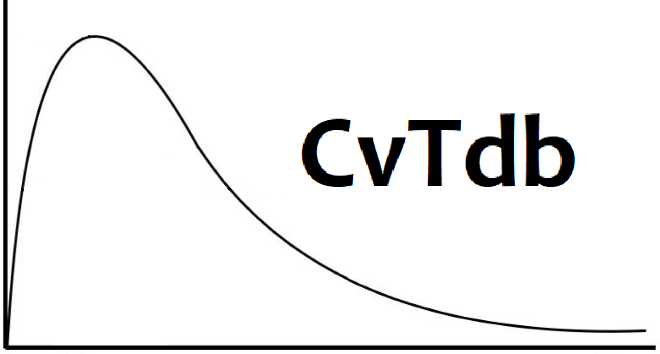
# CvTdb Data Curation Standard Operating Procedure



## Introduction

CvTdb is a database of toxicokinetic tissue concentration vs. time (CvT) data extracted from the scientific literature. The database is described in Sayre et al. (2020)[[1]](#footnote-1), which is available at: <https://doi.org/10.1038/s41597-020-0455-1>

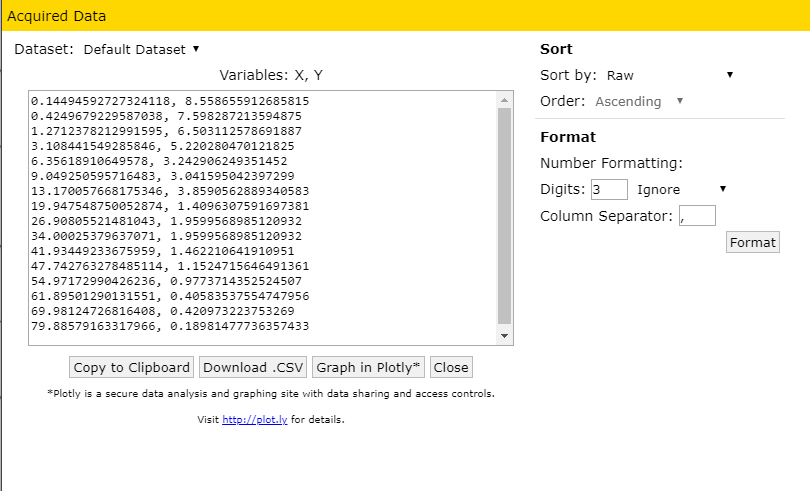
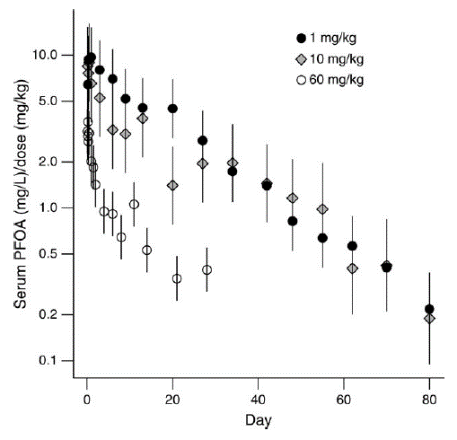


Figure : The CvTdb curation process focuses on extracting chemical concentration vs. time data from peer-reviewed papers and annotating the experiments that led to those data. We often use a tool ([WebPlotDigitizer](https://automeris.io/WebPlotDigitizer/)) to manually extract the data from figures in a paper.

As illustrated in Figure 1, the CvT data curation process is an attempt to extract CvT data from the peer-reviewed scientific literature. These data may be available as supplemental tables, tables in the paper, or, in many cases, only as graphs within the papers. We store the data and document the conditions for each experiment within an Excel file that we call the “CvT Data Template”.

We are primarily interested in organic compounds, including pesticides, industrial chemicals, pharmaceuticals, and any other chemical likely to be found in commerce or the environment. These data will be analyzed statistically with an EPA tool to determine basic pharmacokinetic properties such as volume of distribution and half-life[[2]](#footnote-2). These data also allow for statistical evaluation of generic mathematical models of TK that can be parameterized using *in vitro* data[[3]](#footnote-3), as is illustrated in Figure 2. The uncertainties and biases estimated using *in vivo* data within the CvTdb can be extrapolated to chemicals without *in vivo* data.

Predicted Concentrations

Concentrations from CvTdb

x

x

x

x

x

x

x

y

y

y

y

y

y

y

z

z

z

z

z

z

Figure : We can use in vivo CvT data to evaluate the statistical performance of generic TK models based upon in vitro data, allowing us to considering using those methods for chemicals without in vivo data. (Figure modified from Cohen Hubal, et al., 2018[[4]](#footnote-4))

PubMed ID’s (PMID) are the way we identify papers – all data and other information are linked together with the PMID’s. You can access PubMed at <https://www.ncbi.nlm.nih.gov/pubmed>

The papers for curation have been identified through keyword searches – it is always possible that a given paper does not contain CvT data. We only care about papers that have extractable concentration vs. time data. Therefore, **start with Concentration vs. time values – if you cannot get CvT data then we do not need the meta data describing the experiments**. These data could be multiple series on one graph (different doses, genders, etc.) or different graphs, tables, and/or supplemental material.

There are three ways to get the data:

1. Table in paper
2. Supplemental material
3. Extract from graph in paper (with [WebPlotDigitizer](https://automeris.io/WebPlotDigitizer/))

To keep things simple, we fill out one template file per article.

## CvTdb Definitions

The original inspiration for the set of contextualizing metadata was based on the test guideline for metabolism and pharmacokinetics released by the U.S. EPA Office of Prevention, Pesticides, and Toxic Substances (OPPTS)[[5]](#footnote-5). Identifying the set of parameters necessary to properly annotate the extracted ***CvT*** data was iterative, with improvements evolving after the review of multiple publications (as the consistency and reporting of study details is highly variable in literature). For each paper five linked tables (“Sheets” in the Data Template “Workbook”) must be filled out: “Documents”, “Subjects”, ‘Studies”, “Series”, and “Conc Time Values”.

* A study is the treatment done to the animal.
* A subject is a description of the animal
* A series is a single consistent measurement done over time when a study is applied to a subject.

Each table contains multiple “fields” (columns in the table) which contain the values that describe the experiment. Each field is labeled (for example “analyte\_name” or “dose\_level\_original\_units”). If a field refers to a column in a different table, it is described as a “foreign key” as the prefix “fk” is used (for example “fk\_study\_id” or “fk\_series\_id”).

### Documents

This sheet identifies the reference (data source document). PMID is sufficient as an id.

### Studies

A study is generally equivalent to a single dosing strategy for a test substance. Every different treatment regime (**including dose amount, for example 1 mg/kg versus 10 mg/kg)** is a different study.  If the test substance is different or the dosing regimen changes in any way, it is considered a new study.

The study describes the scenario: test substance, administration route, dose amount, vehicle, and volume, exposure duration, quantity of doses given and their spacing, number of subjects per treatment group, number of treatment groups, fasting status of subjects.

