Database

PK-DB: a pharmacokinetics database for individualized and stratified modelling

Jan Grzegorzewski[1], Dimitra ?, Kathleen Green, and Matthias König[1,*]

[1]Institute for Theoretical Biology, Humboldt University Berlin, Berlin, Germany

[*]Corresponding author

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: We present PK-DB, a database for the representation of pharmacokinetics data from clinical trials and pre-clinical research. Data is either curated from the literature or from accessible raw data. The main focus of PK-DB is to provide high-quality pharmacokinetics data in combination with required meta-information for computational modeling and data integration, i.e., (i) characteristics of studied patient collectives and individuals; (ii) applied interventions (dosing, route, …); and (iii) measured pharmacokinetics information and time courses. Important features are the representation of experimental errors and variation, the representation and normalisation of units, annotation of information to biological ontologies, and calculation of pharmacokinetics information like apparent clearance, half-life, or area under the curve (AUC) from time course data.
We demonstrate the value of PK-DB by (i) a stratified meta-analysis of pharmacokinetics studies for caffeine curated from the literature, thereby integrating pharmacokinetics information from a wide range of sources; (ii) providing examples for the computational modeling of dynamical liver function tests based on PK-DB data.

Database available from: https://develop.pk-db.com/
Latest source code: https://github.com/matthiaskoenig/pkdb
Archive source code as at the time of publication: https://zenodo.org/record/1472242
License: GNU General Public License, version 3 (GPL-3.0)
Contact: konigmatt@googlemail.com

1 INTRODUCTION

Pharmacokinetics Pharmacokinetics is the study on how substances applied to the human body distribute within the body {REF}. A central medical field is the study of pharmacokinetics of drugs and medication. Despite the wealth of reported data in the literature and the requirement for pharma-companies to perform such experiments before bringing drugs on the market almost none of the data is publicly accessible in a machine readable format. A multitude of publications have been written about pharmacokinetics studies, but almost none of the data is accessible so far. The only way to retrieve this treasure is by digitizing and curating this information from the publications.
Standardized representation of data in machine-readable formats is a requirement for computational analysis of data, data science, and the application of methods of machine learning and artificial intelligence {REF}. Open and free access to such information is important. Despite the central role of pharmacokinetics information in the medical and pharma field, or perhaps exactly because of that, no open freely accessible database of pharmacokinetics information exists so far.

- Data integration very challenging because:

    - different clinical protocols and dosing protocols
    - different groups, i.e. very difficult to compare
- Computational models provide a unique opportunity to use this data and provide stratified and individualized information
- Challenging to translate data from one clinical trial to other trials. Computational models accounting for the differences in study protocol and study groups/individuals could provide unique value.
- But currently data model, database as well as curation protocols for data integration missing

So what is pharmacokinetics data? In most pharmacokinetics studies

With exception of PKDB {REF} which reported some information for pharmacokinetic modeling no such database is available so far.
Here we present PK-DB a database for pharamacokinetics information and data which allows

2 DESCRIPTION

Technology and design

Key principle in the design of PK-DB were **(i) accessibility of data for computational modeling and data science**; **(ii) extensibility and generalizability**, i.e., not being too focused on a narrow problem domain, but allowing simple extension to other fields and experimental data sets. A good example are the group and individual characteristics. Additional characteristica can easily be added to cover the important information for a given problem domain. **(iii) tracking changes in curated data** and allowing multi-curator curation.
PK-DB was designed as a web-accessible database via a web frontend. The backend is using django and is written in python. **(iv) unit and data normalisation**; A key challenge in using data for computational modeling are unclear units and data sets from different studies having different units. This requires time consuming retrieval of this information from the literature and error-prone conversion of units and corresping data. **(v) representation of timecourse data**. A key data source in pharmacokinetics studies are time courses of the applied substance and measurements of metabolites after biotransformations. These time courses are crucial for kinetic modeling, e.g., using physiological based models or pharmacokinetics/pharmacodynamics models. Important to be able to calculate secondary pharmacokinetics information from such data sets.

