Source code instructions for manuscript "Evaluating impacts of anatomical, physiological, and biochemical variability on human equivalent doses using PBPK models"

Below are step-by-step instructions to run the code. For more details about each individual folder and file, please see the following section (Code Repository: Complete Description).

Instructions for use (relevant .R code are in **bold**):

1. Install R, Rstudio, the R package "deSolve", Rtools, and GNU MCSim on your system. Instructions can be found in "RMCSim_install_instructions.pdf". Ensure all folders are present: CF, DCM, Figures, PostHocAnalysisCodes, R_data, SharedCodes, and VstatsAndFits.
2. Install required R packages:
    a. For running simulations: `tictoc`. If using parallel processing: `doParallel, foreach, parallel.`
    b. Post hoc analysis: `stats4, MASS, fitdistrplus, lognorm, viridis, ggplot2, moments, ggtext.`
3. Generate chloroform (CF) HED distributions:
    a. Run **build_cf_model.R** to generate pbpk_cf_model.c, pbpk_cf_model.o, pbpk_cf_model.dll (or .so) and pbpk_cf_model_inits.R. This only needs to be run once.
    b. In **ReverseDosimetry_CF_pardist.R**, the user may set the following (or leave as-is):
        i. ifpar = TRUE or FALSE (TRUE for parallel processing or FALSE without parallel processing).
        ii. allom = TRUE or FALSE (TRUE to apply allometric scaling to internal dose POD or FALSE do not apply allometric scaling).
        iii. numsim: number of simulations (default is 5000. Please note that the statistical post hoc analysis does not work for simulations > 5000).
    c. Source **ReverseDosimetry_CF_pardist.R**. Output (numsim number of HEDs per 6 subpopulations/5 parameter distributions for oral and inhaled exposure) will be saved in R_data.
4. Generate dichloromethane (DCM) HED distributions:
    a. Run **build_dcm_model.R** to generate pbpk_dcm_model.c, pbpk_dcm_model.o, pbpk_dcm_model.dll (or .so) and pbpk_dcm_model_inits.R. This only needs to be run once.
    b. In **ReverseDosimetry_DCM_pardist.R,** the user may set the following (or leave as is):
        i. ifpar = TRUE or FALSE (TRUE for parallel processing or FALSE without parallel processing).
        ii. allom = TRUE or FALSE (TRUE to apply allometric scaling to internal dose POD or FALSE do not apply allometric scaling).

iii. numsim: number of simulations (default is 5000. Please note that the statistical post hoc analysis does not work for simulations > 5000).
iv. normalize.tissue = TRUE or FALSE (TRUE to normalize tissue volumes or FALSE not to normalize tissue volumes. We only set normalize.tissue = FALSE with the untruncated lognormal parameter distribution case (case_pardist=5)).

c. Source **ReverseDosimetry_DCM_pardist.R**. Ouput is saved in R_data.
d. To replicate the case in which we do not normalize tissue volumes, the vector case_pardist_vals1 can be modified such that case_pardist_vals1 = 5.

5. Do post hoc analysis.
   a. In PostHocAnalysisCodes folder, run the following codes:
      i. **calc_RoystonVStats.R** (calculates Royston V values for HEDs) and **calc_MOMandFits.R** (parameters for fitted lognormal distributions). Data files from these runs are saved in VstatsAndFits.
   b. Run **CreateFigures_Manuscript.R** to generate figures found in the manuscript. These are saved in Figures/Manuscript.
   c. Run **CreateFigures_Supplementary.R** to generate figures found in the supplementary material. Run as-is to replicate the results discussed in the manuscript. These figures are saved in Figures/Supplementary.
      i. Under section #3 (line 82), the user may specify which fit method they would like to apply (method of moments, maximum likelihood estimation, etc., detailed more below) by changing the variables mu, sigd, method, and meth2. The default method is the geometric mean/sd method of moments (mu = "GMOM.mu" and sigd = "GMOM.sd").

Code Repository: Complete Description

The code repository includes the following folders (relevant .R code are in **bold**):

2. SharedCodes: these codes are shared across multiple files.

- **RMCSim.R**
  - This R source code file defines functions for compiling, loading, and running and ODE model encoded in the GNU MCSim model specification language. The mode.exe utility must be available in the user's PATH for the compile_model function to work properly.
- **human_par_funs.R**
  - Contains functions to generate parameters drawn from physiological parameter distributions shared by both CF and DCM.
- **sampling_funcs.R**
  - Contains functions to convert parameter distributions to either truncated normal, truncated lognormal, untruncated lognormal, or uniform.