The series linked to a study will document all the different measurements that were taken using a particular treatment regime (like different measured analytes, measurements from different tissues, results from different subjects like wt vs compromised rats).

Since this is a concentration vs time database, we do not currently phenotypic changes like body weight.

### Subjects

The test subject details: species, type/strain, sex, age, age category, size, and any other description given by source.

### Series

Series correspond to each line in a PK graph. If everything is the same except the times and concentrations, then all the data are in the same series. Series includes measurement details (the original time and concentration units, analyte, the medium (tissue, circulatory fluid, etc.) in which the analyte was detected and measurement methods (the limits of detection and/or quantification, the analysis method).

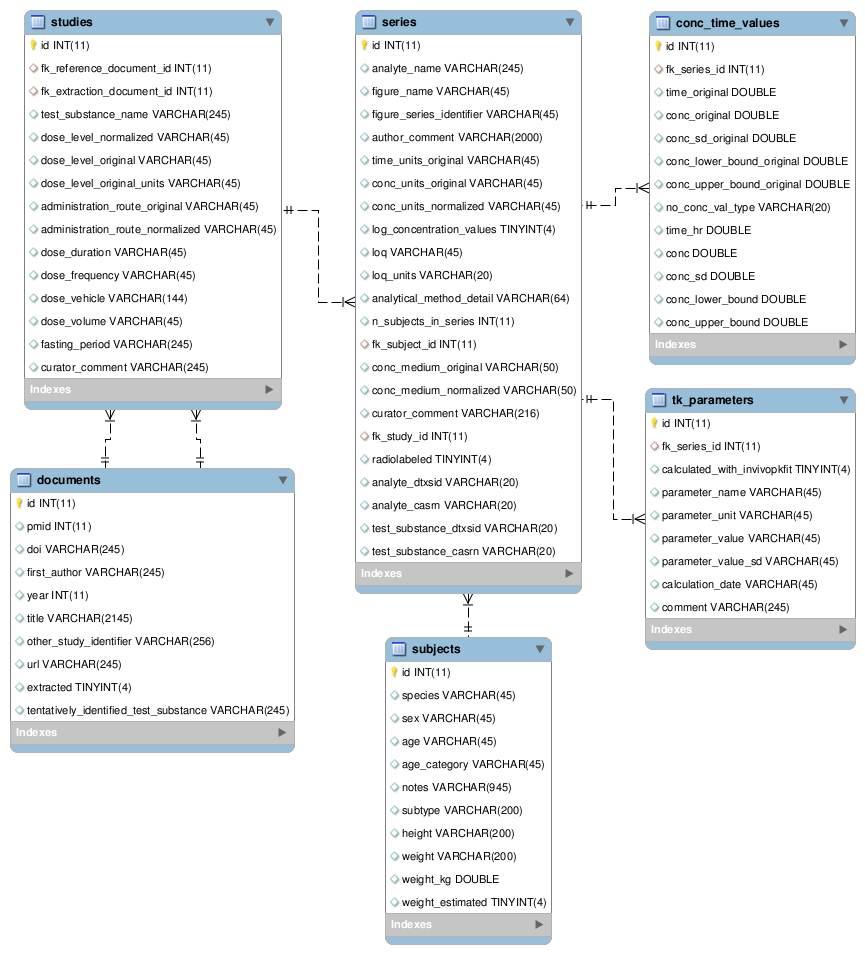
### Conc\_Time\_Values

These are the actual measured values in each series. Each measurement is a pair, with a time and a concentration. Right now, we are not concerned with the standard deviation. If the number is given in a table and can be easily copied, then please do capture it. However, it is currently not worth the extra effort of extracting standard deviations from graphs. **Please do record time points that are non-detects, the curve-fitting software used will make use of these points.** You can enter non-detects as “ND”. Please record the limit of detection (LOD) or limit of quantitation (LOQ) on the Series sheet if available. If not available, leave blank and the curve-fitting software will infer a LOQ.

### Curation Notes

Curator notes are free text fields for noting any assumptions made during the collection process and any additional data. They are meant to be very brief, for example: “Vmax an Km available” or “post-natal data omitted”.

### How the CvTdb Terms Relate



Entity-relationship diagram of CvTdb.

## Curation Process

Visit the CvTdb Data Curation spreadsheet:

<https://docs.google.com/spreadsheets/d/1Uzxw7p_6zlNehtGJlBP-AkTWvydFvt9CeCTtwyeewvE/edit?usp=sharing>

We try to provide most of the necessary links at the top of the curation spreadsheet. When you click on them, they do not immediately open in a web-browser. You can use CTRL-C to copy and CRTL-V to paste the link into a web browser, or click the little button to open URL in browser (it's a small square with an arrow pointing to the upper right)

Select a paper by typing in your name in the “Curator 1” or “Curator 2” column and look up the paper for the PMID on that row. (Our long-range goal is to have each paper separately reviewed twice.)

Obtain the PDF corresponding to the PMID.

Priority PDFs are located here:

<https://www.dropbox.com/sh/o2mt6l3te45fko6/AABsqWSagygml8In-8PEx-2ca?dl=0>

An EndNote library file with many more PDF's is here:

<https://usepa-my.sharepoint.com/:f:/g/personal/beeker_jonathan_epa_gov/EqkrXI5MLtVAq8E8j7v-HIgBpcTCAC79ji7nW4D4eItA-g?e=79hxr0>

Ascertain if there is extractable CvT data in the paper. All the PMIDs were identified by searches that may not always find useful papers. If the paper is not useful, mention that in the note and mark it as done.

If the CvT data is present in table in the paper, then great! If the data is in supplemental information, try to obtain the supplemental information. You can only access most journals when connected to the VPN. You can always email Chris, John, or use the Inter-Library loan if you can't access the paper:

<https://epa.illiad.oclc.org/illiad/>

If the data is only available in graphs, it's still a useful paper, we just have to use [WebPlotDigitizer](https://automeris.io/WebPlotDigitizer/) to extract the data. Follow this link: <https://automeris.io/WebPlotDigitizer/>

Whatever the form of the CvT data, if it's present, then you need to fill out a CvT data template (which is an Excel workbook).

### Downloading the CvT Data Template

[**https://github.com/USEPA/CompTox-PK-CvTdb/blob/master/CvT\_data\_template\_articles.xlsx**](https://github.com/USEPA/CompTox-PK-CvTdb/blob/master/CvT_data_template_articles.xlsx)

1. From the Curation Click the small button to open a URL in browser (it's a small square with an arrow pointing to the upper right).
2. On the GitHub page for the template click the Download button (it's on the lower right next to History)
3. Save the file -- it helps to add the PMID and your initials to the file name.