Access to the database consisting of upload/update of studies and retrieving information from the database is enabled via a REST API based on the django-rest-framework. A key requirement for PK-DB was to be able to interact with the database via different means, i.e. a web frontend and programatically to be able to integrate the data with computational models. A big advantage of using a REST API as central access point to the database is that we are not tying us to a single programming language, but allow users to interact with the database in a way they see fit. We demonstrate the usefulness of the approach, by on the one hand implementing a web frontend in javascript based on vue.js using the API and running analysis scripts against the database in python, or creating overview plots of the database content using R (circos). Direct integration into modelling and analysis workflows is facilitated via the available REST API.
By providing simple programmatic access we enable things like meta-analysis and combined analysis pooling data from multiple studies (we show such an application for stratified analysis of caffeine clearance rates below).
For indexing in the backend we use elasticsearch, which provides useful search endpoints and allows to index the database content for fast access. To allow to run the database locally we provide a docker container to setup the database (easy setup with all dependencies).
For the actual data representation in the curation workflow we decided on a simple JSON format, which after changes will be used to update the database content. This had multiple reasons: (i) we can easily track changes via source control, namely github. Curation is an iterative process with often changing curators over time.

Tracking changes to the curated data is crucial. Instead of implementing such history and change tracking on database level with substantial overhead, we can use the full set of git features out of the box to track changes to our files; (ii) no need for a secondary curation frontend. We can curate data while we develop the database without the need to maintain a full frontend for curation; (iii) we can ensure correctness of the JSON by implementing validation rules in the backend which are applied to the json and accompanying files on save. This allows the curator to perform changes to the json file with direct feedback about allowed choices. Many constraints are added on the validation layer instead of having the data model layer too restrictive.
PK-DB is accessible from https://pk-db.com providing a web interface.

Statistics

The main focus of data curation lies on clinical studies for substances used in dynamical liver function tests as well as for the modelling of whole-body glucose metabolism.
PKDB-v0.7.0 consists of 159 studies containing 424 groups, 1582 individuals, 482 interventions, 14067 outputs and 1105 time courses related to caffeine, glucose, codeine, and paracetamol (see Figure 1, Figure 2, Figure S1, Figure S2 and Table S3).

Curation workflow & minimal information for pharmacokinetics data

The typical workflow for extracting data from the literature is depicted in Figure?. After an initial literature research selecting a corpus of papers to curate (reported data, useful data, minimal quality criteria). Within the curation process the releveant information is manually extracted from the literature and encoded in a study.json format.

Calculation of pharmacokinetics parameters

An important part of PK-DB is the automatic calculation of pharmacokinetics parameters from the reported data. The most prominent is the calculation of pharmacokinetics parameters based on non-compartmental methods (see Figure 3 and Table 1)

Meta-analysis of caffeine

In the following an example application is presented demonstrating the programatic interaction with PK-DB.
The use of our database allowed us for the first time to systematically analyze the effect of lifestyle factors like smoking and oral contraceptive use on pharmacokinetics data like clearance or half-lives. We integrated data from more than ? studies. By curating information about the respective patient characteristics (lifestyle factors), the actual interventions performed in the studies (dosing and route), and important information like the errors on the reported data we could gain a unique view on the large (and consistent effect) of smoking and oral contraceptive use on the clearance of caffeine.
Importantly, the meta-analysis allowed us to directly improve the curation status of many studies by easily visually detecting outliers in the data, which could in most cases directly be backtracked to curation errors, which could subsequently be corrected.
A positive point is that most of the reported studies are consistent. For instance with caffeine, most of the data was in line with each other with a single exception being Balogh1992 {REF}? Here a systematic bias in the data could be observed probably due to an analytic problem; In addition problems existed with reporting the same data set multiple times, overall in four publications {Harder, ….}. A second example is for instance an extreme outlier for smoking? probably due to an incorrect comma in the original data table.
The extreme variability between studies and individuals could be markedly reduced by accounting for lifestyle information.

CONCLUSION

PK-DB is the first publicly available database for pharmacokinetics data. We could demonstrate the value of PK-DB by performing meta-analysis of the available studies for caffeine clearance. The database allows based on the data model and integration of study information like dosing and group information the stratification and individualization of the available data sets.

