An automated approach to extract pharmacokinetic parameters from scientific publications

Frank Kloprogge
Pharmacokinetics-Pharmacodynamics Group
UCL Institute for Global Health

Acknowledgements

- **UCL team**
  - Ferran Gonzalez Hernandez
  - Victoria Smith
  - Waty Lilaonitkul
  - Frank Kloprogge
  - Joseph Standing

- **Contributors and Collaborations**
  - Juha Iso-Sipila
  - Paul Goldsmith
  - James Yates
  - Ahmed Allousa
  - Simon Carter
  - José Antonio Cordero Rigol
  - Maria Rosa Ballester
  - Mario Duran Hortolà
  - Albert Solé Guixeres
  - Palang Chotsiri
  - Thanaporn Wattanakul
  - Gill Mundin

Overall aim

To centralise and structure a repository of PK parameters from scientific literature using text mining and Natural Language Processing.

ID	DRUG/S	DRUG DESCRIPTORS	PARAMETER DATA	DEMOGRAPHICS	DRUG TREATMENT	MODELLING APPROACH	ADDITIONAL INFORMATION
…	…	…	…	…	…	…	…

Rational

1. PK predictions for early drug development
2. Development of population pharmacokinetic models
3. Pop PK models for TDM and personalised medicine

Model pipeline

1. Identify PK papers
2. NER & REX
3. Table extraction

Identify PK papers

1. Search term: “Pharmacokinetics”
2. Training corpus
   - Two independent labelers
   - 3,992 articles
3. Test corpus
   - Three independent labelers
   - 800 articles

Field	Training	Final test
Title	100	100
Abstract	87.17	87.67
Authors	99.44	99.63
Journal	100	100
Publication Type	100	100
Keywords	15.41	16.125
MeSH terms	97.67	98.25
Chemicals	93.86	94.13
Affiliations	80.94	79.125
Label		
Relevant	19.81	20.25
Not Relevant	80.19	79.75

Identify PK papers

- XGBoost
- Unigrams
- Biobert mean pooling
- 200 bootstraps
  - Training 60%
  - Development 20%
  - Test 20%

- **Training data**

Pipeline	Precision (%)	Recall(%)	F_1 (%)	F_1 IQV
Unigr.	80.1 (73.9,86.0)	**82.3 (74.1,88.6)**	80.6 (75.8,85.2)	9.4
Unigr. + mean pool	83.7 (76.7,89.1)	80.4 (74.1,87.3)	**81.7 (77.8,86.0)**	8.2
Unigr. + mean + min&max	**83.8 (75.6,88.8)**	79.1 (73.4,85.4)	81.0 (77.2,85.4)	8.2

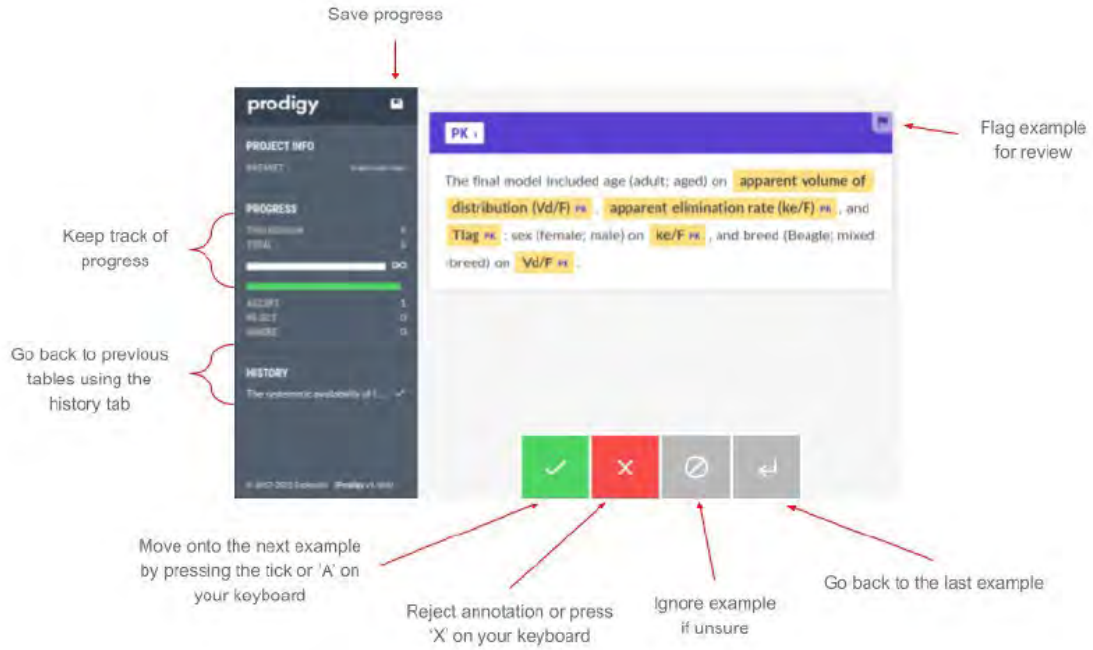
- **Test data**

Precision (%)	Recall(%)	F_1 (%)	Accuracy (%)
84.8%	82.8%	83.8%	93.2%

Gonzalez Hernandez F, Carter SJ, Iso-Sipilä J, Goldsmith P, Almousa AA, Gastine S, Lilaonitkul W, Kloprogge F, Standing JF. An automated approach to identify scientific publications reporting pharmacokinetic parameters. Wellcome Open Res. 2021 Apr 21;6:88. doi: 10.12688/wellcomeopenres.16718.1. PMID: 34381873; PMCID: PMC8343403.

