Automating Pathway Extraction from Clinical Guidelines: A Conceptual Model, Datasets and Initial Experiments

Daniel Grathwol[1], Han van der Aa[2], and Hugo A. López[1]

[1] Technical University of Denmark, DTU Compute, Denmark
hulo@dtu.dk
[2] Faculty of Computer Science, University of Vienna, Austria
han.van.der.aa@univie.ac.at

Abstract. Clinical pathways are structured, multidisciplinary care plans utilized by healthcare providers to standardize the management of specific clinical problems. Designed to bridge the gap between evidence and practice, clinical pathways aim to enhance clinical outcomes and improve efficiency, often reducing hospital stays and lowering healthcare costs. However, maintaining pathways with up-to-date, evidence-based recommendations is complex and time-consuming. It requires the integration of clinical guidelines, algorithmic procedures, and tacit knowledge from various institutions. A critical aspect of updating clinical pathways involves extracting procedural information from clinical guidelines, which are textual documents that detail medical procedures. This paper explores how Large Language Models (LLMs) can facilitate this extraction to support clinical pathway development and maintenance. Concretely, we present a conceptual model for using LLMs in this extraction task, provide a dataset comprising thousands of clinical guidelines for academic research, and share the results of initial experiments demonstrating the efficacy of LLMs in extracting relevant pathway information from these guidelines.

Keywords: Clinical pathways · clinical guidelines · large language models · process extraction · conceptual model.

1 Introduction

Clinical Pathways (CPWs) are structured multidisciplinary care plans used by healthcare providers to describe the care processes with a specific clinical problem [30]. They aim to link evidence to practice, optimize clinical outcomes, and maximize clinical efficiency. Their application can result in reported reductions in length of stay and decreasing hospital costs, among its benefits [30].

Despite their potential to improve the quality and effectiveness of care, it must be recognized that various challenges affect the definition of clinical pathways. Clinical Pathways are established through an interdisciplinary process, including clinical guidelines (CGs) [33], algorithmic processes [35], and tacit knowledge of clinical personnel at each institution [5]. Keeping up-to-date with evidence-based recommendations is a challenge: In just 2023, over 26.000 cancer-related papers were published on PubMed[3]. When new recommendations affect the pathways used in practice, a change-management plan needs to be established. However, it is documented that revisions and change management plans are not implemented for all institutions [5]. Second, its translation is complex: either the process of interpretation and pathway generation is manual, or supported by algorithms that may bias their understanding of the guidelines based on their expected output (e.g. the semantics of a particular process modeling notation). Third, the process lacks transparency as it is unclear which semantic structures get translated from clinical guidelines to clinical pathways. Finally, the representation of the pathways may not be adequate: traditional notations such as workflows or BPMN are imperative notations that do not capture the discretion and observation-based decision exerted by healthcare practitioners, and, in general, by actors in so-called Knowledge-Intensive Processes [9].

This paper explores how Large Language Models (LLMs) can be helpful in the semantic analysis and extraction of clinical pathways from clinical guidelines. LLMs are generally well-suited for information extraction tasks from unstructured text, representing a considerable part of the long-term goal of transforming CG-to-CPWs. We consider this a long-term goal, yet unrealistic to achieve with the current state of the technology due to several factors. First, the domain-specificity of clinical data makes it more challenging for general-purpose LLMs to produce accurate results (even if this is something that can be improved using fine-tuned or specialized LLMs). Second, the lack of clear semantic structures may lead to false positives in information-extraction tasks, which, given the high-risk impact of healthcare applications, need extra attention according to the AI-Act [10]. Third, there is no consensus on what is the best semantic representation of clinical pathways. While flowcharts and imperative process models are regarded as easier to interpret, they are less apt for capturing context-dependency and discretionary decision-making than declarative process models [11]. Declarative Notations, such as Declare [28] or DCR graphs [17], may capture the flexible nature of clinical work, but they may be harder to comprehend than their imperative counterparts [12]. Even this separation may not be sufficient, as general-purpose notations in process modeling might not cover domain-specific aspects in the transformation pipeline. Finally, clinical guidelines are described in natural language, thus prone to multiple sources of ambiguity that will affect the translation processes [1, 14].

Contributions. This paper reports on the initial steps toward the computational support for clinical pathways. First, we present a dataset of clinical guidelines collected from all the regions in Denmark, which can serve for further semantic annotation and information extraction studies. The dataset comprises more

[3] https://pubmed.ncbi.nlm.nih.gov/?term=%28%28%222023%2F01%2F01%22%5BDate+-+Publication%5D+%3A+%222023%2F12%2F31%22%5BDate+-+Publication%5D%29%29+AND+%28cancer%29&sort=

than 90.000 clinical guidelines in 5 regions in Denmark. Second, we develop a conceptual model based on the document analysis of the dataset, to be used for information (e.g. process model) extraction. Third, we document how the conceptual model can be instantiated as a set of guidelines to build an annotated dataset, the challenges in its construction and validation, and the initial results in the application of LLMs for the extraction of pathway extraction components.

Document Structure: In Section 2 we document the process for data collection of the Danish clinical guidelines dataset, and we illustrate the results in Section 3. Section 4 introduces the conceptual model for the identification of process-related information in clinical guidelines. In Section 5 we validate the conceptual model via the construction of different annotation tasks and illustrate its challenges. Section 6 shows the results of process extraction tasks using SOTA LLMs. In Section 7 we present related work and we conclude in Section 8.

2 The MedicalInstruks Dataset

In Section 2.1 we document the selection criteria used to generate a baseline dataset for clinical guidelines. Section 2.2 documents the filtering process used in the baseline in place, starting from 92.695 unstructured and unmarked documents, to 15 fully annotated clinical descriptions.

2.1 Dataset Construction Criteria

We require a baseline dataset of guidelines to quantify the capabilities of LLMs in extracting clinical pathways. Our main criterion is that our dataset should contain realistic information used in clinical processes. Moreover, we considered the following requirements:

R1 The guidelines must be freely accessible.
R2 The documents in the dataset should all refer to service operating procedures (SOPs) in the medical sector.
R3 The documents should have a consistent format to facilitate automated processing.
R4 The corpus should cover various medical specialties to ensure its relevance to a wide audience, including physicians, nurses, and pharmacists.
R5 If possible, the dataset should include documents in English.

