

# Anatomy of a Table – Our Continued Journey to Automating Output Delivery Through Metadata

Edwin van Stein, GSK, The Netherlands

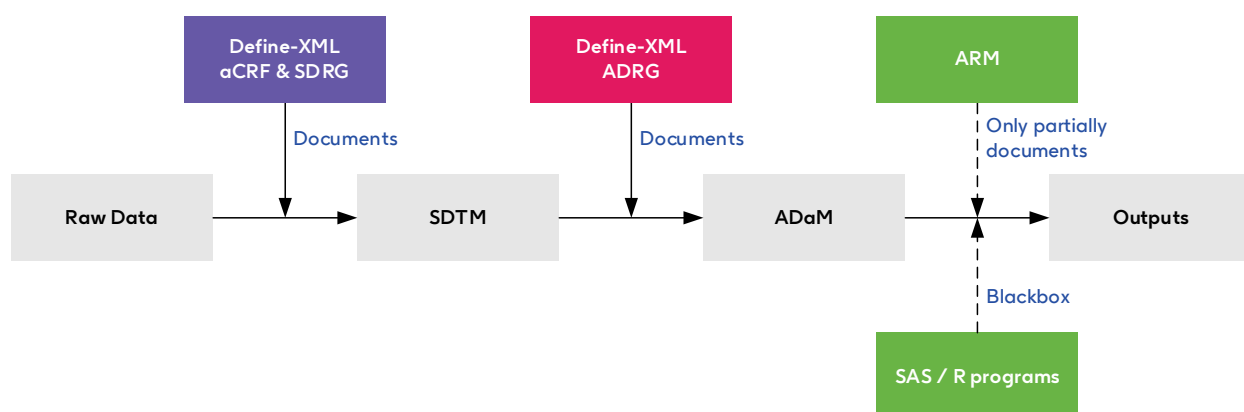
## ABSTRACT

To improve traceability, transparency and consistency, GSK is implementing automation of analyses and output creation through standardised metadata. Tool-agnostic analysis results metadata (A highly-enriched adaptation of CDISC ARM) is used to produce analysis results data. Following the WORM principle (Write Once, Read Many), tool-agnostic analysis output metadata can then be used to render those results into one or more displays (tables, figures and listings, both post-text and in-text), other outputs (e.g. XML files for disclosure) and systems (in whatever format those systems require) without reproducing the numbers. This leads to a process where it's easier to validate and reuse analysis results and to build tools in a multilingual setting.

This poster is a follow-up to my presentation from the PHUSE EU Connect 2023 and focuses on visualising the metadata for a single table and how it all fits together in our journey to standardising our results.

## INTRODUCTION

Traceability and transparency of collected data and analysis data are ensured by using CDISC's SDTM and ADaM standards and are documented in Define-XML, ACRF, SDRG and ADRG. The analysis of primary endpoints can be traced back to ADaM using ARM in Define-XML, but any analysis not defined in ARM has no clear traceability through submission deliverables. SAS or R programs used to create outputs are often part of the submission package and can provide some traceability, but if they use a lot of macros or packages then they become something of a blackbox. In addition, ARM is very often created retrospectively and is not necessarily meant to be used to drive automation.



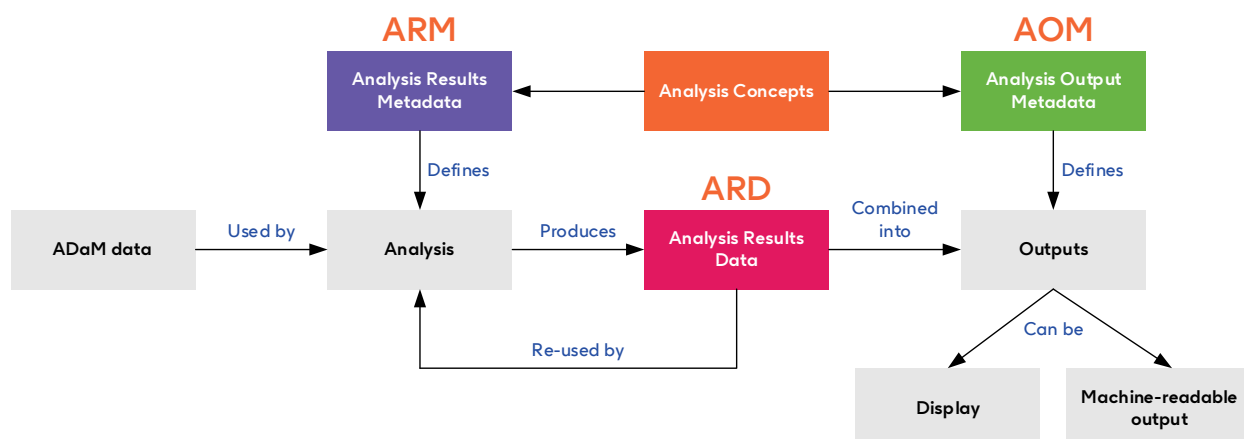
Within the broader e2e (end-to-end) Standards project at GSK, we're looking into defining metadata that describes the path from analysis data to outputs for increased traceability, transparency, automation, and consistency.

## WORM

The central principle behind our implementation is Write Once, Read Many (WORM). Any analysis should be defined, executed, and validated only once and then the numbers should be re-used across outputs and analyses. This also applies to simple analyses like a frequency count analysis on the number of

subjects in the safety analysis set by treatment. This analysis should only be performed once and then be re-used as denominator and in column headers across many outputs.