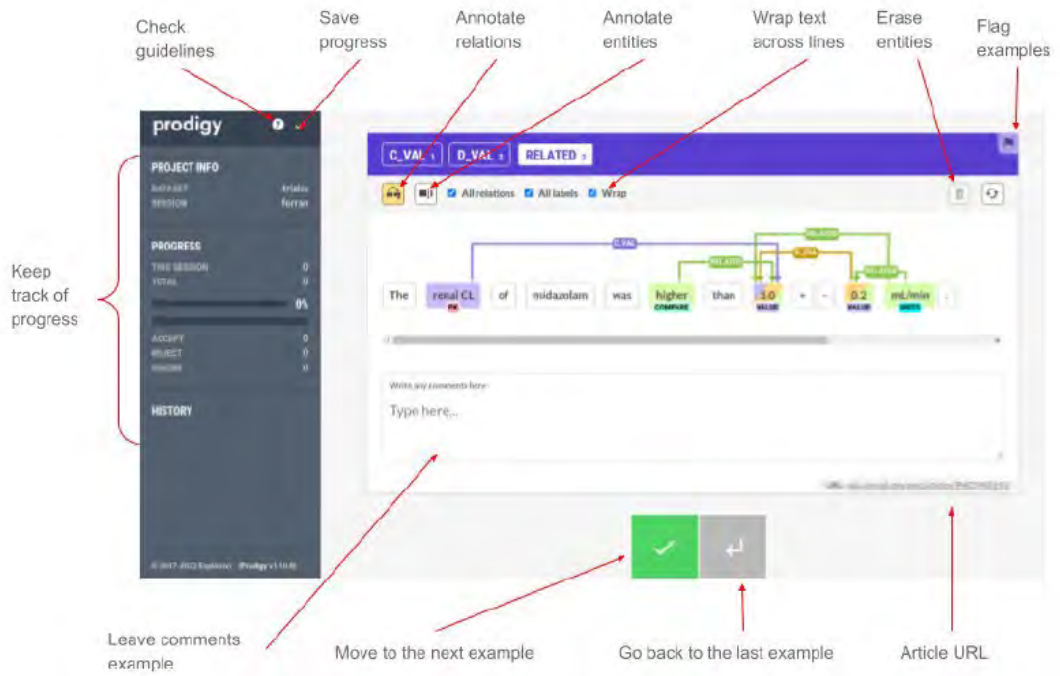
NER & REX

NER



Dataset	# Sentences	# entity mentions	% sentences with PK mentions	% of full-text sentences
Training	2800	3680	64.25	79.46
Development	500	149	16.40	50.8
Test	1500	390	16.40	50.8

REX



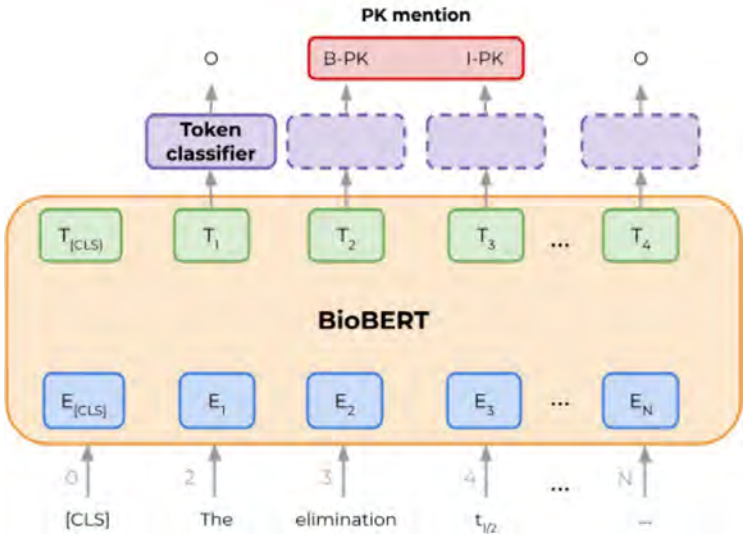
		Training	Development	Test	Total
Sentences	Amount #	2100	500	1000	3600
	with relations (%)	57.05	53.00	56.80	56.42†
	from full-text (%)	48.71	50.00	50.30	49.33†
Entities	PK	1890	394	856	3140
	Units	2286	474	1056	3816
	Value	3524	702	1557	5783
	Range	314	74	174	562
	Compare	51	18	34	103
Relations	$Central_{val}$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