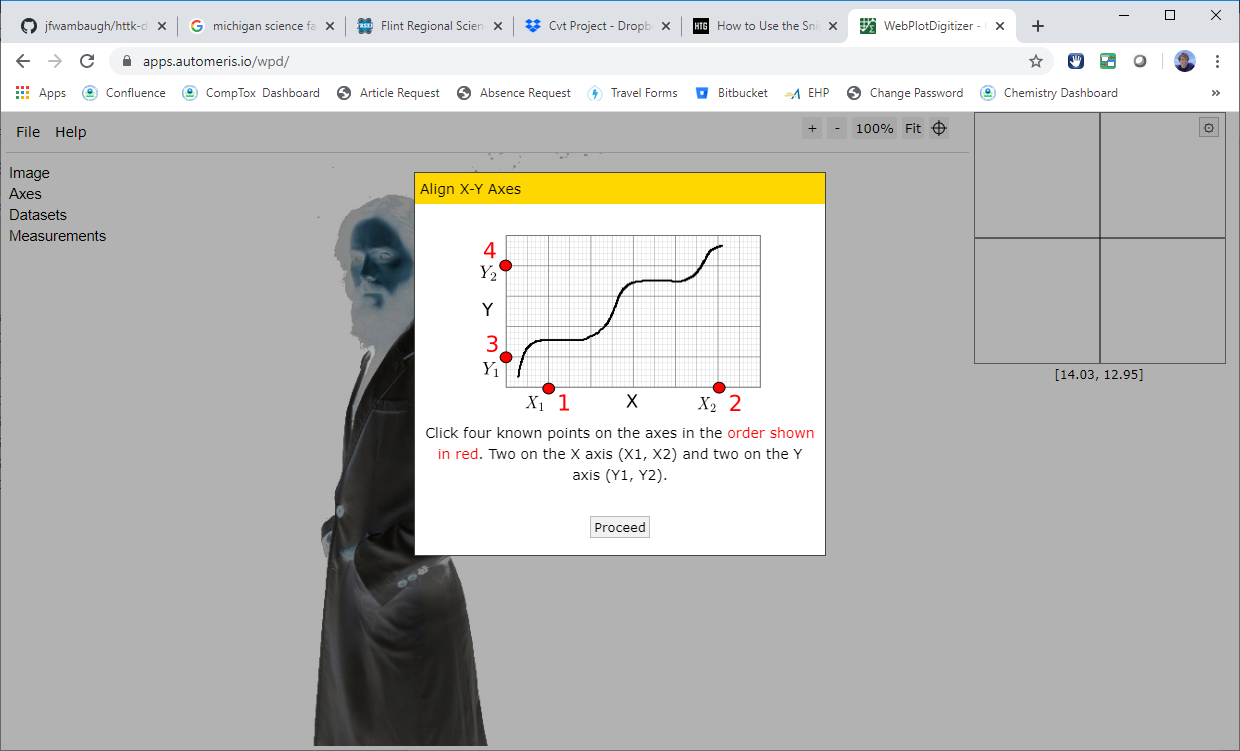
### Using WebPlotDigitizer

WebPlotDigitzer is data extraction software for converting graphs to x-y coordinates. We use the software to convert concentration vs. time graphs into sets of numbers where the first number (the x-axis) is time and the second number (y-axis) is concentration. You can run WebPlotDigitizer through your web browser or have it downloaded and installed on your computer. WebPlotDigitizer is available here:

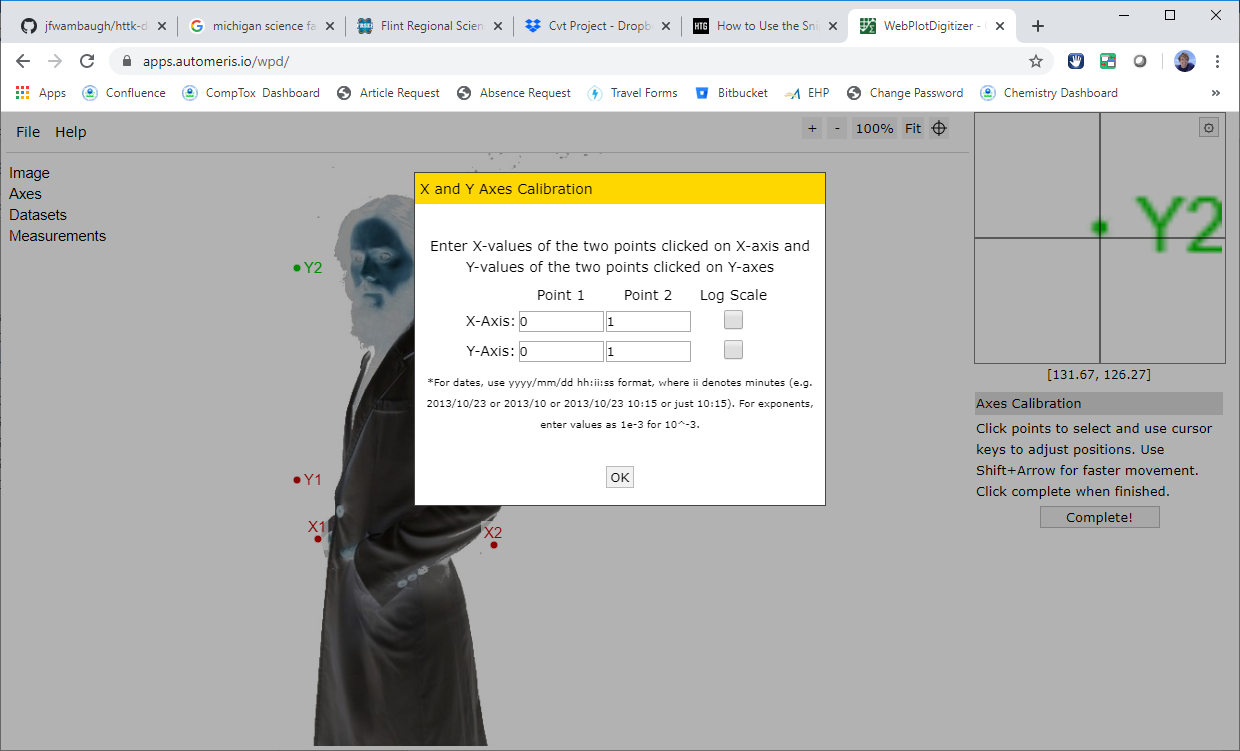
<https://automeris.io/WebPlotDigitizer/>

Click “Launch Now” to run the web-browser version.

1. Create an Image File for a Graph:Each figure containing data must be extracted into a PNG image file (for example MYGRAPH.png). There are many options to do this, including pressing the “Print Screen” or “PrtScr” button on your keyboard when the graph is on the screen. However, one of the better options on Windows 10 is the "Snipping Tool". See: <https://www.howtogeek.com/207754/how-to-use-the-snipping-tool-in-windows-to-take-screenshots/> Save the image somewhere on your computer that you can find it (like a folder named Graphs on your Desktop).
2. Load the Image File into the Data Extractor:Open WebPlotDigitizer and the load image file using the File Menu. Go to “File” then select “Load Image”. Click “Choose File” and select the file you created in the previous step. Click on “Load”. You will be asked to “Choose Plot Type” -- "2D (X-Y) Plot" is the default and you should leave it as that.
3. Calibrate the Axes:Click on “Align Axes”. You will see these instructions:

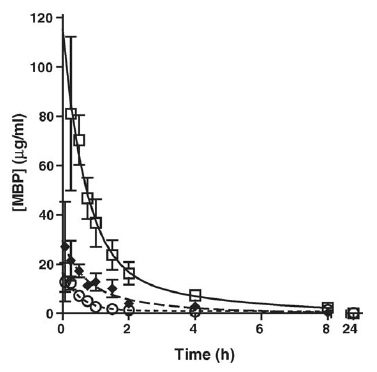


Click “Proceed” to mark two points on the x-axis first, then the y-axis. The farther apart the two points are, the better, but wherever you click, you need to be very sure what the value of the axis is at that point. Click on the necessary points on the image and then select "Complete" on the right side of the interface. You will then see this prompt:



Input the value of the x- and y-values for the points that you clicked. Then select OK.

