|  |
| --- |
| *Database*  **PK-DB: a pharmacokinetics database for individualized and stratified computational modelling**  Jan Grzegorzewski1, Dimitra Eleftheriadou1, Janosch Brandhorst1, Kathleen Green2, and Matthias König1,\*  1Institute for Theoretical Biology, Humboldt University Berlin, Berlin, Germany  2Department of Biochemistry, Stellenbosch University, Stellenbosch, South Africa  \*Corresponding author  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

**ABSTRACT**

**Summary:** Standardized data representation is a prerequisite for data integration and use in computational modeling in systems biology and systems medicine approaches. But so far no open solution for the reproducible and reusable storage of pharmacokinetics data exists. We present PK-DB, a database for pharmacokinetics data and information from clinical trials as well as pre-clinical research. PK-DB provides curated pharmacokinetics data integrated with the corresponding meta-information (1) characteristics of studied patient collectives and individuals (e.g., age, bodyweight, smoking status); (2) applied interventions (e.g., dosing, substance, route of application); and (3) measured pharmacokinetics time courses and pharmacokinetics parameters (e.g., clearance, half-life, area under the curve). Important features are the representation of experimental errors and variation, the representation and normalisation of units, annotation of information to biological ontologies, calculation of pharmacokinetics information from time courses (e.g., apparent clearance, half-life, area under the curve (AUC)), a workflow for collaborative data curation, strong validation rules on data, and simple access via a REST API. PK-DB contains data curated from literature for more than 160 studies. A special focus of PK-DB is the curation of data and meta-data for individualized and stratified computational modelling. We demonstrate the value of PK-DB via a stratified meta-analysis of pharmacokinetics studies for caffeine curated from the literature, which allows us to integrate and standardize pharmacokinetics information from a wide range of studies and sources;

**Database available from:** <https://pk-db.com/>

**Latest source code:** <https://github.com/matthiaskoenig/pkdb>

**Archive source code as at the time of publication:** [10.5281/zenodo.1406979](https://doi.org/10.5281/zenodo.1406979)

**License:** GNU General Public License, version 3 ([GPL-3.0](https://opensource.org/licenses/GPL-3.0))

**Contact:** [konigmatt@googlemail.com](mailto:konigmatt@googlemail.com)

# **INTRODUCTION**

**Pharmacokinetics and Data** A central medical field is the study of pharmacokinetics (PK) of drugs and medication. It describes the effects of the organism on an administered substance. Concentration-time profiles and corresponding PK parameters like elimination half life or clearance rates are the main measures. High inter-individual variability drive the need for individual and population related information such as age, weight, sex or disease status {Thomas2018}. The study of variabilitly in drug exposure due to these covriates is generaly refered to as population pharamcokinetics {Aarons1991}. More modern approaches go beyond classical population information by accounting for genetic variation {Loetsch2006}. This kind of subject information in combination with the main measures are the foundation for an individual approach in drug treatment and dynamical liver function tests (precision dosing/medicine).

A multitude of publications have been written about pharmacokinetics studies but 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 machine readable format and certainly not with FAIR principles in mind. The only way to retrieve this treasure is by digitizing and curating this information from publications. 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. Heterogeneity in clinical study designs, pharmacokinetic measures and individual and population related meta-informations hinder data integration. Currently a data model and database is missing that could represent this data. For computational modelling, meta-analysis and most methods in machine learning and artificial intelligence a standardized and machine readable representation of the data is of major importance {REF}.

Physiological Based Pharamacokinietc Models (PBPK) open a unique opportunity to translate data from one clinical trial to an other. The mechanistic models can accounting for the differences in study protocol, dosing and individuals and population characteristica.

Here we present PK-DB a database for pharmacokinetics information with special focus on storing information for individualization and stratification of computational models.

# **DESCRIPTION & RESULTS**

## 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 or measurement type. Additional characteristics 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. With an intuitive curration work flow.

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 time course 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 version control based ongit. 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 current 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 3. After an initial literature research selecting a corpus of papers to curate (reported data, useful data, minimal quality criteria). Within the curation process the relevant 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 programmatic 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 150 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 {Balogh1992} 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 & DISCUSSION

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 shows, 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.

