Annotation Guidelines PK Named Entity Recognition (NER)

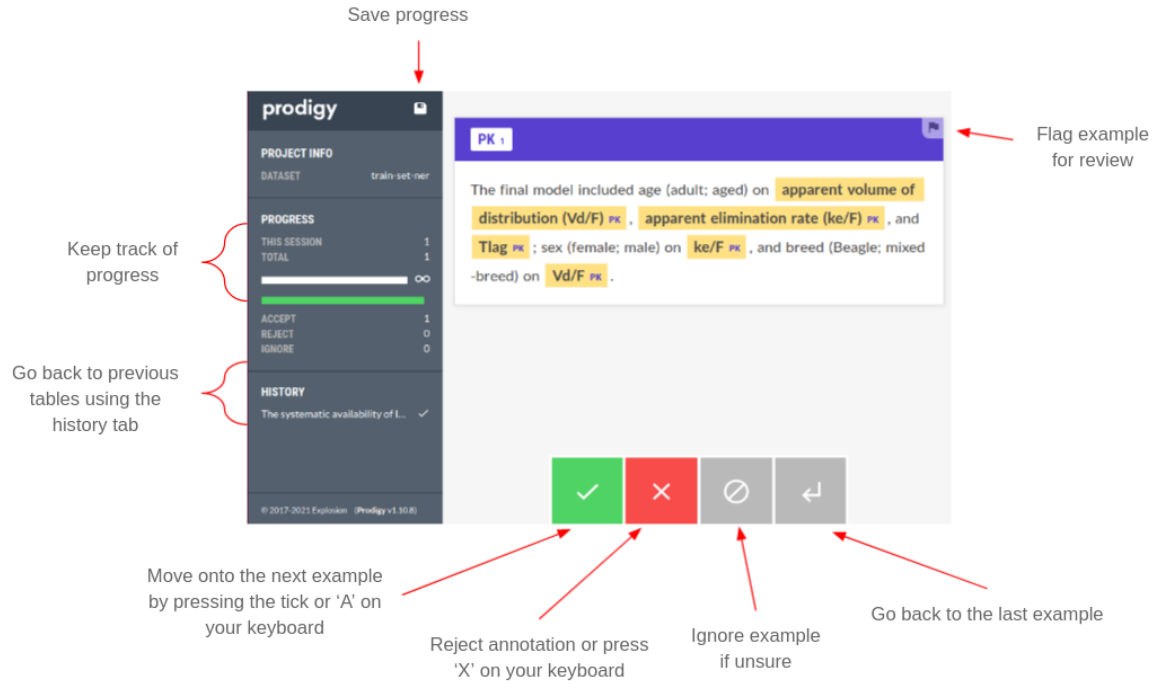
Background

Aim: We aim to develop a named entity recogniser to identify spans of text relating to PK parameter mentions from scientific text.

Method: To do so, we sample sentences within our PK corpus (~120K articles reporting in vivo PK parameters) from the abstract, methods, results and discussion. Then, we ask annotators to label spans of text corresponding to pharmacokinetic parameters.

Task description and interface

The interface displays a single sentence, and the annotator is required to highlight the spans of text relating to PK parameters.



✅ Accept the annotation and move to the following example

🚫 Ignore annotation

🚩 Flag example for review

↩ Go back to the previous example

❌ Reject annotation (only used during active learning)

Questions & Answers

What do we consider "pharmacokinetic parameter" entities? What is included and what is not?

We consider PK entities, those mentions referring to kinetic parameters of pharmacological substances measured either in vivo or in vitro. As a reference, we took the following PK Ontology for the main PK parameters:

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-35/tables/4
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-35/tables/2

However, bear in mind that other PK parameters might appear that are not included in those tables, and some labelling decisions depend on the end-use of the algorithm. We focus on detecting PK parameters for numerical extraction and aiding the characterisation of DDIs, and most labelling decisions are taken with these applications in mind.

This section will try to cover most doubts that appear during the labelling of PK sentences to help annotators resolve their doubts.

Main Cases

Modifiers

Very often, we find noun **modifiers** describing the type of pharmacokinetic parameter, for instance:

1. "**renal** clearance"
2. "**amoxicillin** half-life"
3. "**mean** Vd"
4. "AUC **ratio**", etc.

Answer: When the modifier provides information about the specific **type or subtype of the PK parameter** and can be labelled within a single span, we include it as part of the span to

facilitate subsequent entity-linking efforts. For instance, “renal” and “ratio” would be included since they provide information on the type of parameter (rCL and AUCR, respectively). However, “mean” is not considered to be an entity in this application. If we were to develop an end-to-end information extraction system, “mean” would need to be a different entity type (e.g. type of measurement). “Amoxicillin” refers to a drug, so it is best to keep it outside the PK entity if possible.

1. “renal clearance”
2. “amoxicillin half-life”
3. “mean Vd”
4. “AUC ratio”, etc.

Abbreviations and long-forms

We often see the “full mention” of a pharmacokinetic parameter followed by its abbreviation, for instance:

“The area under the curve from 0 to 24h (AUC0-24h) was calculated for the compound….”
“Oral clearance (CL/F) was 2.3..”

