End-to-End Relation Extraction of Pharmacokinetic Estimates from the Scientific Literature

Ferran Gonzalez Hernandez[1,†,*], **Victoria C. Smith**[2,8†], **Quang Nguyen**[2,†], **José Antonio Cordero**[3], **Maria Rosa Ballester**[3,4], **Màrius Duran**[3], **Albert Solé**[3], **Palang Chotsiri**[5], **Thanaporn Wattanakul**[6], **Gill Mundin**, **Watjana Lilaonitkul**[7], **Joseph F. Standing**[8,9], **Frank Kloprogge**[10]

[1]Department of Computer Science, UCL, UK [2]Institute of Health Informatics, UCL, UK
[3]Blanquerna School of Health Sciences, Ramon Llull University, Spain [4]Institut de Recerca Sant Pau Barcelona, Spain
[5]Clinical Pharmacology, Modelling and Simulation, Parexel International, Thailand
[6]Mahidol Oxford Tropical Medicine Research Unit, Thailand [7]Global Business School for Health, UCL, UK
[8]Great Ormond Street Institute for Child Health, UCL, UK [10]Institute for Global Health, UCL, UK
[9]Department of Pharmacy, Great Ormond Street Hospital for Children, UK
[†] equal contribution [*]{ferran.hernandez.17, f.kloprogge}@ucl.ac.uk

Abstract

The lack of comprehensive and standardised databases containing Pharmacokinetic (PK) parameters presents a challenge in the drug development pipeline. Efficiently managing the increasing volume of published PK Parameters requires automated approaches that centralise information from diverse studies. In this work, we present the Pharmacokinetic Relation Extraction Dataset (PRED), a novel, manually curated corpus developed by pharmacometricians and NLP specialists, covering multiple types of PK parameters and numerical expressions reported in open-access scientific articles. PRED covers annotations for various entities and relations involved in PK parameter measurements from 3,600 sentences. We also introduce an end-to-end relation extraction model based on BioBERT, which is trained with joint named entity recognition (NER) and relation extraction objectives. The optimal pipeline achieved a micro-average F1-score of 94% for NER and over 85% F1-score across all relation types. This work represents the first resource for training and evaluating models for PK end-to-end extraction across multiple parameters and study types. We make our corpus and model openly available to accelerate the construction of large PK databases and to support similar endeavours in other scientific disciplines.

1 Introduction

Pharmacokinetics (PK) aims to quantify drug exposure through the study of drug absorption, distribution, metabolism and excretion (ADME). Drug PK profiles inform the selection of drug candidates and establish therapeutically relevant doses and dosing schedules (Morgan et al., 2012; Reichel and Lienau, 2016). Population PK models, i.e. nonlinear mixed-effects models, have played a significant role over the last decades in characterising PK properties through parameterising PK time series data. This has contributed to improved accuracy of predicting PK profiles across all stages of the drug development process.

Prior data from similar drug compounds are often used to initialise Population PK models and are also relevant for pre-clinical PK predictions for novel compounds (Dearden, 2007; Berellini and Lombardo, 2019; Wang et al., 2019). However, the primary challenge in collating prior PK data is the lack of comprehensive, standardised and open-access databases of PK parameter estimates, which has been recognised as a significant limitation in the drug development pipeline (Kumar et al., 2021; Mould and Upton, 2013; Grzegorzewski et al., 2021; Wang et al., 2009). Existing databases (Grzegorzewski et al., 2021; Wong et al., 2019) are manually curated from scientific literature and are limited to a few drugs. Consequently, researchers must manually compile PK information from the scientific literature (Grzegorzewski et al., 2021; Lombardo et al., 2018). The ability to automatically extract and centralise PK data from the scientific literature is of great interest to solidify existing PK knowledge and improve parameter predictions.

Annotated biomedical datasets have facilitated the development of state-of-the-art models for identifying many biomedical entities and their relationships in free text. However, no such annotated data exists for PK. In this work, we present a new dedicated corpus for Named Entity Recognition (NER) and Relation Extraction (RE) of PK data from scientific articles. This corpus is manually annotated at the sentence level by domain experts and involves entities and relations between PK pa-

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 144–154

rameter names, estimated values, deviation values, units and comparative terms. We also develop an end-to-end relation extraction architecture based on adapting the SpERT model (Eberts and Ulges, 2019) and training it on our corpus to assess the feasibility of automated extraction of PK parameter estimates. Our contributions are as follows:

- The PRED corpus, a publicly available corpus[1] of manually annotated entities and relations between PK parameter names, the central and deviation values, their units and comparative terms. PRED consists of 1764 entity mentions and 2016 relations annotated across 3600 sentences from scientific articles.
- A novel RE pipeline[2], trained and evaluated on PRED, for tackling the extraction of PK parameter estimates from the scientific literature. We compare architectures that model NER and RE jointly against models that optimise for a single task and assess the effect of domain-specific pre-training.

