scientific reports

Check for updates

OPEN

Named entity recognition of pharmacokinetic parameters in the scientific literature

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

The development of accurate predictions for a new drug's absorption, distribution, metabolism, and excretion profiles in the early stages of drug development is crucial due to high candidate failure rates. The absence of comprehensive, standardised, and updated pharmacokinetic (PK) repositories limits pre-clinical predictions and often requires searching through the scientific literature for PK parameter estimates from similar compounds. While text mining offers promising advancements in automatic PK parameter extraction, accurate Named Entity Recognition (NER) of PK terms remains a bottleneck due to limited resources. This work addresses this gap by introducing novel corpora and language models specifically designed for effective NER of PK parameters. Leveraging active learning approaches, we developed an annotated corpus containing over 4000 entity mentions found across the PK literature on PubMed. To identify the most effective model for PK NER, we fine-tuned and evaluated different NER architectures on our corpus. Fine-tuning BioBERT exhibited the best results, achieving a strict F_1 score of 90.37% in recognising PK parameter mentions, significantly outperforming heuristic approaches and models trained on existing corpora. To accelerate the development of end-to-end PK information extraction pipelines and improve pre-clinical PK predictions, the PK NER models and the labelled corpus were released open source at https://github.com/PKPDAI/PKNER.

Bringing a new chemical compound to the market is an extremely costly process, which has been estimated between $161m and $4.5bn[1]. Meanwhile, over 90% of drug candidates fail after entering phase I clinical trials[2,3]. Accurate predictions of candidate drug properties at an early stage are critical for improving the efficiency of this process. To elicit the desired effect, candidate drugs must reach a specific concentration at the target site of the body over a certain time period[4]. Predicting whether candidate drugs will reach the desired concentration over a certain period at the target site requires understanding the processes of absorption, distribution, metabolism and excretion (ADME) of drugs from the human body.

Pharmacokinetic (PK) parameters quantify the ADME processes of chemical compounds through numerical estimates. Accurate estimation of drugs' PK parameters is crucial for drug development research[4]. Mechanistic models have been widely used to predict the PK parameters of candidate drugs before they are tested in humans. However, a significant proportion of those candidates still fail due to PK complications found during the clinical phases[5]. Hence, improving PK predictions of candidate compounds before they are given to humans is crucial for assessing candidate prospects and optimising the drug development pipeline.

One of the main challenges in improving PK predictions for chemical compounds is the lack of comprehensive and standardised PK repositories[6,7]. Although existing open-access databases collect information ranging from chemical structure to a summary of PK publications, they typically only report sparse PK information explicitly[2,6,8]. Consequently, researchers must search and curate PK estimates from scientific literature before pre-clinical predictions can be made[6,9]. The vast and continually increasing number of PK publications, coupled with the extensive amount of PK information locked in scientific articles, limits our ability to efficiently find

[1]Department of Computer Science, University College London, London, UK. [2]Institute of Health Informatics, University College London, London, UK. [3]Blanquerna School of Health Sciences, Ramon Llull University, Barcelona, Spain. [4]Institut de Recerca Sant Pau Barcelona, Barcelona, Spain. [5]Mahidol Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand. [6]Global Business School for Health, University College London, London, UK. [7]Great Ormond Street Institute for Child Health, University College London, London, UK. [8]Department of Pharmacy, Great Ormond Street Hospital for Children, London, UK. [9]Institute for Global Health, University College London, London, UK. ✉email: ferran.hernandez.17@ucl.ac.uk; f.kloprogge@ucl.ac.uk

and curate comprehensive datasets manually[2]. Thus, despite the potential PK data stored in scientific articles, efficiently exploiting this resource remains a significant challenge in drug development.