One common doubt is whether to label the complete mention (or long-form) and the abbreviation separately or together.

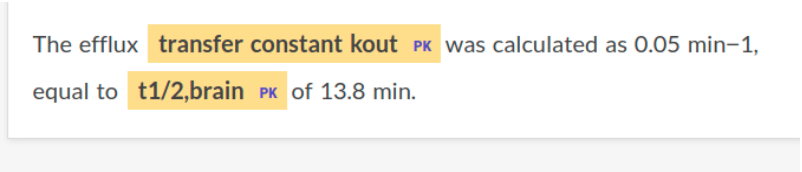
Answer: Since it is easier to relate a single span to a numerical value or a DDI, we always label the longest span that related to a PK mention

“The area under the curve from 0 to 24h (AUC0-24h) was calculated for the compound…”
“Oral clearance (CL/F) was 2.3..”

Also, on some occasions, we found examples in the following forms:

- “TAN possessed a moderate apparent volume of distribution of the central compartment (Vc = 4.2 += 0.82 l/kg), rapid clearance (CL = 94 +- 2 ml/min/kg).”
- “The bioavailability F of compound A was….”

In both cases, we will always try to label the whole span as a single parameter unless it overlaps with other concepts/entities. On other occasions, we find the long-form + abbreviation without parenthesis. In these cases, if possible, we will also label it as a single span:



> The **half-life (t1/2)** PK was calculated as ln2/k and the **elimination rate constant k** PK was estimated as the negative of the slope from a linear regression of log concentration of time.

Creatinine Clearance

Creatinine clearance is often measured as a covariate in PK studies and appears in multiple sentences.

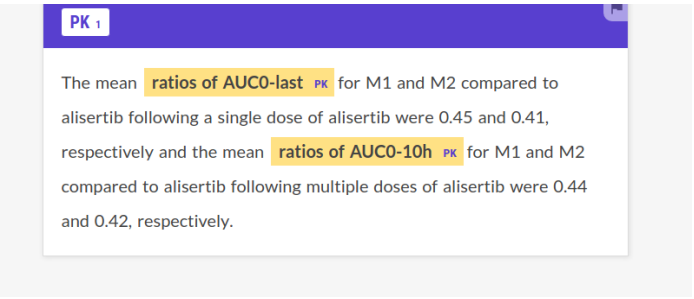
Answer: Creatinine is an endogenous substance, and therefore the clearance never relates to the clearance of an administered compound. Accordingly, creatinine clearance is **not labelled as a PK mention**.

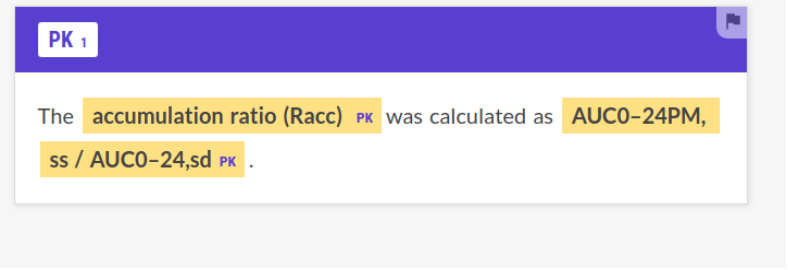
Same with albumin:

> Human serum albumin clearance studies were performed to determine the effect of endogenous MSA on the pharmacokinetic behavior of administered human albumin.

Parameter ratios

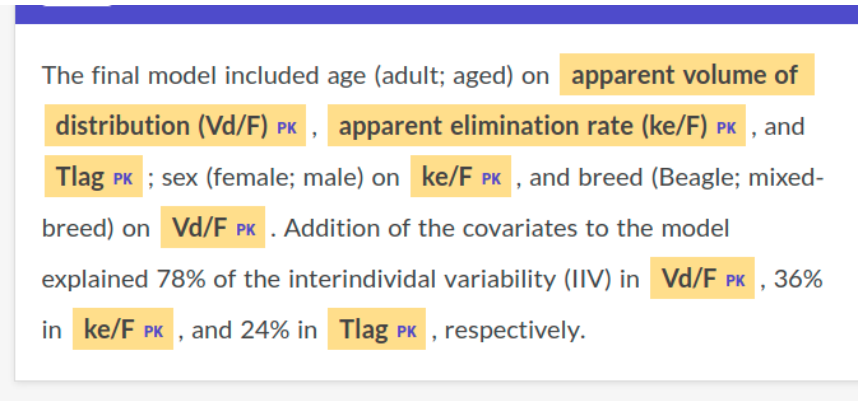
Often we find ratios of the same parameters:





Answer: We will include both parameters of the ratio within the mention. In addition, the term ratio or ratios is also included if present.

On other occasions, we find normalised parameters or different parameters that are divided by each other:



In these cases, we include both parameters in the ratio as part of one single span since the whole span is generally the one that is discussed in numerical estimations or DDI context.