The Danish healthcare sector presents a unique opportunity to create such a dataset. Each of the five regions in Denmark (Syddanmark, Midtjylland, Nordjylland, Sjælland, and Hovedstaden) maintains a centralized guideline document system. These systems aggregate SOPs (locally referred to as *Instruks*) across all regional hospitals and are publicly available. This approach matches selection criteria R1–R4. Concerning criterion R5, the guidelines are only available in Danish. Thus, we needed to add a translation step to benefit from LLMs trained in English. While we considered English alternatives (e.g. [6]), their collection

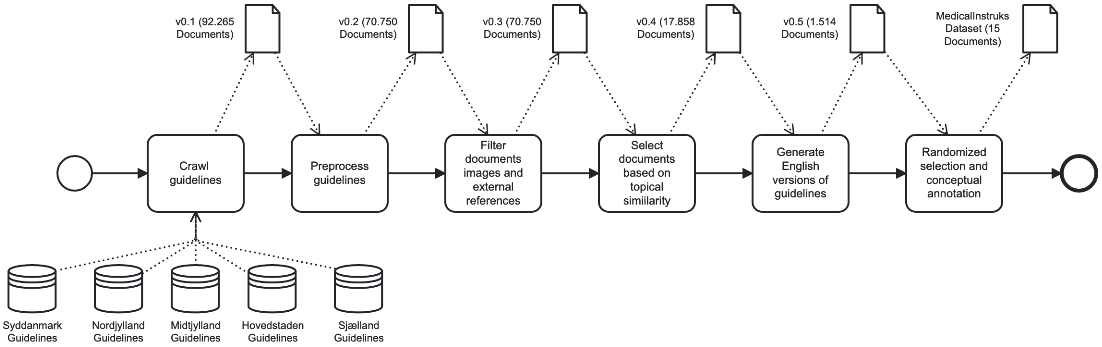


Fig. 1. Process view of the MedicalInstruks dataset construction

included documents not used in practice, such as PubMed abstracts and papers, thus violating R2. In comparison, constructing a dataset directly from the regional sources in Denmark adds ecological validity to our dataset.

Following individual permissions from each region for research use, we compiled a dataset that based on the crawled documents, forms the cornerstone of our study.

2.2 Data Collection and Refinement Process

This section describes the process followed to build the MedicalInstruks dataset departing from the documentation received from the regions. Our process is summarized in a BPMN diagram in Figure 1.

The dataset construction process starts with the crawled documents from the regions (v0.1). This version included several non-medical categories that were removed. The remaining medical guidelines were preprocessed (e.g. converted from HTML to Markdown, identifying titles and contents from the crawled documents, removing table of contents, removing empty, incorrect, or duplicated entries, etc). This constitutes the second version of the dataset and contains 70,750 documents.

Further analysis of the contents in v0.1 of the dataset showed that 12.816 $(18,12\%)$ documents contained references, images, links to internal or external documents, and tables. While process extraction from multimodal artifacts will be exciting research, we proceeded to filter these documents as it will complicate the information extraction task. Multimodal documents were modified to remove references and images. Documents with tables were kept unaltered, as the pilot selection showed them as an important artifact to encode process information. This constituted the v0.2 of the dataset.

The following step included clustering and selection. The 70.750 documents were not evenly distributed among the different areas. For instance, Pediatric Dosage Instructions and Child Healthcare Guidelines included roughly 7.500 documents, while Clinical Practice Standards and Specialist Treatment Guidelines included 90. Moreover, the guidelines included non-medical procedures, such as processes in logistics, human resources, economy, and cleaning. To minimize the risk of undersampling, we used clustering and further refinement based on topic appropriateness for healthcare. Two rounds of clustering (topic/sub-topic) were performed by applying term frequency-inverse document frequency (TF-IDF) transformation, removing terms occurring globally in the dataset. A KMeans clustering algorithm was then applied to the transformation, resulting in a mapping from each document to one cluster in twenty. Clusters obeying inclusion criteria (i.e.: containing clinical process information) were further analyzed via random sampling and analysis of the text contained in the clinical guideline. In addition, we defined minimum and maximum token limits to filter entries that only included links or documents where manual annotation would be unsuitable. Documents below 250 and over 8.000 tokens were removed. As a result, the clustered dataset in v0.3 included 17.858 documents and $21'814.800$ tokens.

Once the documents were clustered and selected, we translated the guidelines into English. We used *DeepL*[4] as it is considered the top-performing translation model [34]. To keep the validation effort feasible for the research team, we randomly selected $1,514$ documents from the included entries in v0.4 and translated them into English. The output of the translation was checked for consistency by two native speakers who revised random selections of the translated documents, confirming the validity of the translations[5]. Thus, v0.4 of the dataset constitutes a baseline dataset ideal as a seed for annotation tasks in guideline extraction from text.

The last version of the dataset included the manual annotation phase according to the conceptual model defined in Section 4. For this step, 31 documents were randomly assigned for annotation. The annotation step included the refinement of the conceptual model and annotation guidelines and involved the three authors of this paper. This sample also removed 16 documents because they contained information not considered part of the conceptual model and thus could only be partially annotated. The final version of the MedicalInstruks dataset included 15 fully annotated documents, with an average of $1,000$ annotated tokens per document.

3 Dataset results

Table 1 renders public the different versions of the dataset to make the research replicable. In particular, we consider that each version of the dataset has individual merit. v0.1 is the raw material, ensuring traceability to the original guidelines [16]. Moreover, it considers multi-modal process descriptions, a topic seldom explored in NLP4BPM tasks. v0.2 enables the exploration of multimodal

[4] `https://www.deepl.com/translator`

[5] Few, negligible cases of mistranslations occurred, for instance, the Danish word “sutten” translates to “Pacifier”, but it could be translated to “hickeys”, “suckle” or “booze”. However, the number of mistranslations was negligible.

