Clinical Evidence Engine: Proof-of-Concept For
A Clinical-Domain-Agnostic Decision Support Infrastructure

BOJIAN HOU and HAO ZHANG, Weill Cornell Medicine, USA

GUR LADIZHINSKY and ALI KAYYAL, Technion, Israel

STEPHEN YANG, Cornell University, USA

VOLODYMYR KULESHOV, Cornell Tech, USA

FEI WANG, Weill Cornell Medicine, USA

QIAN YANG*, Cornell University, USA

Abstruse learning algorithms and complex datasets increasingly characterize modern clinical decision support (CDS) systems. As a result, clinicians cannot easily or rapidly scrutinize the CDSS recommendation when facing a difficult diagnosis or treatment decision in practice. Over-trusting or under-trusting CDS recommendations are frequent, leading to preventable diagnostic or treatment errors. Prior research has explored supporting such assessments by explaining DST data inputs and algorithmic mechanisms. This paper explores a different approach: Providing precisely relevant, scientific evidence from biomedical literature. We present a proof-of-concept system, CLINICAL EVIDENCE ENGINE, to demonstrate the technical and design feasibility of this approach across three domains (cardiovascular diseases, autism, cancer). Leveraging *Clinical BioBERT*, the system can effectively identify clinical trial reports based on lengthy clinical questions (e.g., "risks of catheter infection among adult patients in intensive care unit who require arterial catheters, if treated with povidone iodine-alcohol"). This capability enables the system to identify clinical trials relevant to diagnostic/treatment hypotheses – a clinician's or a CDS's. Further, CLINICAL EVIDENCE ENGINE can identify key parts of a clinical trial abstract, including patient population (e.g., adult patients in intensive care unit who require arterial catheters), intervention (povidone iodine-alcohol), and outcome (risks of catheter infection). This capability opens up the possibility of enabling clinicians to 1) rapidly determine the match between a clinical trial and a clinical question, and 2) understand the result and contexts of the trial without extensive reading. We demonstrate this potential by illustrating two example use scenarios of the system. We discuss future work that can advance the design and ML performances of this system. We discuss the idea of designing DST explanations not as specific to a DST or an algorithm, but as a domain-agnostic decision support infrastructure.

1 INTRODUCTION

Biomedical literature can provide valuable "*decision support*" for clinicians. From best practices to clinical trial reports, the literature contains scientifically proven information that can aid numerous diagnostic and therapeutic dilemmas across all clinical domains. With the rise of Evidence-Based Medicine (EBM), clinicians increasingly turn to literature at point-of-care to inform their decisions [38, 41]; Medical students receive training on how to formulate their clinical questions into good search terms, for example, training on the PICO (Population, Intervention, Comparison, and Outcome) clinical knowledge representation framework [30].

Interestingly, literature is rarely in the spotlight of clinical decision support (CDS) research. With the explosive growth in machine learning (ML), CDS systems are increasingly characterized by patient-data-driven inferences and

---

*Contact author: Email - qy242@cornell.edu

abstruse risk models, each tailored for one clinical decision or one clinical domain. In this context, literature-based systems can appear particularly valuable for clinicians today. They can complement other CDS systems and offer scientific evidence that clinicians can easily understand [8, 14, 47]. They can support many vastly different clinical decisions and domains, including those in data-poor or resource-constrained hospitals.

Document-level literature retrieval systems such as PubMed and Google Scholar have proven their worth in clinical practice. However, at points-of-care, clinicians need far more fine-grained tools to identify information that is *precisely applicable* to their patient case and clinical question at hand [17, 30]. Clinical decision-making is a continuous and iterative process, consisting of a series of micro-decisions [4, 44, 49]. For each micro decision in practice, clinicians could only afford 2-3 minutes on literature search [12]. To be useful, literature systems need to be able to directly answer point-of-care clinical questions such as"*What are the risks of catheter infection if an adult patient in intensive care unit who requires arterial catheters is treated with povidone iodine-alcohol rather than chlorhexidine–alcohol?*" [30] Existing systems (e.g., PubMed, *Trialstreamer* [36]) struggle with processing such complex queries. Searching the query above on PubMed simply returns no results, not to mention eliciting key information from the literature.

This work aims to more precisely address point-of-care clinical questions with biomedical literature. We present CLINICAL EVIDENCE ENGINE, a proof-of-concept system that demonstrates the technical feasibility of achieving this goal. On clinical trial reports of three domains (cardiovascular diseases, autism, cancer), CLINICAL EVIDENCE ENGINE can: **(Capability 1)** Identify relevant clinical trial reports based on complex, long, clinical-question-like search queries (up to 512 words), queries such as "*Would the addition of radiotherapy on top of androgen-deprivation therapy lead to higher risk of bowel toxicity of an adult male patient with locally advanced prostate cancer?*" The literature retrieval model of CLINICAL EVIDENCE ENGINE achieves an accuracy of 99.44%, when evaluated on synthetic queries based on an established expert-annotated literature dataset [35]. **(Capability 2)** Identify critical information in clinical trial report abstracts can serve as clinical evidence, i.e., PICO (Population, Intervention, Comparison, and Outcome) information [30]. The PICO classification model of CLINICAL EVIDENCE ENGINE achieves a F1 score of 0.74. Both models outperform existing state-of-the-art models.

