CCS Explorer: Relevance Prediction, Extractive Summarization, and Named Entity Recognition from Clinical Cohort Studies

Irfan Al-Hussaini
Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, USA
alhussaini.irfan@gatech.edu

Davi Nakajima An
Computer Science
Georgia Institute of Technology
Atlanta, GA, USA
dna@gatech.edu

Albert J. Lee
Machine Learning
Georgia Institute of Technology
Atlanta, GA, USA
albert.jb.lee@gatech.edu

Sarah Bi
Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA, USA
sbi30@gatech.edu

Cassie S. Mitchell
Biomedical Engineering and Machine Learning
Georgia Institute of Technology and Emory University
Atlanta, GA, USA
cassie.mitchell@bme.gatech.edu

***Abstract*—Clinical Cohort Studies (CCS) are a great source of documented clinical research. Ideally, a clinical expert will interpret these articles for exploratory analysis ranging from drug discovery for evaluating the efficacy of existing drugs in tackling emerging diseases to the first test of newly developed drugs. However, more than 100 CCS articles are published on PubMed every day. As a result, it can take days for a doctor to find articles and extract relevant information. Can we find a way to quickly sift through the long list of these articles faster and document the crucial takeaways from each of these articles? In this work, we propose `CCS Explorer`, an end-to-end system for relevance prediction of sentences, extractive summarization, and patient, outcome, and intervention entity detection from CCS. `CCS Explorer` is packaged in a web-based graphical user interface where the user can provide any disease name. `CCS Explorer` then extracts and aggregates all relevant information from articles on PubMed based on the results of an automatically generated query produced on the back-end. `CCS Explorer` fine-tunes pre-trained language models based on transformers with additional layers for each of these tasks. We evaluate the models using two publicly available datasets. `CCS Explorer` obtains a recall of 80.2%, AUC-ROC of 0.843, and an accuracy of 88.3% on sentence relevance prediction using BioBERT and achieves an average Micro F1-Score of 77.8% on Patient, Intervention, Outcome detection (PIO) using PubMedBERT. Thus, `CCS Explorer` can reliably extract relevant information to summarize articles, saving time by $\sim 660\times$.**

***Index Terms*—named entity recognition, pico, relevance prediction, summarization, bert, transformers, language model, evidence based medicine**

I. Introduction

One of the world's largest biomedical publication databases, PubMed, has over 34 million publications. Approximately 2.5 million users perform about 3 million searches and 9 million page views on PubMed every day [1]. Over the past couple of years, 137 articles have been posted per day on PubMed on COVID-19 alone [2]. In particular, clinical cohort studies, which contain information on the specific results of a patient or patient population for a given therapeutic and/or condition, are considered essential for clinical research. Clinical cohort studies (CCS) include randomized clinical trials, prospective cohort studies, retrospective cohort studies, case-control studies, patient case studies, and more. Clinical cohort studies typically describe a patient or patient population, the intervention(s) assessed, and the measured outcome(s).

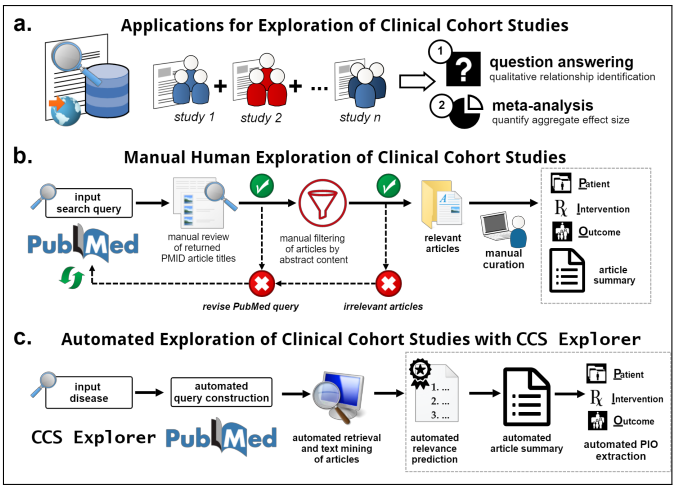


Fig. 1: Why `CCS Explorer`?

The two major applications that require exploration of clinical cohort studies are question answering and meta-analysis (Figure 1a). Exploration of clinical cohort studies is required to answer questions to identify qualitative relationships. Examples of question answering include: What

drugs may be repurposed or used in combination to improve disease outcomes [3]? What comorbidities are most impactful to cardiac disease outcome [4]–[7]? What patient features result in health outcome disparities [8], [9]? Exploration of clinical cohort studies is also required to perform a meta-analysis, which is a quantitative analysis where results of cohort studies are aggregated in order to estimate an overall effect size. Estimating an overall or aggregate effect size, such as the effect of a drug on disease outcome, adjusts for disparity or bias introduced by individual study-specific features (e.g., geography, gender, age, sample size, etc.). Examples of meta-analysis include: determining overall adverse event rates with specific treatments for cancer [10], determining the overall prevalence of comorbidities in a rare neurodegenerative disease population [11], or determining the overall effect size of vaccination on SARS-CoV-2 outcome [12]–[14].

The process for manual exploration of cohort studies is iterative and time-consuming (Figure 1b). The major steps include devising the appropriate advanced PubMed query to find articles in PubMed, reviewing the list of search title results to determine if the query resulted in the expected type or number of studies, examining the abstracts to determine if the journal article contains the desired information, and curating the article to extract the pivotal PIO elements: patient population (disease and/or control population), intervention (what therapy was utilized), and the outcome (what measurement was utilized to determine a result). Depending on the number of studies to be reviewed and included, the exploration process alone can take hours to weeks before final curation and analysis can occur [15]. Moreover, even with a quality control team, there may be some remaining inconsistency between researchers or curators [15]. Critical variations and corresponding delays may occur depending on the researcher's knowledge of constructing an appropriate advanced PubMed query. An appropriate PubMed query must include all relevant synonyms, MESH terms, and appropriate formatting in order to return the most inclusive and relevant list of articles. Additionally, differences in review styles for examining lengthy abstracts or even full-text articles may result in unintended differences in article inclusion or stylistic differences in PIO extraction.

Here we present `CCS Explorer` to automate the process of clinical cohort study exploration (Figure 1c). `CCS Explorer` is an open-source web application that greatly expedites the identification, review, and extraction of data from clinical cohort studies. CCS only requires that the user input a disease name. Using a pre-built list of intervention names (which can also be customized if desired), `CCS Explorer` formulates an advanced query to PubMed. `CCS Explorer` automatically obtains all relevant articles via their unique PubMed identification (PMID) and automatically parses through the text. `CCS Explorer` provides three critical outputs for researchers: 1) a list of all relevant studies along with a relevance prediction score; 2) an abbreviated relevance summary that contains only the most relevant information (or sentences) necessary for the researcher to explore the study; 3) automated extraction of PIO elements. With `CCS Explorer`, question answering or meta-analysis is greatly expedited, streamlined, and optimized. `CCS Explorer` automates all the iterative, front-end work that would normally take a specially trained team of researchers hours to weeks to achieve.