Version	Description	Docs.	Lang.	Filters	Annotated	Availability
v0.1	Original Dataset	92.695	Danish	Raw	No	Zenodo [16]
v0.2	Pre-processed Documents	70.750	Danish	Multi-modalities removed	No	hugginface.co/...v0.2
v0.3	Selection based on topical similarity	17.858	Danish	[250, 8.000] tokens/document Selected according to R2 & R4	No	hugginface.co/...v0.3
v0.4	Danish - English translation	1.514	English		No	hugginface.co/...v0.4
v0.5	Annotated Dataset	15	English		Yes	hugginface.co/...v0.5

Table 1. Dataset versions and availability

extraction of processes including decision and process models. Moreover, it contains process descriptions in classical areas of interest in the BPM community, such as logistics and human resources. v0.4 contains English descriptions and is independent of any annotation scheme, thus facilitating its use for other annotation purposes, thus it can be used as a benchmark for existing annotation schemes such as [2,23]. Finally, v0.5 focuses only on clinical guidelines in English and showcases the application of our annotation guidelines. The MedicalProcessInstruks contains 4.4k tokens, 15 documents, and 270 annotated sentences, with an average token count of 16 per sentence.

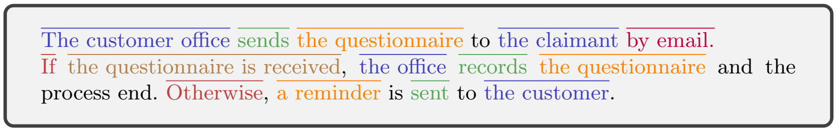
4 Conceptual Model

In this section, we propose CGPET, our conceptual model for the annotation of process elements in clinical guidelines. To establish CGPET, we considered three kinds of artifacts: (1) various notations for Computer-Interpretable Guidelines, which define the information that is necessary to represent a CPW, (2) an existing annotation schema for the annotation of process model elements in textual process descriptions, PET [2], and (3) a metamodel for declarative process models [23]. These are complimentary, given that the first provides insights into the need for a target representation (i.e., the CPW), and the latter two provide a starting point to annotate process information using declarative and imperative process semantics. Our proposed CGPET model combines these three artifacts, particularly by extending the PET schema with several (missing) components critical for the representation of computer-interpretable guidelines.

4.1 Key Components in Clinical and Computer-Interpretable Guidelines

Although there is no standardized notation or structure for the description of clinical guidelines, they generally share certain key components. CGPET captures the following elements commonly seen in analyses of pathways and guidelines [8,13,19]:

- **Goals**: also referred to as *outcomes*, are essential for guiding the direction of clinical care and evaluating their effectiveness.



The customer office sends the questionnaire to the claimant by email. If the questionnaire is received, the office records the questionnaire and the process end. Otherwise, a reminder is sent to the customer.

Fig. 2. Text fragment annotated using the entities of the PET schema (adopted from [2], relations omitted for clarity). Legend: *Activity*, *Activity Data*, *Actor*, *Further Specification*, *Gateway*, *Condition Specification*.

- **Indications and contra-indications**: Information defining the cases where a particular guideline applies (or not) for a patient.
- **Plan**: a description of tasks, decisions, and time conditions. Tasks and decisions may be sequential or concurrent.
- **Classification rules**: convert a patient value into a classification further used in the guideline, for example, a SpO2 of 88% is considered *critical* in a Covid-19 guideline.
- **Decision rules**: logical statements, flowcharts, or tables, defining a set of rules based on input variables coming from the patient's state (e.g., systolic pressure).

4.2 PET Annotation Schema

The PET (Process Extraction from Text) annotation schema was recently proposed [2] as part of an effort to provide a corpus of annotated textual process descriptions that can be used for training and evaluating approaches that extract process information from texts. In this sense, its goal is similar to what we want to achieve in our work, although the input and output formats that it focuses on differ from ours.

The core of the PET schema is the *Activity* entity, which corresponds to (the action of) a task performed in a process, e.g., *sends*, *records*, and *sent* in Figure 2. Each activity can be linked to other kinds of entities, such as *Activity data* to capture objects related to the activity, e.g., the the questionnaire being sent, *Actors* that either perform (The customer office) or are the recipient (the claimant) of the activity, and any *Further specification* that provides details on the execution of the activity, e.g., that the questionnaire is sent by email.

Next to activities, the PET schema defines *Gateway* and *Condition Specification* behavioral entities, which are used, together with the *Flow* relation, to define the control flow of a process, including choices and parallelism. For instance, the words If and Otherwise jointly define a choice (XOR) construct in the process, which PET captures through the *Same Gateway* relation. A *Condition Specification* can be used to annotate a condition that must be satisfied to perform a specific branch of a gateway, for instance, that a the questionnaire is received) before it can be recorded.

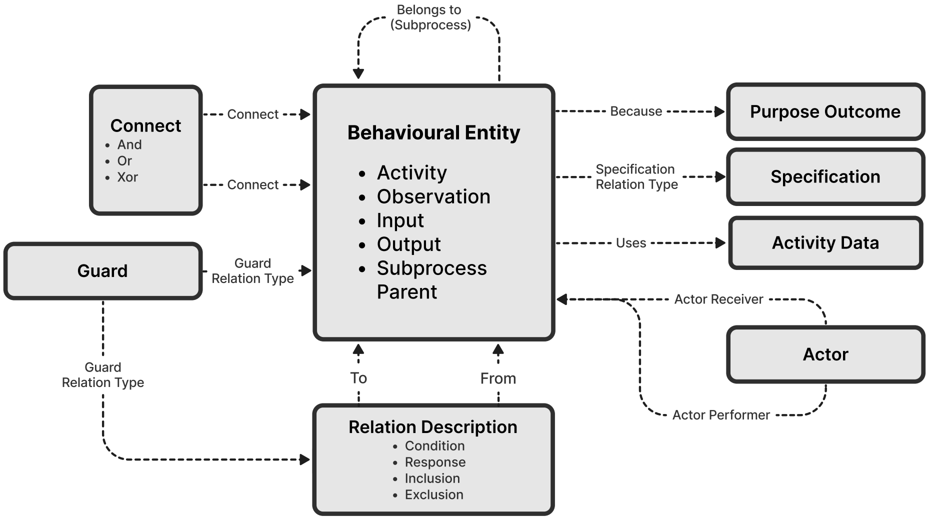


Fig. 3. CGPET Conceptual Model

4.3 Our proposal: CGPET

The PET schema provides a valuable starting point for our work. However, it is important to consider that textual process descriptions used as input for PET are simpler than clinical guidelines and lack corresponding elements for our conceptual model. Due to these differences, we identified the need for several adaptations to the PET schema while annotating several clinical guidelines from the MedicalProcessInstruks dataset. This resulted in CGPET, a proposed conceptual model for annotating process information in clinical guidelines, visualized in Figure 3. The CGPET annotation schema is tailored to how process elements are contained in clinical guidelines. The annotated elements can then be extracted and transformed into a more formalized representation of the procedural information using a language of choice, independent of whether it is an imperative or declarative language, a general-purpose process modeling notation, or a notation specific to CPWs.

