HESTIA\_Report

Oskar Nyberg

2023-03-14

# HESTIA Database construction.

To generate EC20^EC10eq effect endpoints for chemicals i need data from EC10, EC50 as well as LC and NOEC data. see data/CIR\_query\_CAS\_to\_smiles.R for details on the treatment of input CAS numbers, how all available SMILES configurations were gathered and thereafter exported as 4k row long subsets. This information on CAS and SMILES per substance was used as input into OECD QSAR Toolbox software where two distinct operations took place, 1) query for toxicological effect data, and 2) query for substance physicochemical data. When importing the CAS-SMILES input into QSAR toolbox, “multiplying” substances are present, since QSAR toolbox includes all SMILES configurations of any CAS number. hence, 16797 substances as input -> ~22k substances output. Acquired metadata data gives an abundance of test information across 550-660 columns, although the majority of these are redundant for the current purpose. Wrangling of the raw OECD QSAR Toolbox output takes place in data/raw\_data\_read\_and\_wrangle.R and is subsequently imported as a data frame for treatment and filtering.

In several steps of these filtering operations, physicochemical data is required, which is available for **5709** substances, code loaded from file “Physicochemical\_properties.Rmd”. This is imported at the here at the onset of data wrangling, because I need some of the physicochemical data and the pesticide annotations below (for pesticides, ACR annotations are 2.2, not standard 2, when converting EC50\_acute -> EC50\_chronic in USEtoc v2.1 according to USEtox manual!)

## Raw data wrangling

Importing, filtering and wrangling the ecotoxicological effect data from QSAR data output, including relevant metadata that can act as quality control. In this operation the OECD QSAR Toolbox output gets read in and a first step of harmonizing the data set. The files have different lengths and number of columns, which forces me to select a set of defined columns. Additionally, duplicates and completely empty records have been removed to have a neater data set to work with. Available from the 5 databases; ECOTOX, Aquatic OASIS, Aquatic ECETOC,Aquatic Japan MoE, and Food TOX Hazard EFSA with 308708, 2231, 6238, 3238, and 2265 records respectively.

However, anomalies were discovered when visualizing the finished data that are easily fixed at the onset of wrangling. After inspecting source material, it is clear that most of these are incorrect entries into the ECOTOX database. Also, fixing the publication year of source material by merging two columns since different databases have different names for publication year. **Push to export a table of these?**

### Which endpoints to include.

Saouter et al.(2022) defined 6 different conversion coefficients for chronic/acute EC50 -> chronic EC10 etc. Leo Posthuma used several more, all ECx (1-20), records with the endpoints NOEC, LOEC, maximum acceptable toxicant concentration, EC0, EC5, EC10, and EC20 are marked as “chronic NOEC”, records with (EC) or (LC) endpoint ranging from 30 to 70% are marked as “acute EC50” Acute/chronic definitions are available in Posthuma et al., 2019 (Table 1) and Aurisano et al., 2019 for algae, bacteria, unicellular animals, crustaceans, fish, molluscs/worms/etc.

Harmonization and aggregation of endpoints  
[]  
EC0, EL0, IC0, LC0, NOAEC, NOEC, NOEbC, NOErC, NOEL grouped into NOEC; EC10, IC10, LC10, LOEC grouped into EC10eq; EC50, EbC50, EbL50, ErC50, ErL50, IC50, LC50 grouped into EC50.  
“We combined LOEC and EC10 for deriving extrapolation factors based on high uncertainties in the low range of species sensitivity distributions, rendering it difficult to treat LOEC and EC10 as separate metrics in statistical analyses (Iwasaki et al. 2015; King et al. 2017).” [Aurisano et al., 2019, p. 2571]

Table of effect data endpoints included in the construction of the HESTIA toxicological database

Endpoint

n

Harmonized Endpoint(*q*)

LOEC

78564

EC10

EC10

4053

EC10

LC10

1987

EC10

IC10

502

EC10

LD10

33

EC10

ER10

26

EC10

EL10

2

EC10

LC50

103776

EC50

EC50

44910

EC50

IC50

4081

EC50

LD50

1949

EC50

ER50

126

EC50

EL50

22

EC50

NOEC

105703

NOEC

NOEL

13348

NOEC

LC0

1110

NOEC

EC0

527

NOEC

NOER

175

NOEC

LD0

51

NOEC

NOAEC

10

NOEC

### Values reported in range

Effect concentration qualifiers Based on similar work by Saouter et al., 2019, in cases where no mean value is reported, the following rules will be applied: “A large majority of the results have a numeric value in the low range with a qualifier =, ca., >=, or >. In contrast, only a few tests have their results expressed in the higher ranges (5862 test results). The following selections were made to maximize the use of available data:  
1. When there is a lower range value with the descriptors ‘>=, ca., or empty’, the lowest value is selected. If, within this group, a test has also a higher value, this higher value is ignored.  
2. All lower range values described as ‘>’ are ignored (n = 39602), unless the higher value is described as ‘=<’ (n= 80 observations). In case of NOEC > than, the value was kept since it is still representing a concentration with no observed effect.  
3. All higher values described as ‘< than’ are ignored, unless the lower range value is described as ‘>=’. Then the lower value is used.  
4. When a lower range value is missing (0 or blank) and a higher value is available described as ‘<=’, the higher value is used.  
5. When a lower value is described as >= and the higher value is described as <=, the lowest value is used.  
6. Values expressed as ‘<’ are excluded (4397 test results).” [Saouter et al., 2018, p. 47]  
The counts of effect concentration qualifier annotations

Qualifier type

Count across all

Count min qualifier

Count max qualifier

<

3391

999

3842

<=

418

14

18

=

1889

NA

NA

>

13444

4067

107

>=

400

742

6

ca.