Point about availability of data.
By providing a first open database for pharmacokinetics information we provide an important resource which allows to store pharmacokinetics information in a FAIR (findable, accessible, interoperable and reproducible) manner {Wilkinson2016}.
By performing the data curation for commonly apply drugs (codeine and paracetamol), a substance used in liver function tests (caffeine) and a well studied substance (glucose) we could gain insights into how well data is reported in the various fields.
In summary, the reporting of data is very poor despite the main point of the publications being the reporting of the data. Without guidelines on minimal information for studies it is very difficult to compare studies or integrate data from different sources. A good example for this are minimal guidelines about reporting patient characteristics for individuals and groups (which is lacking in most of the studies
Based on our work we have a set of Important suggestions when publishing clinical studies using pharmacokinetics are: (i) publish the actual data in a machine readable format (e.g., a data table in the supplement). (ii) provide the data for individual subjects which is much more informative and allows to calculate all data for the groups). Most studies only report group means and mean time courses (and often not even errors on the data). (iii) provide minimum information on patient characteristics which includes basic anthropometric information like age, bodyweight, sex, height, and the subset of important lifestyle factors known to alter pharmacokinetics, e.g. co-medication, oral contraceptive use, smoking status, alcohol consum or for instance for CYP1A2 substrates like caffeine: methylxanthine consum/abstinence. (iv) Clearly state the study protocol: Which substance was given in which dose, in which route (oral, intravenous), and in which form (tablet, capsule, solution), the more specific the information the better.
As our analysis show, even many of the basic information is not stated in the publication making it impossible to integrate such data. An example is for instance codeine, where often not even the given dose can be retrieved because it is unclear if the dose in [mg] describes the dose of the given codeine(sulphate) or codeine(sulphate) or the actual codeine.

Curation examples via database

- {Wang1985} -> incorrect units identified
- {Seng2008} -> incorrect calculation of per body weight volumes
- check smoking outlier (document study and what is going on)
- Balogh {REF} & Harder {REF} (suspicious, but not obvious what went wrong; assay bias?)
- {Carbo1989} -> Individual No. 4 thalf, thalf very high; No. 3 clearance high
- {Beach1996} (9 smokers & 2 non-smokers) resulting in very high clearance
- {Wu2014} -> timecourse and cmax units ng/ml -> ng/µl

ACKNOWLEDGEMENTS

Funding: JK and MK are supported by the Federal Ministry of Education and Research (BMBF, Germany) within the research network Systems Medicine of the Liver (LiSyM, grant number 031L0054).