    - Includes a function to convert ppm to mg/L.
    - Contains dosing events for inhalation or oral exposures.
- **optim_n_MC.R**
    - Contains functions for optimization runs either with or without parallel processing (which is specified in the ReverseDosimetry codes for each chemical)

3. CF:

- **pbpk_cf.model**
    - This file contains the PBPK model for chloroform. It must be compiled using **build_CFmodel.R**.
- **build_CFmodel.R**
    - This compiles **pbpk_cf.model** to create necessary initial parameter files, pbpk_cf_model.c, pbpk_cf_model.o and pbpk_cf_model.dll (or .so, depending on the OS), which are all needed when you execute load_model(**pbpk_cf.model**) in **RATPOD_CF.R** and **ReverseDosimetry_CF_pardist.R**. This only needs to be sourced once.
- **RATPOD_CF.R**
    - Calculates the human internal dose for CF inhalation and oral exposure by calculating the rat internal dose and applying a scaling factor, which is applied in **ReverseDosimetry_CF_pardist.R.** It uses **RMCSim.R** and **sampling_funcs.R** from the SharedCodes folder.
- **ReverseDosimetry_CF_pardist.R**
    - Calls to **RATPOD_CF.R** to get internal dose, **CFhuman_par.R** for CF human parameter values, and the following codes from the SharedCodes folder: **RMCSim.R**, **human_par_funs.R**, **sampling_funcs.R**, and **optim_n_MC.R**
    - Calculates HEDs for 5 parameter distribution types and 6 subpopulations. The user can set:
        - ifpar = TRUE (parallel computing, requiring doParallel, foreach, and parallel packages) or FALSE (no parallel computing)
        - allom = TRUE (apply allometric scaling) or FALSE (no allometric scaling)
    - Saves the .Rdata in a folder called "R_data"). Each individual .Rdata file contains 5000 simulations pertaining to either inhaled or oral exposure and from one of each of the 5 parameter distribution inputs and 6 subpopulations and the corresponding parameter information. Data files beginning with the prefix "ALL_" contains a 5000x6x5 object with all combined HED data.
- **CFhuman_par.R**
    - Sources **sampling_funcs.R** and **human_par_funs.R** and returns a list of 5000 parameter values for each of the parameters in the model.

4. DCM:

- **pbpk_dcm.model**

    - This file contains the PBPK model for DCM. It must be compiled using **build_DCMmodel.R.**
- **build_DCMmodel.R**
    - This compiles **pbpk_dcm.model** to create necessary initial parameter files, pbpk_dcm_model.c, pbpk_dcm_model.o and pbpk_dcm_model.dll (or .so, depending on the OS), which are all needed when you execute load_model(**pbpk_dcm.model**) in **RATPOD_DCM.R** and **ReverseDosimetry_DCM_pardist.R**. This only needs to be sourced once.
- **RATPOD_DCM.R**
    - Calculates the human internal dose for DCM inhalation and oral exposure by calculating the rat internal dose and applying a scaling factor, which is applied in **ReverseDosimetry_DCM_pardist.R**. It uses **RMCSim.R** from the SharedCodes folder.
- **ReverseDosimetry_DCM_pardist.R**
    - Calls to **DCMhuman_par.R** for DCM parameter distributions, and the following codes from the SharedCodes folder: **RMCSim.R**, **human_par_funs.R**, **sampling_funcs.R**, and **optim_n_MC.R**
    - Calculates HEDs for 5 parameter distribution types and 6 subpopulations. The user can set:
        - ifpar = TRUE (parallel computing, requiring doParallel, foreach, and parallel packages) or FALSE (no parallel computing)
        - allom = TRUE (apply allometric scaling) or FALSE (no allometric scaling)
        - normalize.tissue = TRUE (if fractional tissue volumes are normalized) or FALSE (fractional tissue volumes are not normalized); we only set to FALSE with the UB lognormal case (case_pardist = 5).
    - Saves the .Rdata in a folder called "R_data"). Each individual .Rdata file contains 5000 simulations pertaining to either inhaled or oral exposure and from one of each of the 5 parameter distribution inputs and 6 subpopulations and the corresponding parameter information. Data files beginning with the prefix "ALL_" contains a 5000x6x5 object with all combined HED data.
- **DCMhuman_par.R**
    - Sources **sampling_funcs.R** and **human_par_funs.R** and returns a list of 5000 parameter values for each of the parameters in the model.

5. R_data:

- Contains already generated data. Because the simulations take a long time to run, we include the data if the user wishes to use it for their own post hoc analysis. The folder "R_data" is necessary in the "**ReverseDosimetry_**" codes, so must exist if the users generate their own data.
- Simulations run for DCM in the case when tissue volumes were not normalized have the prefix "ParmsNorNormalized_".

6. PostHocAnalysisCodes: codes used on .Rdata (there are 4 .Rdata files, with prefix "ALL_", which include HED data for each chemical and each exposure scenario). In these codes, the HED data are all normalized around the median.