- **Activities:** Naturally, the *Activity* entity still plays a central role in CGPET, given that activities form the core of any process. Examples of activities in the dataset are *"Check if the patient has a cavity"* and *"Preparing the CRRT device for treatment"*.
- **Activity purpose:** Next to associating activities with *Activity Data*, *Specification*, and *Actor* entities like in PET, we also define an additional entity: *Purpose Outcome*. This entity captures the underlying reason, goal, objective, or anticipated result of a clinical action, procedure, or recommendation. Figure 4 presents an example of the application of the *Purpose Outcome (PO)* entity. It describes how the Activity i.e. *"Postoperative epidural analgesia"* is performed for *"pain management"*.

- **Observations:** An *Observation* can refer to any information or data that is noted or recorded about a patient's health status, including symptoms, diagnoses, test results, risk factors such as smoking, or contextual information such as the patient's age or condition. Multiple observations can be evidenced in one sentence, for instance, the sentence *"The condition should be suspected in* ***seizures with increased heart rate, respiration rate, BP*** *and* ***temperature****, as well as* ***sweating*** *and* ***dystonic movements postures*** *in a* ***patient with severe acquired brain injury****"* contains 7 different observations ranging from specific symptoms to patient's conditions. Observations do not have a corresponding entity in the PET annotation guidelines.
- **Inputs and outputs:** An *Input* refers to any word or phrase that denotes a specific type of clinical measurement, score, or value relevant to performing an activity. These include, but are not limited to, physiological measurements, lab test scores, and specific clinical indices. An *Output* entity is its converse. Note that an Input differs from a Guard (see below), since an Input does not represent specific numerical values or thresholds but rather the type of the measurement or value, e.g., *blood pressure* or *heart rate*. Examples of inputs and outputs are "eGFR" and "study is performed without IV contrast" in *"If eGFR < 45 : The study is performed without IV contrast"*
- **Subprocesses:** Clinical guidelines frequently mention higher-level activities, later described in more detail. For example, the sentence *"It may be indicated to* ***administer morphine*** *to a woman in labor during the dilation phase of labor"* describes the existence of the activity **administer morphine**, which is later detailed: *"Regardless of age and weight: Naloxone 0.2 mg svt 0.5 ml. i.m. It can be repeated if needed."*. To capture this behavior we define a *Subprocess* as an entity to annotate higher-level steps, which can subsequently be linked to more specific activities using the *Belongs to the Subprocess* relation.
- **Control-flow relations:** Contrary to PET, clinical guidelines are discretionary, meaning that flows can be implicit rather than explicit. We annotate the control flow among activities directly using inter-activity relations, in the form of declarative process constraints [21]. CGPET supports *Condition*, *Response*, *No-response*, *Inclusion*, and *Exclusion* relations between activities. A condition relation describes precedents (e.g. *prescribe medicine* cannot be done unless *confirm diagnose* is done). Responses denote the imposition of obligations (e.g. if a *newborn is assessed to be in pain*, *a systematic pain assessment* **must** be performed). Its converse relation is No-response. Inclusions/Exclusions denote contextual information (e.g. if a *diagnosis is observed*, then a specific type of treatment *is pertinent*, or, conversely, should not be offered). The annotation example in Figure 4 uses the *Exclusion* relation to indicate that one activity excludes the other.
- **Guards:** Finally, a key control-flow addition to CGPET is the notion of a *Guard*. A guard refers to a specific type of information that defines conditions, limits, or thresholds in the clinical context. These entities often represent critical values or timeframes that impact clinical decisions, such as dosage limits, duration of treatment, or thresholds for test results. This can

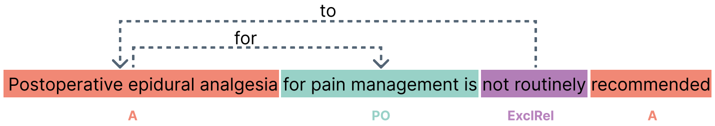


Fig. 4. Annotations of Purpose Outcome (PO) indicating the rationale behind the execution of an activity

include measurements (like volume or concentration), timeframes (like durations or frequencies), or any other quantifiable condition that affects clinical decisions. Note that guards play a similar role as the *Condition Specification* in PET, though with a broader purpose and in line with the terminology used in clinical guidelines (see Section 4.1).

5 Validation via Annotated Dataset Construction

The following section discusses the validation of the metamodel using existing clinical guidelines. The conceptual model was instantiated as an annotation guideline for process extraction as it is commonly done in NLP tasks. The guidelines were applied to a randomized selection of clinical guidelines (c.f. dataset construction v0.5 in Section 2.2). When uncertainties about the application of the guidelines arose, a lead annotator discussed each of the uncertainties with two senior annotators with experience in NLP and BPM. This process resulted in multiple iterations of the annotation guidelines. When ambiguities arose regarding specific terms that could not be inferred via the context, ChatGPT4 was used for clarification. The annotation process can be broken down into four main phases. First, the annotator must classify each document section according to whether it contains process-related information. Guideline documents are organized into definitions, main chapters, subchapters, and paragraphs, but the hierarchical layout does not necessarily correspond to process-centric information. For example, one of the documents describes an *"ordering guide for 18F-FDG PET Examination"*. Its main paragraphs include *target groups*, *definitions*, *procedure*, *indications*, *contraindications*, *patient preparation*, *execution*, *interpretations*, *doses* and *requisitions*. After a closer look at each of the sections it is possible to observe that the *procedure* section is empty, and the process information resides in *indications*, *contraindications*, *patient preparation*, and *execution paragraphs*. Thus, the role of the annotator is to parse the document structure and identify the paragraphs containing process-related sentences. Second, once all sentences containing procedural knowledge have been recognized, the annotator labels individual activities. Third, the remaining elements from the metamodel are then annotated according to their relation to labeled activities. Finally, the control-flow relations between activities are labeled.