These newfound technical capabilities open up new design and research opportunities around literature-based CDS; around designing clinical decision supports as domain-agnostic, intelligent information infrastructure, rather than decision- or domain-specific applications. We concretize these opportunities and questions through two example use scenarios of CLINICAL EVIDENCE ENGINE : aid clinicians in crutinizing (i) their self-derived decision hypothesis and (ii) an abstruse patient-data-based risk model.

This paper makes two contributions. First, it demonstrates novel bioNLP capabilities of harnessing biomedical literature as point-of-care decision. Key to this technical advance is the integrative use of a clinical knowledge representation framework (i.e. PICO [30]) and large pretrained language models (i.e. Clinical BioBERT [2]). Second, CLINICAL EVIDENCE ENGINE offers an initial design exemplar of a domain-agnostic, intelligent information infrastructure. It offers an alternative perspective to the traditional idea of clinical decision supports as decision- or domain-specific applications. It can serve as a valuable point of reference for the research discourse on future AI-infused healthcare.

2 RELATED WORK

2.1 Biomedical Literature Supports Clinical Decision-Making

Clinicians routinely consult biomedical literature for decision-making. When facing diagnostic and prognostic dilemmas, clinicians search the literature – most often on PubMed – for valuable clinical evidence such as clinical trial reports, best

practices, expert-annotated case studies, and biological explanations [17, 21, 27]. Such evidence compliments clinician judgments and patient value, forming the cornerstones of Evidence-Based Medicine [19].

At points of care, clinicians search literature for *highly-specific* information to address the clinical question at hand. Such searches are often very challenging given the overwhelming amount of literature available [17]. To address this challenge, clinicians and medical students receive mandatory trainings on how to translate messy clinical situations into effective literature search queries, for example, by using the PICO framework [18, 30].

Question Type	PICO Template For Formulating Clinical Questions
Therapy/intervention	In _____(**P**opulation), what is the effect of _____(**I**ntervention), compared with _____(**C**omparator) on _____(**O**utcome)?
Diagnosis/assessment	For _____(population), does _____(tool/procedure) yield more accurate or appropriate diagnostic/assessment information than _____(comparator tool/procedure) about _____(outcome)?
Prognosis	For _____(population), does _____ (disease/condition) relative to _____(comparator disease/condition) increase the risk for _____(outcome)?

Table 1. Part of the PICO (Population, Intervention, Comparison, and Outcome), a clinical knowledge representation framework [18, 30]. Clinicians receive training on how to use such frameworks to effectively search biomedical literature for clinical evidence.

Besides specificity, standards of *rigor* also drive clinicians' literature search. Healthcare communities differentiate the levels of evidence in biomedical literature [7]. They consider population-level evidence (e.g., randomized controlled trials and systematic reviews) more rigorous and trustworthy than isolated observations (e.g., peer-reviewed case studies) [38, 41]. The importance of rigor is also evident in clinicians' choice of literature search tools. In research, clinicians built expert-curated-and-maintained literature databases [1, 15]. In practice, more clinicians use PubMed (a keyword-matching-based search engine) than Google Scholar. Despite higher accuracy, the latter is considered less rigorous, as its rankings consider article popularity [39, 40].

Empirical research is sparse on how clinicians use literature in practice. This is in sharp contrast with nonmedical domains, where sense-making and HCI research have extensively studied people's organic information search/foraging and decision-making. They created tools to support sense-making, such as novel visualization techniques and crowd-sourced knowledge representation structure [9, 10, 37].

2.2 Clinical Decision Support and Explainable AI in HCI Research

Computational decision support (CDS) systems are systems that assist point-of-care decision-making. They promise to improve patient outcomes [31], reduce medical error [20], and reduce healthcare disparities [3]. In recent years, with the increasing digitalization of Electronic Health Records (EHR) and the explosive growth in machine learning (ML), clinician-facing CDS research is increasingly characterized by complex algorithms and patient-data-driven inferences [22, 32]. Since the focus of this work is on biomedical literature, we only briefly review this body of work, particularly the challenges it has reported. This is not meant to be a criticism. Instead, these challenges motivated us to use biomedical literature to *complement* non-literature-based CDS systems. More comprehensive reviews of CDS work and their achievements can be found elsewhere [14, 32, 45, 48].

- *Interpretability and accessibility to clinicians.* Abstruse, data-driven algorithms increasingly characterize modern CDS. Clinicians often found these systems too time-consuming to understand, their explanations overly complicated [5, 13, 47]. Over-trusting or under-trusting CDS recommendations are frequent, leading to preventable diagnostic or treatment errors [6, 16, 46, 47].
- *Sense of rigor.* Clinicians did not always trust the diagnostic/treatment predictions, even when they fully understand how the ML correctly generated the prediction [46, 47]. This is because the standards of rigor for clinical evidence are often at odds with the basis on which the rigor of ML is premised. For example, inference-based diagnostic predictions do not qualify as "*empirical evidence*" under the levels of evidence pyramid [7]. Empirical research reported that, when judging the trustworthiness of CDS, clinicians asked whether it has been published in top-tier clinical journals [47].
- *Generalizability challenges.* Researchers most often tailored data-driven CDS models and designed explanations for particular clinical decisions, data types, and/or algorithms [8, 46]. Given the multitude of clinical decisions involved in caring for each patient, it can seem that, at point-of-care, future clinicians will need to make sense of multiple CDS predictions in quick succession (e.g., blood-test-based diagnostic support, computer vision-based X-ray reading, tabular-disease-trajectory prediction, etc.). They are also responsible for scrutinizing each prediction and accounting for its potential biases. This will be extremely difficult.