Ratios defining other parameters

Quite often, we find mentions like:

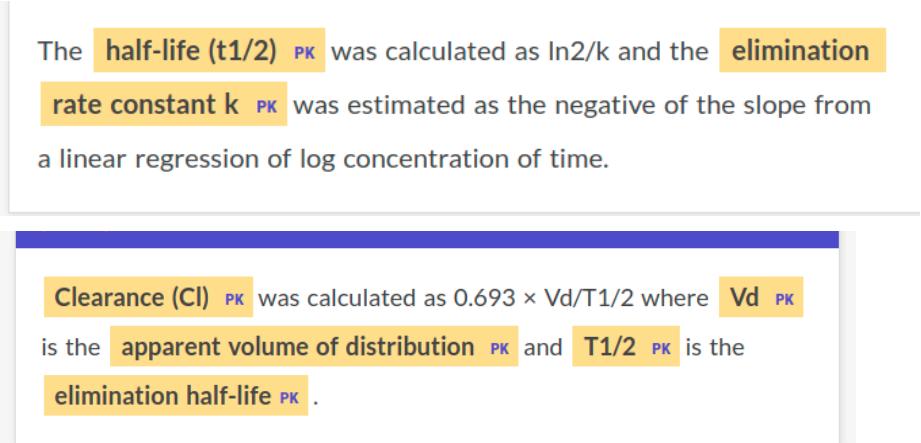
“0.6/Ke”
“Ln2/k”
“Ln2/lambda”

These mentions often refer to how a parameter was calculated and are not helpful entities for PK numerical extraction or DDI.

The half-life (t1/2) PK of the drug was calculated by use of the relationship t1/2 = 0.693/Kel.

The elimination half-life (T1/2) PK was calculated as 0.693/λz.

Terminal elimination t1/2 PK was calculated as ln(2)/λz.

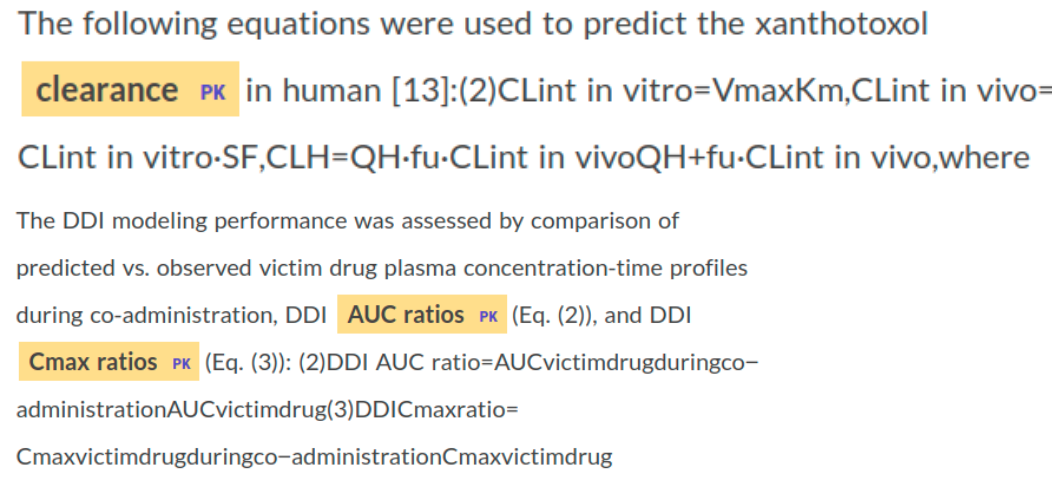


Answer: Only if a numerical estimation is related to this coefficient and the whole parameter mention is not present we will consider it as a PK span. Otherwise, we will not label those as PK mentions.

In the examples above, Vd/T1/2 were not labelled as parameters either since it is part of the definition of CL calculation.

Equations

We often encounter equations that describe how PK parameters were calculated. For instance:



If rapid achievement of steady-state morphine concentration is desired , an intravenous loading dose may be calculated using Equation 7 or simply DL = Ctarget · Vd.

Answer: It is often hard to separate the PK terms within the equation, and, in almost all cases, they do not provide valuable information to extract numerical values or characterise DDIs. **Therefore, we will not label parameters within equations.**

Less-common cases

These are rare mentions. However, to ensure consistency across the dataset, make sure the following cases are always labelled as described:

Drug mentions within parameter mentions

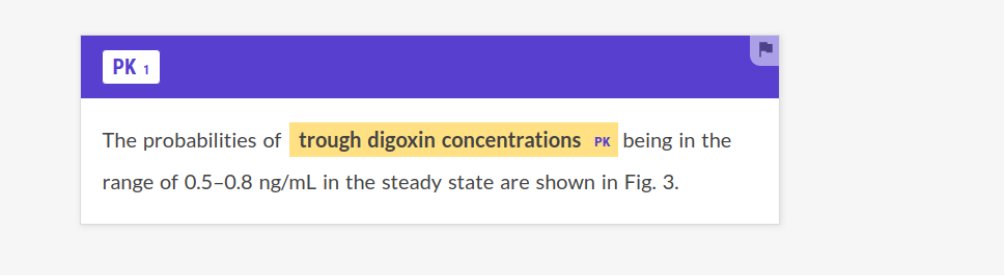
Occasionally, we will find cases in which a drug is mentioned within a PK parameter, e.g.

"The area under the ***midazolam*** concentration-time curve was…."