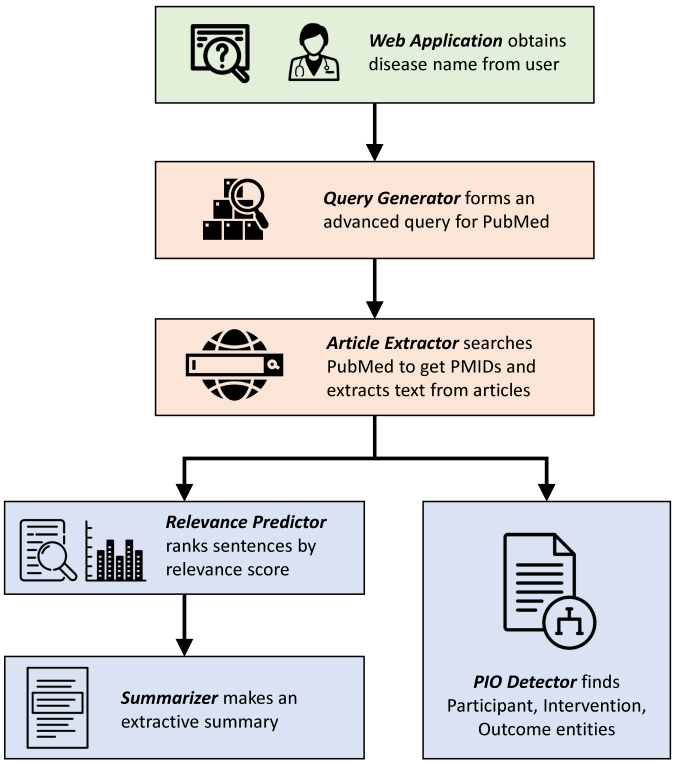


Fig. 2: `CCS Explorer` Framework

There are specially trained groups dedicated to manually synthesizing findings from CCS. However, the rapid publication of new articles makes it impossible to maintain pace [16]. However, Natural Language Processing (NLP) breakthroughs have enabled the automation of many time-consuming tasks related to text exploration in non-biomedical domains. Examples include sentiment analysis of customer reviews [17], language translation [18], [19], ranking search results [20]–[22], abstractive summarization [23]–[25], and extractive summarization [26], [27]. The present work with `CCS Explorer` applies and integrates NLP tasks to automate biomedical text review, namely clinical cohort study exploration.

To this end, we propose `CCS Explorer`, an end-to-end system for exploring clinical cohort studies with PubMed and extracting useful information necessary for tasks like question answering or meta-analysis. 1 highlights the difference between manual exploration by an expert and `CCS Explorer`. It can reduce the time taken to extract relevant information and summarize articles from hours to seconds. For the query demonstrated in Figure 3, it takes 26.32s for `CCS Explorer` to run the query, process the resulting articles, and extract all relevant information to construct a task-specific summary along with the detection of PIO entities.

We introduce two models for this task, one for relevance

prediction of text and another for detecting participant, intervention, and outcome (PIO) entities in CCS articles. We compare the proposed models' performance by initializing the weights using 6-7 different pre-trained BERT [28] or ELECTRA [29] models.

The main contribution of this paper is to design each of the following pieces and combine them to form `CCS Explorer`:

- *Web Application*: front-end designed for taking inputs from the user and displaying outputs
- *Query Generator*: merges MeSH terms to form an advanced query for PubMed to obtain PMIDs
- *Article Extractor*: extracts articles from PubMed and stores the text for subsequent steps
- *Relevance Predictor*: attention-based language model extracts relevant pieces of text from the article along with a relevance score
- *Summarizer*: generates an extractive summary of the article by putting together the most relevant sentences into a coherent body of text
- *PIO Detector*: named entity recognition model finds participants, interventions, and outcomes present in the article along with a score for each entity

This framework is shown in Figure 2.

II. System Design

The goal of `CCS Explorer` is to provide a user-friendly system for researchers to obtain reliable results quickly. For this reason, streamlit [30] was used to design the front-end user interface of `CCS Explorer`. The graphical user interface, GUI, is shown in Figure 3. It takes inputs from the user and displays the results in a user-friendly format for review. It has the following parts:

- Step 1: The query is created during this step. The user has two options: (1) create a manual query by stitching together MeSH terms in the query (2) provide a disease name so that the *Query Generator* can build an advanced query for PubMed.
- Step 2: A query name has to be selected from the options provided consisting of previously formed queries. PMIDs are obtained from PubMed based on the selected query.
- Step 3: The articles are extracted from PubMed using the PMIDs and *Relevance Predictor*, *Summarizer*, and *PIO Detector* are run on each article to obtain aggregated results. It is hidden in Figure 3 because it has already been run and the user has moved on to step 4.
- Step 4: Three tables show the results of *Relevance Predictor*, *Summarizer* and *PIO Detector*.

`CCS Explorer` can be divided into three different pieces which run in the back-end: (1) *Query Generator* and *Article Extractor*, (2) *Relevance Predictor* and *Summarizer*, (3) *PIO Detector*. The details of *Relevance Predictor* is discussed in Section III and *PIO Detector* in Section IV. The framework of `CCS Explorer` is shown in Figure 2.

Query Generator and Article Extraction. `CCS Explorer` provides the users with a graphical user interface to input their