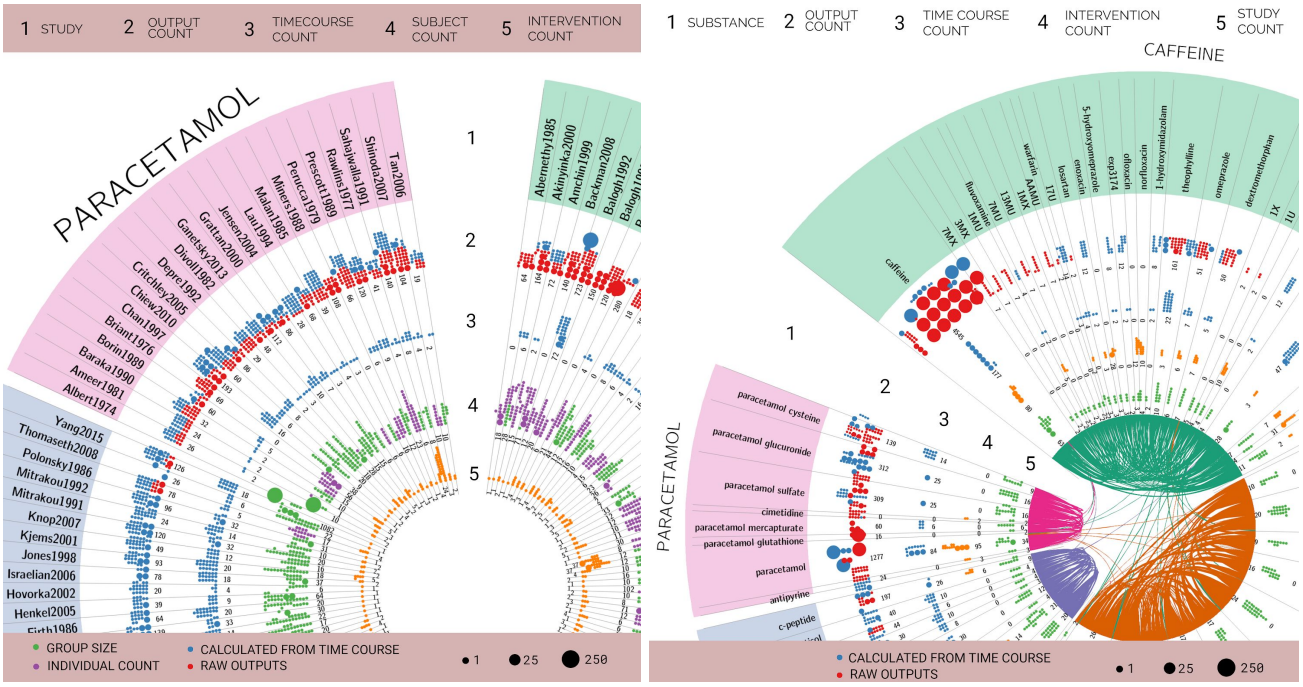
Conflict of Interest: none declared

REFERENCES

{Wang1985}
{Wilkinson2016}

FIGURES

Figure 1 - PK-DB database content



A) Study overview. The figure shows a fraction of the current studies in the database. The complete data corresponding is provided in SUPPLEMENT_TAB1 and visualized in SUPPLEMENT_FIG1. PKDB-v0.7.0 consists of 159 studies containing 424 groups, 1582 individuals, 482 interventions, 14067 outputs and 1105 time courses related to caffeine, glucose, codeine, and paracetamol. Dots represent reported data per study with dot size corresponding to number of entries with the rings listing the following information for the respective study (1) name of the study; (2) number of outputs (pharmacokinetics parameters and other measurements). Red dots represent reported data, blue dots data calculated from time courses reported in the study; (3) number of time courses; (4) number of participants. Purple dots represent participants with individual data, green dots represent participants with are reported as a group; (5) number of interventions applied to the participants in the study.

B) Substance overview. The figure shows a fraction the current substances in the database. Derived substances and substance with few entries are excluded from the plot.
The substances are classified into 5 classes (caffeine, glucose, codeine and paracetamol). The classification is performed by agglomerative clustering of the pair co-occurrence of substances within studies. The names for the classes are chosen to be the name of the most frequent substance within the class. Each co-occurrence of

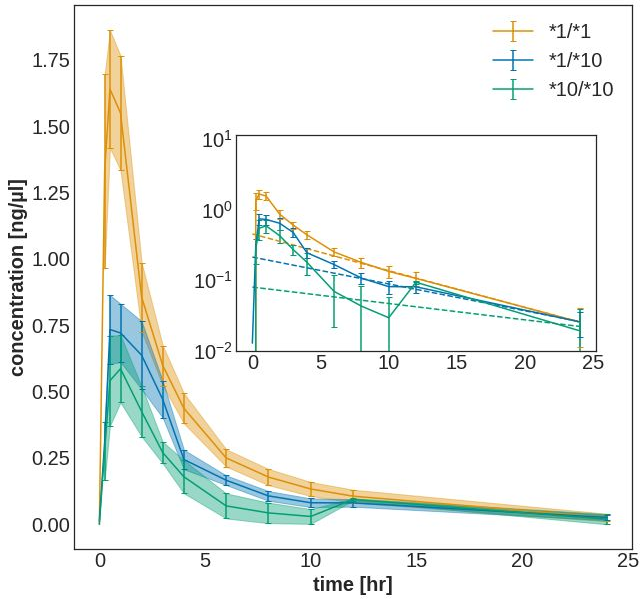
two substances is visualized as by a connecting ribbon (center of figure). The first outer ring (red) displays the number of substances used in outputs, the second ring (blue) displays the number of substances used in time course data, the third ring (green) diplayes the number studies in which the substance referenced (included outputs, time courses and interventions). the orange forth most inner ring displays the number of interactions in which the substance was used.
Dots represent reported data per study with dot size corresponding to number of entries with the rings listing the following information for the respective study (1) name of the substance; (2) number of outputs (pharmacokinetics parameters and other measurements). Red dots represent reported data and blue dots represent data calculated from reported time course; (3) number of time courses; (4) number of applied interventions; (4) number of studies for the given substance.

Figure 2. PK-DB workflow

Overview of steps and extracted information for given studies.