In general, we will try to avoid including drugs into PK spans to avoid nested entities, but in this case, we cannot avoid it, so we will label the whole span as PK mention and rely on separate NER models for drugs and PK or models that can deal with overlapping entities:

"The area under the midazolam concentration-time curve was…."

Other examples:



On other occasions such as:

"Median plasma AG10-AG tmax was 1 hour across all dose levels."

We might wonder whether to label only tmax or with the modifier plasma tmax:

1. "Median plasma AG10-AG tmax was 1 hour across all dose levels"
2. "Median plasma AG10-AG tmax was 1 hour across all dose levels"

Because in this case, including “plasma” would also require including the compound mentioned “AG10-AG”, we will only label the span “tmax” as an entity. The idea is that “plasma” is not an essential term to understand the parameter being discussed, and posterior entity linking systems could take “plasma” into account. So we would select option 1

Modifiers + parameter cue

In some cases, we also encounter the following:

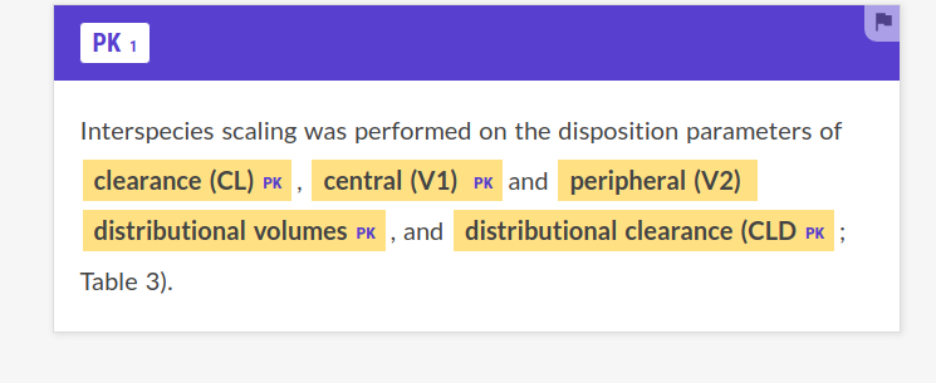
“The **systemic and oral clearance** of midazolam was x and y h/kg, respectively.”

In these cases, we have two parameter mentions that might refer to different numerical estimations. In these cases, we will label in the following manner:

“The systemic and oral clearance of midazolam was x and y h/kg, respectively.”

Even though “systemic” would not refer to a PK parameter, it does refer to systemic clearance if found alone. Therefore, context-aware entity linking systems could still account for this fact, and each mention could still be related to different values.

Other examples:



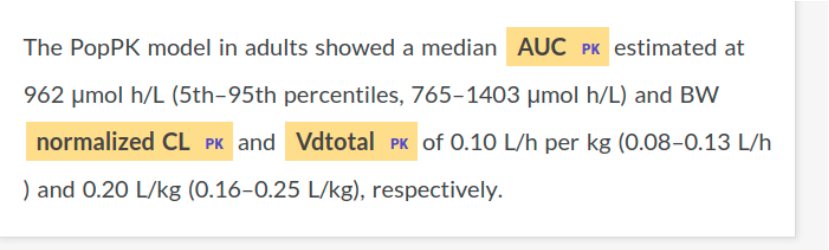
Central and peripheral

Normalised + parameter mention

Answer: Yes if possible:

“Normalized CL was..”

It might help posterior entity linking.



Serum + parameter mention

Answer: Yes if possible:

"Serum CL was.."

It might help posterior entity linking.

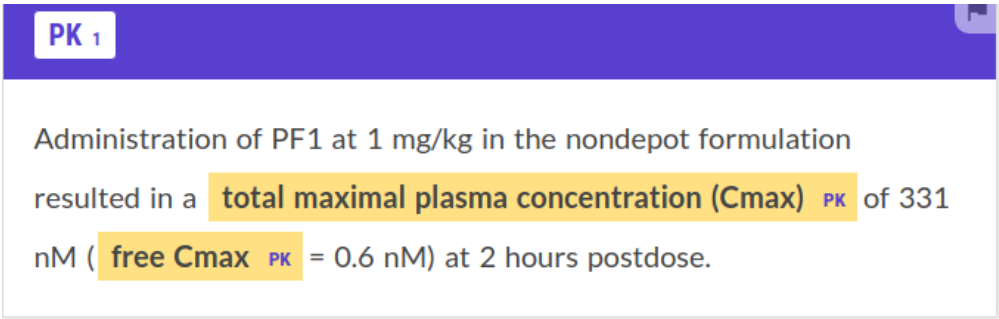
In vitro / in vivo + parameter mention

Answer: No need. It is a contextual entity that might need to be labelled as a separate type for specific information extraction applications.

"*In vitro* CL was x.."
"*In vivo* t1/2 was y.."

Free/total + parameter mention

Answer: Yes



The predicted free AUC/MIC PK level was more sensitive to changes in free fraction levels when compared to the total AUC/MIC PK (Figure6c, right).

Parameters to include

We will often have doubts on whether some parameter mentions are included or not. However, as we mentioned above, any parameter included in the following tables is considered PK according to the PK Ontology.