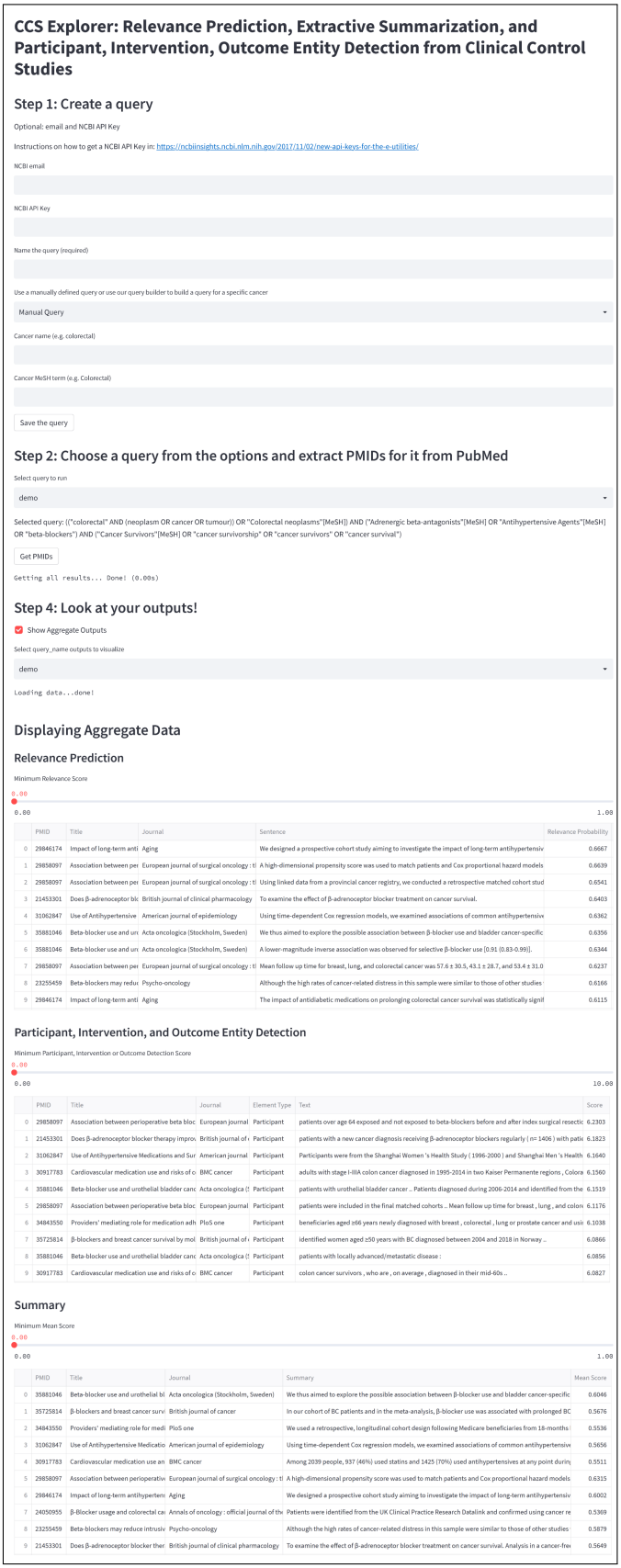


Fig. 3: `CCS Explorer`: Graphical User Interface

National Center for Biotechnology Information (NCBI) email and API key so repeated queries can be sent to PubMed. It also enables the user to manually input a customized advanced query using MeSH terms or to simply provide a cancer type so that an automatically generated query can obtain a baseline result. The query generation and extraction of articles are performed using BioPython [31], [32]. The resulting text is prepared for subsequent steps by splitting it into sentences using SciSpacy [33].

III. Relevance Prediction and Extractive Summarization

A. Data

The data used in building the *Relevance Predictor* and *Summarizer* of `CCS Explorer` originate from an open source dataset named the Evidence Inference dataset [34]–[36]. The dataset contains useful annotations of relevant information in CCS articles.

In this dataset [34]–[36], sections of the text are labeled as *evidence* and *nonevidence*. In designing `CCS Explorer`, we replaced the *evidence* label with *relevant* and *non-evidence* with *irrelevant*. Thus, these *relevant* and *irrelevant* labels were used as ground truth annotations to build the *Relevance Predictor* of `CCS Explorer`. It consists of 4,005 unique articles split across two sets of articles. The selection of articles for training and test set was defined in the Evidence Inference [34], [35] dataset as train_article_ids and validation_article_ids respectively.

The *Summarizer* uses results from the *Relevance Predictor* to formulate summaries along with a Summary Score to denote the quality of the summary.

B. Method

Relevance Prediction. *Relevance Predictor* was designed using BERT-based language models pre-trained on scientific articles obtained from sources such as PubMed, PubMed Central, and UMLS. It was constructed by adding a dense layer to the pre-trained model architecture and fine-tuned on the Evidence Inference dataset described in Section III-A.

The pre-trained BERT models used:

- BioBERT [37]: Initialized using standard BERT [28] model, and then pre-trained on Biomedical domain texts, which includes PubMed abstracts and PubMed Central full-text articles.
- PubMedBERT [38]: Pretrained a BERT [28] model from scratch using 14 million abstracts from PubMed.
- SapBERT [39]: Pre-trained a BERT model on the biomedical knowledge graph of UMLS [40] using self-alignment to cluster synonyms of the same concept.
- BlueBERT [41]: Initialized using standard BERT [28] model and pre-trained on PubMed abstracts (4 Billion words) and clinical notes from MIMIC-III (500 Million words) [42].
- KRISSBERT [43]: Initialized with PubMedBERT [38] parameters, and then pretrained using biomedical entity names from the UMLS ontology [40] to self-supervise entity linking examples from PubMed abstracts.
- SciBERT [44]: Trained a BERT [28] model on scientific papers taken from 1.14 million full papers from Semantic Scholar.

Let $\mathcal{Y}'$ be all the outputs from the model, $\mathcal{Y}$ be all the annotations from the dataset, $y'_i \in [0, 1]$ represent the model prediction and $\boldsymbol{y}_i$ denote the annotation of the i-th sentence. Let $\boldsymbol{h}(\mathcal{X})$ represent the output of the transformer architecture. This is used as input to a fully-connected layer followed by the sigmoid function (σ). So, the output of the model for the i-th sentence is represented by:

$$\begin{aligned} \boldsymbol{z}_i &= \boldsymbol{W}^\top \boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{b} \\ \boldsymbol{y'_i} &= \sigma(\boldsymbol{z}_i) = \frac{1}{1 + \boldsymbol{z}_i} \end{aligned} \tag{1}$$

Binary cross entropy loss is used and is denoted by:

$$\begin{aligned} L(\boldsymbol{y}_i,\ \boldsymbol{y'_i}) = -\ [&\boldsymbol{y}_i \cdot log(\boldsymbol{y'_i}) \\ &+ (1 - \boldsymbol{y}_i) \cdot log(1 - \boldsymbol{y'_i})] \end{aligned} \tag{2}$$

Summarization. The output of the sigmoid function (σ) in Equation (1), $\boldsymbol{y'_i}$, represents the relevance score for the i-th sentence. The sentences are then sorted in descending order by these relevance scores to generate the set of sentences $\mathcal{Y}'_{sorted}$. The first 4 sentences corresponding to the 4 most relevant sentences are joined to form the extractive summary for each article. The summary score is the average of the relevance scores for each of these 4 sentences

$$\text{Summary Score} = \frac{\sum_{i=1}^{4} \boldsymbol{y}'_{i,sorted}}{4} \tag{3}$$