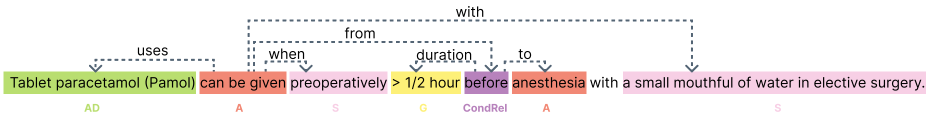


Fig. 5. Ambiguities in Clinical Guidelines: underspecification of timed constraints.

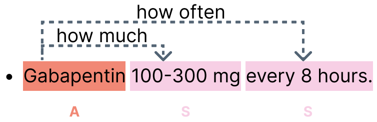


Fig. 6. Example 2 of Guideline Ambiguity.

5.1 Challenges: Textual Ambiguity in Clinical Guidelines

An important finding of the annotation process was how recurrent process-related ambiguities appear inside clinical guidelines. We provide a couple of examples. First, consider the annotations in Figure 5. They show an ambiguous course of drug administration. On the one hand, several time points are specified (preoperative and 0.5 hours before anesthesia). Likewise, the formulations are not exact. Although paracetamol can be administered during this period, it is not obligatory.

The example in Figure 6 shows an excerpt from a list of activities within the guideline. In this case, a connecting verb (e.g. administer) is missing, which does not create an issue for a human annotator given the reading context, but it may lead to problems when applying automated process extraction, as many approaches [2, 15] use verbs as the main identifier for activities.

Very different process models can emerge from parsing one ambiguous sentence in a clinical guideline. For instance, different ways of parsing relations occurring in *"If the procedure is to be performed in LA, blood tests are not required unless the operator has prescribed otherwise."* resulted in three different process graphs as presented in Figure 7 (here illustrated as DCR graphs [17]). This example is especially ambiguous as it contains a double condition with the action of performing a blood test. The process at the bottom left describes that a blood test is no longer possible if the procedure is carried out in room LA (exclude relation). On the other hand, a prescription can reactivate the activity. If both conditions are met, a decision problem will occur. In the process on the right side, the condition relation causes a blood test to be executed only if the procedure takes place in room LA, but at the same time, the exclude relation prevents this.

These examples demonstrate the implications of the ambiguities in the annotation process. As posed forward in [14], an ambiguity will take multiple interpretations as valid. In our annotation process, we discussed each ambiguity encountered and left them unchanged as long as one interpretation was correct.

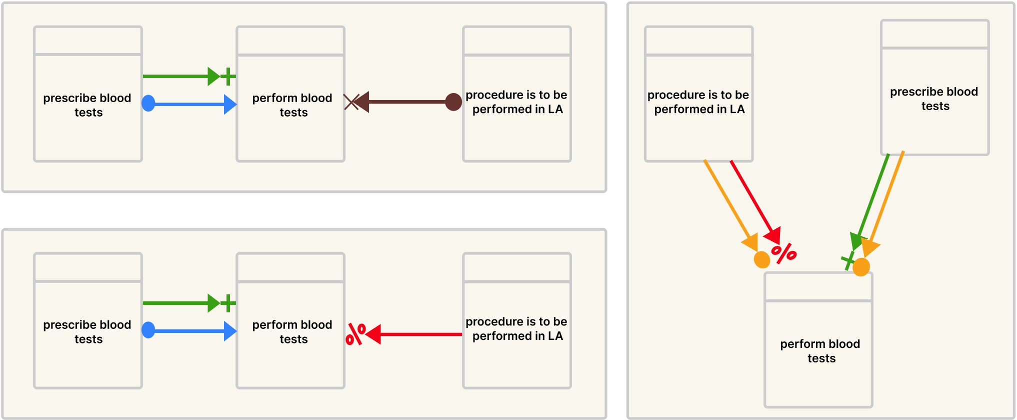


Fig. 7. Ambiguity effects: Three different DCR graphs based on the same description.

The amount of ambiguity and the lack of possibilities for automation made the annotation process extremely difficult and time-consuming. During the beginning of the annotation, one guideline annotation with 70-80 sentences took up to 8 hours (on average 6 min. per sentence). This time included noting ambiguities and looking up medical specialties. The refinement of the guidelines and the learning curve improved the efficiency of the annotation process to about 4 hours/document. Another factor that makes the process very time-intensive is the amount of layers needed to describe the model. For some of the documents, the time could be reduced to about 2 1/2 hours per document, combining manual annotations with a prompting approach where we used Mixtral [18] to semi-automate the annotation process, where a main annotator revised and accepted the outcome. The result of the annotation process became the MedicalInstruks dataset (c.f. v0.5 in Section 2.2) and is the first in-depth process-centered annotated corpus of medical clinical guidelines containing information about the current state of medical practice.

6 Initial Experiments on Process Extraction using LLMs

In this Section, we report on initial experiments using the MedicalProcessInstruks dataset. The process extraction pipeline can benefit from multiple NLP tasks: sentence classification (i.e. does this sentence contain process-related information?), Name-Entity Recognition (i.e. what are the activities in this sentence?) and Relation Extraction (i.e. are these two concepts related?). We focused on Named-Entity-Recognition and relation extraction. In particular, we want to explore whether a pre-trained language model renders process extraction feasible in an automated way. Moreover, we would like to compare its performance against few-shot guideline-based architectures. For the evaluation metrics, we use standard definitions of precision (i.e.: $\mathit{True\ Positives}/(\mathit{True\ Positives} +$

False Positives)), recall (i.e.: *True Positives*/(*True Positives*+*False Negatives*)), and F1 (i.e.: 2 × (*Precision* × *Recall*)/(*Precision* + *Recall*)) scores. For LLM predictions we include hallucination, defined as *Number of wrong predictions/ total number of predictions*.