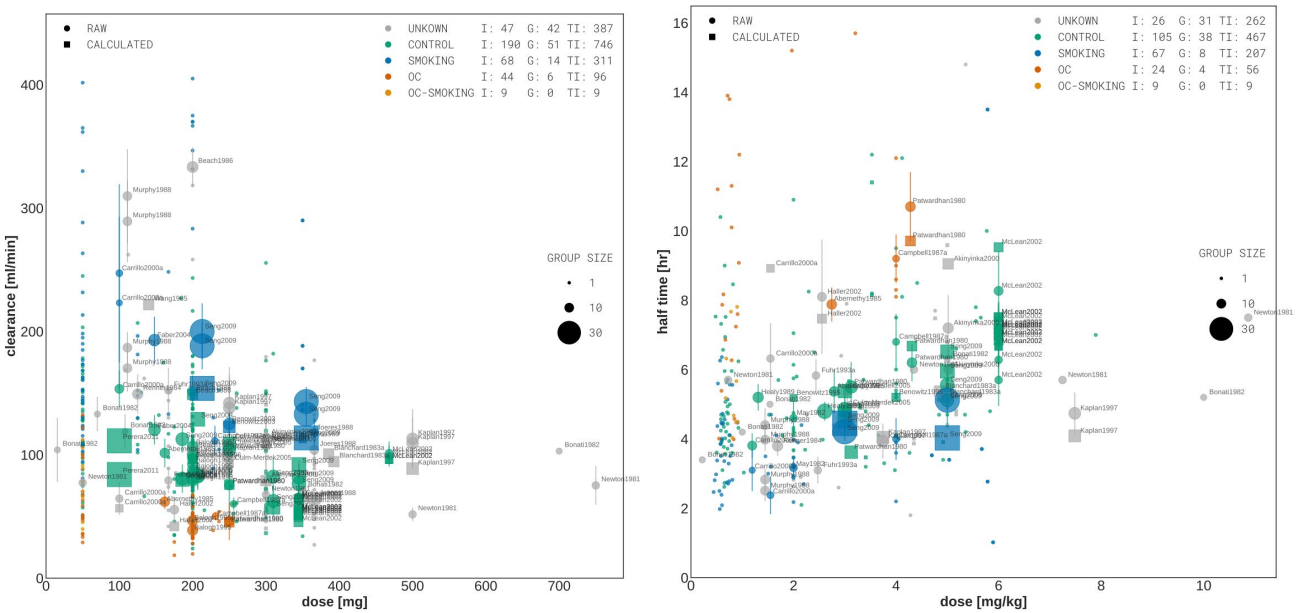
- curation
- validation
- unit normalization
- inference

Figure 3. Calculation of pharmacokinetics data from time courses.



Pharmacokinetics parameters are calculated from reported time courses using non-compartmentalized methods, e.g. here for data curated from {REF}. Information which can be calculated are among others clearance, AUC, half-life, volume of distribution (see Table 1 for the comparison of reported vs. the calculated values in PK-DB). Due to the unavailability of raw data in pharmacokinetics studies parameters are determined on mean time courses.

Figure 4. Caffeine meta-analysis using PK-DB



Meta-analysis of caffeine clearance rates and half-lifes depending on caffeine dose. Data is stratified based on reported smoking and oral contraceptive (OC) use. UNKNOWN (grey) data corresponds to unreported smoking and OC, CONTROL (green) are non-smokers, not taking OC, SMOKING (blue) are smokers and not taking OC, OC (dark orange) are non-smokers taking OC, and OC-SMOKING (light orange) to smokers taking oral contraceptives. Data is individual and group data with group size depicted as dot size. Reported data is depicted as circles, calculated data ... , inferred data ...

TABLES

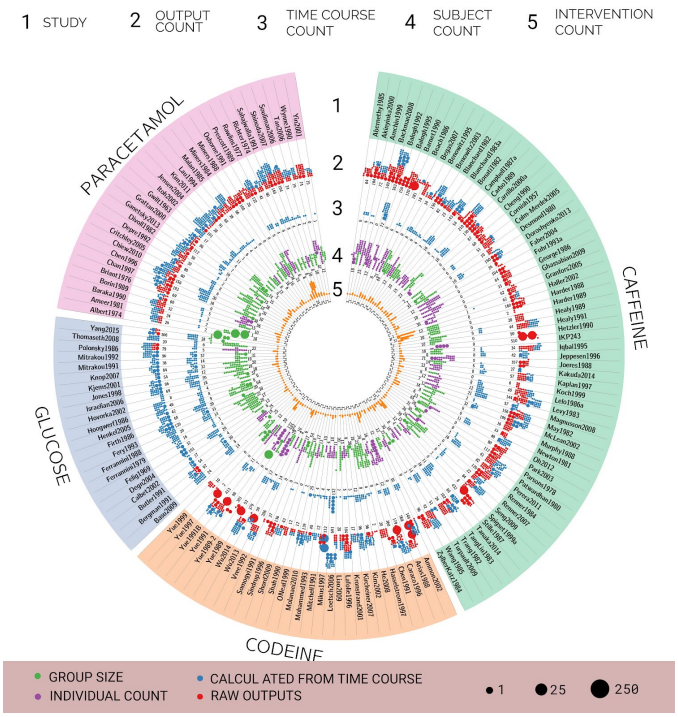
Table 1 - Comparison of calculated and reported pharmacokinetics parameters.