Metrics. The following metrics were used to evaluate the performance of the relevance prediction model:

$$\begin{aligned} \text{Accuracy} &= \frac{|\mathcal{Y} \cap \mathcal{Y}'|}{N} \\ \text{Recall}, R &= \frac{|\mathcal{Y} \cap \mathcal{Y}'|}{|\mathcal{Y}|} \\ \text{Precision}, P &= \frac{|\mathcal{Y} \cap \mathcal{Y}'|}{|\mathcal{Y}'|} \\ \text{F1 score} &= \frac{2\ *\ P\ *\ R}{P\ +\ R} \end{aligned} \tag{4}$$

where the annotated relevance labels of the entire dataset are denoted by $\mathcal{Y}$ and the model predictions by $\mathcal{Y}'$; $|\mathcal{Y}|$ and $|\mathcal{Y}'|$ represent the number of annotated tokens and the number of model predictions. In addition to the above metrics, the area under receiver operating characteristics curve (AUC-ROC) is used for comparison.

Implementation Details. We implemented *Relevance Predictor* using PyTorch [45], [46] and transformers [47]. The model was trained using a machine equipped with Intel Xeon Gold 6136 Processor, 376GB RAM, an Nvidia V100 GPU and CUDA 11.4. While training *Relevance Predictor*, we used a batch size of 16 and ADAM [48] as the optimization method. The learning rate was set at 10^{-5} and it was trained for 4 epochs. 3,562 total articles defined in train_article_ids are

TABLE I: CCS Explorer: Relevance Prediction Model Performance

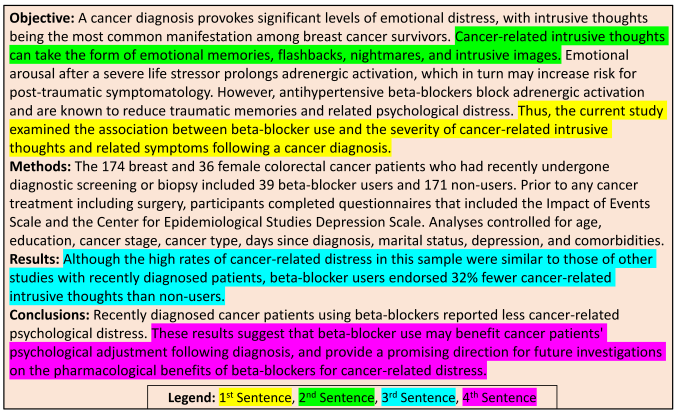
Model	Accuracy	Precision	Recall	AUC-ROC	F1-Score
BioBERT [37]	0.883	0.083	0.802	0.843	0.150
PubMedBERT [38]	0.880	0.080	0.801	0.841	0.145
SapBERT [39]	0.887	0.083	0.776	0.832	0.150
BlueBERT [41]	0.875	0.078	0.817	0.846	0.143
KRISSBERT [43]	0.884	0.082	0.792	0.839	0.149
SciBERT [44]	0.877	0.080	0.814	0.846	0.145

used as the training set and the 443 articles defined in validation_article_ids list are used as the test set of the Evidence Inference Dataset [34], [35]. To ensure consistent performance and fair comparison, we use the same hyperparameters for all models.

C. Result

A high recall is essential for relevance prediction as we want to be sure that we are detecting all the relevant sentences. It is acceptable for an automated system to include some irrelevant sentences as long as the significant ones appear at the top of the list. Prior research in machine translation show alignment with human expectation is highest when the optimization focuses on recall [59]. User evaluation of interactive information retrieval performance [60] indicates recall is significantly more correlated with the users' expectation of success. Similarly, recall is more important than precision for downstream tasks such as summarization [61]. Most of the evaluated models for relevance prediction displayed a recall above 80%, AUC-ROC above 84%, and accuracy above 88%. The low F1-score is due to the low precision, which is less critical for tasks such as relevance prediction [59]–[61]. Due to the highest average metrics among all methods, BioBERT [37] was selected as the model used to make predictions in the back-end of the web-based interface of CCS Explorer.

Case Study. An example of the relevant sentence prediction and subsequent summary formulation using CCS Explorer for a PubMed query targeting colorectal cancer articles is demonstrated in Figure 4. The article obtains a Summary Score of 0.588 using *Summarizer*. In this article, titled *Beta-blockers may reduce intrusive thoughts in newly diagnosed cancer patients* by Lindgren et al. [56], the highest scoring sentence perfectly summarizes the goal of the study. The second sentence provides an example of potential problems faced by the cohort. The third sentence focuses on the results of the study and the fourth sentence draws conclusions from the study. The summary score is obtained by averaging the relevance score of each sentence forming the summary. The summary scores of all the articles resulting from the query are shown in Table II. It shows that the model is consistent and obtains a good summary score for all articles, with a maximum score of 0.631 and a minimum score of 0.537.



(a) Article with the 4 most relevant sentences according to *Relevance Predictor* highlighted

Thus, the current study examined the association between beta-blocker use and the severity of cancer-related intrusive thoughts and related symptoms following a cancer diagnosis. Cancer-related intrusive thoughts can take the form of emotional memories, flashbacks, nightmares, and intrusive images. Although the high rates of cancer-related distress in this sample were similar to those of other studies with recently diagnosed patients, beta-blocker users endorsed 32% fewer cancer-related intrusive thoughts than non-users. These results suggest that beta-blocker use may benefit cancer patients' psychological adjustment following diagnosis, and provide a promising direction for future investigations on the pharmacological benefits of beta-blockers for cancer-related distress.

(b) Extractive summary using most relevant sentences

Fig. 4: Extractive Summarization of PMID 23255459 [56] titled *Beta-blockers may reduce intrusive thoughts in newly diagnosed cancer patients* by Lindgren et al. with a summary score of 0.588 using *Summarizer* of CCS Explorer

IV. Patient, Intervention, Outcome Detection

A. Data

The final piece of CCS Explorer is aimed at named entity recognition of Patient, Intervention, and Outcome in articles that describe the conduct and results of clinical cohort studies. To train *PIO Detector* for this task, we used the EBM-NLP corpus [62]. The dataset includes 4,970 medical article abstracts with annotations indicating sequences of text that describe the Participants, Interventions, and Outcome elements of the respective CCS. 4,782 of these abstracts contain crowd-sourced labels. 188 articles among the 4,970 abstracts contain annotations from domain experts with medical training. This test set is held-out while training the models and only used to test the performance of the final *PIO Detector* models.