2 Related Work

Automated text mining approaches have been extensively explored regarding drugs and chemicals (Krallinger et al., 2015; Lee et al., 2020; Sung et al., 2022), drug-drug interactions (Herrero-Zazo et al., 2013; Segura Bedmar et al., 2013; Kolchinsky et al., 2013, 2015), and biochemical kinetics (e.g. enzyme kinetics) (Hakenberg et al., 2004; Spasić et al., 2009; Tsay et al., 2009). However, little research has been conducted on automatically extracting PK data from text.

Wang et al. (2009) explored pattern-based approaches for a single PK parameter for one drug. However, extending this approach to other PK parameters, drugs, and study designs becomes unfeasible due to the high diversity of surface forms. Instead, approaching PK information extraction with machine learning approaches has the potential to model a higher variability of PK parameters and relations effectively. Previously, Hernandez et al. (2021) presented an automated pipeline to identify scientific publications reporting PK parameter estimates measured *in vivo*. Subsequently, Hernandez et al. (2024) released a large annotated dataset of PK parameter mentions in the scientific literature and fine-tuned BioBERT (Lee et al., 2020) to perform NER of PK parameters. However, to our knowledge, no study has yet tackled the task of end-to-end relation extraction of PK parameter estimates, which represents a crucial step to automatically construct PK databases useful for drug development.

3 Methods

3.1 Corpus construction

The PRED corpus was developed to train and evaluate end-to-end pipelines that extract PK measurements from sentences and can be found at https://zenodo.org/records/11187303. All the relations tackled in this task appeared between entities within the same sentence.

Data Source

The following pipeline was applied to create a candidate pool of sentences. A PubMed search for *"pharmacokinetics"* was initially conducted in June 2020 to retrieve articles. The pipeline from Gonzalez Hernandez et al. (2021) retrieved 114,921 relevant publications reporting PK parameters. Out of these, 10,132 articles (8.82%) were accessible in full text from the PMC OA subset[3], while only abstracts were available for the rest. Both, abstracts and full-text articles were downloaded in XML format from PubMed[4] and PMC[5] FTP sites. The PubMed Parser (Titipat and Acuna, 2015) was used to parse the XML files, and paragraphs from the introduction section were excluded. The scispaCy sentence segmentation algorithm (Neumann et al., 2019) split abstracts and paragraphs into sentences. The resulting sets were randomly sampled to produce a candidate pool of 1,443,044 sentences, with a balanced proportion of sentences from the abstract and full-text. Noticeably, 16.4% of sentences from the initial candidate pool mentioned PK parameters. Therefore, a filtering protocol was applied to promote the development of a corpus with a wide variety of PK mentions and relation instances. The PK NER model from Hernandez et al. (2024) was first applied to all the candidate pool sentences. Then, we selected sentences that at least had (1) one PK mention detected by the NER model and (2) a numerical value. From the resulting pool of sentences, 3600 instances were randomly sampled

[1] https://zenodo.org/records/11187303
[2] https://github.com/PKPDAI/PKRelations
[3] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[4] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[5] https://ftp.ncbi.nlm.nih.gov/pub/pmc/

without replacement and divided into 2100, 500 and 1000 instances for the training, development and test sets, respectively.

Annotation

The annotation team comprised 12 individuals with extensive PK expertise and familiarity with the different parameters and study types in PK literature. Annotation guidelines were developed and distributed to the annotators before labelling and updated as new complex cases emerged during annotation. To ensure consistency, annotations were performed in batches of 200 sentences, following a three-step procedure: (1) initial annotation by one PK expert, (2) review by another PK annotator and (3) final check focusing on span boundary consistency by an annotator with bio-NLP experience. After the second step in each batch, comments from the first and second steps were reviewed, and feedback regarding incorrect annotation patterns was given to the annotators. Inter-annotator agreement was examined using the pair-wise F_1 score on 200 sentences, and the mean was computed across each pair of annotators. For further details on the annotation guidelines and interface, please see Appendix A.

Task Definition

End-to-end RE aims to identify named entities and extract relations between them. Given some input text X, the output of any end-to-end RE system is a list of triplets in the form of (s_i, s_j, r) where $s_i, s_j \in S$ and $r \in R$ and S denote all the possible spans in X and R the set of pre-defined relation types (Zhong and Chen, 2020). Hence, the annotated data was represented as a list of sentences, each with their corresponding list of relation triplets and compared to model predictions in the same format. Because end-to-end RE systems need to (1) identify candidate spans and (2) predict relation classes for pairs of spans, this task is often decomposed into two sub-tasks:

1. Named Entity Recognition: which attempts to detect the list of entity mentions (i.e. spans) and their type $\mathcal{E} = \{PK, Units, Value, Range, Compare\}$ from the input text X.
2. Relation Extraction: which compares all pairs of spans in X and outputs a relation class for each pair $R = \{Central_{val}, Deviation_{val}, Related\}$.

For step 1, the following entities were considered and annotated at the sentence level:

1. **PK**: Mentions of parameters. This entity refers to spans mentioning PK parameters, and it is the same concept as the entity described by Hernandez et al. (2024).
2. **Units**: Spans of text corresponding to units of numerical PK estimations.
3. **Value**: Spans encapsulating numerical estimations related to PK parameters (i.e. central and deviation values).
4. **Range**: Two values defining the boundaries of a PK estimation.
5. **Compare**: Textual mentions that provided information about whether a specific value/range mention was the extreme of an estimated parameter. This entity appeared with low frequency, but it was important for detecting extracted measurements that were not central estimations of a certain parameter.