NER & REX

NER

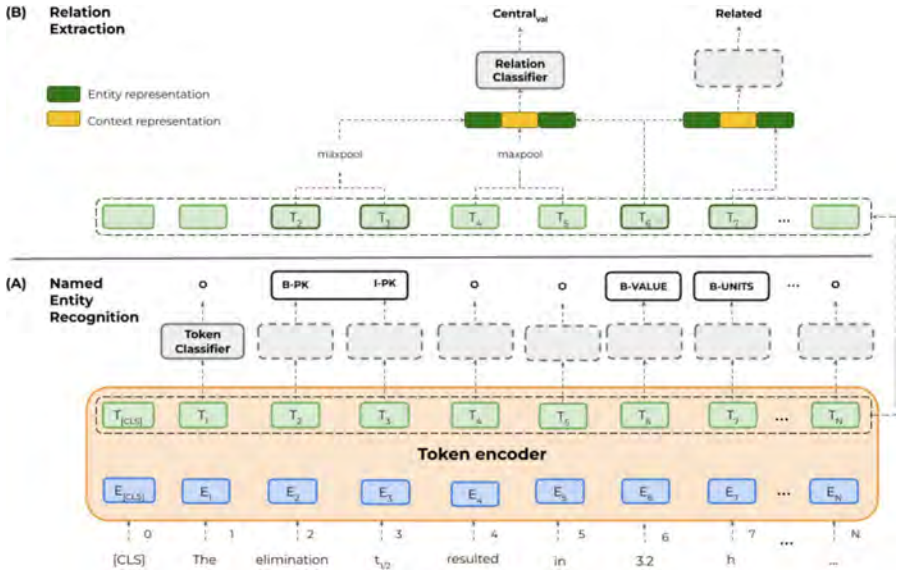
- Neural network
  - Spacy (Convolutional Neural Network)
  - Best performing BioBert (pre-training with PubMED abstracts)

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{w=1}^{j} e^{x_w}}$$



REX

- Multi-task learning
  - Architecture the same
  - NER for PK, units etc simultanously
  - Tweaked to do relation extraction simultanously



NER & REX

NER

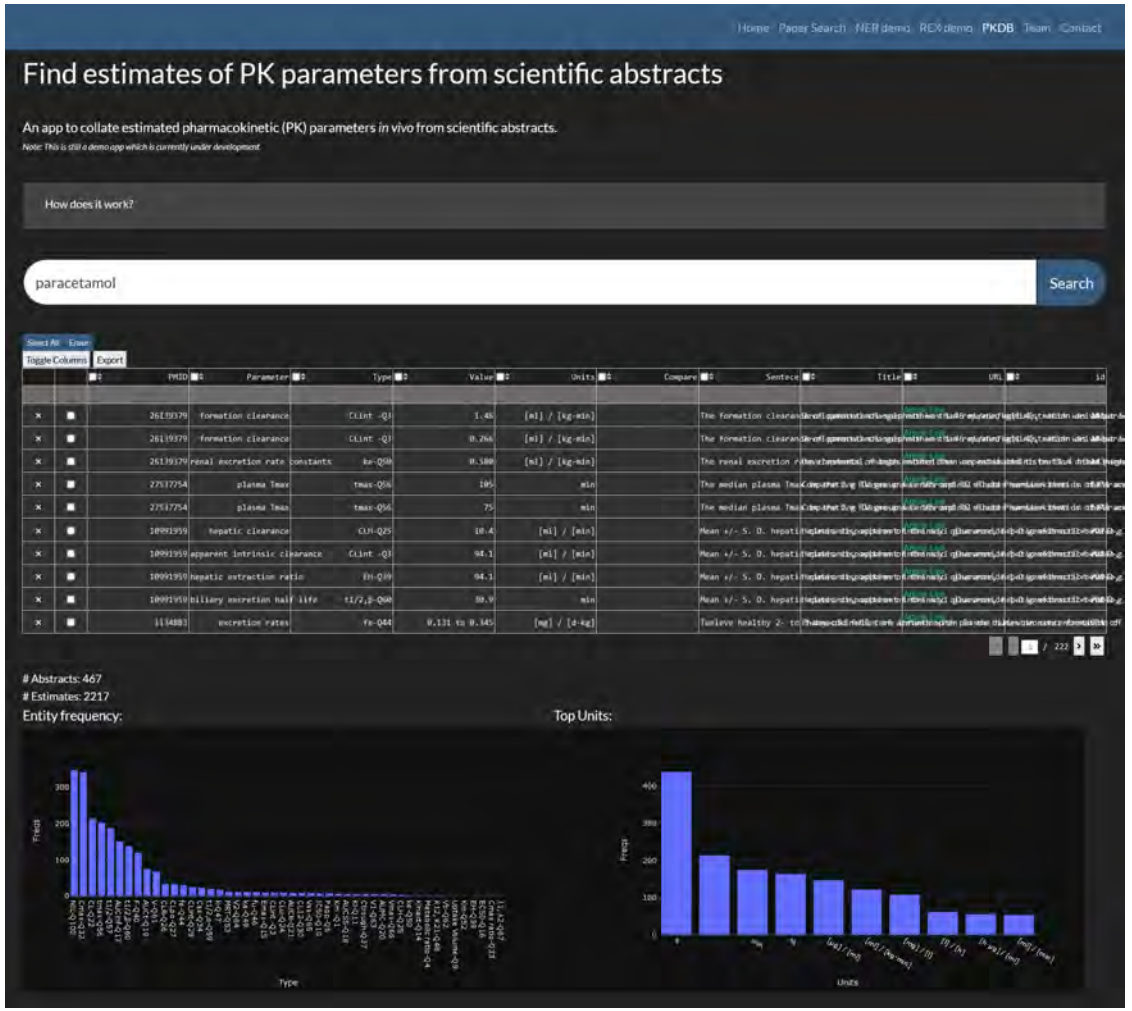
	Strict			Partial		
Model	P	R	F_1	P	R	F_1
Rule-based	52.8	43.59	47.75	69.25	57.18	62.64
ScispaCy	77.09	82.82	79.85	80.91	86.92	83.81
BERT	81.47	87.72	84.48	84.92	91.43	88.05
BioBERT	**90.49**	**90.26**	**90.37**	**92.54**	**92.31**	**92.43**