Within such a milieu, the inherent characteristics of literature – *scientifically-proven, domain-agnostic, accessible* – can be particularly valuable for clinicians today [8, 14, 47].

2.3 Mining Biomedical Literature in BioNLP Research

Algorithmically retrieving and processing biomedical literature are challenging tasks at the frontier of NLP research. In comparison to other documents, biomedical literature includes complex terminologies, concepts, and relationships that even scientists may not fully understand [33]. Popular literature mining systems, such as PubMed, utilize keyword matching techniques. They can struggle with many search queries that clinicians need at point-of-care [21].

Large pretrained language models promise substantial advances on these fronts. For example, BERT [26] can perform literature mining tasks such as documents retrieval and key information extraction. *Clinical BioBERT* [25] can even more effectively capture biomedical and clinical knowledge as it has been fine-tuned with biomedical and clinical documents. Literature datasets created for enhancing Evidence-Based Medicine (EBM) can also catalyze novel literature mining and information retrieval capabilities. One such dataset is the EBM-NLP dataset [34], a corpus of 4991 clinical trial report abstracts annotated with PICO elements. The abstracts are extracted from PubMed and focus on cardiovascular diseases, cancer, and autism. Medical experts and crowd-workers collaboratively annotated the PICO elements. These PICO annotations can serve as the ground truth for many EBM-related NLP research efforts.

Researchers have started to leverage these large language models and datasets in creating novel biomedical literature mining systems. For example, a COVID-19 literature mining system that can surface emergent research directions on the topic [23] and a question answering systems also on scholarly COVID literature [42]. The most closely related to point-of-care is TrailStreamer [28], which uses both ML and rule-based methods to extract PICO information from human-subject Randomized controlled trials (RCTs) reports. While advancing on classification performance, Trialstreamer has limited success in retrieving literature with long or highly-specific queries. For example, Trialstreamer can only define a patient population based on a single clinical condition (e.g., all patients with diabetes). This limits

the system from identifying RCTs that match more meaningfully with patient cases, e.g., according to their medical histories, commodities, and demographics.

3 SYSTEM DESIGN GOALS AND OVERVIEW

We set out to create a literature-based system that can support clinicians' point-of-care decision-making. Drawing up prior work, we had two goals: (1) To process long, complex clinical questions as literature search queries, e.g., "*what are the risks of catheter infection if an adult patient in intensive care unit who requires arterial catheters is treated with povidone iodine-alcohol rather than chlorhexidine–alcohol?*"; (2) To identify not only relevant literature documents but the clinically-relevant information within. These orientations require state-of-the-art bioNLP capabilities. They also differ from the convention of CDS designs, in which each system is tailored for particular clinical decisions or domains.

Towards these goals, we designed a system architecture that integrated both large pre-trained language models (i.e. Clinical BioBERT [2]) and a clinical knowledge representation framework (i.e. the PICO framework and annotations [2, 30].) The former offers the capabilities to process long, complex biomedical texts across domains (up to 512 words); The latter allows us to identify clinically relevant information from the literature. Figure 1 illustrates the system architecture design.
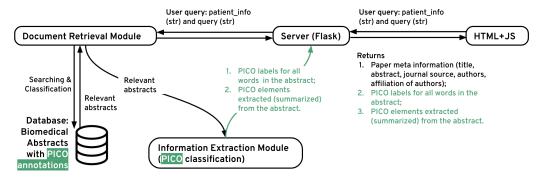

Fig. 1. Clinical Evidence Engine system architecture. Key to its design is the organic integrative use of large pretrained language models (i.e. Clinical BioBERT [2]) and a clinical knowledge representation framework (i.e. PICO [30], highlighted in green.

Clinical Evidence Engine 's backend includes two modules. One is the *Document Retrieval Module*, which retrieves relevant biomedical literature articles according to long, complex clinical questions as search queries. To the best of our knowledge, no prior work has created such retrieval models. The other is the *Information Extraction Module*, which identifies and extracts the PICO elements within the article's abstract. Prior research reported challenges in balancing such models' precision and recall, thus causing relatively low F1 scores [34]. We aim to address this challenge.

Given our focus on biomedical literature, we chose to train our models using the EBM-NLP dataset [34], a clinical trial report dataset with expert annotations of PICO elements for each report. It focuses on three clinical domains: cardiovascular disease, autism, and cancer. As a result, our system will also focus on trial reports on these domains.

4 DEVELOPING A PROOF-OF-CONCEPT SYSTEM

4.1 Point-of-Care Biomedical Literature Retrieval

We set out to train a document retrieval model by fine-tuning the Clinical BioBERT models on concatenated *(query, abstract)* pairs:

$$[CLS] \ h(\text{Query}) \ [SEP] \ h(\text{Abstract}) \ [SEP] \tag{1}$$

where h is the BERT model, "[CLS]" and "[SEP]" are the special signs for the input of BERT. At test time, we used this model to predict the probability that a query is associated with each abstract in the dataset. The model then returns a literature document ranking based on the probability scores. Fig. 2 illustrates this process in detail.
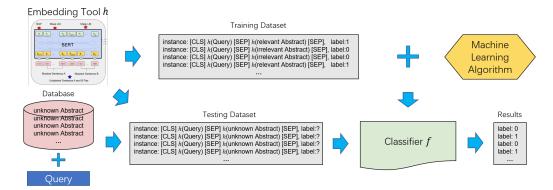

Fig. 2. Workflow of the document retrieval module.