1684

64

70

NA

301454

316794

318637

### Test media (Freshwater & culture media filter)

Assuming tests without defined media type is “freshwater”. Especially considering the majority of organisms are Daphnia magna, Pseudokirchneriella subcapitata, Pimephales promelas and Oryzias latipes (after inspection), just as Saouter 2018.

Media.type

n

Culture

7468

Freshwater

253938

Nosubstrate

15

Saltwater

51465

NA

9794

### Effect Criterion selection

Posthuma et al., 2019, selected few effect criterion where: “… records with the endpoints NOEC, lowest-observed-effect concentration, maximum acceptable toxicant concentration, EC0, EC5, EC10, and EC20 are marked as “chronic NOEC” when they have an appropriate taxon-dependent test duration (see Table 1) and population-relevant effect criterion (e.g., reproduction, growth, population growth, and development, next to mortality and immobility); and records with a sublethal (EC) or lethal endpoint ranging from 30 to 70% are marked as “acute EC50” when they have an appropriate taxon-dependent test duration (see Table 1) and effect criterion (e.g., mortality and immobility). However, Posthuma et al. also clustered NOEC, LOEC, EC 0-20 into one “Chronic NOEC”-category and all EC 30-70 into one Acute EC50 category.

We decide to select endpoint criterions based on expert consultation with dr. Andreu Rico

Effect

n

Mortality

86593

Population

31102

Growth

14753

Intoxication

9052

Reproduction

8398

Behavior

5779

Development

4904

Cell(s)

3803

Feeding Behavior

1391

Acute

850

Growth Rate

832

Chronic

686

Biomass

273

Immobilisation

116

Frond Number

98

Mobility

78

Body Weight

10

Seedling Emergence

10

Behaviour

7

### Defining Acute/chronic conditions

Operation based on taxonomic group First step is to harmonize test durations and test concentration units and transform test concentrations into mg/L measurements. This is performed according to table step 4: Filter out data that is not convertable into correct time units or concentration formats.

Step 5: Define Acute or Chronic test conditions:

effect concentration conversions needs a revision: - #For toxic endpoints, are milligrams per liter (mg / L) equivalent to parts per million (ppm)? The endpoint concentrations for regulated toxic substances under the risk management program rule (40 CFR Part 68 Appendix A) are listed in units of milligrams per liter (mg/L). Is this equivalent to parts per million (ppm)? No, mg/L is not always equivalent to ppm. Whereas ppm is a volume-to-volume or mass-to-mass ratio, mg/l is a mass-to-volume relationship. To convert from units of mg/L to ppm, use the following equation. Endpoint (ppm) = [Endpoint (mg/L) x 1000 x 24.5] / [Molecular Weight] EPA has included the RMP toxic endpoints in both ppm and mg/L in Appendix B of the Risk Management Program Guidance for Offsite Consequence Analysis (EPA550-B-99-009, April, 1999).#

Hence, to convert ppm or ppb to mg/L, I need to redo the conversions. Endpoint (ppm) = [Endpoint (mg/L) x 1000 x 24.5] / [Molecular Weight] –> Endpoint (ppm) x Molecular Weight / 1000 x 24.5 = Endpoint (mg/L)

### EC10eq conversions

With endpoints harmonized according to EC10eq conversions

Harmonized endpoint (q)

Acute or chronic exposure (a)

Taxonomy group (t)

extrapolation factor (g)

High CI (95%)(h)

Low CI (95%)(l)

EC50

Acute

Fish

7.44

18.95

2.92

EC50

Acute

Crustacean

3.38

5.34

2.14

EC50

Acute

Algae

4.00

6.10

2.60

EC50

Acute

Others

4.00

6.10

2.60

EC50

Chronic

Fish

1.55

3.66

0.67

EC50

Chronic

Crustacean

1.94

2.41

1.56

EC50

Chronic

Algae

2.24

2.65

1.90

EC50

Chronic

Others

2.00

2.50

1.80

EC10

Acute

Fish

1.00

1.00

1.00

EC10

Acute

Crustacean

1.00

1.00

1.00

EC10

Acute

Algae

1.00

1.00

1.00

EC10

Acute

Others

1.00

1.00

1.00

EC10

Chronic

Fish

1.00

1.00

1.00

EC10

Chronic

Crustacean

1.00

1.00

1.00

EC10

Chronic

Algae

1.00

1.00

1.00

EC10

Chronic

Others

1.00

1.00

1.00

NOEC

Acute

Fish

3.97

17.39

0.90

NOEC

Acute

Crustacean

1.55

2.64

0.91

NOEC

Acute

Algae

1.80

2.70

1.00

NOEC

Acute

Others

1.80

2.70

1.00

NOEC

Chronic

Fish

0.60

0.70

0.40

NOEC

Chronic

Crustacean

0.95

1.16

0.77

NOEC

Chronic

Algae

0.44

0.49

0.39

NOEC

Chronic

Others

0.60

0.70

0.40

### Summarising the chemical properties work

How many different types of substances can be ascribed a certain use-property? By looking in several data repositories I have collected data on