Automated text mining approaches can aid researchers in extracting information from the scientific literature more efficiently. Recognising entities of interest is a crucial step in information extraction pipelines that enables subsequent downstream tasks such as relation extraction or entity linking. In this study, we focus on the initial step towards automated extraction of PK parameter estimates from the scientific literature, Named Entity Recognition (NER). Developing systems that can identify mentions of PK parameters in scientific text is crucial for end-to-end PK extraction as well as characterising drug-drug interactions (DDIs), as many interactions are reported by mentioning their effect on specific PK parameters[10]. However, PK NER remains a challenging task since there are multiple PK parameter types and their mentions are often highly variable across the scientific literature, involving the frequent use of acronyms and long textual spans[11]. Additionally, the scarcity of annotated resources limits the development of effective NER models that can deal with this diversity. In this work, we tackle these challenges by developing annotated corpora and machine-learning models for effective PK NER.

Methodology

Corpus construction

A protocol was established to generate corpora of labelled sentences that allowed training and evaluation of PK NER models. The final corpus is referred to as the PK-NER-Corpus and can be found at https://zenodo.org/records/4646970[12].

Source

To create a candidate pool for sentence annotation, the pipeline described in Fig. 1 was applied. A PubMed search for *"pharmacokinetics"* was initially conducted to retrieve articles using the default search parameters in PubMed. No additional filters were applied. The pipeline from Gonzalez Hernandez et al.[7] was used to identify 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 (https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/), while only abstracts were available for the rest. Both, abstracts and full-text articles were downloaded in XML format from PubMed (https://www.nlm.nih.gov/databases/download/pubmed_medline.html) and PMC (https://ftp.ncbi.nlm.nih.gov/pub/pmc/) FTP sites. The PubMed Parser[13] was used to parse the XML files, and paragraphs from the introduction section were excluded. The scispaCy sentence segmentation algorithm[14] split abstracts and paragraphs into sentences. The resulting sets were the abstract pool with over a million sentences and the full-text pool with 721,522 sentences. To create a balanced candidate pool for ML model training and evaluation, 721,522 instances were randomly sampled from the abstract pool and combined with full-text sentences, resulting in a balanced pool of 1,443,044 sentences, referred to as the candidate pool. All labelled sentences in the corpus construction were sampled from the candidate pool.

Annotation

The team responsible for the annotation involved twelve annotators with extensive PK expertise and familiarity with the different parameters and study types in the PK literature. To ensure consistency in the annotation process, each annotator initially labelled a small set of 200 examples to identify sources of disagreement. The

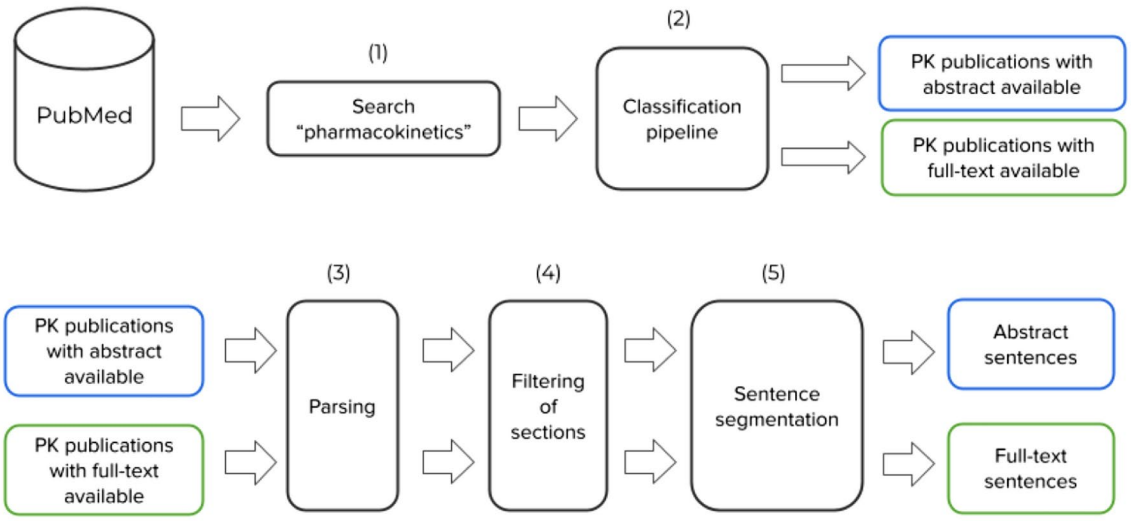

Figure 1. Flow diagram showing the main processes involved to generate a pool of candidate sentences for NER labelling. (1) Search for "pharmacokinetics" in PubMed and (2) run binary classification pipeline to filter abstracts containing PK parameters. (3) Parse XML abstract and full-text documents, and (4) filter out introduction sections. Finally, (5) segment each paragraph into sentences to generate the final corpus of PK sentences.