CYP_G	pktype	Unit	Raw	Calculated	Difference	Difference %
*1/*1	auc_end	µg/l*hr	6630.00 +- 2070.00	6240	390	5.88
	auc_inf	µg/l*hr	8520.00 +- 4100.00	-3650000	3658520	42940.38
	clearance	1/hr	5.08 +- 3.39	7.97	2.89	56.89
	cmax	ng/µl	2.06 +- 0.89	1.64	0.42	20.39
	kel	hr		0.12		
	thalf	hr	9.40 +- 11.70	5.8	3.6	38.3
	tmax	hr	0.64 +- 0.28	0.5	0.14	21.88
	vd	l		66.7		
*1/*10	auc_end	µg/l*hr	3770.00 +- 1930.00	3800	30	0.8
	auc_inf	µg/l*hr	5050.00 +- 3300.00	-11800000	11805050	233763.37
	clearance	1/hr	9.7 +- 8.72	12.3	2.58	26.54
	cmax	ng/µl	0.96 +- 0.42	0.73	0.23	23.54
	kel	hr		0.09		
	thalf	hr	11.50 +- 11.10	7.92	3.58	31.13
	tmax	hr	0.86 +- 0.52	0.5	0.36	41.86
	vd	l		141		
*10/*10	auc_end	µg/l*hr	2650.00 +- 1950.00	2730	80	3.02
	auc_inf	µg/l*hr	3260.00 +- 2430.00	-49300000	49303260	1512369.94
	clearance	1/hr	16.20 +- 12.30	20	3.8	23.46
	cmax	ng/µl	0.68 +- 0.50	0.59	0.09	13.82
	kel	hr		0.05		
	thalf	hr	6.84 +- 5.46	13	6.16	90.06
	tmax	hr	0.86 +- .52	1	0.14	16.28

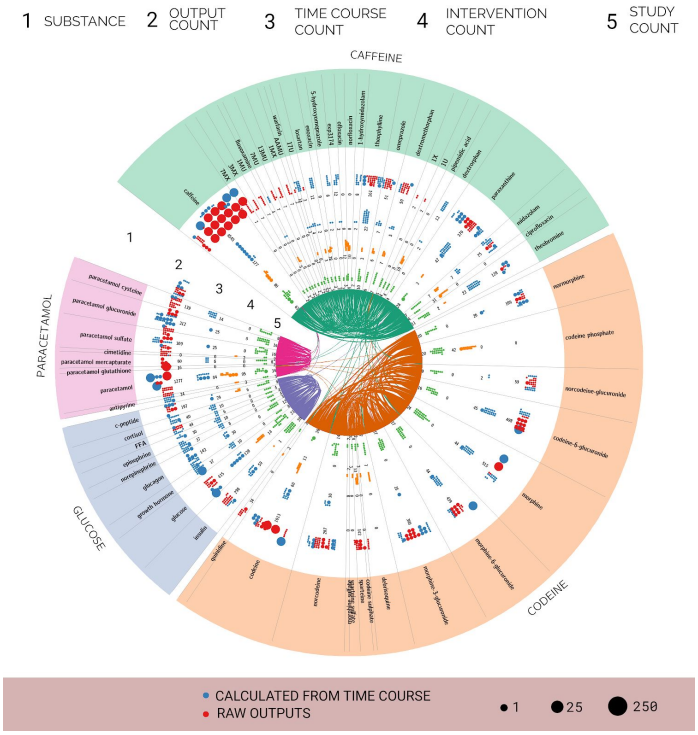
For one of studies {REF} we show the comparison of the reported vs. the PK-DB calculated values.

SUPPLEMENTS

Supplementary Figure 1 (S1) - PK-DB study overview



Supplementary Figure 2 (S2) - PK-DB substance overview



Supplementary Table 1 (S3) - Overview table content PK-DB (data underlying circos plots)

Supplementary Table 2 (S4) - Data underlying caffeine meta-analysis