One critical challenge, however, is that the EBM-NLP dataset does not include search queries for such training. To address this challenge, we generated synthetic search queries by concatenating the expert-annotated PICO elements in literature abstracts. As a result, each literature abstract represents a perfect match with one synthetic search query. The match represents one positive training instance "[CLS] h(Query) [SEP] h(**relevant** Abstract) [SEP]", while others represent negative instances "[CLS] h(Query) [SEP] h(**irrelevant** Abstract) [SEP]." These synthetic queries can effectively simulate clinician search queries, because the PICO framework is the best practice with which clinicians formulate point-of-care clinical questions into search questions [18, 30].

Clinical questions in practice do not always include all PICO elements, for example, some diagnostic questions do not have a comparator [18]. In this light, we generated synthetic queries that included all PICO elements as well as those that only included a subset. This data generation process (Fig. 3) generated 4 positive instances and 4 negative instances for each abstract.

Next, we randomly selected 4000 abstracts for training and used the rest 991 for testing. We trained the retrieval model on the $4000 \times 8 = 32000$ training instances and evaluate it on the $991 \times 8 = 7928$ testing instances. We run 5 times with different splittings of the training and testing datasets.

This evaluation process revealed a F1 score of 0.9945 for positive document relevance, and 0.9944 for negative document relevance. It also shows that quantitatively, our model outperforms the best results of the keyword matching approach [1], a common strategy used by popular literature retrieval systems. Table 2 details the performance comparison

---

[1]In our experiment, keyword-based approaches achieved their best performance when the system retrieves only the documents that include 40% of the keywords appeared in a search query. This performance is inferior to our models, as detailed in Table 2.
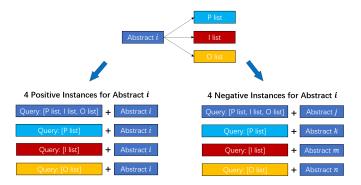
Fig. 3. The generation process of the retrieval model dataset, where $i \neq j \neq k \neq m \neq n$. For abstract i, we first extract its PICO elements and then generate the positive and negative instances with queries of different lengths.

	Accuracy	F1 for Negative Relevance	F1 for Positive Relevance
This work	0.9944±.00089	0.9945±.00091	0.9944±.00087
Best Baseline	0.9535	0.9549	0.9519

		Predicted Negative Relevance	Predicted Positive Relevance
This work	True Negative	3940	24
	True Positive	21	3943
Best Baseline	True Negative	3905	59
	True Positive	310	3654

Table 2. Top: Retrieval model accuracy and F1 with standard deviation over five runs. Bottom: Retrieval model confusion matrix (one run).

between this work and the best results from keyword matching approaches, in terms of accuracy, F1 score, confusion matrix. One limitation of this evaluation is that it focused solely on whether the system possesses the ability to identify all relevant documents. Confusion matrix, accuracy and F1 score are sufficient to measure this ability. However, it did not assess the model's ranking abilities (e.g., calculating the NDCG or AUC values). Unfortunately, there exists no ground truth datasets that could enable such evaluation.

4.2 Salient Information (PICO) Extraction Module

The information extraction module aims to identify and summarize the salient elements of clinical evidence from literature abstracts. Based on the PICO framework, we consider Population (P), Intervention/Comparator (I/C), and Outcome (O) as salient information [18, 30]. We used Clinical BioBERT tokenizer to tokenize the words into tokens that are expressed as numerical vectors. Using "(token, label)" pairs as training data, we trained a linear four-class classifier with the EBM-NLP dataset; a classifier that predicts whether each token in the abstract describes P, I/C, or O. After obtaining the classification results, the system groups adjacent words with same annotations and remove duplicates to generate the final PICO phrases. Fig. 4 summarizes the information extraction module workflow.
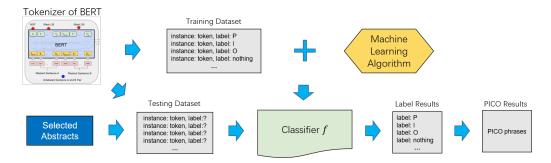
Fig. 4. Illustration of information extraction module.

Model	F1	Precision	Recall
LogReg	0.45	0.31	0.82
LSTM-CRF	0.68	0.70	0.66
LSTM-CRF-BERT	0.68	0.69	0.66
This Work	0.73	0.70	0.76

Table 3. Key information (PICO) extraction model performance. Leveraging Clinical BioBERT, our model outperforms state-of-the-art methods on F1 score.

We evaluated this model and compared it to the three state-of-the-art PICO classifiers in [34]. These include logistic regression (LogReg), LSTM-CRF [24] and LSTM-CRF-BERT, which uses BERT as the embedding tool for LSTM-CRF. For fair comparison, we tested all four models on the extra 200 withheld testing data from EBM-NLP [34]. Our model outperforms other methods in F1 score, achieving balanced precision and recall (Figure 3.)

5 EXAMPLE INTERACTION SCENARIOS: DESIGN AND RESEARCH OPPORTUNITIES IN HARNESSING LITERATURE AS DECISION SUPPORT

We have so far described how Clinical Evidence Engine demonstrates two novel technical capabilities necessary for harnessing biomedical literature as point-of-care decision support: (1) Identifying relevant literature documents based on long, complex search queries and (2) extracting precisely relevant information (i.e., PICO elements) from the identified documents. The newfound technical capabilities mark a clear design space in supporting biomedical literature sense-making. They also open up new HCI research opportunities around supporting clinical decision-making *across domains* with literature. To concretize these design and research opportunities, below we discuss them via two example use scenarios of Clinical Evidence Engine : helping clinicians to (i) answer their self-derived clinical questions and to (ii) scrutinize a deep-learning-based risk model.