team then discussed which parameters to include and how to define span boundaries using the PK ontology from Wu et al.[11] as a reference. Annotation guidelines were provided to annotators before they began the labelling task, and were updated as new challenging examples were resolved during the annotation process. Details about the annotation interface and guidelines can be found in Supplementary Information: Appendix A. Training, development and test sets were developed to train and evaluate different NER pipelines.

Training set

Training effective NER models for new entity types often requires a large number of annotated samples with diverse spans to account for the variability of surface forms and contexts of use[15]. However, the sampling strategy followed by the test and development sets resulted in a low proportion of sentences containing PK entities (16.4%). To generate an effective training dataset, two main approaches were sequentially applied to selectively sample informative sentences for PK NER while reducing annotation efforts:

1. *Heuristic labelling.* The rule-based model described in Supplementary Information: Appendix B was applied to all the sentences within the candidate pool. From those sentences that contained matches from the rule-based model, 300 were randomly chosen to form an initial training set with a substantial number of entity mentions. To enhance the quality of this set, the annotators corrected the labels generated by the rule-based model. Following correction, 86.67% of the sentences retained PK mentions, although adjustments were often needed for their span boundaries. Subsequently, an initial scispaCy NER model[14] was trained on this dataset.
2. *Active learning.* After training the initial scispaCy model, it was used to identify spans from the candidate pool that were most informative for model training. Utilising the active learning interface from Prodigy[16], which presents candidate spans to annotators based on model uncertainty, annotators provided binary labels denoting the correctness of suggested spans. During the active learning process, the model underwent updates in a loop after every set of 10 annotated sentences. After obtaining binary labels, a final round of annotation was conducted to label any additional spans present in the sentences and correct span boundaries. Following this protocol, a total of 2800 sentences with a large number of PK entity mentions were labelled. Further details on the Active Learning protocol can be found in Supplementary Information: Appendix C.

Test and development sets

The development and test sets were generated by randomly sampling sentences from the candidate pool without replacement, to preserve the distribution of sentences found in PK articles. In total, 1,500 and 500 sentences were selected for the test and development sets, respectively. Then, each sentence in the development and test sets followed a two-stage procedure of (1) initial annotation by one expert and (2) review and standardisation of span boundaries by at least two additional experts (similar to[17]). This process was carried out in batches of 200 sentences. After each batch, sources of disagreement were discussed, and annotation guidelines were updated.

Inter-annotator agreement (IAA)

We selected pair-wise F1 as the main metric for measuring IAA in NER[18,19]. IAA was computed for each pair of annotators and F1 was obtained by treating the labels of one annotator as ground truth and the other as the system prediction. All annotators independently labelled a total of 200 sentences from the test set, used to derive the IAA. This exercise was done with the last batch of the test set when guidelines had already been updated multiple times, but no corrections were performed before computing the IAA.

External dataset validation

We utilised the PK Ontology and its corresponding corpus developed by Wu et al.[11] for external validation. This corpus, referred to as PK-Ontology-Corpus, comprises 541 abstracts manually labelled, encompassing the annotation of key terms, sentences related to Drug-Drug Interactions (DDI), and annotated DDI pairs. The abstracts originated from four study types: clinical PK, clinical pharmacogenetics, *in vivo* DDI, and *in vitro* DDI studies. One of the annotated key terms in the PK-Ontology-Corpus was PK parameters. The NER models developed in this study were also evaluated in the PK-Ontology-Corpus, which allowed for assessing model performance in different study types, including several DDI sentences and detecting differences in the annotation criteria.