For step 2, three relations classes were considered between entities to extract structured information from raw sentences in a usable format. Please note the directionality of relations is not considered in this work as it is not necessary for the desired tabular output (see Figure 1):

1. **Central**$_{val}$[6]: This relation type happened between PK parameter mentions and their estimated values or ranges. This involved central measurements of the parameter but not measures of deviation or % of increase concerning other experimental conditions. The entities between which this relation could happen were:
   - PK $\leftrightarrow$ Value/Range
2. **Deviation**$_{val}$[7]: This relation type informed whether a specific measurement was the deviation of a central measurement and only happened between the entities:
   - Value/Range $\leftrightarrow$ Value/Range (involved in a $Central_{val}$ relation)
3. **Related**: This relation type complemented values/ranges with their units or compare terms and only happened between the following entities:

[6] Abbreviated as C_VAL in the annotation interface.
[7] Abbreviated as D_VAL in the annotation interface.

- Compare ↔ Value/Range
- Units ↔ Value/Range

3.2 Pipeline

Recent work has shown that sharing token representations and modelling NER and RE tasks simultaneously in a multi-task setting can enhance performance in both tasks (Bekoulis et al., 2018; Luan et al., 2019; Eberts and Ulges, 2019). This might be especially relevant in our corpus, where spans were only considered entities if they were part of a relation. For this reason, we propose an architecture to model NER and RE jointly to share encoded knowledge from both tasks.

Multi-task Architecture

Our multi-tasking architecture (illustrated in Figure 2), was inspired by the architecture in the SpBERT model developed by Eberts and Ulges (2019). The main modification was using sequential BIO labelling (Palen-Michel et al., 2021; Gu et al., 2021) instead of a span-based approach, as the PRED data does not contain overlapping spans. There was also no need to predict the directionality of relations for our work, so entity pairs were arranged in order of appearance in the original text. Finally, due to only one relation type existing between entity pairs in PRED, a softmax activation was used instead of a sigmoid activation.

Using the BERT tokenizer, an input sentence is initially tokenised into a sequence of sub-words. Then, tokens are passed through an encoder that aims to incorporate contextual information in each token's representation. The output embeddings from the encoder ($T_1, T_2, ...T_N$) are then used to (1) recognise entities through the token classifier using the BIO scheme, (2) generate candidate pairs of predicted entities and (3) classify all pairs of recognised entities with a relation classifier. NER and RE use the same encoder to generate contextual representations of input tokens and have one task-specific classification layer for each sub-task. We assessed the effect of domain-specific pretraining by comparing BERT$_{BASE}$ (Devlin et al., 2018) and BioBERT v1.1 (Lee et al., 2020) as the encoder.

Named Entity Recognition Task

NER was treated as a sequential labelling problem where each output token representation from the encoder (T_i) was classified into one unique BIO scheme class using a feed-forward layer with a sigmoid activation function. The model was trained with cross-entropy loss over token-level labels $\mathcal{L}_{NER}$.

Relation Extraction Task

After NER is performed in a specific sentence, all potential pairs of predicted spans are arranged and filtered before going to the relation classifier. Then, each candidate entity pair was classified into one relation class $[Central_{val}, Deviation_{val}, Relation, No\ Relation]$. Following Taillé et al. (2020), the representation of those spans composed of multiple tokens was generated by max-pooling their contextual token embeddings. Given the effective results of the max-pooling strategy presented by Eberts and Ulges (2019), no other fusion functions were analysed. The input to the relation classifier $x(s_1, s_2)$ was the concatenation of the two-span representations e(s_1) and e(s_2) with their context representation c(s_1, s_2):

$$x(s_1, s_2) = [e(s_1); c(s_1, s_2); e(s_2)] \quad (1)$$

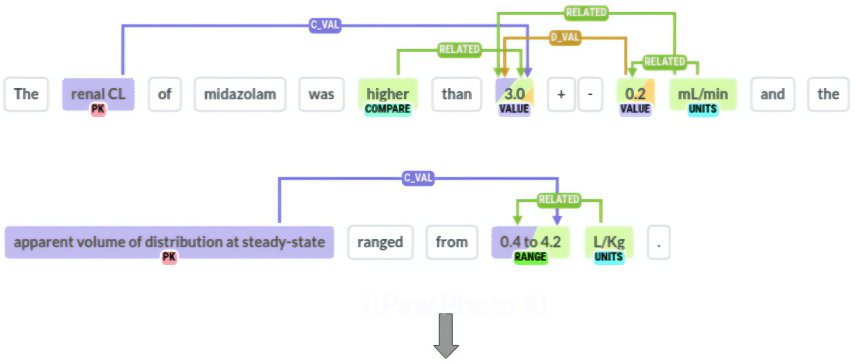
The context representation for two spans was generated by max-pooling all tokens strictly between them. If there were no tokens present between two spans $c(s_1, s_2) = 0$. Relations between entities were symmetric (non-directional) in the PRED corpus, and no overlapping spans were annotated. As a consequence, $e(s_1)$ and $e(s_2)$ were arranged according to their relative position in the sentence from left to right. Analogous to the token classifier, a single-feed forward layer was used to classify each candidate span pair. Since only one relation class could be associated between two entities, a softmax operation was used as an activation function. The model was trained with cross-entropy loss over relation classes, $\mathcal{L}_{RE}$.