5.1   Use Scenario 1: Supporting Clinicians Questions as Search Queries

> **Scenario:** The clinician team just diagnosed a male adult patient with prostate cancer. At that point, the cancer cells have only been developing locally. Based on their clinical acumen and standard practice, the clinicians have decided to deploy androgen-deprivation therapy (ADT). However, they are unsure, for this particular patient, whether the addition of radiation therapy (RT) would further improve their chance of survival. This is a critical decision, as radiation has substantial side effects and should not be used casually.

Facing this therapeutic dilemma, clinicians start searching for relevant trials on CLINICAL EVIDENCE ENGINE using a nuanced description of the scenario as search term.

> **Clinician's search query:** *Would the addition of "radiotherapy" on top of "androgen-deprivation therapy" help improve the survival of an adult male patient with "locally advanced prostate cancer"?*

Technical capabilities.   Based on this query, CLINICAL EVIDENCE ENGINE can identify a list of relevant randomized controlled trial reports. The highest ranking result is the report "*Final report of the intergroup randomized study of combined androgen-deprivation therapy plus radiotherapy versus androgen-deprivation therapy alone in locally advanced prostate cancer*" [29]. It precisely matches the clinical question in terms of clinicians' population, interventions, and outcome of interest. Further, the PICO classifier of CLINICAL EVIDENCE ENGINE extracts the following information and can concisely answer the clinical question raised.

- **P**opulation: "*patients with locally advanced prostate cancer*", and more specifically, "*Patients with T3-4, N0/Nx, M0 prostate cancer or T1-2 disease with either prostate-specific antigen (PSA) of more than 40 µg/L or PSA of 20 to 40 µg/L plus Gleason score of 8 to 10*";
- **I**ntervention and **C**omparator: "*Combined Androgen-Deprivation Therapy Plus Radiotherapy Versus Androgen-Deprivation Therapy Alone*". specifically, "*lifelong ADT alone*";
- **O**utcome: "*overall survival*", "*deaths from prostate cancer*", and "*frequency of adverse events related to bowel toxicity*".

Design opportunities.   How can literature-based CDS systems best support clinicians' literature sense-making and clinical decision-making with the retrieved PICO elements? Extensive prior HCI work has studied how to support information foraging, sense-making, and decision-making [9, 10]. CLINICAL EVIDENCE ENGINE 's novel technical abilities illuminate a clear design space for expanding this research into biomedical domains. For example, prior research has shown that scaffolding search process and results can reduce users' cognitive efforts, enabling them to build up a deeper understanding of the decision being made [9]. PICO information extraction capabilities enable such scaffolding, for example, by allowing clinicians to compare populations and outcomes on comparable interventions across studies (Figure 5 bottom right). Prior research has also built tools that allow users to create a collection of composable and reusable "lenses" to reflect their different latent interests [10]. Such tools were effective in improving users' depth of information understanding. Future literature-based CDS tools can explore enabling clinicians to create dynamic and reusable lenses. In this particular search scenario, clinicians may find the patient-population-match lens valuable, as they can use it to identify the trials that most closely align with the patient at hand in terms of cancer type, severity, and spread. Later in the search process, clinicians can shift to a temporal lens, rapidly and effectively examining the temporal progression of ADT-plus-radiotherapy treatment effects (Figure top right). Finally, visualizing the key clinical
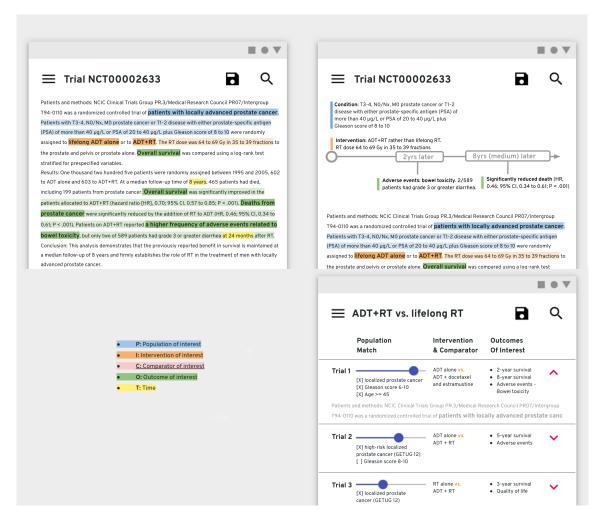
Fig. 5. Example interface designs of CLINICAL EVIDENCE ENGINE as it directly addresses clinicians' points-of-care clinical questions. Each embodies a sense-making support design lesson in prior HCI research; All are enabled by the ML capabilities described in section 4.

evidence in the literature exemplifies near-term, pragmatic design solutions. It can offer perceptual cues for clinicians and expedite information foraging at points of care [11] (Figure top left).

Open research opportunities. PICO represents only one way of summarizing salient clinical evidence in biomedical literature. Future research should advance on the machine learning models described in this work by systematically investigating other ontological clinical information frameworks (e.g., levels of evidence) and assessing their effectiveness in point-of-care decision support. Other summarization or knowledge extraction techniques beyond supervised learning can also be valuable. There is a real need for curating HCI- and CDS-oriented biomedical literature datasets. It is critical for such human-centered ML research to advance further.