Models

Rule-based system

Given the PK expertise of the annotation team, a set of rules was generated to develop a rule-based model covering well-known PK parameters and their primary surface forms and acronyms. The model was implemented using the entity ruler from spaCy, which requires a set of token-level patterns and can incorporate rules regarding part-of-speech (POS) and dependency labels. ScispaCy[14] was used as a base tokeniser, POS tagger and dependency parser to incorporate the token-level patterns into the model. Developing the list of terms and rules was an iterative process performed together with the annotation team, and rules were updated by assessing their performance on the development set. Supplementary Information: Appendix B describes the iterative process followed to develop the rule-based system.

BERT

The Transformer architecture has emerged as state-of-the-art for NLP tasks[20]. In this study, pre-trained BERT models were fine-tuned to perform PK NER[21]. We added a task-specific layer (fully-connected + softmax) to map output token embeddings from BERT models to BIO labels[22]. Two pre-trained models were compared:

Dataset	Sentences	Entity mentions	Sentences with PK mentions (%)	Full-text sentences (%)
Training	2800	3680	64.25	79.46
Development	500	149	16.40	50.8
Test	1500	390	16.40	50.8

Table 1. Corpus statistics of the PK-NER-Corpus stratified by the training, development and test sets.

Dataset	Sentences	Entity mentions	Sentences with PK mentions (%)
Training	4008	1478	23.68
Test	1021	377	25.27

Table 2. Corpus statistics of the PK-Ontology-Corpus stratified by the training and test sets.

$BERT_{BASE}$[21] which was pre-trained on general-domain English text, and BioBERT v1.1[23] which was further pre-trained on PubMed articles. Models were implemented in PyTorch[24] using the Transformers library[25].

BERT tokenizers split each input sentence into sub-word tokens, each associated with a BIO label. The model was trained to minimise categorical cross-entropy loss. Both BERT and classification layer parameters were fine-tuned during 20 epochs. The model's performance was evaluated on the development set at the end of each epoch, saving the state with the highest entity-level F1 score. We used the Adam optimizer with a linear weight decay of 0.05 and a dropout probability of 0.1 on all layers. We used a batch size of 16 and the learning rate was grid-searched, with $\mu = 3e^{-5}$ yielding the best performance. The maximum sequence length was set to 256 to cover most training instances. During inference, sentences with over 256 tokens were split, and predictions were re-joined after BIO label assignments. Experiments ran on a single NVIDIA Titan RTX (24GB) GPU.

ScispaCy

The scispaCy model was also fine-tuned to perform NER of PK parameters. ScispaCy is built on top of spaCy but focuses on biomedical and scientific text processing[14]. In this work, all components from the scispaCy pipeline were reused, and the NER layer was trained from scratch. Analogous to the BERT pipelines, models were trained for 20 epochs and the state of the model with the best performance on the development set was saved. The rest of the hyperparameters were kept identical to Neumann et al.[14].

Evaluation

We computed precision and recall, and derived F1 score for comparing model performance. To determine true positives we used both, strict and partial matching. Strict matching requires complete overlap in entity boundaries between predictions and annotations while partial matching considers instances where system predictions partially overlap with annotated entities. Both strict and partial matching metrics were computed using the *nervaluate* library (https://github.com/MantisAI/nervaluate).

Results and discussion

Corpus statistics

The main statistics for the PK-NER-Corpus are shown in Table 1. Since the evaluation sets randomly sampled sentences from PK articles, the proportion of sentences containing PK parameter mentions was only 16.40%. Despite preserving the distribution of sentences in which PK NER algorithms might be applied, fewer entity mentions were present in the evaluation sets. On the other hand, 64.25% of sentences in the training set contained mentions of PK parameters, resulting in many entity mentions. This difference in the distribution of parameter mentions was due to the active learning sampling protocol selecting sentences with a higher proportion of entity mentions. Additionally, while we randomly sampled sentences from the abstract or full-text section in the evaluation sets, the active learning protocol selected a higher proportion of sentences from the full text (79.56%).