Naming is not consistend, reporting often incomplete {Kanji2015}. Standards exit but most reports fail the standards and no standard advocates open data with individual participant data (timecourse data for each meausred individual). Mean (median) and standard deviations are not the raw data. Assumption of identically and independently distriubuted samples do not hold {Hanin2017} and oppertunities for data pooling are with data integration are less possible.

Curation examples via database

* {Wang1985} -> incorrect units identified
* {Seng2009} -> incorrect calculation of per body weight volumes
* Balogh {Balogh1992} & 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

*Contributions:*

# REFERENCES

* {AARONS1991}
* {Balogh1992} Balogh, A.; Harder, S.; Vollandt, R. & Staib, A. H. **Intra-individual variability of caffeine elimination in healthy subjects.** International journal of clinical pharmacology, therapy, and toxicology, 1992, 30, 383-387
* {Benet1984} LZ. Pharmacokinetic parameters: which are necessary to define a drug substance? Eur J Respir Dis Suppl. 1984;134 45-61. PMID: 6586486.
* {Carbo1989} Carbó, M.; Segura, J.; De la Torre, R.; Badenas, J. M. & Camí, J. **Effect of quinolones on caffeine disposition.** Clinical pharmacology and therapeutics, 1989, 45, 234-240
* {Hanin2017} Hanin, L. (2017). Why statistical inference from clinical trials is likely to generate false and irreproducible results. BMC Medical Research Methodology, 17(1). https://doi.org/10.1186/s12874-017-0399-0
* {Ioannidis2019} Ioannidis, J. P. A. **Reproducible pharmacokinetics.** Journal of pharmacokinetics and pharmacodynamics, 2019, 46, 111-116
* {Kanji2015} Kanji, S., Hayes, M., Ling, A., Shamseer, L., Chant, C., Edwards, D. J., … Moher, D. (2015). Reporting Guidelines for Clinical Pharmacokinetic Studies: The ClinPK Statement. Clinical Pharmacokinetics, 54(7), 783–795. https://doi.org/10.1007/s40262-015-0236-8
* {Seng2009} Seng, K.-Y.; Fun, C.-Y.; Law, Y.-L.; Lim, W.-M.; Fan, W. & Lim, C.-L. **Population pharmacokinetics of caffeine in healthy male adults using mixed-effects models.** Journal of clinical pharmacy and therapeutics, 2009, 34, 103-114
* {Thomas2018} Thomas M. Polasek, Sepehr Shakib & Amin Rostami-Hodjegan (2018) **Precision dosing in clinical medicine: present and future, Expert Review of Clinical Pharmacology**, 11:8, 743-746,
* {Wang1985} Wang, T.; Kleber, G.; Stellaard, F. & Paumgartner, G.**Caffeine elimination: a test of liver function.** Klinische Wochenschrift, 1985, 63, 1124-1128
* {Wilkinson2016} Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J. & Mons, B. **The FAIR Guiding Principles for scientific data management and stewardship.**Scientific data, 2016, 3, 160018
* {Wu2014}