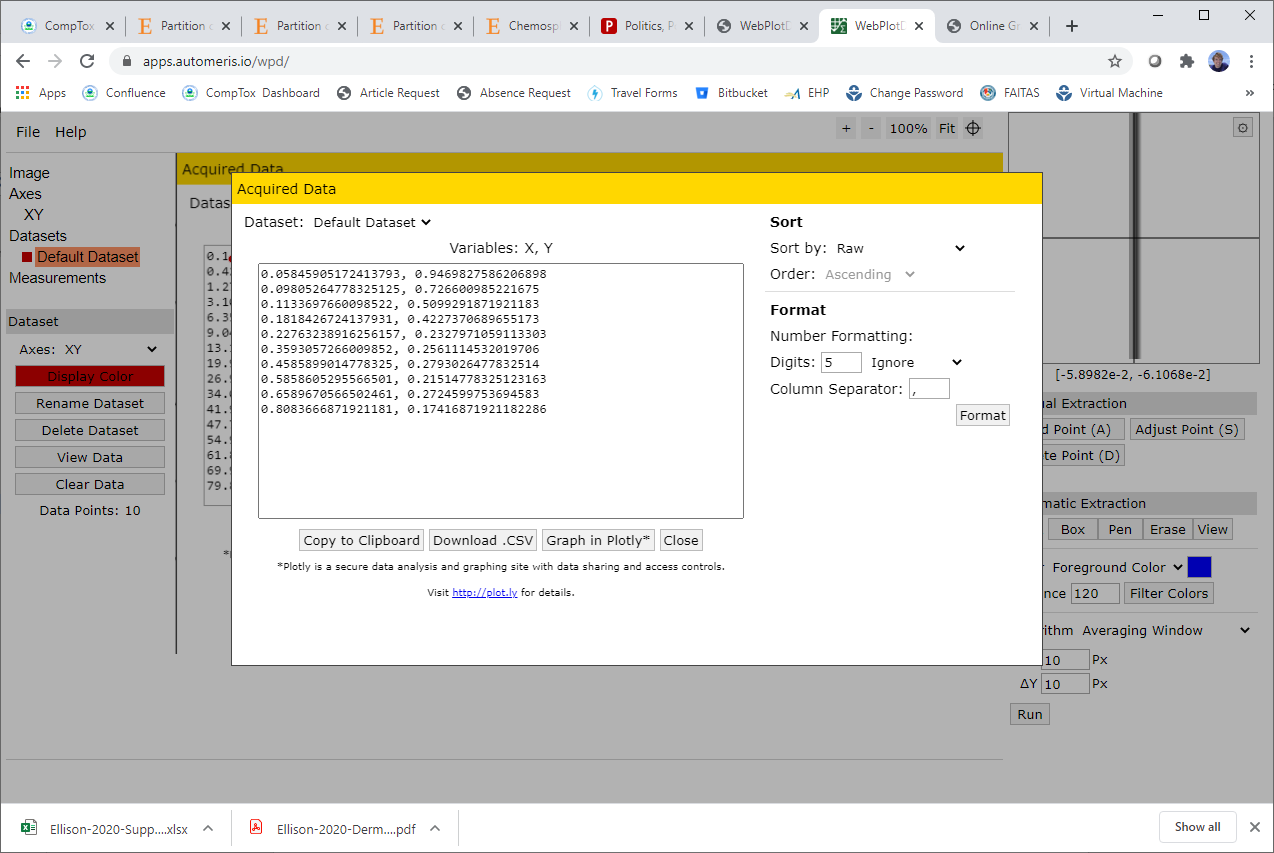
#### What to Do If There is a Break in the Axis

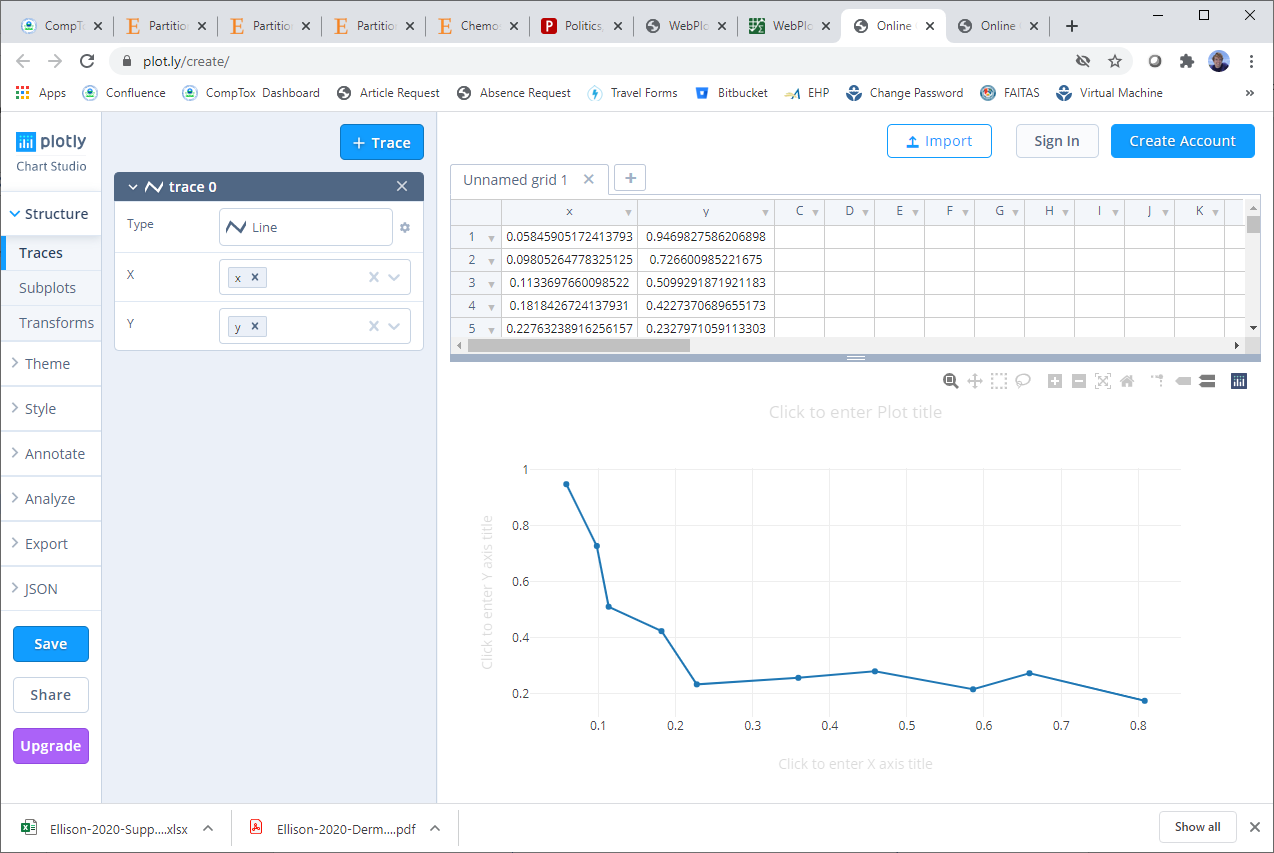


When there is a break in the axis you can try two different things. In this case, since there is just one time point, you might just click the 24 hour points and then edit the *x*-values to be “24” instead of whatever the digitizer tool comes up with.