The statistics of our external evaluation corpus (PK-Ontology-Corpus) from Wu et al.[11] are shown in Table 2

Effects of active learning

To evaluate the effects of active learning we performed the following experiment. The development set ($n = 500$) was used as an example of an annotated set randomly sampled while 500 sentences from the training set collected with active learning were randomly sampled to perform a fair comparison. Ten separate runs with different random seeds were performed. The active learning experiment randomly sampled a different subset of sentences from the training set and randomly initialised the classification layer parameters in each run. The BioBERT model was trained for five epochs with a learning rate of $3e^{-5}$, and the final model was applied to the test set at the end of each run.

Figure 2 show the results of these experiments. Training the BioBERT model with the active learning dataset resulted in over 7% increase in the median F1 score for strict matching compared to training with randomly sampled sentences. These results suggest that the protocol used to generate the training set highly benefited the model performance compared to randomly sampling sentences. Most of this benefit is the consequence of an improved recall, suggesting that the active learning dataset contains a wide variety of PK spans not covered by

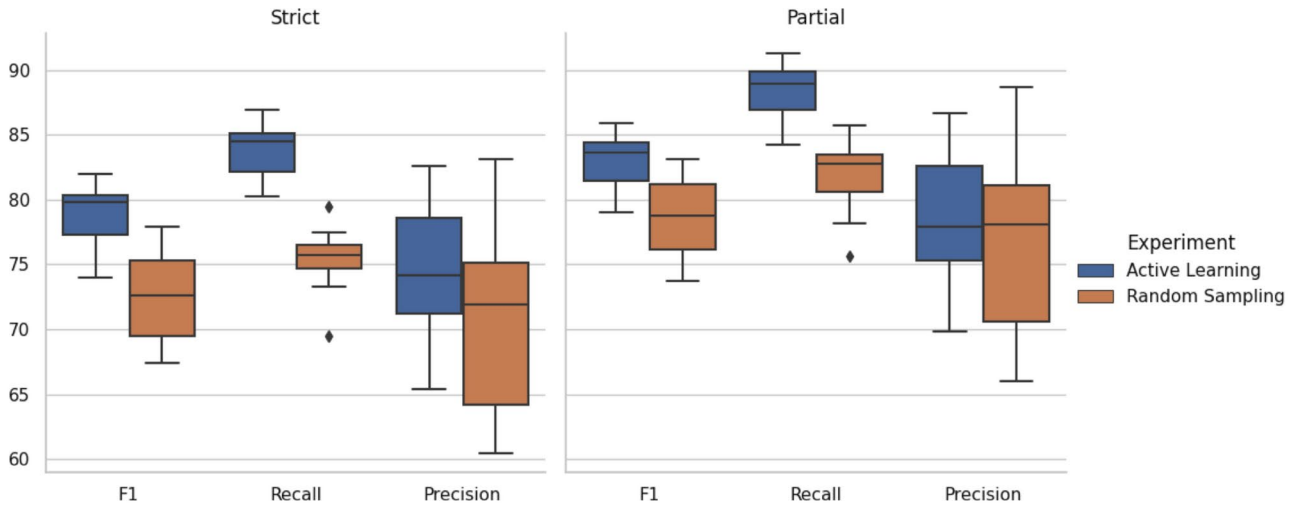

Figure 2. Distribution of F1, Recall and Precision scores for the *Active Learning* and *Random Sampling* datasets (n=500 sentences) after 10 runs with different random seeds. The left and right panels display the scores considering strict and partial matching of entities, respectively.

	Strict			Partial		
Model	P	R	F_1	P	R	F_1
Rule-based	52.8	43.59	47.75	69.25	57.18	62.64
ScispaCy	77.09	82.82	79.85	80.91	86.92	83.81
BERT	81.47	87.72	84.48	84.92	91.43	88.05
BioBERT	90.49	90.26	90.37	92.54	92.31	92.43

Table 3. Results on the test set for different NER models. Metrics are reported at the entity level using strict and partial matches.

the random sampling dataset. Considering the frequency of named entities in each dataset (i.e. only 16.4% of sentences mentioned PK parameters in the randomly sampled datasets), it is likely that the selective sampling approach implemented for this task was particularly beneficial for covering a wider variety of relevant spans.