Training and Optimisation

All the parameters from the encoder, the token and the relation classifier were fine-tuned during the training phase. Given sentences with annotated entities and relations, the loss was computed jointly by adding the NER and RE losses:

$$\mathcal{L} = \mathcal{L}_{NER} + \mathcal{L}_{RE} \quad (2)$$

Both losses were averaged over each batch's samples. Each batch consisted of B sentences from which samples were drawn for both classifiers. For the token classifier, the loss was computed for all



Parameter	Central measure	Units	Deviation	Deviation Units	Compare
renal CL	3.0	mL/min	0.2	mL/min	higher
apparent volume of distribution at steady-state	0.4 to 4.2	L/Kg			

Figure 1: The top panel shows a sentence where all entities and relations have been annotated. The bottom panel shows how the annotated entities and relations can be mapped into a tabular format that can be integrated into a database of PK measurements.

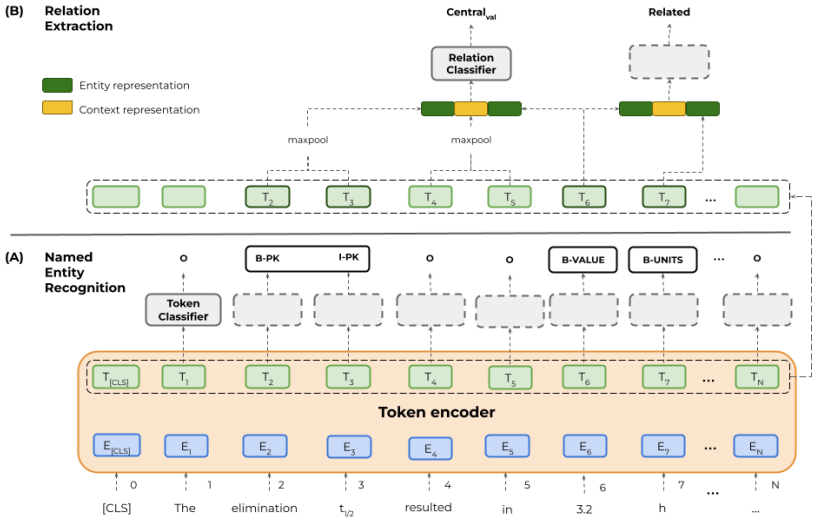


Figure 2: The model first receives a sequence of token embeddings (blue boxes, E_i) and goes through the encoder layers to generate a sequence of contextual token embeddings (green boxes, T_i), which are shared in both tasks. Then, (A) contextual token embeddings go through the token classifier (feed-forward layer) to output BIO labels that will allow recognising entities. (B) Entities and contexts (span between two entities) are represented by max-pooling their contextual token embeddings. Finally, pairs of entities are concatenated with their context representation and passed through the relation classifier (feed-forward layer).

tokens in the batch using the BIO labels. For the relation classifier, ground truth (annotated) entities were used to generate candidate pairs at training time. Negative samples (*No Relation* class) were generated with all candidate entity pairs not labelled with a relation during the annotation phase. At inference time, only those entities predicted by the NER module were passed to the RE classifier instead of using ground truth entities.

Models were trained for 50 epochs and evaluated on the development set after each epoch, saving the model state with the highest $Central_{val}$ F_1 score. The maximum sequence length for all experiments was set to 256, the batch size to 8, and the learning

rate to $\mu = 2e^{-5}$. The Adam Optimizer with a linear weight decay of 0.05 was used, and a dropout probability of 0.1 was applied on all layers. All experiments were run on a single GPU, NVIDIA Titan RTX (24GB).

Evaluation

Precision, Recall, and F_1 scores were computed for NER and RE. For NER, scores were based on strict matching of entity boundaries and types. F_1 scores were calculated per entity, with macro and micro-averages across entity types for overall system evaluation. In RE, focus was on F_1 score of $Central_{val}$ relations, as predicting $Deviation_{val}$ or $Related$ relations without $Central_{val}$ renders extracted data useless. Micro-averaged F_1 scores for NER and $Central_{val}$ relations for RE served as the main metrics for comparing different architectures on the PRED corpus.

4 Results and Discussion

4.1 Corpus Statistics

The main statistics for the PRED dataset are presented in Table 1. A total of 3,600 sentences were annotated, from which 56.42% contained annotated entities and relations. Sentences were evenly sampled from full-text and abstract sections. A total of 13,404 entity mentions were annotated. 12,411 relations were annotated, most coming from the $Related$ and $Central_{val}$ classes. The number of annotated $Central_{val}$ relations was over 2.5 times the number of $Deviation_{val}$, indicating that measures of deviation are not often reported along with central measures of PK parameters (only in 35.8% of cases).