The models selected for this evaluation were sourced from a public repository of pre-trained models[6]. Our initial selection included BioLink-BERT [38] and XLM-Roberta in the `large` model configuration [7], with 355 million parameters. For few-shot learning, we used three models: first, OpenAI NER Model GPT4-0125 Preview with Top_p : 0.9, and $Temperature$: 0.7. Second, Mixtral 8x7b instruct-v0.1 NER. Setup: Top_p : 0.9, temperature: 0.7. Finally, we compared with Guideline Prompting using GoLLIE-34B, an entity and relation extraction finetuned model based on CodeLlama using few-shot prompting combined with a guideline approach [31]. The same parameters were used for the Relation Extraction task. The experiments were carried out on an HPC cluster on two *Tesla A100-PCIE* graphics cards with 40GB memory each and the smaller models (BERT & Roberta) on a *Tesla V100-SXM2* with 32GB memory. Initially, our approach involved document segmentation without contextual information (that is, one sentence at a time), using a 5-fold cross-validation strategy and iterating over 3 epochs. Unfortunately, this strategy did not yield significant accuracy improvements for either BioLinkBERT or XLMRoberta. Subsequently, the models were subjected to an extended training regime of 6 additional epochs with the same 5-fold cross-validation. This phase demonstrated a marginal improvement, particularly with XLM-Roberta, which began to show the first signs of measurable performance. In our final and best approach, we used 10-fold cross-validation.

6.1 Results

We report the best results for both the pre-trained and the guideline approach in Tables 2, 3, and 4. Our initial expectation was that fine-tuned domain-specific models would better extract process elements from clinical guidelines against general-purpose models. Still, the poor performance of BioLinkBert challenged this. In contrast, XLM-Roberta, a model with a significantly larger general corpus, provided the best results for pre-trained models. Independently of the model and the class, the results of our experiments show an accuracy below random guesses, highlighting process extraction from clinical data as a complex task for pre-trained large language models. They also contrast against the high F1 scores for activity, actor, and specification classes in the PET dataset (0.81, 0.76, 0.19, respectively). The results for the few-shot approaches in Tables 3 and 4 show a difference on the performance against different tasks. While the relation extraction evidenced satisfactory results (particularly in the case of OpenAI), the NER task evidence that most of the entity extraction is below the level of random classifiers, and thus not ready to be used in production for full process extraction pipeline.

[6] Huggingface Transformers Library

	Precision	Recall	F1
Activity	0.5556	0.2170	0.3125
Activity Parent	0.2500	0.5000	0.3300
Actor	0.2600	0.6250	0.3700
Specification	0.4000	0.1100	0.1700
Input	1.0000	0.2500	0.4000
Observation	0.2720	0.3300	0.3000
Overall	**0.3220**	**0.1700**	**0.2260**

Table 2. NER results using XLMRobertaLarge

	OpenAI			**Mixtral**			**Gollie**		
Class	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Activity	0.4000	0.4819	**0.4372**	0.3125	0.2703	0.2899	0.2266	0.1576	0.1859
Activity Data	0.3592	0.4568	**0.4022**	0.2899	0.2740	0.2817	0.2273	0.0314	0.0552
Actor	0.3077	0.4000	0.3478	0.4500	0.4737	**0.4615**	0.0909	0.0286	0.0435
And	0.3571	0.4545	**0.4000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Condition Entity	0.3889	0.3889	**0.3889**	0.0556	0.0667	0.0606	0.0000	0.0000	0.0000
Exclusion Entity	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	0.0909	**0.1538**
Guard	0.3182	0.5385	**0.4000**	0.3333	0.0909	0.1429	0.2500	0.1429	0.1818
Inclusion Entity	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Input	0.1667	0.0909	0.1176	0.0000	0.0000	0.0000	0.2333	0.2800	**0.2545**
Observation	0.3780	0.3827	**0.3804**	0.2907	0.3247	0.3067	0.3158	0.2264	0.2637
Or	0.1250	0.2000	**0.1538**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Output	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Parent Activity	0.1154	0.1875	**0.1429**	0.0323	0.0714	0.0444	0.0000	0.0000	0.0000
Purpose Outcome	0.2308	0.2500	**0.2400**	0.1667	0.0833	0.1111	0.0000	0.0000	0.0000
Response Entity	0.3333	0.1818	**0.2353**	0.1667	0.0909	0.1176	0.0000	0.0000	0.0000
Specification	0,4444	0.2295	**0.3027**	0.3500	0.0648	0.1094	0.1765	0.0147	0.0271
Overall	**0.3426**	**0.3516**	**0.3470**	0.2299	0.1937	0.2103	0.2409	0.0839	0.1245

Table 3. Results on NER tasks against different LLMs. Hallucination: 0.0% (OpenAI), $3,61\%$ (Mixtral). Bold means best results.

6.2 Threats to Validity

We are aware of the following limitations of our research. Regarding the annotation process: the annotations were carried out by the annotators with computer science backgrounds. The absence of prior knowledge posed a risk and increased the annotation time. Our mitigation strategy included consultations with medical material, as well as disambiguation using LLMs. Concerning translation errors. The original guidelines were received in Danish, and an additional translation step was required to perform process extraction activities. While mitigation strategies such as random sampling with native speakers were in place, there is the risk that some of the texts may have been unnecessarily altered in the translation phase. Regarding the size of the annotated dataset: the annotated data set may be small to train an encoder-only model, thus having an influence in the low f1 scores reported. As a mitigation strategy, an annotation guideline was developed and every case of ambiguity was discussed among the authors. Further strategies may include inter-annotator agreement and data augmentation strategies for some of the underrepresented classes.