Clinicians not only need concise and precisely relevant clinical evidence, but they also need enough contextual integrity to interpret and act on this evidence. So far, little to no empirical research has studied how clinicians forage and utilize clinical evidence from literature *in practice.* This should be a critical research question in CDS design and research. It can help illuminate clinicians' sequential search behaviors and high-level strategies, further improving the retrieval and extraction performances. It can also inform summarization or visualization design strategies and ensure the correct interpretation of extractive literature evidence for point-of-care decision-making.

5.2 Use Scenario 2: Harnessing Literature Along with Other Decision Support Tools

---

Scenario: A nurse practitioner in a small, rural hospital is assessing whether a 15-month-old girl with disruptive behavior history has Autism Spectrum Disorder (ASD). She does so by conducting the standard Autism Diagnostic Interview (ADI), Autism Diagnostic Observation Schedule (ADOS), as well as observing her interaction with her family members and strangers in the hospital. While the standard tests (ADI and ADOS) indicate autism, the nurse practitioner notices that the girl has a secure attachment relationship with her parents. Does that mean this girl does not have autism, considering that autistic children most often exhibit pervasive deficits in social, affective, and communicative behaviors? The nurse practitioner wonders.

Other decision support systems at play. In this context, the nurse practitioner looks up her decision support systems. One autism detection CDS she uses[43] analyzes a 3-minute video of the girl interacting with her family. It predicts that the girl is 54% likely to have ASD. This system is trained on large video databases of children with and without ASD and has a 92% accuracy. However, the system deploys eight complex ML models to make a diagnostic prediction collectively. The nurse practitioner struggles to decide whether to trust this prediction or not.

Facing this diagnostic dilemma, the nurse practitioner opens up Clinical Evidence Engine and gives it permission to use this patient's medical industry and the featurized video data that the autism detection CDS has produced.

> **Auto-generated search query:** *Female child, "disruptive behaviors", "secure parental attachment", "Autism Spectrum Disorder", "social behavior after separation from parents."*

Technical capabilities. Based on this query, Clinical Evidence Engine identifies a list of relevant biomedical literature documents. The highest ranking result, a publication named "*A Parent-Mediated Intervention That Targets Responsive Parental Behaviors Increases Attachment Behaviors in Children with ASD: Results from a Randomized Clinical Trial*", which largely matches the clinical question in terms of clinicians' population, interventions, and outcome of interest. More importantly, the system extracted the P,I, and O elements that together read "*[...]the attachment behaviors of children with Autism Spectrum Disorder (ASD) show striking similarities to those of typically developing children.*" – A sentence that addresses the nurse practitioner's question and confirms a positive diagnosis.

- **P**opulation: "*children aged 12 and 24 months diagnosed with ASD*", "children with Autism Spectrum Disorder (ASD) ".
- **I**ntervention: "*Focused Playtime Intervention (FPI)*", a type of "*parent-mediated intervention*"
- **C**omparator: n/a
- **O**utcome: "*parental perceptions of child attachment*", "*attachment related outcomes*", "*attachment-related behaviors*", "*similarities to those of typically developing children*"

Design opportunities. Abstruse algorithms and complex patient-data-driven inferences increasingly characterize modern CDS systems. Clinicians have frequently reported challenges in understanding the trustworthiness of such

systems and their predictions, especially under the time constraints of busy clinical work [5, 13, 47]. The multi-learning-algorithm, video-based autism detection system offers merely one example.

Via Clinical Evidence Engine , this work proposed biomedical literature as an alternative approach to providing "explainability" to these complex CDS systems, aiding clinicians in scrutinizing the correctness of each prediction. As shown in the example scenario, *to clinicians*, evidence from biomedical literature can be much more easily understandable and intuitively convincing of algorithmic inner-workings. This work opens up a clear design space around supporting otherwise-abstruse CDS predictions with evidence from clinical literature, for example, helping clinicians scrutinizing ML predictions by surfacing the clinical-trial-proven casual relations between ML features and its predicted diagnoses (Figure 6.)
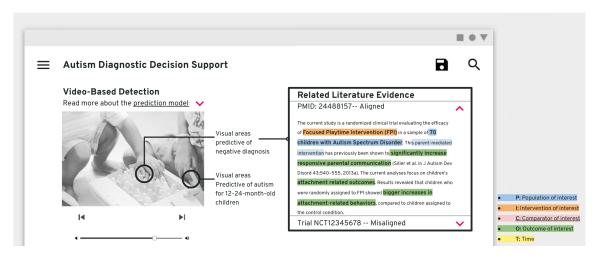

Fig. 6. Example interface designs of Clinical Evidence Engine as it aids clinician decision-making alongside other diagnostic or prognostic CDS predictions. The literature contains clinical evidence that can validate or invalidate many CDS predictions, therefore helping clinicians scrutinizing them. This functionality can be particularly valuable for complex systems such as deep-learning-based video analysis because their predictions and predictive mechanism are difficult to explain to clinicians.

Open research opportunity: Designing a better blend of heterogeneous clinical decision supports. A key premise of this work is to explore what an AI-*infused* clinical decision-making process might look like in future healthcare. While research has created numerous CDS systems, most focused on one system, one clinical decision, and one clinical domain. However, clinical decision-making is a continuous and iterative process; It consists of a series of micro-decisions. These micro-decisions are often cross-modal and cross-disciplinary, therefore involving distinct CDS systems and potential risks. We encourage future research to investigate how heterogeneous AI decision supports (e.g., literature-based and EHR-based) can best collaborate with clinician teams, forming an effective multi-AI, multi-clinician team. This example scenario offers a small first step towards this ambitious goal.