4.2 Annotator Agreement

The average micro and macro-F_1 scores for NER were 88.74% and 92.36%, respectively, exhibiting high agreement on entity surfaces on the first annotation phase. For RE, the average pair-wise scores were 93.02%, 94.47% and 83.2% for $Related$, $Deviation_{val}$ and $Central_{val}$, respectively. A lower agreement was obtained between central values and their PK parameter mentions, mostly caused by disagreement on parameter span boundaries.

4.3 Multitask Model Performance

The effect of using a multi-task (MT) learning approach, jointly optimising NER and RE, was compared against a model only optimising for NER. BioBERT was used as an encoder in both cases. The MT architecture saved the model with the best $Central_{val}$ $F1$ on the development set, while micro-averaged $F1$ was used as a metric to select the best model for the no-MT experiment. Table 2 shows the NER performance on the test set for each entity type and the macro and micro-averaged $F1$ scores after ten runs of each experiment. Higher performance was obtained when using the MT architecture for all entities in the PRED corpus. Although the performance gain was relatively low ($\approx$+$\Delta F1$ 0.5%), the consistency of this gain across all entity types suggests that having the RE objective combined with NER helped the model perform better on NER. Finally, we noted higher interquartile variance for $Range$ and $Compare$ entities, which were the ones with the least number of annotations.

Although the performance gain of the MT architecture was small, such an approach also helped reduce the number of parameters required to model the task by only having one encoder. These results indicate that sharing token representations and optimising a single loss for NER and RE is beneficial for extracting PK measurements from the scientific literature compared to treating both tasks independently.

The MT solution's performance on the RE task is summarized in Table 3. Results show successful linking of deviation measurements and units in most cases. Notably, when values and units are correctly detected, their relation often requires minimal context, especially with a short distance between them. Additionally, the context between units and values typically lacked other units, simplifying extraction. Similarly, the context between central and deviation values often lacked other value entities. Therefore, with high NER performance for $Value$ and $Units$, few errors were observed for $Deviation_{val}$ and $Related$ relations. $Central_{val}$ relation showed relatively high performance, indicating consistency in dataset annotation and effective end-to-end modeling. Errors in $Central_{val}$ predictions mostly stemmed from incorrect NER predictions and sentences mentioning multiple parameters and values. However, some incorrect predictions of PK entities partially matched PK parameters, suggesting $Central_{val}$ performance could be a lower bound for PK measurement extraction. F_1 scores in Table 3 were close to or exceeded inter-annotator agreement: 93.02% vs. 93.66% for $Related$, 94.47% vs.

Table 1: Corpus statistics summarising the sentences, entities and relations in the dataset stratified by the training, development and test sets.

		Training	Development	Test	Total
Sentences	Amount #	2100	500	1000	3600
	with relations (%)	57.05	53.00	56.80	56.42†
	from full-text (%)	48.71	50.00	50.30	49.33†
Entities	PK	1890	394	856	3140
	Units	2286	474	1056	3816
	Value	3524	702	1557	5783
	Range	314	74	174	562
	Compare	51	18	34	103
Relations	$Central_{val}$	2794	571	1312	4677
	$Deviation_{val}$	1049	207	419	1675
	$Related$	3643	764	1652	6059

† Weighted average across datasets.

Table 2: Named Entity Recognition results on the test set for the model using multi-task (MT) learning, NER + RE, against a model only optimising for NER (no-MT). The metrics reported consider strict matching over entity mentions. Results are displayed as the median over ten runs with their interquartile variance in subscript.

	Precision		Recall		F1	
Entity	MT	no-MT	MT	no-MT	MT	no-MT
PK	**90.82**$_{4.02}$	89.98$_{3.86}$	**90.57**$_{3.76}$	90.09$_{3.05}$	**90.39**$_{2.1}$	90.02$_{1.72}$
Units	95.49$_{1.87}$	**95.79**$_{1.66}$	**96.17**$_{2.07}$	95.69$_{3.85}$	**95.65**$_{0.68}$	95.56$_{1.52}$
Value	94.83$_{2.78}$	**94.96**$_{2.87}$	**96.18**$_{3.17}$	95.21$_{5.94}$	**95.54**$_{2.53}$	95.04$_{2.02}$
Range	**93.49**$_{4.9}$	93.28$_{6.24}$	**90.26**$_{8.22}$	87.39$_{10.33}$	**91.66**$_{4.41}$	90.4$_{3.71}$
Compare	**88.23**$_{6.81}$	88.23$_{16.99}$	66.67$_{9.09}$	**68.18**$_{11.44}$	**76.53**$_{5.82}$	75.64$_{8.12}$
Micro-average					**94.03**$_{1.63}$	93.69$_{1.60}$
Macro-average					**90.02**$_{2.23}$	89.56$_{2.45}$

93.53% for $Deviation_{val}$, 83.2% vs. 86.1% for $Central_{val}$, for inter-annotator and MT model cases, respectively. These results imply that posterior reviews and standardization of span boundaries significantly improved dataset consistency, and the model developed competes well with the expected agreement between pharmacometricians.