REX

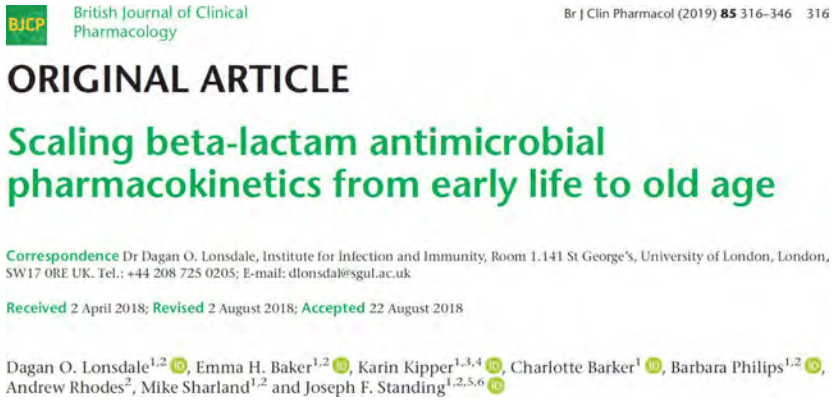
	Precision		Recall		F1	
Entity	MT	no-MT	MT	no-MT	MT	no-MT
PK	$\mathbf{90.82}_{4.02}$	$89.98_{3.86}$	$\mathbf{90.57}_{3.76}$	$90.09_{3.05}$	$\mathbf{90.39}_{2.1}$	$90.02_{1.72}$
Units	$95.49_{1.87}$	$\mathbf{95.79}_{1.66}$	$\mathbf{96.17}_{2.07}$	$95.69_{3.85}$	$\mathbf{95.65}_{0.68}$	$95.56_{1.52}$
Value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

PKPDai.com

Initial papers #	Papers with abstract # (%)	Abstracts with PK estimates # (%)	PK estimates #
120,913	113,415 (93.80%)	65,377 (57.64%)	274,277



Case study



British Journal of Clinical Pharmacology

Br J Clin Pharmacol (2019) 85 316–346 316

ORIGINAL ARTICLE

Scaling beta-lactam antimicrobial pharmacokinetics from early life to old age

Correspondence Dr Dagan O. Lonsdale, Institute for Infection and Immunity, Room 1.141 St George's, University of London, London, SW17 0RE UK. Tel.: +44 208 725 0205; E-mail: dlonsdal@sgul.ac.uk

Received 2 April 2018; Revised 2 August 2018; Accepted 22 August 2018

Dagan O. Lonsdale[1,2], Emma H. Baker[1,2], Karin Kipper[1,3,4], Charlotte Barker[1], Barbara Philips[1,2], Andrew Rhodes[2], Mike Sharland[1,2] and Joseph F. Standing[1,2,5,6]