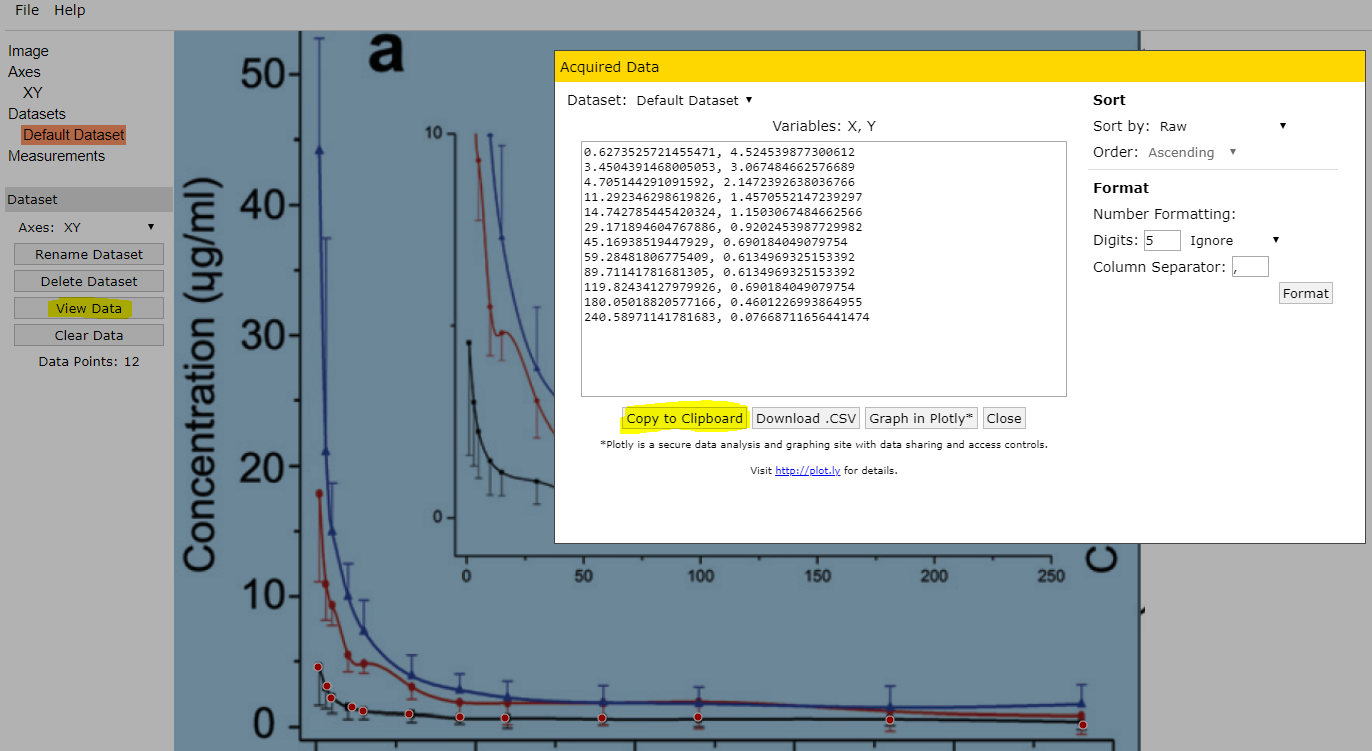
If there were many points after the break then it is probably easiest to digitize it twice – once for the part of the graph before the break, and once for the part after. When you calibrate the x-axis only do one part of the graph at a time. Then paste both values together into the template.

1. Extract Data Series: Once calibrated, the WebPlotDigitizer automatically creates a default dataset. Each dataset in WebPlotDigitizer corresponds to a “Series” in the CvTdb. There might be multiple series in one graph (such and males and females, or different doses). Each series should have a separate dataset in WebPlotDigitizer. When a dataset is selected in the upper-left portion of the interface, a "Manual Extraction" panel appears on the right side of the interface with "Add Point" selected by default. This allows points to be added to the dataset by clicking on them in your image. These points can then be adjusted to be as close to accurate using the "Adjust Point" function and using the keyboard arrow and viewing the image in the zoomed in window in the top right corner of the interface. Once all data points for a single series are clicked and adjusted,
2. Check Data: You can visualize the data that you have extracted by first selecting/clicking on ”View Data” in the left and then selecting/clicking on “Graph in Plotly” from the buttons underneath the data (labelled “Variables: X, Y”. After a moment a graph of the data will appear in a second window.

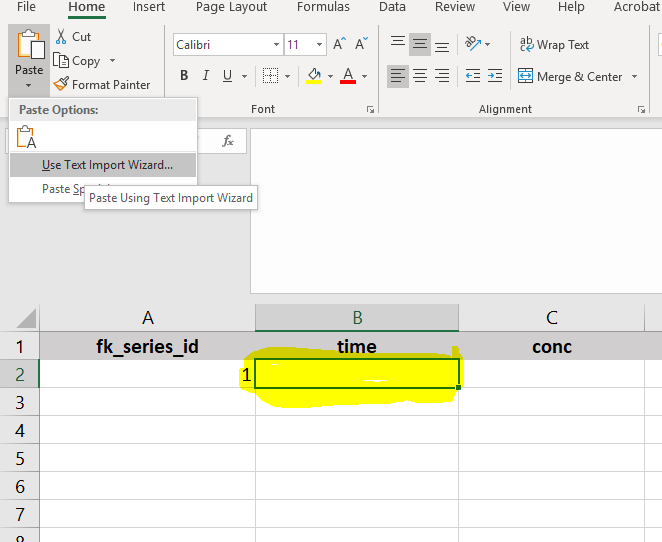




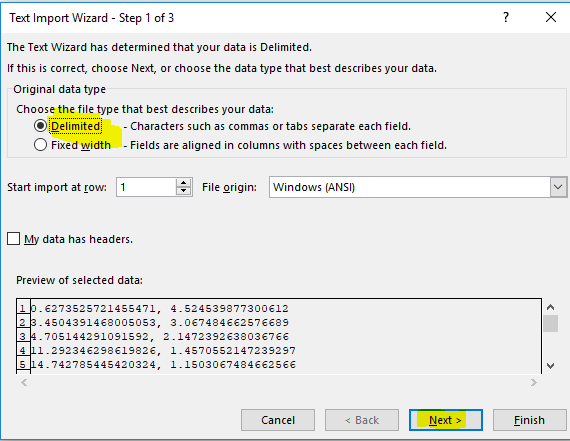
1. Copy Data from WebPlotDigitizer to Data Template: Once you have selected and adjusted all the datapoints in WebPlotDigitizer, you click “View Data” in the left panel and then “Copy to Clipboard” in the resulting pop-up:

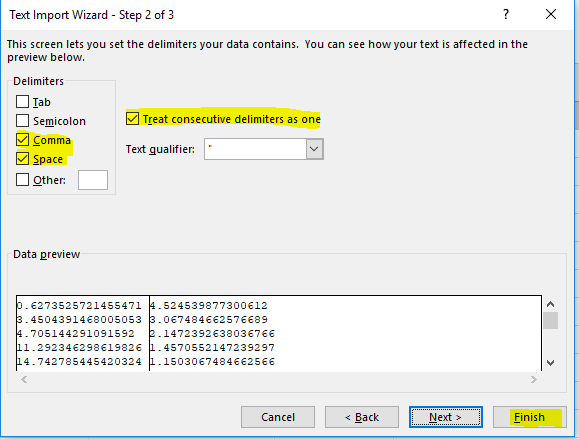


The content is then in your clipboard and you go back to your spreadsheet and put your cursor where the data should be then use the “Paste>Test Import Wizard” feature.

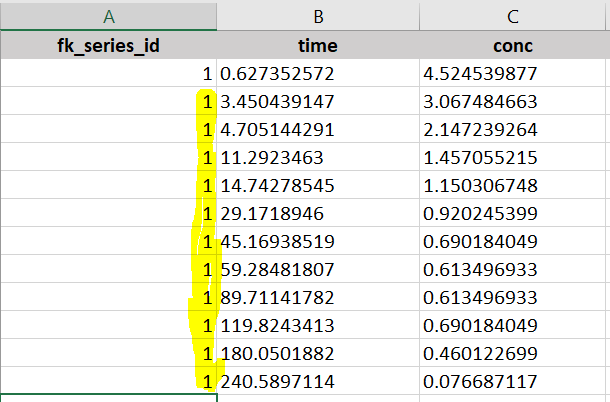


That will take you into the wizard where you should select “Delimited”, click next, then both “comma” and “space” along with “Treat consecutive delimiters as one” and then click Finish.





The data should appear as desired in the worksheet and you just need to put in the right fk\_series\_id for each row.



### Time-Zero Points

If a value (or plot point) is given for time zero, please do record it.

### Non-Detects

Please do record time points where the concentration is below the limit of detection (LOD) or limit of quantitation (LOQ). The toxicokinetic curve-fitting software that we use will make use of these points. IF the non-detect points are indicated on the graph, then click on the x-axis at each time point. Once you have exported the data you can change non-detects to “ND”. If the time points are described in the text but not graphed, manually add rows with the times and “ND” to the Conc\_Time\_Values table. Please make sure to record the limit of detection (LOD) or limit of quantitation (LOQ) on the Series sheet if available. If not available, leave the LOQ blank and the curve-fitting software will infer a LOQ.

### Secondary Data

In some cases (especially modeling papers and review articles), the paper being examined is a “secondary source” for data. That is, the data originally appeared in a different paper that we describe as the “primary source”. In that case, please make a reasonable effort to obtain the primary data, for instance through Inter-Library Loan (<https://epa.illiad.oclc.org/illiad/>). One exception would be a case where the secondary paper has already done extraction/curation and has made the CvT data available as a table. Then it makes sense to use their table as is, noting the primary references where possible.

In general, if you must digitize the data then it makes more sense to digitize the original figures in the primary source. However, if the primary source is very old and cannot be obtained (not having a PMID is a bad sign) or the secondary source indicates that the data were obtained from the original author but not from the paper, it is perfectly acceptable to digitize the data in the secondary source article. Again, note the original reference for the data if possible.

### Handling Metabolites and Other Analytes

We handle metabolites and other related analytes on the basis of the chemical to which the test subjects were exposed. If the original exposure is to a parent molecule and metabolites or other progeny were monitored, then those should be noted using different values of “analyte\_name” on the Series tab of the data template. If the test subjects were originally exposed to a metabolite, however, that should be noted using “test\_substance\_name” on the Studies tab.

### Submitting Completed CvT Data Templates

When you have completed a template, add any notes you have (such as “Vmax and Km data also available”) to the Curation Coordination spreadsheet (<https://docs.google.com/spreadsheets/d/1Uzxw7p_6zlNehtGJlBP-AkTWvydFvt9CeCTtwyeewvE/edit#gid=0>) and enter the date in the “Curator [1/2] Finish Date” column. Then submit the template. There are multiple options, please choose whichever one works for you:

1. Email them to John (wambaugh.john@epa.gov) and Chris (grulke.chris@epa.gov)
2. L:\Lab\NCCT\_ExpoCast\ExpoCast2020\CvT-CompletedTemplates
3. <https://www.dropbox.com/home/Cvt%20Project/Completed%20Templates>

## Example

Populating the CvT data template (which is an Excel workbook) requires expert judgement and is highly manual due to the heterogeneity in PK articles. For each paper five linked tables must be filled out: Documents, Subjects, Studies Series, and Conc Time Values

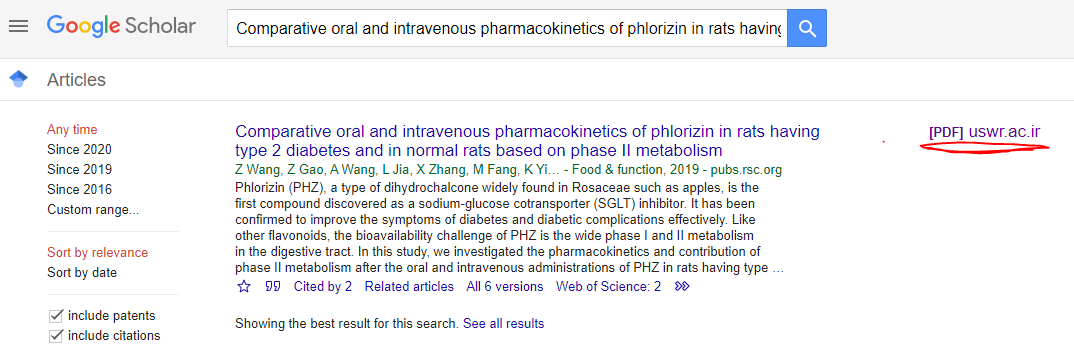
This example using article "Comparative oral and intravenous pharmacokinetics of phlorizin in rats having type 2 diabetes and in normal rats based on phase II metabolism" (<https://doi.org/10.1039/C8FO02242A>) with a pmid of 30806398 represents a typical case and, while informative, will not necessarily be reflective of some more complicated extractions. The pdf and filled in template are available in <https://github.com/USEPA/CompTox-PK-CvTdb/blob/master/WikiImages/Example1>