4.4 Encoders and Context

To analyse the effect of domain-specific pre-training in the encoder, the BioBERT model was replaced with BERT$_{BASE}$, which was pre-trained on general-domain English text. As shown in Table 4, there was a significant benefit of pre-training in biomedical text, with BioBERT exhibiting over 3% gains in all metrics compared to BERT$_{BASE}$. The largest gain ($\approx \Delta 6\%$) was observed in the $Central_{val}$ relation, indicating that pre-training on biomedical text highly improved PK NER and the understanding between parameter mentions and their measurements. These results are in line with previous findings from Wadden et al. (2019) and Eberts and Ulges (2019). Previous work on end-to-end relation extraction showed improvements between 1.1-4.4% on the SciERC and GENIA datasets with in-domain pre-training (Wadden et al., 2019; Eberts and Ulges, 2019). However, 5.9% improvement was obtained in this task for $Central_{val}$, suggesting that in-domain pre-training is particularly useful. Hence, it is likely that further pre-training on PK literature helps the model performance, and it might be a promising area for future work.

The effect of removing the local context between entities was studied. For this, the input to the RE layer was simplified to the entity embeddings. In other words, the yellow vector from Figure 2 B was removed. Table 5 shows the results of this experiment. Surprisingly, it was observed that the local context improved not only RE but also NER. Both micro and macro-F1 scores were slightly improved, suggesting that explicitly encoding local

Table 3: End-to-end relation extraction results on the test set for the MT model configuration. Results are displayed as the median over ten runs with their interquartile variance in subscript.

Relation	P	R	F_1
$Central_{val}$	$85.77_{5.04}$	$85.46_{5.07}$	$86.1_{3.49}$
$Deviation_{val}$	$92.33_{1.9}$	$94.39_{6.27}$	$93.53_{3.01}$
$Related$	$93.83_{1.69}$	$94.08_{2.51}$	$93.66_{1.52}$

Table 4: Results on the test set when using different encoder models. Results are displayed as the median over ten runs with their interquartile variance in subscript. NER metrics are the micro- and macro-averaged F_1 scores over all entities, and RE metrics are the F_1 scores for each relation class.

	NER		RE		
Encoder	macro-F_1	micro-F_1	$Related$	$Deviation_{val}$	$Central_{val}$
BERT$_{BASE}$	$85.82_{4.07}$	$90.81_{1.77}$	$89.44_{1.69}$	$90.27_{2.2}$	$80.16_{4.14}$
BioBERT	$\mathbf{90.02}_{2.23}$	$\mathbf{94.03}_{1.63}$	$\mathbf{93.66}_{1.52}$	$\mathbf{93.53}_{3.01}$	$\mathbf{86.1}_{3.49}$

Table 5: Results on the test set when using different representations as input to the relation classifier. Local context is the max-pooling of all tokens strictly between two entities. No context only used the concatenation of each entity representation in a specific relation. Results are displayed as the median over ten runs with their interquartile variance in subscript. NER metrics are the micro- and macro-averaged F_1 scores over all entities, and RE metrics are the F_1 scores for each relation class.

	NER		RE		
RE layer representaiton	macro-F_1	micro-F_1	$Related$	$Deviation_{val}$	$Central_{val}$
Local context	$\mathbf{90.02}_{2.23}$	$\mathbf{94.03}_{1.63}$	$\mathbf{93.66}_{1.52}$	$\mathbf{93.53}_{3.01}$	$\mathbf{86.1}_{3.49}$
No context (E1E2)	$89.47_{2.16}$	$93.69_{1.0}$	$91.61_{1.84}$	$90.52_{4.44}$	$81.04_{2.96}$

context between entities in RE layers can also help recognise entities better.

For relation extraction, local context seemed to provide a significant improvement for all relation types, and especially for the $Central_{val}$. This result suggests that entity embeddings might capture local information around the entity mentioned while failing to incorporate longer-range dependencies. The results obtained in this experiment are in-line with Eberts and Ulges (2019). Although recurrent and Transformer models have improved the detection of long-range dependencies in sequential inputs, the noise introduced with long context still represents a challenge in relation extraction (Eberts and Ulges, 2019; Zhong and Chen, 2020). Using this local context, the model can focus on those tokens that might be more informative about the dependencies between both entities. Nonetheless, future studies might benefit from further exploring different contextual representations for RE of PK measurements.

5 Conclusion and Future work

We introduce the PRED corpus, a large and comprehensive public corpus consisting of PK entities and relations annotated in sentences from the scientific literature. This dataset facilitates training and benchmarking models for extracting PK measurements from the scientific literature. We also train and release a new end-to-end RE model based on a BioBERT encoder. This model initially performs NER to identify spans of interest in text, followed by predicting relations between spans. Our benchmark results on the PRED dataset are promising, achieving a micro-average F1-score of 94% for NER and over 85% F1-score across all PK relation types. Our dataset and model can accelerate the construction of ADME datasets from the scientific literature, which can benefit drug development and off-label dosing.

Acknowledgements

VS and QN acknowledge support from a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare studentship (EP/S021612/1). VS acknowledges support from a studentship from the NIHR Biomedical Research Centre at University College London Hospital NHS Trust. FG would like to thank Pontus Stenetorp for helpful feedback and technical input.