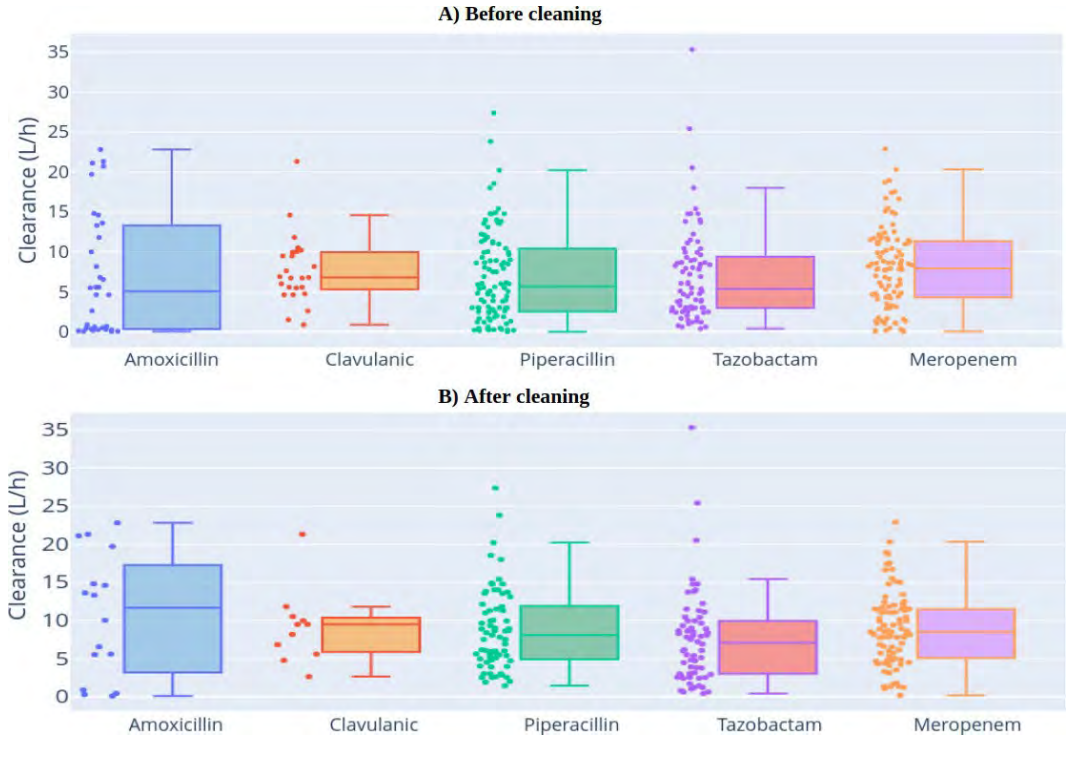
Presence of parameters in abstract

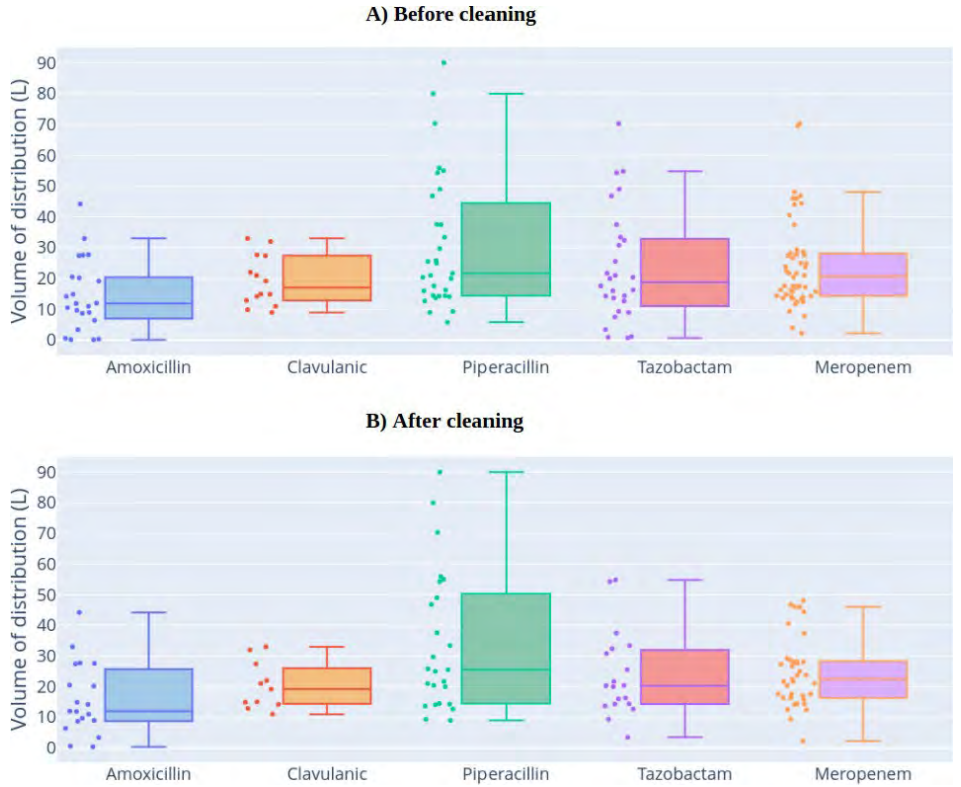
1. Amoxicillin ± clavulanate
   - Clearance 43.48 %
   - Volume of distribution 43.48 %
2. Piperacillin-tazobactam
   - Clearance 42.49 %
   - Volume of distribution 33.33 %
3. Meropenem
   - Clearance 45.28 %
   - Volume of distribution 41.51 %

Recall rates

1. Amoxicillin ± clavulanate
   - Clearance 100 %
   - Volume of distribution 91.67 %
2. Piperacillin-tazobactam
   - Clearance 97.05 %
   - Volume of distribution 87.50 %
3. Meropenem
   - Clearance 91.67 %
   - Volume of distribution 94.45 %