### Documents

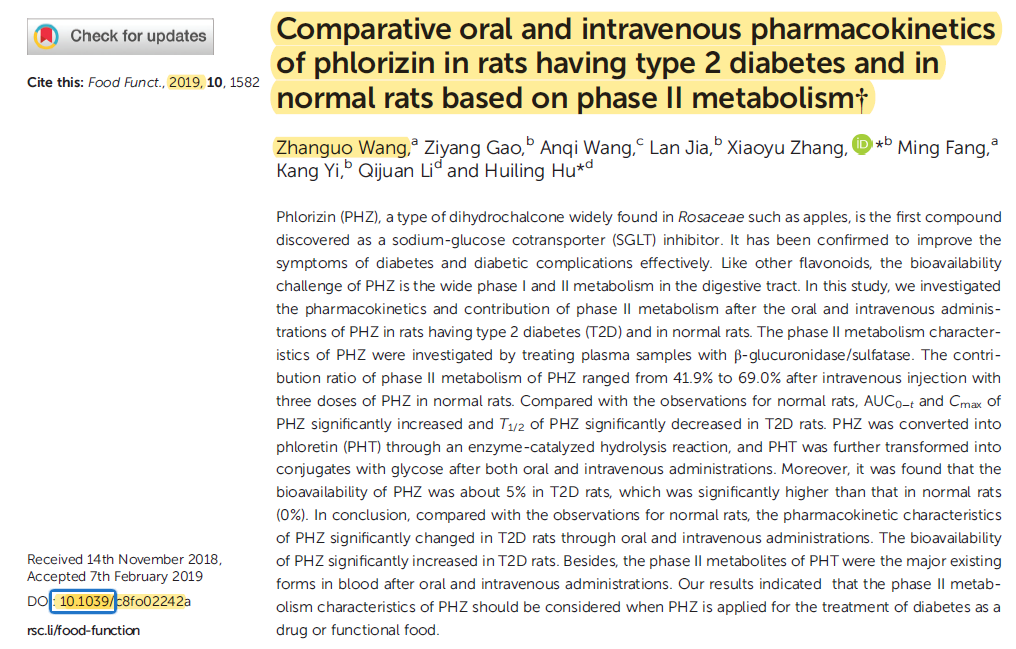
Fill in a new copy of xlsx template for each article to facilitate mapping during the load into the database and to reduce possible errors during template creation and parsing. Therefore, the Documents sheet should only contain a single row of data. For PMID\_30806398, the title was found using the search from the pmid on pubmed (<https://www.ncbi.nlm.nih.gov/pubmed/>). This title was used to search Google Scholar (<https://scholar.google.com/> which provided a link to the pdf (see Fig 1). Once a pdf has been sourced, the extraction of the fields is quite straightforward with all information typically occurring on the first page (See Fig 2).

PMID is sufficient as an id.

**Fig 1.** Finding a pdf using Google Scholar



**Fig 2.** Document information in PMID\_30806398 highlighted

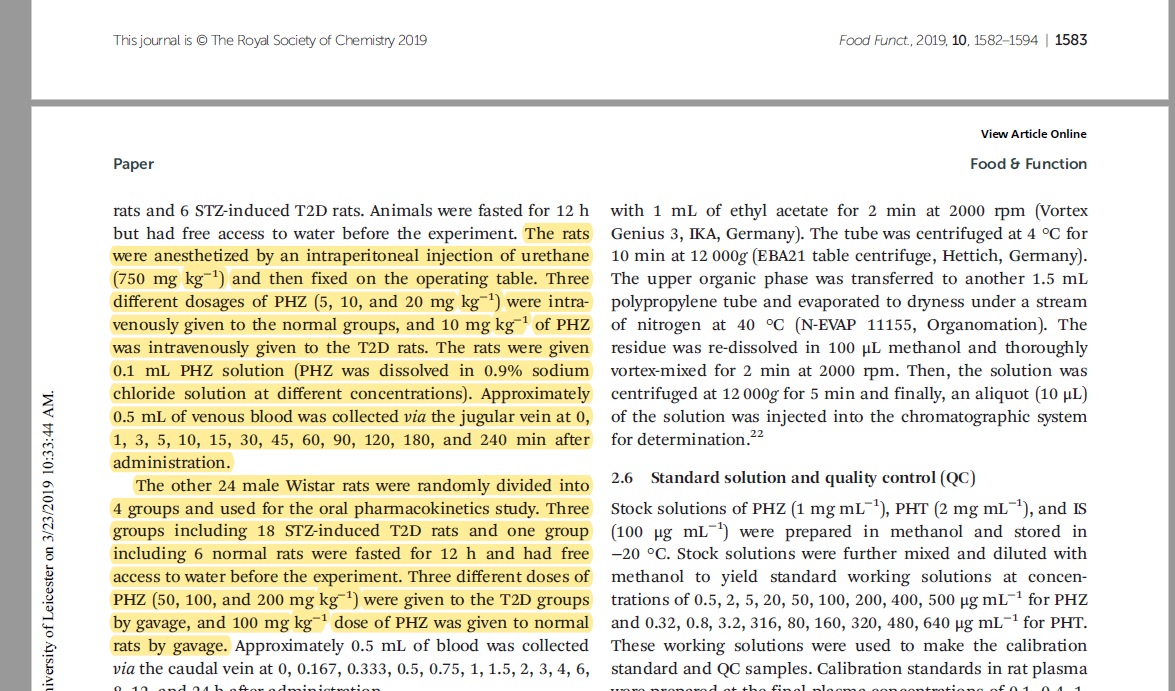


### Studies

While extraction for Documents is pretty self explanatory, the extraction of Studies is where the complexity begins and requires a pretty thorough understanding of the experiments described in the article. A study is generally equivalent to a single dosing strategy for a test substance. If the test substance is different or the dosing regime changes in any way, it is considered a new study.

In this publication there are 6 different dosing regimens: 3 intravenous doses of PHZ at 5, 10, and 20 mg/kg; and 3 oral gavage doses at 50, 100, and 200 mg/kg. This content can be found in the paper in the "Methods>Animal experiments" section 2.3 (See Fig 3). An important note when reading the information is that performs a 10 mg/kg dosing on normal animals is the same study as performing a 10 mg/kg dosing on T2D animals. It is only the subject that has been changed.