Table 2:

Experiment types	Parameters	Description	Unit	References
Single Drug Metabolism Experiment	K_m	Michaelis-Menten constant.	mg L^{-1}	Segel p28.
	V_{max}	Maximum velocity of the enzyme activity.	mg h^{-1} mg^{-1} protein	Segel p19
	CL_{int}	Intrinsic metabolic clearance is defined as ratio of maximum metabolism rate, Vmax, and the Michaelis-Menten constant, Km.	ml h^{-1} mg^{-1} protein	RT p165
	Metabolic ratio	Parent drug/metabolite concentration ratio	NA	
	fm_{enzyme}	Fraction of drug systemically available that is converted to a metabolite through a specific enzyme.	NA	RT xiii
Single Drug Transporter Experiment	Papp	The apparent permeability of compounds across the monolayer cells.	cm/sec	Transport Consortium
	Re	Re is the ratio of basolateral to apical over apical to basolateral.	NA	Transport Consortium
	Radioactivity	Total radioactivity in plasma and bile samples is measured in a liquid scintillation counter	dpm/mg protein	Transport Consortium
	Uptake Volume	The amount of radioactivity associated with the cells divided by its concentration in the incubation medium.	ul/mg protein	Transport Consortium
Drug Interaction Experiment	IC_{50}	Inhibitor concentration that inhibits to 50% of enzyme activity.	mg L^{-1}	
	K_i	Inhibition rate constant for competitive inhibition, noncompetitive inhibition, and uncompetitive inhibition.	mg L^{-1}	Segel p103
	K_{deg}	The natural degradation rate constant for the Enzyme.	h^{-1}	Rostami-Hodjegan and Tucker
	K_I	The concentration of inhibitor associated with half maximal Inactivation in the mechanism based inhibition.	mg L^{-1}	Rostami-Hodjegan and Tucker
	K_{inact}	The maximum degradation rate constant in the presence of a high concentration of inhibitor in the mechanism based inhibition.	h^{-1}	Rostami-Hodjegan and Tucker
	E_{max}	Maximum induction rate	Unit free	Rostami-Hodjegan and Tucker
	EC_{50}	The concentration of inducer that is associated with the half maximal induction.	mg L^{-1}	Rostami-Hodjegan and Tucker

Table 4:

Name	Description	Unit	reference
AUC_{inf}	Area under the drug concentration time curve.	mg h L^{-1}	RT p37
AUC_{SS}	Area under the drug concentration time curve within a dosing curve at steady state.	mg h L^{-1}	RT pxi
AUC_t	Area under the drug concentration time curve from time 0 to t.	mg h L^{-1}	RT p37
AUMC	Area under the first moment of concentration versus time curve.	mg^2 h L^{-2}	RT p486
AUCR	AUC ratio (drug interaction parameter).	Unit free	
CL	Total clearance is defined as the proportionality factor relating rate of drug elimination to the plasma drug concentration.	ml h^{-1}	RT p23
CL_b	Blood clearance is defined as the proportionality factor relating rate of drug elimination to the blood drug concentration.	ml h^{-1}	RT p160
CL_u	Unbound clearance is defined as the proportionality factor relating rate of drug elimination to the unbounded plasma drug concentration.	ml h^{-1}	RT p163
CL_H	Hepatic portion of the total clearance.	ml h^{-1}	RT p161
CL_R	Renal portion of the total clearance.	ml h^{-1}	RT p161
CL_{po}	Total clearance of drug following an oral dose.	ml h^{-1}	
CL_{IV}	Total clearance of drug following an IV dose.	ml h^{-1}	
CL_{int}	Intrinsic metabolic clearance is defined as ratio of maximum metabolism rate, Vmax, and the Michaelis-Menten constant, Km.	ml h^{-1}	RT p165
CL_{12}	Inter-compartment distribution between the central compartment and the peripheral compartment.	ml h^{-1}	
CL ratio	Ratio of the clearance (drug interaction parameter).	Unit free	
C_{max}	Highest drug concentration observed in plasma following administration of an extravascular dose.	mg L^{-1}	RT pxii
C_{max} ratio	The ratio of C_{max} (drug interaction parameter).	Unit free	
C_{ss}	Concentration of drug in plasma at steady state during a constant rate intravenous infusion.	mg L^{-1}	RT pxii
C_{ss} ratio	The ratio of C_{ss} (drug interaction parameter).	Unit free	
E	Extraction ratio is defined as the ratio between blood clearance, CL_b, and the blood flow.	Unit free	RT p159
E_H	Hepatic extraction ratio.	Unit free	RT p161
F	Bioavailability is defined as the proportion of the drug reaches the systemic blood.	Unit free	RT p42
F_G	Gut-wall bioavailability.	Unit free	
F_H	Hepatic bioavailability.	Unit free	RT p167
F_R	Renal bioavailability.	Unit free	RT p170
fe	Fraction of drug systemically available that is excreted unchanged in urine.	Unit free	RT pxiii
fm	Fraction of drug systemically available that is converted to a metabolite.	Unit free	RT pxiii
fu	Ratio of unbound and total drug concentrations in plasma.	Unit free	RT pxiii
k	Elimination rate constant.	h^{-1}	RT pxiii
K_{12}, k_{21}	Distribution rate constants between central compartment and peripheral compartment.	h^{-1}	
ka	Absorption rate constant.	h^{-1}	RT pxiii
ke	Urinary excretion rate constant.	h^{-1}	RT pxiii
km	Rate constant for the elimination of a metabolite.	h^{-1}	RT pxiii
Km	Michaelis-Menten constant.	mg L^{-1}	RT pxiii
MRT	Mean time a molecular resides in body.	h	RT pxiv
Q	Blood flow.	L h^{-1}	RT pxiv
Q_H	Hepatic blood flow.	L h^{-1}	RT pxiv
t_{max}	Time at which the highest drug concentration occurs following administration of an extravascular dose.	h	RT pxiv
$t_{1/2}$	Half-life of the drug disposition.	h	RT pxiv
$t_{1/2}$ ratio	Half-life ratio (drug interaction parameter).	Unit free	
$t_{1/2,\alpha}$	Half-life of the fast phase drug disposition.	h	
$t_{1/2,\beta}$	Half-life of the slow phase drug disposition.	h	
V	Volume of distribution based on drug concentration in plasma.	L	RT pxiv
V_b	Volume of distribution based on drug concentration in blood.	L	RT pxiv
V_1	Volume of distribution of the central compartment.	L	RT pxiv
V_2	Volume of distribution of the peripheral compartment.	L	
V_{ss}	Volume of distribution under the steady state concentration.	L	RT pxiv
Vmax	Maximum rate of metabolism by an enzymatically mediated reaction.	mg h^{-1}	RT pxiv
λ_1, λ_2	Disposition rate constants in a two-compartment model.	h^{-1}	GP p84