Case study

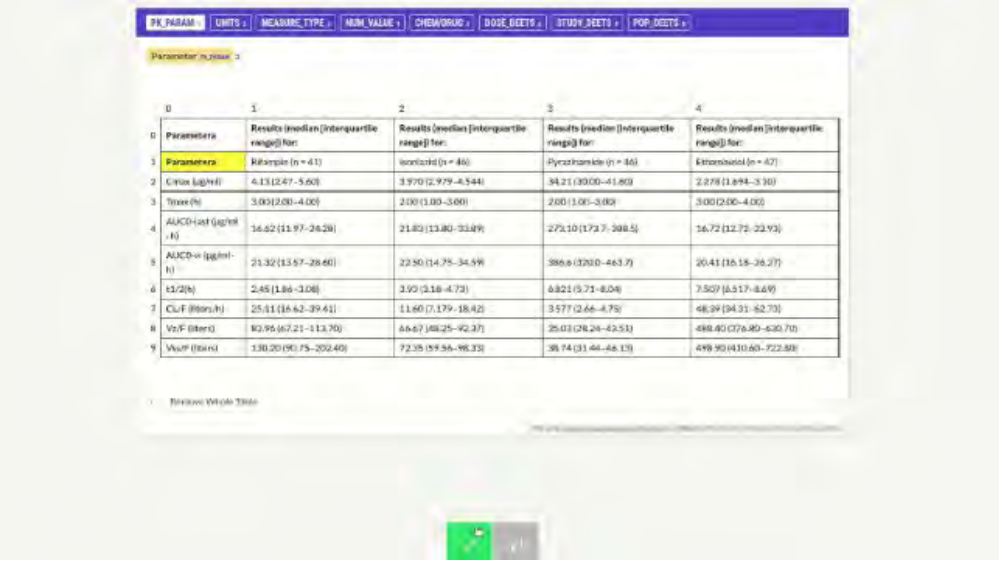

A) Before cleaning
Clearance (L/h)
0
5
10
15
20
25
30
35
Amoxicillin
Clavulanic
Piperacillin
Tazobactam
Meropenem
B) After cleaning
Clearance (L/h)
0
5
10
15
20
25
30
35
Amoxicillin
Clavulanic
Piperacillin
Tazobactam
Meropenem


A) Before cleaning
Volume of distribution (L)
0
10
20
30
40
50
60
70
80
90
Amoxicillin
Clavulanic
Piperacillin
Tazobactam
Meropenem
B) After cleaning
Volume of distribution (L)
0
10
20
30
40
50
60
70
80
90
Amoxicillin
Clavulanic
Piperacillin
Tazobactam
Meropenem

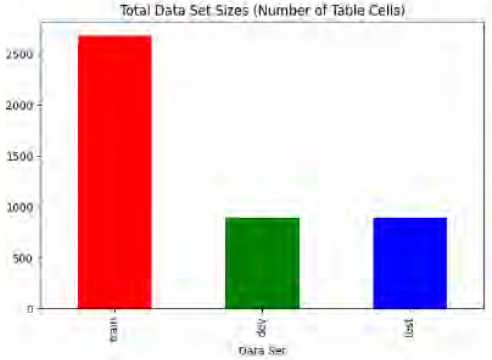
Work in progress: Full text machine reading

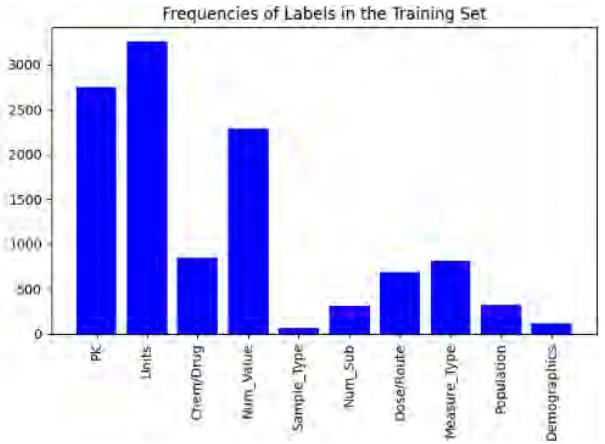
- **Parsed XML PK papers:**
  - Total Number of XML PK Papers: 7,353
  - Total Number of Sentences in the PK Papers (excluding the abstract): 1,159,341
  - Total Number of Results Sentences: 263,092
  - Total Number of Numeric Sentences in Results Sections: 200,880
- **Machine reading ongoing**