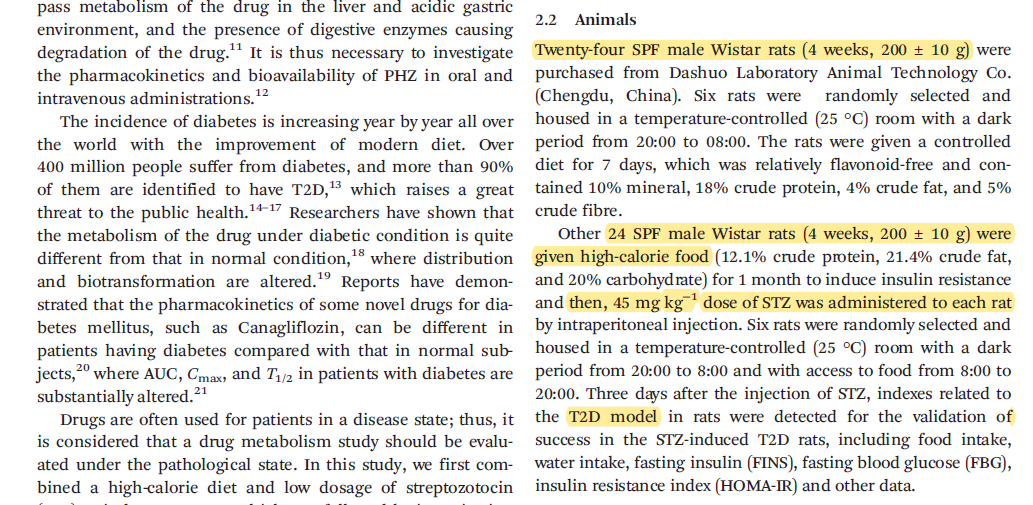
**Fig 3.** Study information in PMID\_30806398 highlighted



### Subjects

The animals that are tested are generally pretty simple to describe in the Subjects sheet, unless complex preparation of the animals is done to simulate a disease state. In this case there are 2 different types of subjects: the normal mice and the T2D mice. Occasionally, publication provide detailed information on the subjects (particular for human PK trials) in which case the complexity of the subjects can be high. However, in this case the information contained "Methods>Animals" section 2.2 (see Fig 4) is sufficient to populate the two subject records in the template.

**Fig 4.** Subject information in PMID\_30806398 highlighted



### Series

Series correspond to each line in the PK graph, and it can sometimes be difficult to understand the information that is being presented in that graph. Also, the same information can be presented in more than one figure. This example has some complexity as the analytical methods are being used to detect (1) parent PHZ (PHZ that exists as PHT in the plasma) (2) total PHZ (parent PHZ + conjugated PHZ) and (3) total PHT (parent PHT + conjugated PHT). Additionally, some of the graphs have zoomed in panels, meaning the same data is displayed more than once. In total, 19 distinct series were extracted. In this case, it was decided to extract the data from the following figures in the paper: Fig 4 (9 series), Fig 6 (2 series), Fig 7 (6 series), and Fig 8 (2 series). It should be noted that Fig 6 replicates its normal curves from Fig 4 and Figure 8 replicates its T2D curves from Figure 7.

Most of the metadata information from the series come from the figure captions. The difficulty comes in properly mapping the series to its study and subjects. This is easier (provided you have enough monitor space) if you open each sheet in its own window using the "View>New Window" capabilities in excel. That way, the series tab can be viewed at the same time as the study and subject tabs to ensure the id mapping is preserved.

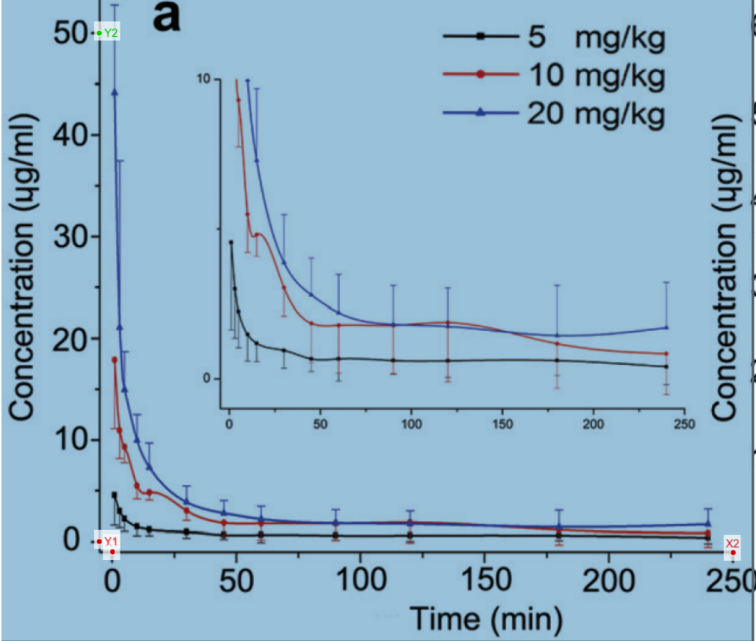
### Conc\_Time\_Values

Finally, it is time to extract the actual data. This is the most laborious part of curating CvT data.

First each figure containing data is put into a png (in my case using the Windows 10 native "Snipping Tool", though there are many options for getting an image out of the pdf). Then data was extracted using WebPlotDigitizer (<https://automeris.io/WebPlotDigitizer/>).

Using Fig 4a from the publication as an example, the extracted image was uploaded into WebPlotDigitizer using File > Load Image. "2D (X-Y) Plot" was selected as the plot type and then axes were aligned by clicking on the necessary points on the image (see Fig 5) followed by clicking "Complete" on the right side of the interface and inputting the x and y-values for those calibration points. Once calibrated, the web application automatically created a default dataset. When a dataset is selected in the upper-left portion of the interface, a "Manual Extraction" panel appears on the right side of the interface with "Add Point" selected by default. This allowed points to be added to the dataset by clicking on them in your image. These points were adjusted to be as close to accurate using the "Adjust Point" function and using the keyboard arrow and viewing the image in the zoomed in window in the top right corner of the interface. Once all data points for a single series were clicked and adjusted, the data was exported by using the "View Data" button on the left side of the interface and then copying/pasting into the ConcTimeValues sheet in the template. The appropriate series id was added to that extracted dataset. Since multiple series were available in Fig 4a, after copying the data into the template, the data was cleared the next series was extracted. Then this was repeated for each figure that contained extractable data.

**Fig 6.** ConcTimeValue information being extracted using WebPlotDigitizer



1. Sayre, R.R., Wambaugh, J.F. & Grulke, C.M. Database of pharmacokinetic time-series data and parameters for 144 environmental chemicals. *Sci Data* **7,**122 (2020). <https://doi.org/10.1038/s41597-020-0455-1> [↑](#footnote-ref-1)
2. <https://github.com/USEPA/CompTox-ExpoCast-invivoPKfit> [↑](#footnote-ref-2)
3. Pearce, Robert G., et al. "Httk: R package for high-throughput toxicokinetics." *Journal of Statistical Software* 79.4 (2017): 1. <https://dx.doi.org/10.18637%2Fjss.v079.i04> [↑](#footnote-ref-3)
4. Cohen Hubal, Elaine A., et al. "Advancing internal exposure and physiologically-based toxicokinetic modeling for 21st-century risk assessments." *Journal of Exposure Science & Environmental Epidemiology* 29.1 (2019): 11-20. <https://doi.org/10.1038/s41370-018-0046-9> [↑](#footnote-ref-4)
5. U.S. Environmental Protection Agency. (1998). Health Effects Test Guidelines OPPTS 870.7485 Metabolism and Pharmacokinetics. Washington, D.C. Retrieved from http://nepis.epa.gov [↑](#footnote-ref-5)