Model performance

Table 3 summarises the main results on the test set. The results showed that the rule-based model could not efficiently cover the diversity of PK parameter mentions annotated by field experts, achieving a strict F1 score below 50%. Some of the main challenges of the rule-based approach were (1) the great variety of PK parameter types, which limited the pipeline's recall, (2) the presence of complementary terms within PK spans (e.g. total body clearance) and (3) acronyms highly dependent on context (e.g. "F" for bioavailability). Notably, there was a large difference in precision between strict and partial matches (over 15%). This is a consequence of challenge (2), where rules often detected the primary PK term, but complementary terms determining the parameter sub-type were missed. The machine learning pipelines significantly outperformed the heuristic model with over 30% gain on the strict F1 score, mostly driven by substantial improvements in recall.

We observed distinct patterns in the true positives and errors produced by the rule-based and LLM models. The machine learning models, particularly those fine-tuned with pre-trained transformers, demonstrated much higher F1 scores by effectively capturing a wider variety of PK parameter mentions not explicitly covered by the rules defined by PK experts, and they were also more flexible at encapsulating complementary terms that often vary in PK parameter mentions. Nonetheless, this flexibility also introduced a few extra false positives, where the model would occasionally overgeneralise, predicting incorrect PK entities in similar contexts such as mentions of pharmacodynamic parameters, which the rule-based models avoided (e.g. Area Under the Effect Curve (AUEC), Maximum Tolerated Dose (MTD)).

As it has been previously reported[26], it was observed that the models based on BERT provided substantial performance benefits in comparison to the scispaCy model. The test set predictions showed that the scispaCy pipeline was x10 faster at inference time on CPU than running BERT models on a single GPU. Therefore, we also released the fine-tuned scispaCy pipelines open-source (https://github.com/PKPDAI/PKNER). The BioBERT model outperformed the BERT model pre-trained on general-domain English text, especially on strict entity matching. Specifically, BioBERT provided a large gain (+ 9%) on the pipeline precision in comparison to all the other models. This result suggests that domain-specific pre-training is crucial for effective PK NER.

Performance on external corpus

To assess the generalisability and robustness of the BioBERT model fine-tuned on the PK-NER-Corpus, we conducted external validation using the PK-Ontology-Corpus developed by Wu et al.[11]. This corpus comprises 541 manually annotated abstracts, focusing on key terms and sentences related to PK parameters and drug-drug interactions (DDIs). Importantly, the annotated abstracts were selected with more specific filtering criteria regarding study types and focusing on specific drugs (e.g. midazolam).

Our model, fine-tuned on the PK-NER-Corpus, was directly applied to the PK-Ontology-Corpus test set without any additional training, achieving a competitive strict F1 score of 74.52% and a partial matching F1 score of 81.10% (see Table 4). The substantial increase from strict to partial matching indicates that the main PK terms were often identified, although discrepancies in annotated span boundaries between the two corpora impacted strict matching. These discrepancies might be due to different annotation criteria used in the development of the PK-NER-Corpus and the PK-Ontology-Corpus. However, the competitive performance on a different dataset demonstrates our model's robustness in identifying PK parameters across varied contexts, indicating that it is not overfitted to specific features of the PK-NER-Corpus, thus enhancing its applicability to other PK studies.

Conversely, when the BioBERT model was fine-tuned on the PK-Ontology-Corpus and evaluated on the PK-NER-Corpus, the strict matching F1 score was 66.13%, highlighting the limitations of training models on narrowly focused datasets. This cross-dataset validation underscores the necessity of training on diverse datasets to capture a wide range of PK parameters and contexts. By demonstrating that models trained on a broad corpus covering multiple PK study types and drugs (PK-NER-Corpus) perform well on an externally developed dataset, we illustrate the importance of comprehensive and varied training data for developing robust PK NER models.