Work in progress: Table extraction

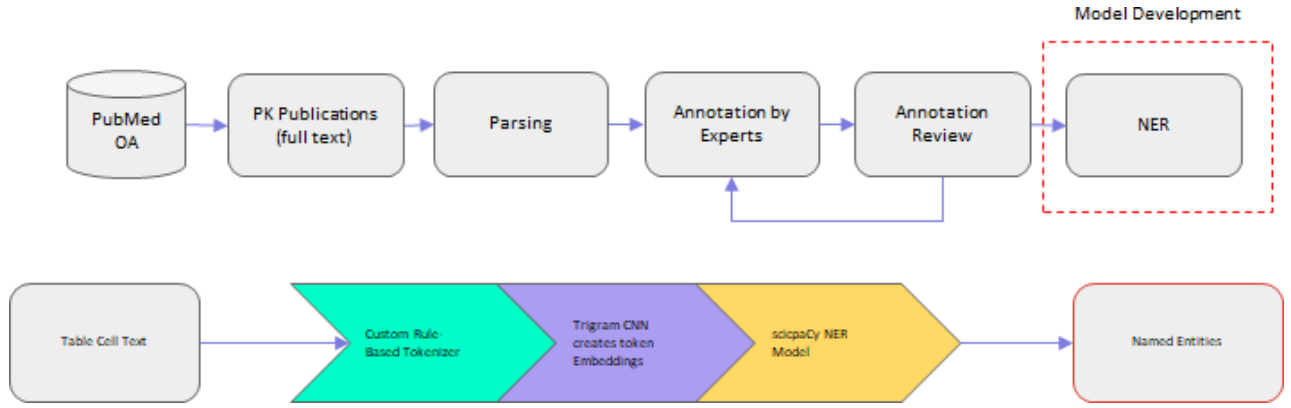


	0	1	2	3	4
0	Parameters	Results (median [interquartile range]) for:	Results (median [interquartile range]) for:	Results (median [interquartile range]) for:	Results (median [interquartile range]) for:
1	Parameters	Rifampin (n = 41)	Isoniazid (n = 46)	Pyrazinamide (n = 46)	Ethambutol (n = 47)
2	Cmax (µg/ml)	4.13 (2.47–5.60)	3.970 (2.979–4.544)	34.21 (30.00–41.80)	2.278 (1.894–3.30)
3	Tmax (h)	3.00 (2.00–4.00)	2.00 (1.00–3.00)	2.00 (1.00–3.00)	3.00 (2.00–4.00)
4	AUC0-last (µg/ml · h)	16.62 (11.97–24.28)	21.80 (13.80–33.69)	273.10 (173.7–388.5)	16.72 (12.73–22.93)
5	AUC0-∞ (µg/ml · h)	21.32 (13.57–28.66)	22.50 (14.75–34.59)	386.6 (320.0–463.7)	20.41 (16.18–26.27)
6	t1/2(h)	2.45 (1.86–3.08)	3.93 (3.18–4.73)	6.821 (5.71–8.04)	7.507 (6.517–8.69)
7	CL/F (liters/h)	25.11 (16.62–39.41)	11.60 (7.179–18.42)	3.577 (2.66–4.75)	48.39 (34.31–62.73)
8	Vz/F (liters)	82.96 (67.21–113.70)	66.67 (48.25–92.37)	25.03 (20.24–42.51)	488.40 (376.80–630.70)
9	Vss/F (liters)	130.20 (90.75–202.40)	72.35 (59.56–98.33)	38.74 (31.44–46.19)	498.90 (410.60–722.80)

