural language processing to identify problem usage of prescription opioids. *Int. J. Med. Inform.* **84**, 1057–1064 (2015).

40. Afshar, M. et al. External validation of an opioid misuse machine learning classifier in hospitalized adult patients. *Addict. Sci. Clin. Pract.* **16**, 1–11 (2021).

41. Afshar, M. et al. Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (smart-ai): a retrospective deep learning study. *Lancet Digit. Health* **4**, e426–e435 (2022).

42. Lingeman, J. M., Wang, P., Becker, W. & Yu, H. Detecting opioid-related aberrant behavior using natural language processing. In *AMIA Annual Symposium Proceedings*, vol. 2017, 1179 (American Medical Informatics Association, 2017).

43. Blackley, S. V. et al. Using natural language processing and machine learning to identify hospitalized patients with opioid use disorder. In *AMIA Annual Symposium Proceedings*, vol. 2020, 233 (American Medical Informatics Association, 2020).

44. Zhu, V. J. et al. Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy. *Health Inform. J.* **28**, 14604582221107808 (2022).

45. Poulsen, M. N., Freda, P. J., Troiani, V., Davoudi, A. & Mowery, D. L. Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing. *Front. Public Health* **10**, 850619 (2022).

46. Ward, P. J. et al. Enhancing timeliness of drug overdose mortality surveillance: a machine learning approach. *PloS ONE* **14**, e0223318 (2019).

47. Badger, J. et al. Machine learning for phenotyping opioid overdose events. *J. Biomed. Inform.* **94**, 103185 (2019).

48. Hazlehurst, B. et al. Using natural language processing of clinical text to enhance identification of opioid-related overdoses in electronic health records data. *Pharmacoepidemiol. Drug Saf.* **28**, 1143–1151 (2019).

49. Harris, D. R., Eisinger, C., Wang, Y. & Delcher, C. Challenges and barriers in applying natural language processing to medical examiner notes from fatal opioid poisoning cases. In *2020 IEEE International Conference on Big Data (Big Data)*, 3727–3736 (IEEE, 2020).

50. Goodman-Meza, D. et al. Development and validation of machine models using natural language processing to classify substances involved in overdose deaths. *JAMA Netw. Open* **5**, e2225593–e2225593 (2022).

51. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.* **34**, 301–310 (2001).

52. Lee, J. et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

53. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, 2019).

54. Mahbub, M., Srinivasan, S., Begoli, E. & Peterson, G. D. BioADAPT-MRC: adversarial learning-based domain adaptation improves biomedical machine reading comprehension task. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac508 (2022).

55. Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392 (Association for Computational Linguistics, Austin, Texas, 2016). https://aclanthology.org/D16-1264.

56. Joshi, M., Choi, E., Weld, D. & Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611 (Association for Computational Linguistics, Vancouver, Canada, 2017). https://aclanthology.org/P17-1147.

57. Harkema, H., Dowling, J. N., Thornblade, T. & Chapman, W. W. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. Biomed. Inform.* **42**, 839–851 (2009).

58. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task.* (eds Demner-Fushman, D., Cohen, K. B., Ananiadou, S. & Tsujii, J.) 58–65 (Association for Computational Linguistics, Florence, Italy, 2019).

59. Alsentzer, E. et al. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (eds Rumshisky, A., Roberts, K., Bethard, S. & Naumann, T.) 72–78 (Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019).

60. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Adv. neural Inf. Process. Syst.* **32**, 8026–8037 (2019).

61. Wolf, T. et al. Transformers: State-of-the-Art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, 2020).

62. Yasunaga, M., Leskovec, J. & Liang, P. LinkBERT: pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers* (eds Muresan, S., Nakov, P. & Villavicencio, A.) 8003–8016 (Association for Computational Linguistics, Dublin, Ireland, 2022).

63. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc. (HEALTH)* **3**, 1–23 (2021).

64. Raj Kanakarajan, K., Kundumani, B. & Sankarasubbu, M. Bioelectra: Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, 143–154 (2021).

65. Beltagy, I., Lo, K. & Cohan, A. SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620 (Association for Computational Linguistics, Hong Kong, China, 2019). https://aclanthology.org/D19-1371.

66. Alsentzer, E. et al. Publicly available clinical Bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78 (Association for Computational Linguistics, 2019).
67. Liu, Y. et al. Roberta: a robustly optimized Bert pretraining approach. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1907.11692 (2019).
68. UzZaman, N. et al. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 1–9 (2013).
69. Gong, J. J., Soleimani, H., Murray, S. G. & Adler-Milstein, J. Characterizing styles of clinical note production and relationship to clinical work hours among first-year residents. *J. Am. Med. Inform. Assoc.* **29**, 120–127 (2022).
70. Nadapana, V. & Kommanti, H. B. Investigating the role of named entity recognition in question answering models. In *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, 1–7 (IEEE, 2022).
71. Liu, A. T. et al. Qaner: prompting question answering models for few-shot named entity recognition. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.01543 (2022).
72. Mahbub, M. qa-system-for-injection-drug-use. https://doi.org/10.5281/zenodo.10428212 (2023).

Acknowledgements

Author contributions

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-024-00470-6.

Correspondence and requests for materials should be addressed to Maria Mahbub.

Peer review information *Communications Medicine* thanks John Osborne, Alvin Jeffrey, and David Goodman-Meza for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Cyber Resilience and Intelligence Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [2]Information Technology Services Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [3]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [4]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. [5]Program Evaluation and Resource Center, Office of Mental Health and Suicide Prevention, Department of Veterans Affairs, Menlo Park, CA, USA. [6]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [7]Veterans Affairs Great Lakes Health Care System, Westchester, IL, USA. [8]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Knoxville, TN, USA. ✉e-mail: mahbubm@ornl.gov