The observed differences in transferability highlight the importance of corpus diversity when training and evaluating NLP models for PK applications. Future work could involve creating and annotating additional datasets that bridge the gap between general and specific corpora. For instance, incorporating clinical trial reports, which contain a large number of PK parameter estimates, or other relevant contexts, could provide a more comprehensive training and evaluation ground for PK NER models.

Potential applications and implications

The NER models developed in this study can now be used to characterise DDIs by identifying PK parameters involved in those interactions and performing downstream bio-NLP tasks such as extending knowledge graphs with PK-related entities. Additionally, they provide a fundamental step to achieve end-to-end extraction of PK parameter estimates and automatically construct comprehensive databases used for pre-clinical drug development. However, further work is required to develop subsequent relation extraction systems that extract numerical values and related entities. This step is crucial for accurately capturing numerical estimates and their contexts. Such a system will facilitate the creation of extensive, high-quality PK databases while minimising human effort. These databases can serve as valuable resources for literature reviews, extraction of parameter distributions for (semi-)mechanistic and physiological-based models, and machine learning-based predictions of PK parameters for new molecules.

Conclusion and future work

This work presented a new corpus to train and evaluate NER models to detect mentions of PK parameters in the scientific literature. A variety of models were compared, and fine-tuning BioBERT resulted in the best performance on PK NER with over 90% F1 score on strict entity matching. Domain-specific pre-training with transformers was crucial to obtain optimal performance. Machine learning models largely outperformed the rule-based model, potentially due to the high diversity in PK parameter surface forms and the importance of context to determine PK entities.

The active learning protocol helped accelerate the curation of PK data while improving the information provided by labelled sentences compared to random sampling. A variety of approaches have been applied for active learning in NER[27–29]. For instance, bayesian approaches have recently shown promising results[29], although their application comes with computation costs. It is still unclear which active learning approaches are most beneficial to make efficient use of a model in the loop. In this study, many approaches are left for exploration. For instance, using transformer-based models in the loop instead of scispaCy, using diversity sampling or applying other criteria to estimate model uncertainty. However, the framework developed with Prodigy allowed for fast annotations that reduced the labelling load, and the samples selected for annotations provided diverse and challenging spans that resulted in larger information gains than samples randomly sampled.

Finally, the best-performing model showed good generalisation to various study types when applied to external annotated corpora and validated its potential application to improve the characterisation of DDIs. The experiment results indicate that NER models trained on the PK-NER-Corpus generalise better to unseen PK publications than those trained on existing corpora. Overall, we believe that these resources can become crucial

	Strict			Partial		
Training corpus	P	R	F_1	P	R	F_1
PK-NER-Corpus	77.05	72.15	74.52	83.85	78.51	81.10

Table 4. Results obtained on the external PK-Ontology-Corpus test set after training BioBERT on the PK-NER-Corpus.

in developing end-to-end PK information extraction pipelines, improving the characterisation of drug-drug interactions, and ultimately helping to improve PK pre-clinical predictions.

Data availability

The PK NER models and the labelled corpus have been released open source at https://github.com/PKPDAI/PKNER.

Received: 16 April 2024; Accepted: 16 September 2024
Published online: 08 October 2024