B. Method

The pre-trained models are used for *PIO Detector*:

- BioELECTRA [63]: Pre-trained an ELECTRA model on full text articles from PubMed and PubMed Central.
- PubMedBERT [38]: Pretrained a BERT model from scratch using 14 million abstracts from PubMed.
- SciBERT [44]: Pre-trained a BERT model trained on scientific papers taken from 1.14 million full papers from Semantic Scholar.
- BioBERT [37]: Initialized using standard BERT [28] model, and then pre-trained on Biomedical domain texts which includes PubMed abstracts and PubMed Central full-text articles.

TABLE II: CCS Explorer generated extractive summaries of the following query: (("colorectal" AND (neoplasm OR cancer OR tumour)) OR "Colorectal neoplasms" [MeSH]) AND ("Adrenergic beta-antagonists" [MeSH] OR "Antihypertensive Agents" [MeSH] OR "beta-blockers") AND ("Cancer Survivors" [MeSH] OR "cancer survivorship" OR "cancer survivors" OR "cancer survival")

PMID	Title	Journal	Summary Score
24050955 [49]	β-Blocker usage and colorectal cancer mortality: a nested case-control study in the UK Clinical Practice Research Datalink cohort.	Annals of oncology ...	0.537
35881046 [50]	Beta-blocker use and urothelial bladder cancer survival: a Swedish register-based cohort study.	Acta oncologica (Stockholm, Sweden)	0.605
29858097 [51]	Association between perioperative beta blocker use and cancer survival following surgical resection.	European journal of surgical oncology ...	0.631
29846174 [52]	Impact of long-term antihypertensive and antidiabetic medications on the prognosis of post-surgical colorectal cancer: the Fujian ...	Aging	0.600
34843550 [53]	Providers' mediating role for medication adherence among cancer survivors.	PloS one	0.554
31062847 [54]	Use of Antihypertensive Medications and Survival Rates for Breast, Colorectal, Lung, or Stomach Cancer.	American journal of epidemiology	0.566
35725814 [55]	β-blockers and breast cancer survival by molecular subtypes: a population-based cohort study and meta-analysis.	British journal of cancer	0.568
23255459 [56]	Beta-blockers may reduce intrusive thoughts in newly diagnosed cancer patients.	Psycho-oncology	0.588
30917783 [57]	Cardiovascular medication use and risks of colon cancer recurrences and additional cancer events: a cohort study.	BMC cancer	0.551
21453301 [58]	Does β-adrenoceptor blocker therapy improve cancer survival? Findings from a population-based retrospective cohort study.	British journal of clinical pharmacology	0.565

- BlueBERT [41]: Initialized using standard BERT [28] model and pre-trained on PubMed abstracts (4 Billion words) and clinical notes from MIMIC-III (500 Million words) [42].
- KRISSBERT [43]: Initialized with PubMedBERT [38] parameters, and then pretrained using biomedical entity names from the UMLS ontology [40] to self-supervise entity linking examples from PubMed abstracts.
- SapBERT [39]: Pre-trained a BERT model on the biomedical knowledge graph of UMLS [40] using self-alignment to cluster synonyms of the same concept.

The labels provided in the dataset for each token is mapped onto the following 4 labels, where 3 represent the target named entities Patient, Intervention, and Outcome, while the 4^{th} is denoted by None and represents tokens which are not any of these 3 entities.

Let $\mathcal{Y}'$ be all the outputs from the model, $\mathcal{Y}$ be all the annotations from the dataset, $\boldsymbol{y}_i'$ represent the model prediction and $\boldsymbol{y}_i$ denote the annotation of the i-th token. Let $\boldsymbol{h}(\mathcal{X})$ represent the output of the transformer architecture. This is used as input to a fully-connected layer. So, the output of the i-th token is represented by $\boldsymbol{y_i'} = \boldsymbol{W}^\top \boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{b}$.

To train the model, we used cross entropy loss Eq. 5:

$$L(\boldsymbol{y}_i,\ \boldsymbol{y_i'}) = -\sum_{j=1}^{4} \boldsymbol{y}_i[j]\ log(\boldsymbol{y_i'}[j]) \tag{5}$$

where $L(\boldsymbol{y}_i,\ \boldsymbol{y_i'})$ is the estimated cross entropy loss for the i-th token between annotations $\boldsymbol{y} \in \mathbb{R}^4$ and the predicted probabilities $\boldsymbol{y'} \in \mathbb{R}^4$, $\boldsymbol{y_i'}[j]$ represents the model predictions for the i-th token and j-th entity.

Metrics. The following metrics were used to evaluate the performance of the NER models for PIO detection:

$$\begin{aligned} \text{Recall}, R^{(k)} &= \frac{\left|\mathcal{Y}^{(k)} \cap \mathcal{Y}'^{(k)}\right|}{\left|\mathcal{Y}'^{(k)}\right|} \\ \text{Precision}, P^{(k)} &= \frac{\left|\mathcal{Y}^{(k)} \cap \mathcal{Y}'^{(k)}\right|}{\left|\mathcal{Y}^{(k)}\right|} \\ \text{F1 score} &= \frac{2\ *\ P\ *\ R}{P\ +\ R} \end{aligned} \tag{6}$$

Given annotations $\mathcal{Y}$, model predictions $\mathcal{Y}'$, k ={Patient, Intervention, Outcome, None} indicating the entity, $\left|\mathcal{Y}^{(k)}\right|$ and $\left|\mathcal{Y}'^{(k)}\right|$ represent the number of annotations and model predictions with the label k.

Implementation Details. The was implemented using PyTorch [45], [46] and transformers [47]. We trained the model using a machine equipped with Intel Xeon Gold 6136 Processor, 376GB RAM, an Nvidia V100 GPU and CUDA 11.4. While training *PIO Detector*, we use batch size of 6 and AdamW [64] as the optimization method and a learning rate of 10^{-4}. *PIO Detector* is trained for 2 epochs.

To train the model, we randomly split the data by subjects into training and validation set in a 9:1 ratio. We train *PIO Detector* using the training set and use the validation set for hyperparameter optimization. The held-out test set is used to evaluate *PIO Detector* and compare different baselines. To ensure consistent performance and fair comparison, we use the same model hyperparameters for all models.