References

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Giuliano Berellini and Franco Lombardo. 2019. An Accurate In Vitro Prediction of Human VDss Based on the Øie-Tozer Equation and Primary Physicochemical Descriptors. 3. Analysis and Assessment of Predictivity on a Large Dataset. *Drug metabolism and disposition: the biological fate of chemicals*.

John C. Dearden. 2007. In silico prediction of ADMET properties: How far have we come? *Expert Opinion on Drug Metabolism and Toxicology*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

Ferran Gonzalez Hernandez, Simon J Carter, Juha Iso-Sipilä, Paul Goldsmith, Ahmed A Almousa, Silke Gastine, Watjana Lilaonitkul, Frank Kloprogge, and Joseph F Standing. 2021. An automated approach to identify scientific publications reporting pharmacokinetic parameters. *Wellcome Open Research*, 6:88.

Jan Grzegorzewski, Janosch Brandhorst, Kathleen Green, Dimitra Eleftheriadou, Yannick Duport, Florian Barthorscht, Adrian Köller, Danny Yu Jia Ke, Sara De Angelis, and Matthias König. 2021. Pk-db: pharmacokinetics database for individualized and stratified computational modeling. *Nucleic acids research*, 49(D1):D1358–D1364.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Jörg Hakenberg, Sebastian Schmeier, Axel Kowald, Edda Klipp, and Ulf Leser. 2004. Finding kinetic parameters using text mining. *OMICS A Journal of Integrative Biology*, 8(2):131–152.

Ferran Gonzalez Hernandez, Simon J Carter, Juha Iso-Sipilä, Paul Goldsmith, Ahmed A Almousa, Silke Gastine, Watjana Lilaonitkul, Frank Kloprogge, and Joseph F Standing. 2021. An automated approach to identify scientific publications reporting pharmacokinetic parameters. *Wellcome Open Research*, 6.

Ferran Gonzalez Hernandez, Quang Nguyen, Victoria C Smith, Jose Antonio Cordero, Maria Rosa Ballester, Marius Duran, Albert Sole, Palang Chotsiri, Thanaporn Wattanakul, Gill Mundin, et al. 2024. Named entity recognition of pharmacokinetic parameters in the scientific literature. *bioRxiv*, pages 2024–02.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

A. Kolchinsky, A. Lourenço, L. Li, and L. M. Rocha. 2013. Evaluation of linear classifiers on articles containing pharmacokinetic evidence of drug-drug interactions. *Pacific Symposium on Biocomputing*, pages 409–420.

Artemy Kolchinsky, Anália Lourenço, Heng-Yi Wu, Lang Li, and Luis M Rocha. 2015. Extraction of pharmacokinetic evidence of drug–drug interactions from the literature. *PloS one*, 10(5):e0122199.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Vikas Kumar, Mohammad Faheem, Keun Woo Lee, et al. 2021. A decade of machine learning-based predictive models for human pharmacokinetics: Advances and challenges. *Drug discovery today*.

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

Franco Lombardo, Giuliano Berellini, and R. Scott Obach. 2018. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 1352 drug compounds. *Drug Metabolism and Disposition*, 46(11):1466–1477.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. *arXiv preprint arXiv:1904.03296*.

Matthew Montani, Ines and Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear.

Paul Morgan, Piet H Van Der Graaf, John Arrowsmith, Doug E Feltner, Kira S Drummond, Craig D Wegner, and Steve DA Street. 2012. Can the flow of medicines be improved? fundamental pharmacokinetic and pharmacological principles toward improving phase ii survival. *Drug discovery today*, 17(9-10):419–424.

Diane R Mould and Richard Neil Upton. 2013. Basic concepts in population modeling, simulation, and model-based drug development—part 2: introduction to pharmacokinetic modeling methods. *CPT: pharmacometrics & systems pharmacology*, 2(4):1–14.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. Seqscore: Addressing barriers to reproducible named entity recognition evaluation. *arXiv preprint arXiv:2107.14154*.

Andreas Reichel and Philip Lienau. 2016. Pharmacokinetics in drug discovery: an exposure-centred approach to optimising and predicting drug efficacy and safety. *New approaches to drug discovery*, pages 235–260.

Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Irena Spasić, Evangelos Simeonidis, Hanan L. Messiha, Norman W. Paton, and Douglas B. Kell. 2009. KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways. *Bioinformatics*, 25(11):1404–1411.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. Bern2: an advanced neural biomedical named entity recognition and normalization tool. *arXiv preprint arXiv:2201.02080*.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's stop incorrect comparisons in end-to-end relation extraction! *arXiv preprint arXiv:2009.10684*.

Achakulvisut Titipat and Daniel Acuna. 2015. Pubmed Parser: A Python Parser for PubMed Open-Access XML Subset and MEDLINE XML Dataset.

Jyh Jong Tsay, Bo Liang Wu, and Chang Ching Hsieh. 2009. Automatic extraction of kinetic information from biochemical literatures. *6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009*, 5:28–32.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Yuchen Wang, Haichun Liu, Yuanrong Fan, Xingye Chen, Yan Yang, Lu Zhu, Junnan Zhao, Yadong Chen, and Yanmin Zhang. 2019. In silico prediction of human intravenous pharmacokinetic parameters with improved accuracy. *Journal of chemical information and modeling*, 59(9):3968–3980.