	OpenAI			Mixtral			Gollie		
Class	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Activity Actor-Performer rel.	0.8571	0.8000	**0.8276**	0.2273	0.8333	0.3571	0.0000	0.0000	0.0000
Activity Actor-Receiver rel.	0.6000	0.5000	**0.5455**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Activity Data rel.	0.9063	0.8788	**0.8923**	0.6667	0.7200	0.6923	0.8913	0.1646	0.4481
Activity Guard rel.	0.9231	1.0000	**0.9600**	0.7778	0.6364	0.7000	0.2857	0.1429	0.1905
Activity Parent rel.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Activity Purpose Outcome rel.	0.7273	1.0000	**0.8421**	0.6667	0.6667	0.6667	0.5000	0.1333	0.2105
Activity Specification rel.	0.8144	0.8587	**0.8360**	0.6727	0.5211	0.5873	0.0000	0.0000	0.0000
Actor rel.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7500	0.3243	**0.4528**
Condition rel.	0.6842	0.8667	**0.7647**	0.4167	0.6250	0.5000	0.0000	0.0000	0.0000
Condition Response rel.	0.1429	0.3333	**0.2000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Exclusion rel.	0.8571	0.8571	**0.8571**	0.8333	0.8333	0.8333	0.0000	0.0000	0.0000
Response rel.	1.0000	0,5000	**0.6667**	1.0000	0.4000	0.5714	0.0000	0.0000	0.0000
Inclusion rel.	0,0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Overall	**0.7817**	**0.8383**	**0.8090**	0.5320	0.5902	0.5596	0.6087	0.2090	0.3111

Table 4. Benchmarking LLMs on RE tasks. Hallucination: 4.1825% (OpenAI), 16.11% (Mixtral). Bold means best results.

7 Related Work

We can divide the related work into three streams: works in datasets for clinical guidelines, works in annotation schemes, and overarching works in process extraction in BPM.

Regarding datasets for clinical guidelines, Pedersen et al. [26] published a set of word embeddings using Standard Operating Procedure (SOP) documents from five Danish regions. Their dataset, MeDa-Bert, includes non-guideline-specific information such as books and Wikipedia entries. To focus on guideline-specific information, we replicated the crawling process and compiled a new dataset using permissions from each region. Compared to [26], we increase the ecological validity of this work by only considering guideline-specific information.

Concerning annotation schemes for clinical guidelines, a broad range of notations for knowledge extraction from Computer-interpretable guidelines has been proposed, for instance, Asbru [24], EON [25], GLIF [27], and PROforma [37]. Each notation uses specific labels and categories for each aspect, with high variability in the covered details. For example, Asbru offers the most developed syntax to formally express a wide range of intentions that can be used to define goals and purposes, whereas the same aspects are left optional in GLIF. By proposing a language-agnostic model, we want to generate as much structured text as possible while being neutral to one specific modeling language.

Concerning general works at the intersection of NLP and BPM, we can report some works on process discovery from unstructured texts. While process discovery from text artifacts has been explored earlier, the inputs have considered business process descriptions [2, 15, 21, 22, 32], e-mail communications [36], cooking recipes [29], or legal documents [20]. To the best of our knowledge, this is the first work exploring the extraction of process elements from the complexity of clinical guidelines. From these works, only [2] provides both a set of annotation guidelines and a dataset for verification, being the closest to our aims (even though the domains are different). For the experimental setup, [3] evaluated few-shot extraction using GPT-3.5, being a good point of comparison. In the medical domain, the recent work by Bombieri et al. [4] studied the sentence classification task to identify procedural knowledge in robotic-assisted surgical text using LLMs. Their approach may be complementary to our work, being sentence classification a pre-condition to the process extraction phase in our extraction workflow.

8 Concluding Remarks

This work is the first of its nature aiming at exploring process extraction from unstructured documents in a highly complex environment, the medical domain. The dataset collected is a legitimate representation of the complex decisions and documents that a knowledge-worker need to deal with. In itself, the dataset (and the different versions from v0.1 to v0.5) evidence multiple challenges when extracting process information, including multiple modalities, long documents, non-standardized layouts, and semantic ambiguities. Moreover, our paper presents a conceptual model of the process-related information found in the documents. Such a conceptual model can be instantiated as a set of clinical guidelines, and it is independent of a specific modeling language, thus allowing hybrid combinations of declarative and imperative approaches. Our validation via annotation of a smaller dataset evidenced the complexity of the tasks for manual annotators, and our experiments using LLMs show ample room for improvement for this complex task. In future work, we would like to enrich the semantic annotations and explore the impact of multi-modal information in process extraction.

Acknowledgments This work was supported by the research grant “Center for Digital CompliancE (DICE)” (VIL57420) from VILLUM FONDEN.

References

1. Van der Aa, H., Leopold, H., Reijers, H.A.: Dealing with behavioral ambiguity in textual process descriptions. In: BPM. pp. 271–288. Springer (2016)
2. Bellan, P., van der Aa, H., Dragoni, M., Ghidini, C., Ponzetto, S.P.: Pet: an annotated dataset for process extraction from natural language text tasks. In: BPM workshops. pp. 315–321. Springer (2022)
3. Bellan, P., Dragoni, M., Ghidini, C.: Leveraging pre-trained language models for conversational information seeking from text (2022)
4. Bombieri, M., Rospocher, M., Dall’Alba, D., Fiorini, P.: Automatic detection of procedural knowledge in robotic-assisted surgical texts. International Journal of Computer Assisted Radiology and Surgery **16**, 1287 – 1295 (2021)
5. Burgers, J.S., Grol, R., Klazinga, N.S., Mäkelä, M., Zaat, J.: Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs. International Journal for Quality in Health Care **15**(1), 31–045 (2003)
6. Chen, Z., Cano, A.H., et al., A.R.: Meditron-70b: Scaling medical pretraining for large language models (2023)