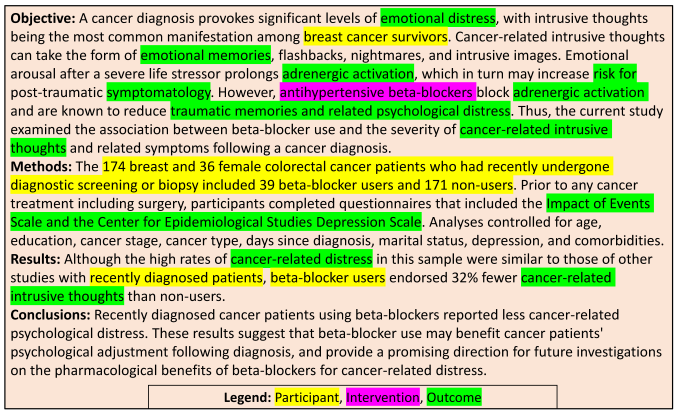
C. Result

Table III compares the results of *PIO Detector* used in CCS Explorer using different pre-trained BERT [28] and ELECTRA [29] models. The average Mirco-F1 score shows the efficacy of *PIO Detector* in detecting all the 3 entities: Participants, Intervention, and Outcome. The pre-trained states of these models do not affect the performance after fine-tuning, as highlighted by a difference $< 1\%$ in the average micro F1-score. *PIO Detector* performs particularly well in detecting Participants resulting in the highest Recall and Micro-F1 Score among the 3 entities detected. Due to the highest average F1-Score among all methods, PubMedBERT [38] was selected as the model used to make predictions in the back-end of the web-based interface of CCS Explorer.

Case Study. Figure 5 shows the Participants, Interventions, and Outcomes detected along with the respective scores for the same paper expanded upon in Section III-C titled *Beta-blockers may reduce intrusive thoughts in newly diagnosed cancer patients* by Lindgren et al. [56]. In this paper, participant entities obtain much higher prediction scores on average compared with other entities. The higher performance metrics for participant entities are also evident in Table III, where participant entities obtain the highest recall and F1-scores. Overall, all the PIO entities detected align well with a manual review.

TABLE III: CCS Explorer: Participant, Intervention, Outcome Detection Model Performance

Model	Precision			Recall			Micro F1-Score			Average Micro F1-Score
	Participant	Intervention	Outcome	Participant	Intervention	Outcome	Participant	Intervention	Outcome	P/I/O
BioELECTRA [63]	0.738	0.609	0.851	0.923	0.763	0.619	0.820	0.677	0.717	0.776
PubMedBERT [38]	0.744	0.636	0.849	0.920	0.758	0.602	0.823	0.692	0.705	0.778
SciBERT [44]	0.743	0.609	0.854	0.910	0.750	0.607	0.818	0.673	0.710	0.773
BioBERT [37]	0.743	0.635	0.853	0.915	0.765	0.591	0.820	0.694	0.698	0.776
BlueBERT [41]	0.724	0.635	0.852	0.916	0.749	0.593	0.809	0.687	0.700	0.771
KRISSBERT [43]	0.760	0.613	0.852	0.918	0.756	0.601	0.832	0.677	0.705	0.776
SapBERT [39]	0.740	0.619	0.860	0.920	0.757	0.601	0.820	0.681	0.708	0.775



(a) Article with PIO Elements highlighted

Element Type	Text	Score
Participant	174 breast and 36 female colorectal cancer patients who had recently undergone diagnostic screening or biopsy included 39 beta-blocker users and 171 non-users .. Prior	5.98
Participant	breast cancer survivors ..	5.88
Participant	recently diagnosed patients	5.61
Participant	beta-blocker users	5.18
Intervention	antihypertensive beta-blockers	2.84
Outcome	cancer-related intrusive thoughts	2.88
Outcome	emotional memories	2.92
Outcome	Impact of Events Scale and the Center for Epidemiological Studies Depression Scale ..	2.54
Outcome	traumatic memories and related psychological distress ..	2.44
Outcome	cancer-related distress	2.35
Outcome	emotional distress	2.28
Outcome	risk for	1.89
Outcome	symptomatology ..	1.82
Outcome	adrenergic activation	1.82

(b) PIO Elements Ranked

Fig. 5: Participant, Intervention, Outcome (PIO) Detection of PMID 23255459 [56] titled *Beta-blockers may reduce intrusive thoughts in newly diagnosed cancer patients* by Lindgren et al. using CCS Explorer

V. Comparison with Manual Exploration

To provide context, the goal of the query used to illustrate the capabilities of CCS Explorer and explore studies to answer the following question: How do antihypertensive drugs impact the outcome of colorectal cancer survival? The advanced PubMed query automatically constructed by CCS Explorer shown in Figure 3 is: (("colorectal" AND (neoplasm OR cancer OR tumor)) OR "colorectal neoplasms" [MeSH]) AND ("Adrenergic beta-antagonists" [MeSH] OR "Antihypertensive Agents" [MeSH] OR "beta-blockers") AND ("Cancer Survivors" [MeSH] OR "cancer survivorship" [MeSH] OR "cancer survivors" OR "cancer survival"). The formatting of the query is critical to finding the most relevant clinical cohort studies. The aforementioned query returned 11 studies. Entering a more general PubMed query of "colorectal cancer" at the time of this writing returned 281,217 studies. A query of "colorectal cancer AND hypertension" returned 1,617 results. CCS Explorer automatically formats the anti-hypertensive drug names and all synonymous versions of the outcome "cancer survival" to insure maximal coverage while still restricting the output to the most relevant studies.

Explicitly comparing CCS Explorer to manual exploration by a trained human curator is informative. Even if the human curator does appropriately format the advanced PubMed query, there is still a substantial time saving with CCS Explorer. Here we compared the exploration time once relevant articles have been selected. Based on timed trained curator studies [15], the average exploration time per relevant article is 29 minutes with a range of 24 to 42 minutes. The variability in manual exploration is based both on the innate skill of the curator and how difficult it is to find the relevant PIO parts in the article (based on changes in article structure, length, etc.). Thus, even if a curator only explored the 10 relevant articles, the process would take on average 290 minutes compared to the 26.32 seconds required by CCS Explorer.

Beyond time savings, CCS Explorer also provides critical context that is not provided during the equivalent manual process. CCS Explorer provides the quantitative relevance rankings of each study. The relevance ranking is extremely helpful for prioritizing the review of large sets of returned relevant articles. The relevance ranking is also helpful to evaluate how well the advanced PubMed query returns results the curator deems relevant to the exploratory objective. CCS Explorer also provides its own extractive summary, which takes in only the most relevant sentences from each study. Here, we chose to use only the 4 most relevant sentences to construct the extractive summary. However, the number of sentences included in each extractive summary can be adjusted by the user. The extractive summary allows for fast and efficient exploration by the human curator. Finally, beyond critical context, the automated PIO detection and extraction expedites the formation of study inclusion criteria and preliminary curation steps for a subsequent meta-analysis.