# **FIGURES**

## Figure 1 - PK-DB database content

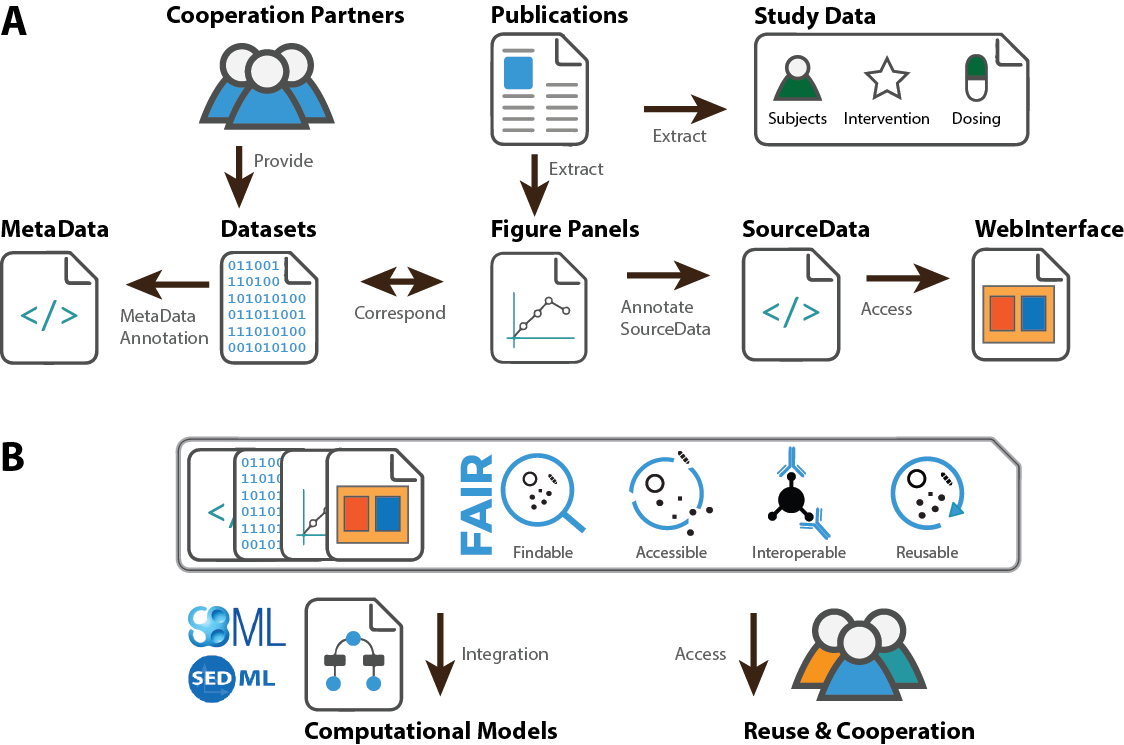
## ￼

A) Study overview. The figure shows a fraction of the current studies in the database. The complete corresponding data 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.

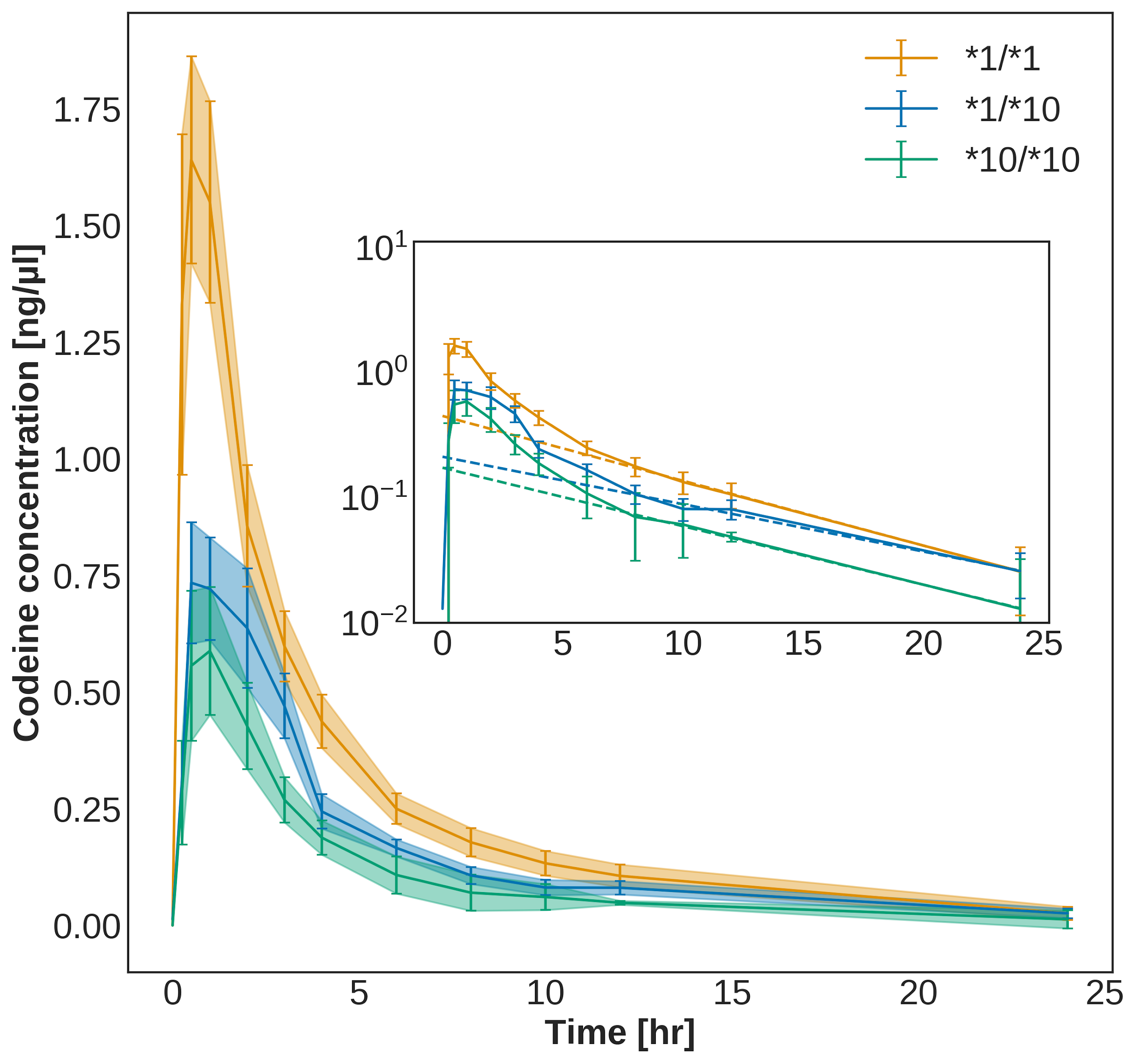
The substances are classified into 5 classes (caffeine, glucose, codeine and paracetamol). The classification is performed by agglomerative clustering of the pair co-occurrence of substances within studies. The names for the classes are chosen to be the name of the most frequent substance within the class. Each co-occurrence of two substances is visualized by a connecting ribbon (center of figure). (2) The first outer ring (red,blue) displays the number of substances used in outputs (blue dots are calculted from conenctration-time profiles), (3) the second ring (blue) displays the number of substances used in time course data, (4) the third ring (orange) diplayes the number internvetions in which the study was administrted and (5most inner ring displays the number of studies in which the substance was referenced (included outputs, time courses and interventions). 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