Apart from those, other parameters also appeared. Here, we will cover most cases that we have come across:

MIC

Answer: MIC alone is not considered as a PK parameter

- "The MIC of MDZ was…"

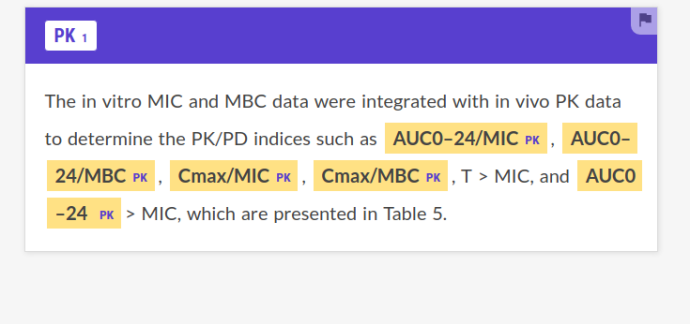
PK/MIC

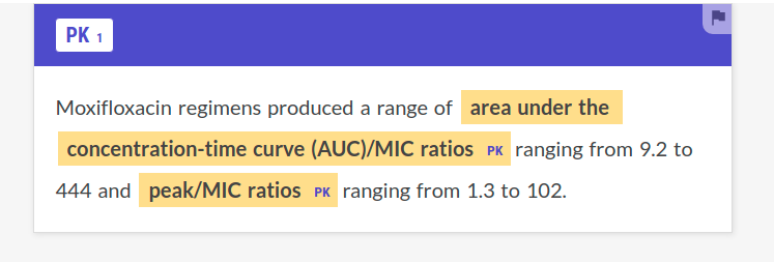
We often encounter rations of PK/PD parameters. For instance:

- “The AUC/MIC was…”
- “The CL/MIC was…”

Answer: We do consider this whole span, as PK mentions

- “The AUC/MIC was…”
- “The CL/MIC was…”





Bioavailability

Answer: Always include with modifier if present. Careful with the abbreviation F, since it can sometimes refer to other concepts:

Both values of F were less than F0.05(2,2), with the value 19.00 indicating that the regression equations were not statistically significant, and the arranged ethosome compositions had no significant effect on EE or DSD.

Figure 2E,F illustrates the intestinal stability of PMX53 and PMX205 when incubated at 37 °C for up to 60 min.

Test conditions were the same as described in the “Fluorescence properties of compound F” section.

Absorption

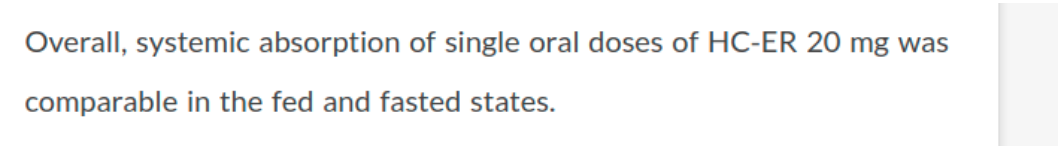
Absorption is mentioned very frequently as a general property of chemical compounds. For instance:

“The **absorption** of drug X (F=45%) increased when co-administered with drug Y (F=68%)

However, it often does not refer to numerical values of parameters but is derived from other parameters (most often **bioavailability (F)**).