7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F.e.a.: Unsupervised cross-lingual representation learning at scale. In: Procs. of the 58th Annual Meeting of the ACL. pp. 8440–8451 (2020)
8. De Bleser, L., Depreitere, R., De Waele, K., Vanhaecht, K., Vlayen, J., Sermeus, W.: Defining pathways. J Nurs Manag **14**(7), 553–563 (Oct 2006)
9. Di Ciccio, C., Marrella, A., Russo, A.: Knowledge-intensive processes: characteristics, requirements and analysis of contemporary approaches. J. on Data Sem. **4**, 29–57 (2015)
10. Edwards, L.: The EU AI Act: a summary of its significance and scope. Artificial Intelligence (the EU AI Act) **1** (2021)
11. Fahland, D., Lübke, D., Mendling, J., Reijers, H., Weber, B., Weidlich, M., Zugal, S.: Declarative versus imperative process modeling languages: The issue of understandability. In: BPMDS 2009. pp. 353–366. Springer (2009)
12. Figl, K., Di Ciccio, C., Reijers, H.A.: Do declarative process models help to reduce cognitive biases related to business rules? In: ER. pp. 119–133. Springer (2020)
13. openEHR Foundation: openehr specification proc, `https://specifications.openehr.org/releases/PROC/Release-1.6.0/overview.html#_clinical_practice_guidelines_cpgs`, accessed: 2024-01-15
14. Franceschetti, M., Seiger, R., López, H.A., Burattin, A., García-Bañuelos, L., Weber, B.: A characterisation of ambiguity in bpm. In: ER. pp. 277–295. Springer (2023)
15. Friedrich, F., Mendling, J., Puhlmann, F.: Process model generation from natural language text. In: CAiSE. pp. 482–496. Springer, Berlin, Heidelberg (2011)
16. Grathwol, D., van der Aa, H., López, H.A.: Instruks Dataset - A dataset of Clinical Guidelines in Denmark (Oct 2024). `https://doi.org/10.5281/zenodo.11396622`, `https://doi.org/10.5281/zenodo.11396622`
17. Hildebrandt, T.T., Mukkamala, R.R.: Declarative event-based workflow as distributed dynamic condition response graphs. In: Places (2011)
18. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
19. Latina, R., Salomone, K., D'Angelo, D., Coclite, D., Castellini, G., Gianola, S., Fauci, A., Napoletano, A., Iacorossi, L., Iannone, P.: Towards a new system for the assessment of the quality in care pathways: An overview of systematic reviews. Int J Environ Res Public Health **17**(22) (Nov 2020)
20. López, H.A.: Challenges in legal process discovery. In: ITBPM@ BPM. pp. 68–73 (2021)
21. López, H.A., Strømsted, R., Niyodusenga, J.M., Marquard, M.: Declarative process discovery: Linking process and textual views. In: CAiSE Forum. pp. 109–117. Springer International Publishing, Cham (2021)
22. López, H.A., Marquard, M., Muttenthaler, L., Strømsted, R.: Assisted declarative process creation from natural language descriptions. In: EDOC Workshops. pp. 96–99 (2019). `https://doi.org/10.1109/EDOCW.2019.00027`
23. López A., H.A., Simon, V.: How to (re)design declarative process notations? a view from the lens of cognitive effectiveness frameworks. In: 15th IFIP Working Conference on the Practice of Enterprise Modeling (POEM). pp. 81–97 (11 2022)
24. Miksch, S., Shahar, Y., Johnson, P.: Asbru: a task-specific, intention-based, and time-oriented language for representing skeletal plans. In: Procs. of KEML-97. pp. 9–19. Milton Keynes, UK (1997)

25. Musen, M.A., Tu, S.W., Das, A.K., Shahar, Y.: EON: a component-based approach to automation of protocol-directed therapy. J Am Med Inform Assoc **3**(6), 367–388 (Nov 1996)
26. Pedersen, J.S., Laursen, M.S., Vinholt, P.J., Savarimuthu, T.R.: Meda-BERT: A medical danish pretrained transformer model. In: The 24rd Nordic Conference on Computational Linguistics (2023)
27. Peleg, M., Boxwala, A.A., Ogunyemi, O., Zeng, Q., Tu, S., Lacson, R., Bernstam, E., Ash, N., Mork, P., Ohno-Machado, L., et al.: Glif3: the evolution of a guideline representation format. In: Proceedings of the AMIA Symposium. p. 645. American Medical Informatics Association (2000)
28. Pesic, M., Schonenberg, H., Van der Aalst, W.M.: Declare: Full support for loosely-structured processes. In: EDOC. pp. 287–287. IEEE (2007)
29. Qian, C., Wen, L., Kumar, A., Lin, L., Lin, L., Zong, Z., Li, S., Wang, J.: An approach for process model extraction by multi-grained text classification. In: CAiSE 2020. pp. 268–282. Springer (2020)
30. Rotter, T., Kinsman, L., James, E.L., Machotta, A., Gothe, H., Willis, J., Snow, P., Kugler, J.: Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane database of systematic reviews (3) (2010)
31. Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O.L., Rigau, G., Agirre, E.: Gollie: Annotation guidelines improve zero-shot information-extraction (2023)
32. Sànchez-Ferreres, J., Burattin, A., Carmona, J., Montali, M., Padró, L.: Formal reasoning on natural language descriptions of processes. In: BPM. pp. 86–101. Springer International Publishing, Cham (2019)
33. Schnabel, M., Kill, C., El-Sheik, M., Sauvageot, A., Klose, K., Kopp, I.: From clinical guidelines to clinical pathways: development of a management-oriented algorithm for the treatment of polytraumatized patients in the acute period. Der Chirurg; Zeitschrift fur Alle Gebiete der Operativen Medizen **74**(12), 1156–1166 (2003)
34. Sebo, P., de Lucia, S.: Performance of machine translators in translating french medical research abstracts to english: A comparative study of DeepL, google translate, and CUBBITT. PLoS One **19**(2), e0297183 (Feb 2024)
35. Sitter, H., Prünte, H., Lorenz, W.: A new version of the programme algo for clinical algorithms. In: Medical Informatics Europe'96, pp. 654–657. IOS Press (1996)
36. Soares, D.C., Santoro, F.M., Baião, F.A.: Discovering collaborative knowledge-intensive processes through e-mail mining. Journal of Network and Computer Applications **36**(6), 1451–1465 (2013)
37. Sutton, D.R., Fox, J.: The syntax and semantics of the PROforma guideline modeling language. J Am Med Inform Assoc **10**(5), 433–443 (Jun 2003)
38. Yasunaga, M., Leskovec, J., Liang, P.: LinkBERT: Pretraining language models with document links. In: Procs. of the 60th Annual Meeting of the ACL. pp. 8003–8016. ACL, Dublin, Ireland (May 2022)