6 CLOSING NOTES

Abstruse learning algorithms and complex datasets increasingly characterize modern decision support system. As a result, clinicians cannot easily or rapidly scrutinize the CDSS recommendation when facing a difficult diagnosis or treatment decision in practice. Over-trusting or under-trusting CDSS recommendations are frequent, leading to

preventable diagnostic or treatment errors. Prior research has explored supporting such assessments by explaining DST data inputs and algorithmic mechanisms. This paper explores a different approach: By providing precisely relevant, scientific evidence from biomedical literature. We present a proof-of-concept system, Clinical Evidence Engine, to demonstrate the technical and design feasibility of this approach across three domains (cardiovascular diseases, autism, cancer). It can effectively identify clinical trial reports based on lengthy clinical questions (e.g., "risks of catheter infection among adult patients in intensive care unit who require arterial catheters, if treated with povidone iodine-alcohol"). This capability enables the system to identify clinical trials relevant to diagnostic/treatment hypotheses – a clinician's or a DST's. Further, Clinical Evidence Engine can identify key parts of a clinical trial abstract, including patient population (e.g., adult patients in intensive care unit who require arterial catheters), intervention (povidone iodine-alcohol), and outcome (risks of catheter infection). Through two example use scenarios of the system, we have demonstrated the many design opportunities and open research questions that this capability opens up.

At a higher level, this work proposes a future where intelligent literature tools can serve as a decision support infrastructure and support many clinical decisions across domains. Such an information infrastructure should be valuable both independently (as illustrated in use scenario 1) and when supporting other intelligent systems, particularly for practitioners in rural or low-resource hospitals where data-intensive CDS is less available (scenario 2). A decision-support infrastructure – because it operates at a PubMed scale – can have an outsized impact on clinical practice and improving the quality of patient care.

REFERENCES

[1] [n. d.]. Journal Club for iPhone/Android. https://wikijournalclub.org/app/
[2] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *CoRR* abs/1904.03323 (2019).
[3] Sam Amirfar, John Taverna, Sheila Anane, and Jesse Singer. 2011. Developing public health clinical decision support systems (CDSS) for the outpatient community in New York City: our experience. *BMC Public Health* 11, 1 (2011), 1–8.
[4] Louise Bate, Andrew Hutchinson, Jonathan Underhill, and Neal Maskrey. 2012. How clinical decisions are made. *British journal of clinical pharmacology* 74, 4 (2012), 614–620.
[5] Michael Z. Bell. 1985. Why Expert Systems Fail. *The Journal of the Operational Research Society* 36, 7 (1985), 613–619. http://www.jstor.org/stable/2582480
[6] A Billis. 2011. Identification of Gleason pattern 5 on prostatic needle core biopsy: frequency of underdiagnosis and relation to morphology. *International braz j urol: official journal of the Brazilian Society of Urology* 37, 6 (2011), 790.
[7] Patricia B Burns, Rod J Rohrich, and Kevin C Chung. 2011. The Levels of Evidence and their role in Evidence-Based Medicine. *Plastic and reconstructive surgery* 128, 1 (2011), 305.
[8] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.
[9] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. [n. d.]. Mesh: Scaffolding Comparison Tables for Online Decision Making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2020-10-20) *(UIST '20)*. Association for Computing Machinery, 391–405. https://doi.org/10.1145/3379337.3415865
[10] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. [n. d.]. SearchLens: composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2019-03-17) *(IUI '19)*. Association for Computing Machinery, 498–509. https://doi.org/10.1145/3301275.3302321
[11] Ed H. Chi, Peter Pirolli, Kim Chen, and James Pitkow. 2001. Using Information Scent to Model User Information Needs and Actions and the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 490–497. https://doi.org/10.1145/365024.365325
[12] Guilherme Del Fiol, T Elizabeth Workman, and Paul N Gorman. 2014. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA internal medicine* 174, 5 (2014), 710–718.
[13] Srikant Devaraj, Sushil K Sharma, Dyan J Fausto, Sara Viernes, Hadi Kharrazi, et al. 2014. Barriers and facilitators to clinical decision support systems adoption: A systematic review. *Journal of Business Administration Research* 3, 2 (2014), 36.
[14] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Legare, Trudy van der Weijden, Carmen L Lewis, Richard M Wexler, et al. 2013. " Many miles to go...": a systematic review of the implementation of patient decision support interventions into

routine clinical practice. *BMC Medical Informatics and Decision Making* 13 (2013).

[15] Gary N Fox and Nashat S Moawad. 2003. UpToDate: a comprehensive clinical database. *Journal of family practice* 52, 9 (2003), 706–710.

[16] Michael Goodman, Kevin C. Ward, Adeboye O. Osunkoya, Milton W. Datta, Daniel Luthringer, Andrew N. Young, Katerina Marks, Vaunita Cohen, Jan C. Kennedy, Michael J. Haber, and Mahul Amin. 2012. Frequency and determinants of disagreement and error in gleason scores: A population-based study of prostate cancer. *Prostate* 72, 13 (15 9 2012), 1389–1398. https://doi.org/10.1002/pros.22484

[17] Gordon Guyatt, Drummond Rennie, Maureen Meade, Deborah Cook, et al. 2002. *Users' guides to the medical literature: a manual for evidence-based clinical practice.* Vol. 706. AMA press Chicago.