Answer: We will label it as PK if the mention refers to a kinetic parameter (e.g. absorption rate) but not if it refers to a general property. Therefore in the example above, absorption would not be labelled as PK, but F would:

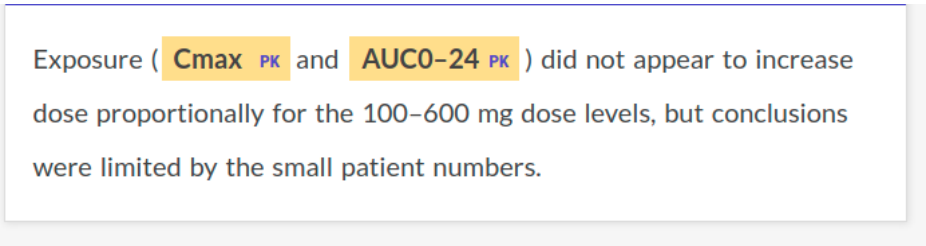
“The absorption of drug X (F=45%) increased when co-administered with drug Y (F=68%)



Exposure

Like absorption, sometimes we find sentences describing the “exposure” of a drug, often measured through the AUC or Cmax.

We will not label exposure as a parameter but the actual parameter used to describe the exposure (often AUC).



Concentrations

Multiple concentrations are mentioned across PK literature. However, here we will **include**:

- Cmin
- Cmax
- Cthourgh
- Cavg
- Css
- Tissue-to-blood concentration ratio
- X-to-Y concentration ratio

We will **not include:**

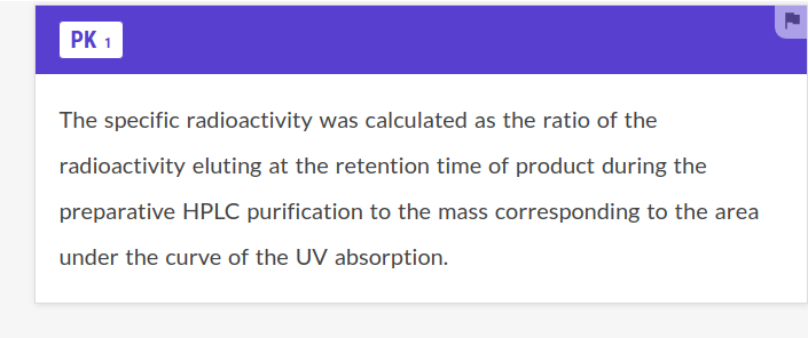
- Clast
- C0
- Minimum quantifiable C
- Cin
- Cout
- Cp (concentration in compartment x)
- MIC
- Other C

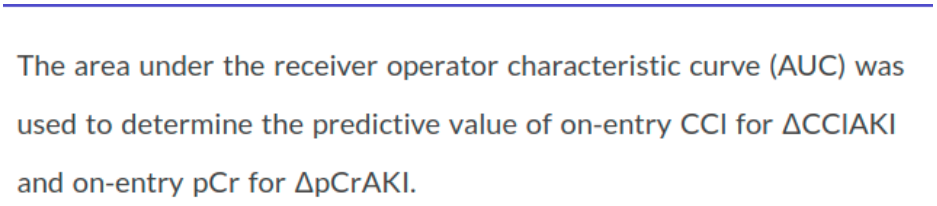
Plasma concentration-time curve

Not as such.

AUC

Most often refers to the area under the concentration-time curve. Then we would label it as a PK parameter. However, on some occasions, it refers to other, non-PK relevant curves (e.g. chromatography). In those cases, they will not be labelled:





AUMC

The area under the moment curve: Yes, always include

AUEC

The area under the effect curve -> No, considered PD parameter.

IC50

Yes, always → present in Table 2

Bmax

No, this refers to the number of binding sites

Dissociation and association constants

No, physicochemical properties

Permeability and radioactivity

We found many mentions across the literature. Mainly:

- Apparent permeability (Papp), Ratio of the basolateral to apical permeability and Apical to basolateral permeability (Re) Radioactivity, permeability

These mentions are present in table 2. These are less relevant for this application since they are in vitro parameters, but we will label them if present.

Kdeg

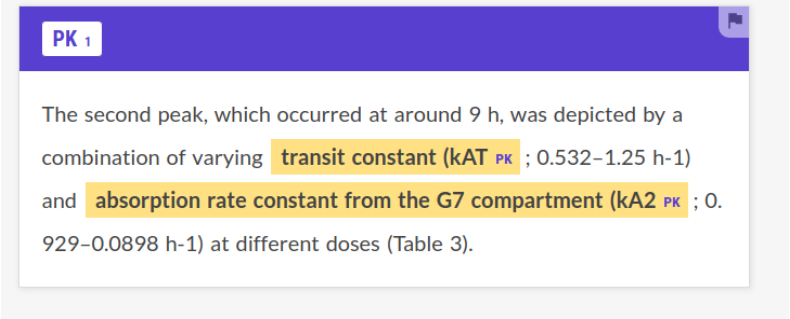
Yes → Table 2

KI

Yes → Table 2

Transit constant

Yes, if mentioned in PK context



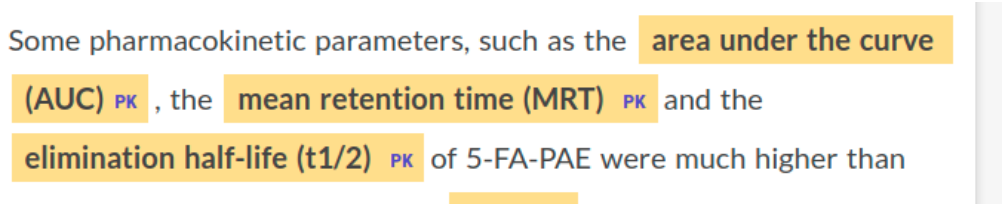
Similarly, the parameters for disposition and gastrointestinal transit absorption process (kA PK , kA2 PK , kAT PK , and F PK) were obtained by simultaneously fitting the IV and oral data from dogs with Eqs 4–9 (Fig 2B).