- **royston_sf.R**
  - Contains a function to calculate the Royston V index.
- **calc_RoystonVStats.R**
  - Sources **royston_sf.R** to calculate Royston's V for all chemical/exposures subpopulation/parameter distribution combination HEDs. Saves the V values to a folder "VstatsandFits".
  - Calculates V values for HED distributions from both allometric scaling and no allometric scaling.
  - Data is saved with prefix "CF_Vstats" or "DCM_Vstats" in the folder "VstatsAndFits"; if the HEDs were generated without allometric scaling, the prefix begins with "unscaled_".
- **calc_MOMandFITS.R**
  - Contains a function to use the method of moments (MOM) to calculate parameters for a lognormal distribution and then calculates the fit for each HED distribution.
  - Also uses Fitdistrplus to fit distributions for MGE (Maximum goodness of fit) for right tail (ADR), left tail (ADL), Cramer-von Mises distance (CVM), Kolmogorov-Smirnov (KS); MLE (maximum likelihood estimation), MME (moment matching estimation using fitdistrplus), and QME (quantile matching estimation) (in supplementary material); techniques outlines in https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.pdf.
  - Saves the mean and standard deviation for each HED fit in the folder "VstatsandFits" with prefix "muSigFits_".
    - Each list gives the mean and standard deviation for each fit method
      - Columns are labelled: c("mle.mu", "mle.sd", "ADL.mu", "ADL.sd", "ADR.mu", "ADR.sd", "CVM.mu", "CVM.sig", "KS.mu", "KS.sd", "mme.mu", "mme.sd", "qme.mu", "qme.sd", "GMOM.mu","GMOM.sd","ARITHMOM.mu","ARITHMOM.sd")
      - Ex: Open "VstatsAndFits/muSigFits_DCMCONC.RData". allFit[,c("GMOM.mu", "GMOM.sd"),"UBlognorm"] contains means and standard deviations for all subpopulations for DCM inhalation when using the original parameter set using the method of moments, where matched moments are the geometric mean and geometric standard deviation* (in manuscript).

- **CreateFigures_Supplementary.R**
  - Generates figures for the supplementary material for CF oral and inhalation and DCM oral and inhalation.
  - Loads the .RData (previously generated from **calcMOMandFITS.R** and **calc_RoystonVStats.R)** which is located in the folder VstatsAndFits.

        - Density plots for only results with allometric scaling applied.
        - QQ plots of HED distributions for results generated using both allometric scaling and no allometric scaling.
        - Error plots showing relative errors between the HED distributions and their fits using method of moments (for only results with allometric scaling applied)
            - Mean and standard deviation parameters for the fits are loaded from the files in VstatsandFits with prefix "muSigFits_", which contains means and sd for each fit method.
            - User can specify which fit method they would like to use (if not method of moments) by setting mu, sigd, method, and meth2 on lines 82-85.
        - QQ plots comparing HEDs when tissues were normalized vs not normalized (only DCM, 1 y.o. children and 70 y.o. females from untruncated lognormal parameter distributions)
    - Figures are saved to Figures/Supplementary.
- **CreateFigures_Manuscript.R**
    - Generates figures for the manuscript
        - A heatmap of all of the V values.
            - Data is loaded from VstatsAndFits (V value .RData has prefix "DCM_Vstats" or "CF_Vstats" or "unscaled_CF_Vstats" or "unscaled_DCM_Vstats")
        - QQ plots of arbitrary distributions to show difference in V values between lognormal and not lognormal (using **royston_sf.R**)
        - QQ plots comparing HEDs for each chemical (oral DCM, normal parameter distribution, 30 y.o. male; inhaled CF, normal distribution, 30 y.o. male), using **royston_sf.R.**
    - Figures are saved to Figures/Manuscript.

7. VstatsAndFits: this folder contains output from PostHocAnalysisCodes/**calc_RoystonVStats.R** and PostHocAnalysisCodes/**calc_MOMandFits.R**

- Royston V values for each chemical/exposure are in each file (CF_VstatsCONC.RData, CF_VstatsSTOM.RData, DCM_VstatsCONC.RData, DCM_VstatsSTOM.RData), where "CONC" refers to inhalation and "STOM" refers to oral exposure. Each .RData file has 30 V values, one for each subpopulation/parameter distribution case.
    - Files with prefix "unscaled_" are V values for the HED distributions generated from simulations not using allometric scaling.
- Mean and standard deviations for the lognormal fits for each HED distribution are labeled: muSigFits_CFSTOM.RData, muSigFits_CFCONC.RData, muSigFits_DCMSTOM.RData, muSigFits_DCMCONC.RData. These values are only generated with the scaled HED data.

- Note that if the users wish to generate their own data, this folder must still exist because the PostHocAnalysisCodes folder codes save to it.

8. Figures: This folder contains subfolders “Manuscript” and “Supplementary”. These folders must exist because the codes PostHocAnalysisCodes /**CreateFigures_Supplementary.R** and PostHocAnalysisCodes/**CreateFigures_Manuscript.R** call to them when saving figures.