[18] Judith Haber. 2018. PART II Processes of Developing EBP and Questions in Various Clinical Settings. *Evidence-Based Practice for Nursing and Healthcare Quality Improvement-E-Book* (2018), 31.

[19] C Harris and T Turner. 2011. Evidence-Based Answers to Clinical Questions for Busy Clinicians. In *Centre for Clinical Effectiveness.* Monash Health, 1–32.

[20] Peng Li Jia, Pei Fang Zhang, Han Dong Li, Long Hao Zhang, Ying Chen, and Ming Ming Zhang. 2014. Literature review on clinical decision support system reducing medical error. *Journal of evidence-based medicine* 7, 3 (2014), 219–226.

[21] S Kendall. 2017. PubMed, Web of Science, or Google Scholar? A behind-the-scenes guide for life scientists. *Research Guides* (2017).

[22] Gilad J Kuperman, Anne Bobb, Thomas H Payne, Anthony J Avery, Tejal K Gandhi, Gerard Burns, David C Classen, and David W Bates. 2007. Medication-related clinical decision support in computerized provider order entry systems: a review. *Journal of the American Medical Informatics Association* 14, 1 (2007), 29–40.

[23] D Lahav, JS Falcon, B Kuehl, S Johnson, S Parasa, N Shomron, DH Chau, D Yang, E Horvitz, DS Weld, et al. 2021. A Search Engine for Discovery of Scientific Challenges and Directions. (2021).

[24] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

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

[26] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations.*

[27] Zhiyong Lu. 2011. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011 (2011).

[28] Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association* 27, 12 (2020), 1903–1912.

[29] Malcolm D Mason, Wendy R Parulekar, Matthew R Sydes, Michael Brundage, Peter Kirkbride, Mary Gospodarowicz, Richard Cowan, Edmund C Kostashuk, John Anderson, Gregory Swanson, et al. 2015. Final report of the intergroup randomized study of combined androgen-deprivation therapy plus radiotherapy versus androgen-deprivation therapy alone in locally advanced prostate cancer. *Journal of Clinical Oncology* 33, 19 (2015), 2143.

[30] Rasoul Masoomi. 2012. What is the Best Evidence Medical Education? *Research and Development in Medical Education* 1, 1 (2012), 3–5.

[31] Sharare Taheri Moghadam, Farahnaz Sadoughi, Farnia Velayati, Seyed Jafar Ehsanzadeh, and Shayan Poursharif. 2021. The effects of clinical decision support system for prescribing medication on patient outcomes and physician practice performance: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 1–26.

[32] Zlatana Nenova and Jennifer Shang. 2021. Chronic Disease Progression Prediction: Leveraging Case-Based Reasoning and Big Data Analytics. *Production and Operations Management* (2021).

[33] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics* 9, 1 (2018), 1–13.

[34] Benjamin E. Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain James Marshall, Ani Nenkova, and Byron C. Wallace. 2018. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 197–207.

[35] Benjamin E. Nye, Ani Nenkova, Iain J. Marshall, and Byron C. Wallace. [n. d.]. Trialstreamer: Mapping and Browsing Medical Evidence in Real-Time. ([n. d.]). http://arxiv.org/abs/2005.10865v1

[36] Benjamin E Nye, Ani Nenkova, Iain J Marshall, and Byron C Wallace. 2020. Trialstreamer: mapping and browsing medical evidence in real-time. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, Vol. 2020. NIH Public Access, 63.

[37] Peter Pirolli and Stuart Card. [n. d.]. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. ([n. d.]), 6.

[38] Anthony L Rosner. 2012. Evidence-based medicine: revisiting the pyramid of priorities. *Journal of Bodywork and Movement Therapies* 16, 1 (2012), 42–49.

[39] Salimah Z Shariff, Shayna AD Bejaimal, Jessica M Sontrop, Arthur V Iansavichus, R Brian Haynes, Matthew A Weir, and Amit X Garg. 2013. Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches. *Journal of medical Internet research* 15, 8 (2013), e2624.

[40] Mary Shultz. 2007. Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association: JMLA* 95, 4 (2007), 442.

[41] Richard Smith. 1996. What clinical information do doctors need? *Bmj* 313, 7064 (1996), 1062–1068.

[42] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J. Barezi, and Pascale Fung. 2020. CAiRE-COVID: A Question Answering and Query-focused Multi-Document Summarization System for COVID-19 Scholarly Information Management. arXiv:2005.03975 [cs.CL]

[43] Qandeel Tariq, Jena Daniels, Jessey Nicole Schwartz, Peter Washington, Haik Kalantarian, and Dennis Paul Wall. 2018. Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLoS medicine* 15, 11 (2018), e1002705.

[44] Jennifer Tiffen, Susan J Corbridge, and Lynda Slimmer. 2014. Enhancing clinical decision making: development of a contiguous definition and conceptual framework. *Journal of professional nursing* 30, 5 (2014), 399–405.

[45] Jeremy C Wyatt and Douglas G Altman. 1995. Commentary: Prognostic models: clinically useful or quickly forgotten? *Bmj* 311, 7019 (1995), 1539–1541.

[46] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[47] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[48] Qian Yang, John Zimmerman, and Aaron Steinfeld. 2015. Review of Medical Decision Support Tools : Emerging Opportunity for Interaction Design. In *Proceedings of the 6th IASDR (The International Association of Societies of Design Research Congress* (Brisbane, Australia) *(IASDR '15)*. IASDR (The International Association of Societies of Design Research), Australia, 2366–2382.

[49] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 4477–4488. https://doi.org/10.1145/2858036.2858373