Mean residence time

Yes

Retention times

Not if mentioned in the context of chromatography. On some occasions, retention is used interchangeably with residence:



Formation constant

No, this is often a physicochemical property

Solubility

No, this is often a physicochemical property.

Rf

Very rare, but it has eventually been referred to as Retention Fraction. Not considered PK entity.

Hill coefficient

Yes

Kapp

Depending on the context, it can refer to different concepts. If PK parameter is in vivo or in vitro, yes. It is not a frequent mention but if it appears, check the context.

Dose

No, not considered PK parameter

MTD

No, considered PD property

Perfusion rate

No, most often not reported in PK context

Emax

Yes, included in table 2

Imax (maximum Inhibition)

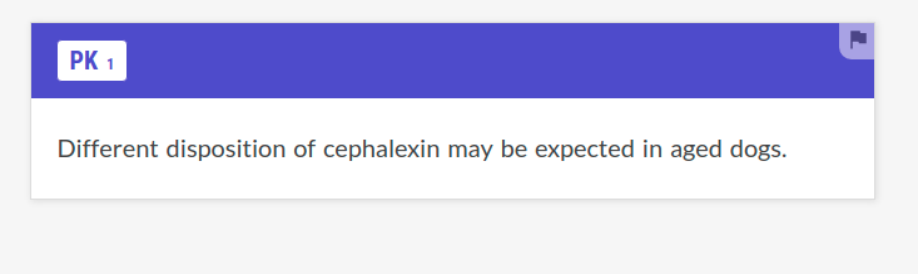
Yes, to be consistent with labelling Emax

IC50

Yes

Disposition constants

Disposition rate constants are considered PK parameters. However, when discussing the general concept of disposition, we do not label it as PK, for instance:

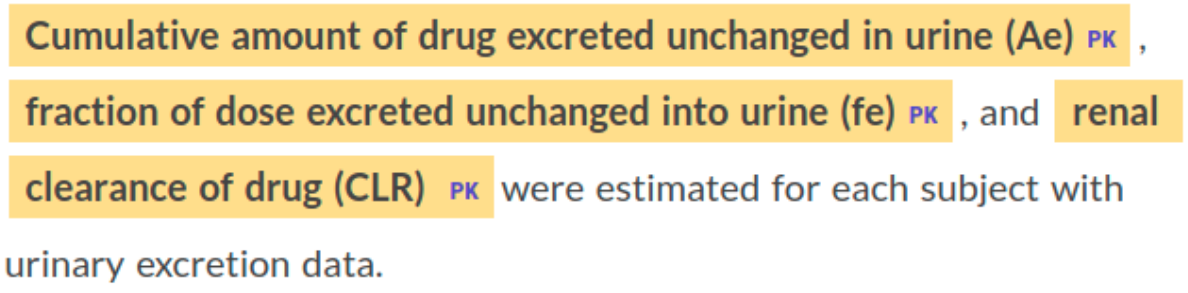


A0

Yes

AE

In general, it refers to adverse effects, so we do not label it as PK. However, it can occasionally refer to PK parameters:



Bleeding rate

No, endogenous and not related to compound.

Peak areas

Often found in chromatography studies but not relevant as a PK parameter:

> Peak areas in the chromatograms for the spiked plasma samples containing the above lowest concentrations were compared with the signal-to noise ratio ≥ 10.

Flow rate

In general, no, unless is the blood flow rate (which is included in table 4)

> Acetonitrile and water containing 0.1% (v/v) formic acid was adopted as the mobile phase, the flow rate was 0.3 mL/mi.

> However, the mobile phase was a mixture of acetonitrile/water/tert-butyl methyl ether/phosphoric acid (525/425/50/1, v/v/v/v) at a flow rate of 1.5 mL/minute for in vivo determination.16

Cout Cin

Influent and effluent drug concentrations. Not relevant, often reported for in silico models.

Glomerular filtration rate

No, often mentioned as a covariate

Diffusion coefficient

No

Bound/unbound fraction

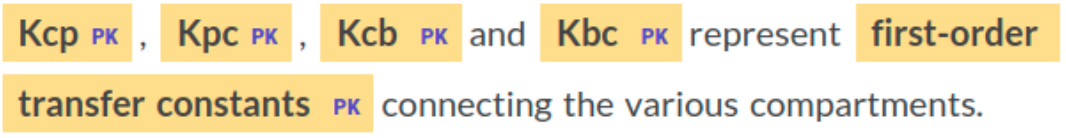
Yes

Times

Tmax and tlag yes, others no (e.g. t0, tlast). We can apply the same rules as for the concentrations.

Transfer constants between compartments

Yes



Biotransformation

No