Zhiping Wang, Seongho Kim, Sara K Quinney, Yingying Guo, Stephen D Hall, Luis M Rocha, and Lang Li. 2009. Literature mining on pharmacokinetics numerical data: a feasibility study. *Journal of biomedical informatics*, 42(4):726–735.

Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. 2019. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286.

Zexuan Zhong and Danqi Chen. 2020. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*.

A Appendix: Corpus Construction

A.1 Annotation Guidelines

The annotation guidelines for annotating entities and relations of PK estimations from scientific sentences can be found at `https://github.com/PKPDAI/PKRelations/blob/master/docs/Annotation_Guidelines_PKRelations.pdf`. Annotators were asked to base their labelling decisions on these guidelines, which were updated accordingly as new cases appeared.

Final Annotation Check. After multiple expert annotators had annotated the development and test sets, a final check involved comparing model predictions against their annotated version. This allowed for identifying potentially missed entities and relations during the annotation.

A.2 Annotation Interface

The annotation interface (see Figure 3) was developed in Prodigy (Montani, Ines and Honnibal, 2018) and allowed annotation of both entities and relations at the sentence level. The annotators were presented with a single sentence at a time and could swap between the entity and relation annotation modes. The annotations of named entities were represented at the character level, and relations were defined with the unique identifiers of each entity and their relation class. Candidate values and ranges were pre-highlighted in the interface using a rule-based system. PK terms were pre-highlighted using the NER model from Hernandez et al. (2024), and a list of dictionary terms was used to pre-annotate Compare entities.

A.3 Corpus Limitations

The main limitation of PRED is the potential bias in selecting candidate sentences. The sampled sentences went through two filtering stages that involved model predictions: (1) selection of PK-relevant documents identified by the document classifier from Hernandez et al. (2021) and (2) selection of sentences that at least had one PK entity recognised by our PK RE model. As a result, if the document classifier missed specific types of documents, these would not appear on this dataset. Using the trained PK NER model from Hernandez

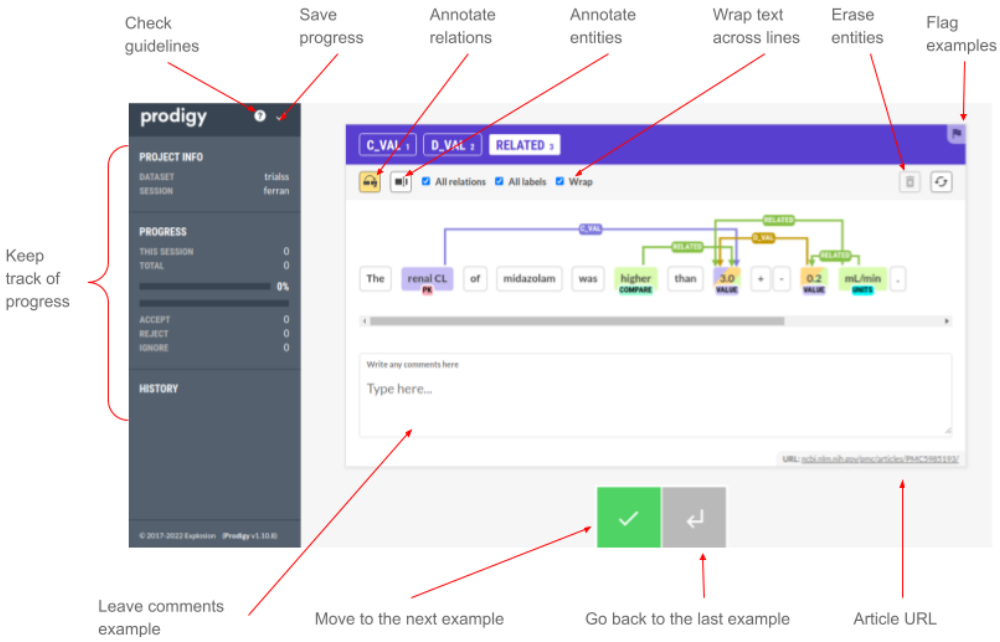


Figure 3: Screenshot of the interface used to annotate entities and relations from scientific text. The example displays a single sentence after entities and relations were annotated.

et al. (2024) for filtering instances with PK parameter mentions might exclude sentences where the NER model missed a single PK mention. Furthermore, if a specific sentence mentioned more than one parameter and only one match (partial or not) was detected by the NER model, the sentence was included in the candidate pool, and these incorrect predictions were later corrected during the annotation process. Overall, it is important to consider that training RE models on this dataset and directly applying them to sentences in the literature without additional filtering might result in the extraction of non-PK measurements due to the filtering approach performed in the sampling stage. For this reason, when deploying systems in production, it is important to combine models trained on this dataset with filtering approaches to discard irrelevant measurements (e.g. pre-tagging PK parameters or posterior EL of PK mentions recognised with RE models).