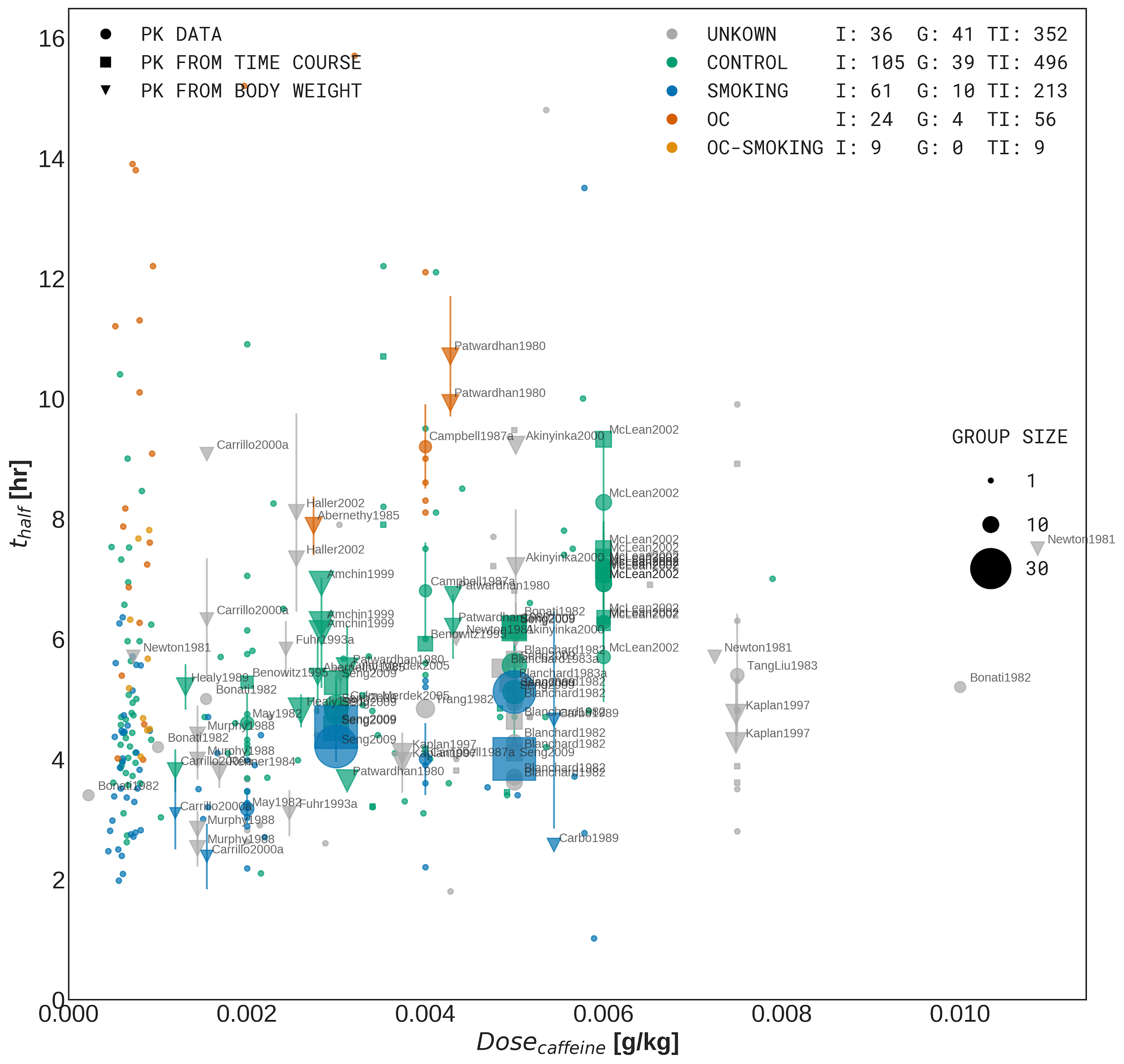
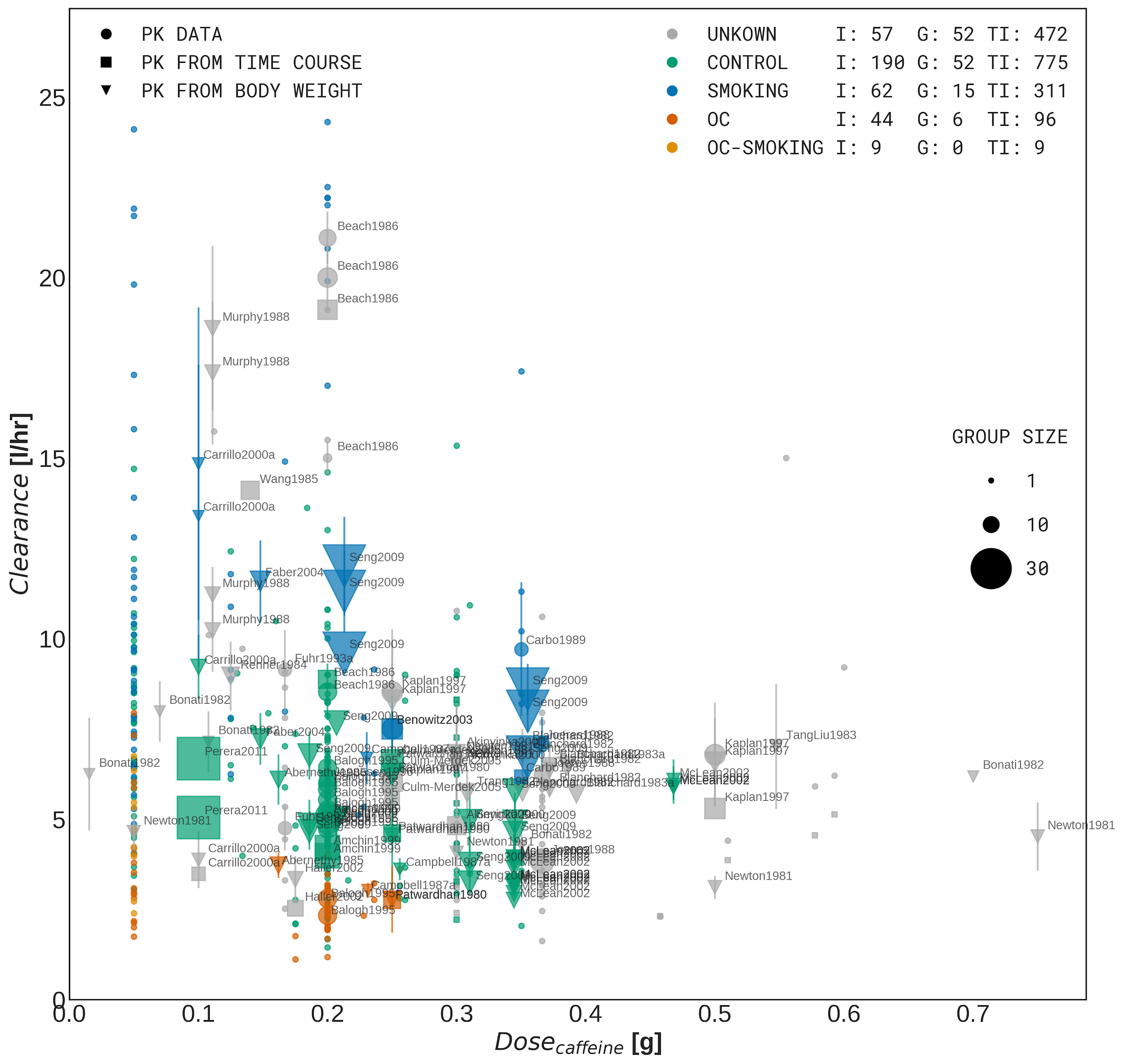
Overview of steps and extracted information for given studies.