VI. Conclusion

Recently, there has been an explosion of articles on clinical cohort studies (CCS), which are readily available through PubMed. However, the sheer number of articles published every day makes it impossible to read through them to extract relevant information manually. In this paper, we propose an end-to-end system with a user-friendly graphical interface called `CCS Explorer`, which makes this accessible to anyone. `CCS Explorer` can take a disease as input, generate an advanced query for PubMed, and extract the text from all the resulting articles. It then proceeds to rank each sentence based on a relevance score, creates an extractive summary of the article along with a summary score, and extracts all Participant, Intervention, and Outcome (PIO) entities in the article. The *Relevance Predictor*, *Summarizer*, and *PIO Detector* are evaluated quantitatively and case studies are performed to demonstrate their effectiveness. Thus, `CCS Explorer` makes the difficult task of performing large-scale meta-analysis and review feasible by drastically reducing the time required.

Acknowledgment

We would like to thank the wonderful team at Morningside Center for Innovative and Affordable Medicine, Emory University for consultation during the study. This research was funded by National Science Foundation CAREER grant 1944247 to C.M, National Institute of Health grant U19-AG056169 sub-award to C.M., and the McCamish Parkinson's Disease Innovation Program at Georgia Institute of Technology and Emory University to C.M.

References

[1] J. White, "Pubmed 2.0," *Medical Reference Services Quarterly*, vol. 39, no. 4, pp. 382–387, 2020.

[2] N. S. L. Yeo-Teh and B. L. Tang, "An alarming retraction rate for scientific publications on coronavirus disease 2019 (covid-19)," *Accountability in research*, vol. 28, no. 1, pp. 47–53, 2021.

[3] K. McCoy, S. Gudapati, L. He, E. Horlander, D. Kartchner, S. Kulkarni, N. Mehra, J. Prakash, H. Thenot, S. V. Vanga *et al.*, "Biomedical text link prediction for drug discovery: a case study with covid-19," *Pharmaceutics*, vol. 13, no. 6, p. 794, 2021.

[4] M. A. Burke and W. G. Cotts, "Interpretation of b-type natriuretic peptide in cardiac disease and other comorbid conditions," *Heart failure reviews*, vol. 12, no. 1, pp. 23–36, 2007.

[5] A. Cavaillès, G. Brinchault-Rabin, A. Dixmier, F. Goupil, C. Gut-Gobert, S. Marchand-Adam, J.-C. Meurice, H. Morel, C. Person-Tacnet, C. Leroyer *et al.*, "Comorbidities of copd," *European Respiratory Review*, vol. 22, no. 130, pp. 454–475, 2013.

[6] J. Listerman, V. Bittner, B. K. Sanderson, and T. M. Brown, "Cardiac rehabilitation outcomes: impact of comorbidities and age," *Journal of cardiopulmonary rehabilitation and prevention*, vol. 31, no. 6, p. 342, 2011.

[7] C. C. Lang and D. M. Mancini, "Non-cardiac comorbidities in chronic heart failure," *Heart*, vol. 93, no. 6, pp. 665–671, 2007.

[8] M. Arnold, M. Halpern, N. Meier, U. Fischer, T. Haefeli, L. Kappeler, C. Brekenfeld, H. P. Mattle, and K. Nedeltchev, "Age-dependent differences in demographics, risk factors, co-morbidity, etiology, management, and clinical outcome of acute ischemic stroke," *Journal of neurology*, vol. 255, no. 10, pp. 1503–1507, 2008.

[9] M. Ashwell, P. Gunn, and S. Gibson, "Waist-to-height ratio is a better screening tool than waist circumference and bmi for adult cardiometabolic risk factors: systematic review and meta-analysis," *Obesity reviews*, vol. 13, no. 3, pp. 275–286, 2012.

[10] P. Mohanavelu, M. Mutnick, N. Mehra, B. White, S. Kudrimoti, K. Hernandez Kluesner, X. Chen, T. Nguyen, E. Horlander, H. Thenot *et al.*, "Meta-analysis of gastrointestinal adverse events from tyrosine kinase inhibitors for chronic myeloid leukemia," *Cancers*, vol. 13, no. 7, p. 1643, 2021.

[11] C. S. Mitchell, S. K. Hollinger, S. D. Goswami, M. A. Polak, R. H. Lee, and J. D. Glass, "Antecedent disease is less prevalent in amyotrophic lateral sclerosis," *Neurodegenerative Diseases*, vol. 15, no. 2, pp. 109–113, 2015.

[12] M. Makhoul, H. H. Ayoub, H. Chemaitelly, S. Seedat, G. R. Mumtaz, S. Al-Omari, and L. J. Abu-Raddad, "Epidemiological impact of sars-cov-2 vaccination: Mathematical modeling analyses," *Vaccines*, vol. 8, no. 4, p. 668, 2020.

[13] E. Pritchard, P. C. Matthews, N. Stoesser, D. W. Eyre, O. Gethings, K.-D. Vihta, J. Jones, T. House, H. VanSteenHouse, I. Bell *et al.*, "Impact of vaccination on new sars-cov-2 infections in the united kingdom," *Nature medicine*, vol. 27, no. 8, pp. 1370–1378, 2021.

[14] A. J. Shattock, E. A. Le Rutte, R. P. Dünner, S. Sen, S. L. Kelly, N. Chitnis, and M. A. Penny, "Impact of vaccination and non-pharmaceutical interventions on sars-cov-2 dynamics in switzerland," *Epidemics*, vol. 38, p. 100535, 2022.

[15] C. S. Mitchell, A. Cates, R. B. Kim, and S. K. Hollinger, "Undergraduate biocuration: developing tomorrow's researchers while mining today's data," *Journal of Undergraduate Neuroscience Education*, vol. 14, no. 1, p. A56, 2015.

[16] G. Tsafnat, A. Dunn, P. Glasziou, and E. Coiera, "The automation of systematic reviews," 2013.

[17] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware BERT for cross-domain sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4019–4028. [Online]. Available: https://aclanthology.org/2020.acl-main.370

[18] K. Chen, R. Wang, M. Utiyama, and E. Sumita, "Neural machine translation with reordering embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1787–1799. [Online]. Available: https://aclanthology.org/P19-1174

[19] T. Nishihara, A. Tamura, T. Ninomiya, Y. Omote, and H. Nakayama, "Supervised visual attention for multimodal neural machine translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4304–4314.

[20] K. Anyanwu, A. Maduko, and A. Sheth, "Semrank: ranking complex relationship search results on the semantic web," in *Proceedings of the 14th international conference on World Wide Web*, 2005, pp. 117–127.

[21] C. W. Belter, "A relevance ranking method for citation-based search results," *Scientometrics*, vol. 112, no. 2, pp. 731–746, 2017.