References

1. Schlander, M., Hernandez-Villafuerte, K., Cheng, C. Y., Mestre-Ferrandiz, J. & Baumann, M. How much does it cost to research and develop a new drug? a systematic review and assessment. *PharmacoEconomics***39**, 1243 (2021).
2. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics***20**(2), 273–286 (2019).
3. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R &D costs. *J. Health Econ.***47**, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012 (2016) (**ISSN 18791646.**).
4. Morgan, P. et al. 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 (2012).
5. Palmer, A. M. New horizons in drug metabolism, pharmacokinetics and drug discovery. *Drug News Perspect.***16**(1), 57–62 (2003).
6. Grzegorzewski, J. et al. Pk-db: Pharmacokinetics database for individualized and stratified computational modeling. *Nucleic Acids Res.***49**(1D), D1358–D1364 (2021).
7. Hernandez, F. G. et al. An automated approach to identify scientific publications reporting pharmacokinetic parameters. *Wellcome Open Res.***6**, 88 (2021).
8. Hernandez, F.G. *Structuring the Unstructured: Unlocking pharmacokinetic data from journals with Natural Language Processing.* PhD thesis, UCL (University College London), (2022).
9. Lombardo, F., Berellini, G. & Obach, R. S. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 1352 drug compounds. *Drug Metab. Dispos.***46**(11), 1466–1477. https://doi.org/10.1124/dmd.118.082966 (2018) (**ISSN 1521009X.**).
10. Kolchinsky, A., Lourenço, A., Wu, H.-Y., Li, L. & Rocha, L. M. Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. *PLoS ONE***10**(5), e0122199 (2015).
11. Wu, H.-Y. et al. An integrated pharmacokinetics ontology and corpus for text mining. *BMC Bioinf.***14**(1), 1–15 (2013).
12. Hernandez, F.G. Pharmacokinetic named entity recognition benchmark (0.0.1), (2024). URL https://doi.org/10.5281/zenodo.4646970.
13. Titipat, A., & Acuna, D. Pubmed Parser: A python parser for pubmed open-access XML subset and MEDLINE XML Dataset, (2015). URL https://github.com/titipata/pubmed_parser.
14. Neumann, M., King, D., Beltagy, IZ & Ammar, W (2019) ScispaCy: Fast and robust models for biomedical natural language processing. https://doi.org/10.18653/v1/w19-5034
15. Wang, X., Yang, C. & Guan, R. A comparative study for biomedical named entity recognition. *Int. J. Mach. Learn. Cybern.***9**(3), 373–382 (2018).
16. ExplosionAI. Prodigy: An annotation tool powered by active learning, (2021). URL https://prodi.gy/.
17. Hope, T., Amini, A., Wadden, D., van Zuylen, M., Parasa, S., Horvitz, E., Weld, D., Schwartz, R. & Hajishirzi, H. Extracting a knowledge base of mechanisms from covid-19 papers. arXiv preprint arXiv:2010.03824, (2020).
18. Hripcsak, G. & Rothschild, A. S. Agreement, the f-measure, and reliability in information retrieval. *J. Am. Med. Inform. Assoc.***12**(3), 296–298 (2005).
19. Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K. *et al.* Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association, (2012).
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp 5998–6008, (2017).
21. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, (2018).
22. Campos, D., Matos, S. & Oliveira, J. L. Biomedical named entity recognition: A survey of machine-learning tools. *Theory Appl. Adv. Text Min.***11**, 175–195 (2012).
23. Lee, J. et al. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics***36**(4), 1234–1240 (2020).
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. Automatic differentiation in pytorch. (2017).
25. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, Rémi, F., Morgan *et al.* Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, (2019).
26. Weber, L. et al. Hunflair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics***37**(17), 2792–2794 (2021).
27. Shen, Y., Yun, H., Lipton, Z. C., Kronrod, Y., & Anandkumar, A. Deep active learning for named entity recognition. arXiv preprint arXiv:1707.05928, (2017).
28. Shen, D., Zhang, J., Su, J., Zhou, G., Tan, & Chew L. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 589–596, (2004).
29. Siddhant, A. & Lipton, Z. C. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. arXiv preprint arXiv:1808.05697, (2018).

Author contributions

Conception: F.H., W.L., J.S., and F.K. Research: F.H. Labelling: F.H., V.S., J.C., M.B., M.D., A.S., P.C., T.W., G.M., J.S., and F.K. Writing manuscript: F.H. and Q.N. reviewing manuscript: all.

Funding

FK conducted this research as part of a Wellcome Trust/Royal Society sir Henry Dale fellowship (grant number 220587/Z/20/Z). VS acknowledges support from a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare studentship (EP/S021612/1). VS ac-knowledges support from a studentship from the NIHR Biomedical Research Centre at University College London Hospital NHS Trust.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-73338-3.

Correspondence and requests for materials should be addressed to F.G.H. or F.K.

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

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