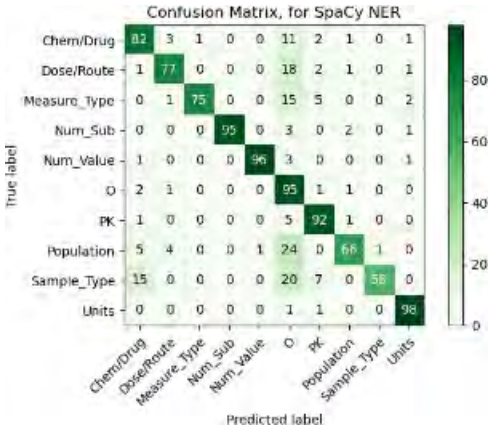




Work in progress: Table extraction



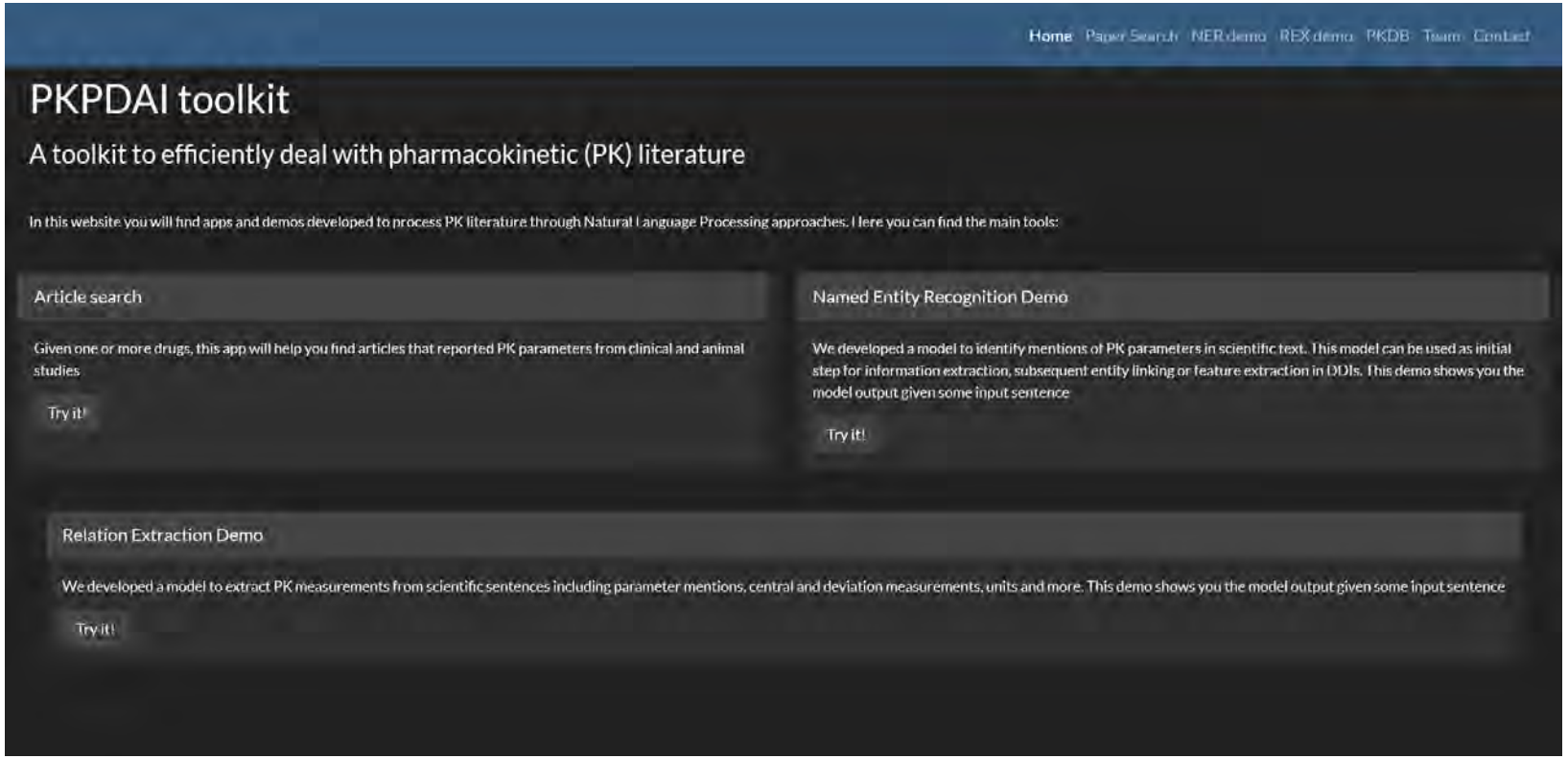
- **Confusion matrix**

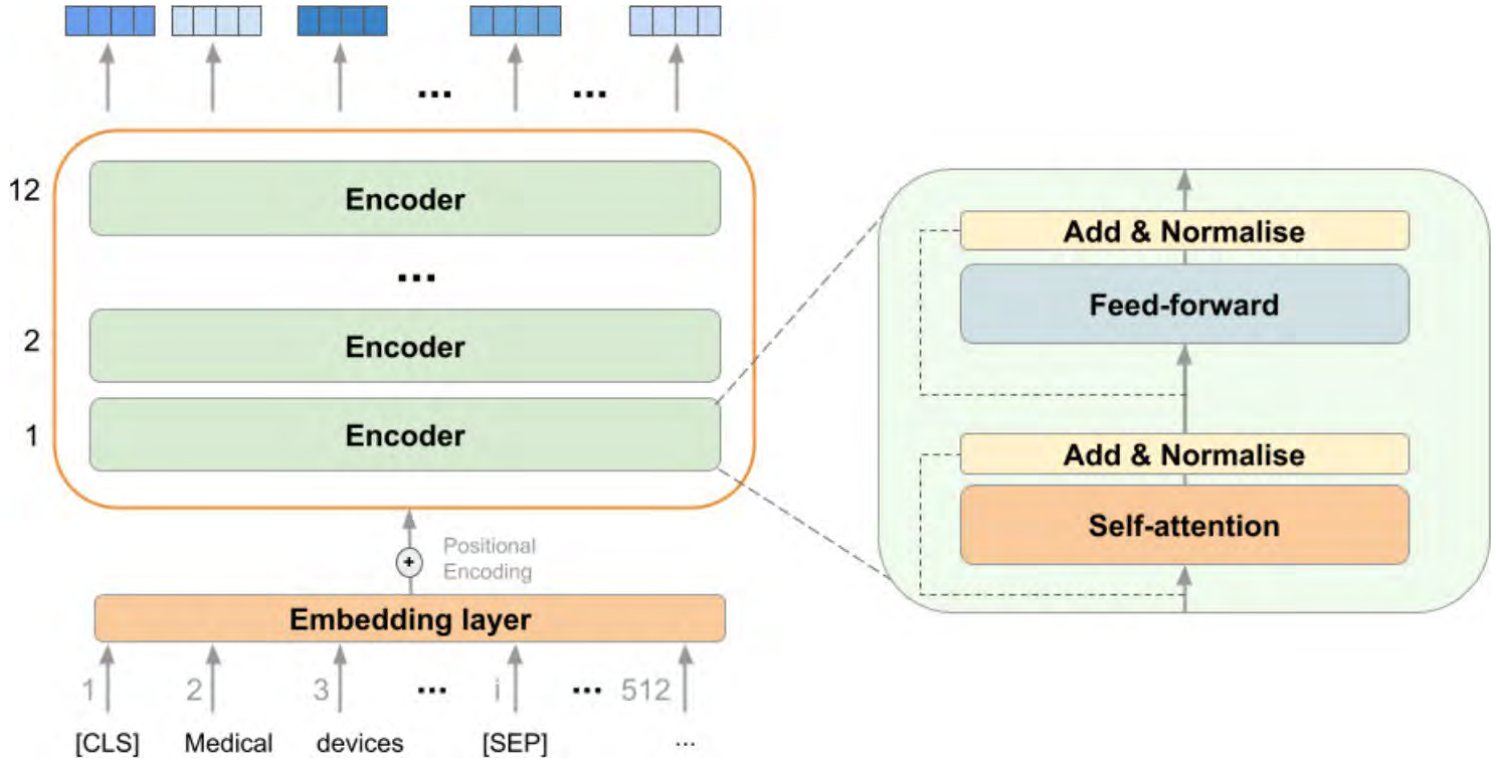


- **Performance**

Entity Label	Annotator F1-score	Model F1-score
Units	0.95	0.98
Numeric Value	0.95	0.97
PK Parameter	0.97	0.96
Number of Subjects	0.95	0.94
Measure Type	0.90	0.94
Chem/Drug	0.94	0.92
Dose/Route	0.82	0.85
Sample Type	0.87	0.83
Population	0.74	0.81
Macro-F1	**0.90**	**0.91**

Please use PKPDai!



info@pkpdai.com | https://pkpdai.com/

Backup-slides