[22] R. Gao and C. Shah, "Toward creating a fairer ranking in search engine results," *Information Processing & Management*, vol. 57, no. 1, p. 102138, 2020.

[23] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[24] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," *arXiv preprint arXiv:2005.00661*, 2020.

[25] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," *arXiv preprint arXiv:1808.10792*, 2018.

[26] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive summarization as text matching," *arXiv preprint arXiv:2004.08795*, 2020.

[27] Y. Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[30] "Streamlit: the fastest way to build and share data apps." [Online]. Available: https://streamlit.io/

[31] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski *et al.*, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[32] B. Chapman and J. Chang, "Biopython: Python tools for computational biology," *ACM Sigbio Newsletter*, vol. 20, no. 2, pp. 15–19, 2000.

[33] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. [Online]. Available: https://www.aclweb.org/anthology/W19-5034

[34] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace, "Inferring which medical treatments work from reports of clinical trials," in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 3705–3717.

[35] J. DeYoung, E. Lehman, B. Nye, I. J. Marshall, and B. C. Wallace, "Evidence inference 2.0: More data, better models," 2020.

[36] B. E. Nye, J. DeYoung, E. Lehman, A. Nenkova, I. J. Marshall, and B. C. Wallace, "Understanding clinical trial reports: Extracting medical entities and their relations," *AMIA Summits on Translational Science Proceedings*, vol. 2021, p. 485, 2021.

[37] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[38] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020.

[39] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 4228–4238. [Online]. Available: https://www.aclweb.org/anthology/2021.naacl-main.334

[40] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[41] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," in *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, 2019, pp. 58–65.

[42] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[43] S. Zhang, H. Cheng, S. Vashishth, C. Wong, J. Xiao, X. Liu, T. Naumann, J. Gao, and H. Poon, "Knowledge-rich self-supervised entity linking," *arXiv preprint arXiv:2112.07887*, 2021.

[44] I. Beltagy, K. Lo, and A. Cohan, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[46] W. Falcon *et al.*, "Pytorch lightning," *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, vol. 3, no. 6, 2019.

[47] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[49] B. Hicks, L. Murray, D. Powe, C. Hughes, and C. Cardwell, "β-blocker usage and colorectal cancer mortality: a nested case–control study in the uk clinical practice research datalink cohort," *Annals of oncology*, vol. 24, no. 12, pp. 3100–3106, 2013.

[50] R. Udumyan, E. Botteri, T. Jerlstrom, S. Montgomery, K. E. Smedby, and K. Fall, "Beta-blocker use and urothelial bladder cancer survival: a swedish register-based cohort study," *Acta Oncologica*, pp. 1–9, 2022.

[51] R. P. Musselman, S. Bennett, W. Li, M. Mamdani, T. Gomes, C. van Walraven, R. Boushey, O. Al-Obeed, M. Al-Omran, and R. C. Auer, "Association between perioperative beta blocker use and cancer survival following surgical resection," *European Journal of Surgical Oncology*, vol. 44, no. 8, pp. 1164–1169, 2018.

[52] F. Peng, D. Hu, X. Lin, B. Liang, Y. Chen, H. Zhang, Y. Xia, J. Lin, X. Zheng, and W. Niu, "Impact of long-term antihypertensive and antidiabetic medications on the prognosis of post-surgical colorectal cancer: the fujian prospective investigation of cancer (fiesta) study," *Aging (Albany NY)*, vol. 10, no. 5, p. 1166, 2018.

[53] J. G. Trogdon, K. Amin, P. Gupta, B. Y. Urick, K. E. Reeder-Hayes, J. F. Farley, S. B. Wheeler, L. Spees, and J. L. Lund, "Providers' mediating role for medication adherence among cancer survivors," *PloS one*, vol. 16, no. 11, p. e0260358, 2021.

[54] Y. Cui, W. Wen, T. Zheng, H. Li, Y.-T. Gao, H. Cai, M. You, J. Gao, G. Yang, W. Zheng *et al.*, "Use of antihypertensive medications and survival rates for breast, colorectal, lung, or stomach cancer," *American Journal of Epidemiology*, vol. 188, no. 8, pp. 1512–1528, 2019.

[55] L. L. Løfling, N. C. Støer, E. K. Sloan, A. Chang, S. Gandini, G. Ursin, and E. Botteri, "β-blockers and breast cancer survival by molecular subtypes: a population-based cohort study and meta-analysis," *British Journal of Cancer*, pp. 1–11, 2022.

[56] M. E. Lindgren, C. P. Fagundes, C. M. Alfano, S. P. Povoski, D. M. Agnese, M. W. Arnold, W. B. Farrar, L. D. Yee, W. E. Carson, C. R. Schmidt *et al.*, "Beta-blockers may reduce intrusive thoughts in newly diagnosed cancer patients," *Psycho-oncology*, vol. 22, no. 8, pp. 1889–1894, 2013.

[57] E. J. Bowles, O. Yu, R. Ziebell, L. Chen, D. M. Boudreau, D. P. Ritzwoller, R. A. Hubbard, J. M. Boggs, A. N. Burnett-Hartman, A. Sterrett *et al.*, "Cardiovascular medication use and risks of colon cancer recurrences and additional cancer events: a cohort study," *BMC cancer*, vol. 19, no. 1, pp. 1–12, 2019.

[58] S. M. Shah, I. M. Carey, C. G. Owen, T. Harris, S. DeWilde, and D. G. Cook, "Does β-adrenoceptor blocker therapy improve cancer survival? findings from a population-based retrospective cohort study," *British journal of clinical pharmacology*, vol. 72, no. 1, pp. 157–161, 2011.

[59] A. Lavie, K. Sagae, and S. Jayaraman, "The significance of recall in automatic metrics for mt evaluation," in *Conference of the Association for Machine Translation in the Americas*. Springer, 2004, pp. 134–143.

[60] L. T. Su, "The relevance of recall and precision in user evaluation," *Journal of the American Society for Information Science*, vol. 45, no. 3, pp. 207–217, 1994.

[61] A. Nenkova, "Summarization evaluation for text and speech: issues and approaches," in *Ninth International Conference on Spoken Language Processing*, 2006.

[62] B. Nye, J. J. Li, R. Patel, Y. Yang, I. Marshall, A. Nenkova, and B. Wallace, "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 (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 197–207. [Online]. Available: https://aclanthology.org/P18-1019

[63] K. r. Kanakarajan, B. Kundumani, and M. Sankarasubbu, "Bio-ELECTRA:pretrained biomedical text encoder using discriminators," in *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, Jun. 2021, pp. 143–154. [Online]. Available: https://aclanthology.org/2021.bionlp-1.16

[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.