## Figure 3. Calculation of pharmacokinetics data from time courses.



Pharmacokinetics parameters are calculated from reported concentration-time profiles using non-compartmentalized methods, e.g. here for data curated from {Wu2014}. 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 ofCochraneDatabase of Systematic Reviews individual participant data in pharmacokinetics studies parameters are determined on the 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. Individual and group data with group size depicted as dot size is displayed. In the legend (I) stands for number of Individuals , (G) for number of groups and (TI) Total participants reported either as individuals or group members . Data is depicted as circles if originally reported in the study, as squares if calculated from concentration-time profiles and as triangles if inferenced from corresponding pharmacokineitic data and body weights. Typically dosing and pharmacokineitic parameters are reported normal units or in units per bodyweight.

## Table 1 - Comparison of calculated and reported pharmacokinetics parameters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Measurement Type** | **CYP G** | **Original** | **Calculated** | **Difference** | **Difference %** |
|  | Genotype CYP2D6 |  |  |  |  |
|  | \*1/\*1 | 6.63 ± 2.07 | 6.24 | 0.39 | 5.88 |
| \*1/\*10 | 3.77 ± 1.93 | 3.8 | -0.03 | -0.8 |
| \*10/\*10 | 2.65 ± 1.95 | 2.59 | 0.06 | 2.26 |
|  | \*1/\*1 | 8.52 ± 4.10 | 6.4 | 2.12 | 24.88 |
| \*1/\*10 | 5.05 ± 3.30 | 3.97 | 1.08 | 21.39 |
| \*10/\*10 | 3.26 ± 2.43 | 2.68 | 0.58 | 17.79 |
|  | \*1/\*1 | 5.08 ± 3.39 | 5.24 | -0.16 | -3.15 |
| \*1/\*10 | 9.72 ± 8.72 | 8.02 | 1.7 | 17.49 |
| \*10/\*10 | 16.20 ± 12.30 | 13.1 | 3.1 | 19.14 |
|  | \*1/\*1 | 2.06 ± 0.89 | 1.64 | 0.42 | 20.39 |
| \*1/\*10 | 0.96 ± 0.42 | 0.73 | 0.23 | 23.96 |
| \*10/\*10 | 0.68 ± 0.50 | 0.59 | 0.09 | 13.24 |
|  | \*1/\*1 | 9.40 ± 11.70 | 4.15 | 5.25 | 55.85 |
| \*1/\*10 | 11.50 ± 11.10 | 4.76 | 6.74 | 58.61 |
| \*10/\*10 | 6.84 ± 5.46 | 4.66 | 2.18 | 31.87 |
|  | \*1/\*1 | 0.64 ± 0.28 | 0.5 | 0.14 | 21.88 |
| \*1/\*10 | 0.86 ± 0.52 | 0.5 | 0.36 | 41.86 |
| \*10/\*10 | 0.86 ± 0.52 | 1 | -0.14 | -16.28 |
|  | \*1/\*1 | NR | 31.3 | NR | NR |
| \*1/\*10 | NR | 55 | NR | NR |
| \*10/\*10 | NR | 88.1 | NR | NR |
|  | \*1/\*1 | NR | 0.17 | NR | NR |
| \*1/\*10 | NR | 0.15 | NR | NR |
| \*10/\*10 | NR | 0.15 | NR | NR |

For one of studies {Wu2014} 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