## VISION



Analysis Results Metadata (ARM, a highly-enriched adaptation of CDISC ARM) defines a single analysis, for instance the calculation of analysis set counts by treatment. Standard tools consume the ARM and ADaM data to produce Analysis Results Data (ARD). As mentioned before the ARD can be re-used by another analysis, for instance by using the analysis set counts by treatment as denominator in another analysis. Analysis Output Metadata (AOM) defines an output, for instance a demography table. It defines which sets of ARD should be included, how to combine them and how to format them. Standard tools consume the AOM and ARD to then produce the outputs. Creating a slightly different in-text version of an already defined end-of-text table is then just a matter of defining the AOM to produce that table without redoing the entire analysis. The same Analysis Results Data can also be used to feed into other systems, for instance as an XML file for disclosure purposes or as JSON file for CSR automation.

All metadata that can be re-used across analyses is captured in Analysis Concepts. For instance:

- Safety analysis set is always the same WhereClause (SAFFL EQ "Y") with the same label (Safety)
- Actual treatment may be defined by TRTAN, TRTA and a corresponding codelist, where TRTAN is used for sorting and TRTA (or the codelist) for display purposes

## SHELL & INPUT DATA

The table shell used to illustrate the metadata examples in this paper is shown below and is a very basic demography table. The numbers in the orange and purple circles refer to specific parts of the metadata.

1

2

Protocol: 123456

Page x of y

Analysis Set: Safety

(Data as of: ddMMMyyyy)

2

3

Table 1.1 Summary of Demographic and Baseline Characteristics

4

Treatment A

5

Treatment B

Total

10

(N= 11)

(N=xxx)

(N=xxx)

6

Sex

12

8

n

xxx

xxx

xxx

13

Female

xx (xx%)

xx (xx%)

xx (xx%)

14

Male

xx (xx%)

xx (xx%)

xx (xx%)

7

Age (YEARS) at First Dose [1]

9

n

xxx

xxx

xxx

Mean

xx.x

xx.x

xx.x

SD

x.xx

x.xx

x.xx

Median

xx.x

xx.x

xx.x

Min.

xx

xx

xx

Max.

xx

xx

xx

3

[1] Age is imputed when full date of birth is not provided.

1

Path/program.R ddMMMyyyy hh:mm

The input data is a very basic ADSL data set as shown below.

| ADSL        |       |      |     |       |         |        |
|-------------|-------|------|-----|-------|---------|--------|
| USUBJID     | AAGE1 | SEXN | SEX | SAFFL | TRT01AN | TRT01A |
| 123456-1001 | 42    | 1    | F   | Y     | 1       | TRTA   |
| 123456-1002 | 57    | 2    | M   | Y     | 2       | TRTB   |
| ...         | ...   | ...  | ... | ...   | ...     | ...    |

## CONVENTIONS USED

In the examples below shading is used to illustrate metadata that links to other metadata or input data:

- Blue: links to input data sets and variables
- Green: links to AOM
- Purple: links to ARM
- Orange: links to Analysis Concepts

There are no direct links to ARD, this is always via ARM and analysisID.

## SYSTEM AND ENVIRONMENT METADATA

Some metadata applies to every single output produced in the system or every single output produced in the environment (e.g. a reporting effort on a study). This information is stored in TitleFootSystem and EnvironmentVariable respectively.

### TitleFootSystem

TitleFootSystem contains the definition of the first two titles and last footnote of every output. Values between curly brackets are environment variables and values derived at output production.

1

| TitleFootSystem |         |           |                             |
|-----------------|---------|-----------|-----------------------------|
| type            | ordinal | alignment | text                        |
| Title           | 1       | Left      | Protocol: {Study}           |
| Title           | 1       | Right     | {PageOf}                    |
| Title           | 2       | Left      | Analysis Set: {AnalysisSet} |
| Title           | 2       | Right     | {DataAsOf}                  |
| Footnote        | 99      | Left      | {Path} {DateTime}           |

### EnvironmentVariable

EnvironmentVariable contains the definition of certain environment variables used in titles.

2

| EnvironmentVariable |           |
|---------------------|-----------|
| variable            | value     |
| Study               | 123456    |
| DataAsOf            | ddMMMyyyy |

### ANALYSIS OUTPUT METADATA

The AOM contains the definition of outputs and is spread across multiple sets of metadata. The example display does not have spanning column headers, so SpanningHeaderAOM is not shown in the poster or paper.

### DeliverableAOM

DeliverableAOM defines a deliverable, its purpose and output number and links to the definition of the actual output.

| DeliverableAOM |           |
|----------------|-----------|
| ID             | dm_t01    |
| purpose        | Post text |
| outputNumber   | 1.1       |
| outputID       | dm_t01    |

### MainAOM

MainAOM defines an output and links all other metadata that defines the contents. bigNAnalysisID links to the analysis that defines the analysis set counts, bigNLocation defines that those analysis set counts should be printed wherever the actual treatment is printed.

| MainAOM         |                           |
|-----------------|---------------------------|
| outputID        | dm_t01                    |
| outputLabel     | Demography Table          |
| outputType      | Table                     |
| titlesFootnotes | dm_t01_t1 * dm_t01_fl     |
| columnGroups    | dm_t01_s * trt_r          |
| sections        | dm_t01_sex * dm_t01_aage1 |
| bigNAnalysisID  | as_saf                    |
| bigNLocation    | trta_full                 |
| stubType        | Indented                  |

## TitleFootAOM

TitleFootAOM defines the titles and footnotes. Values between curly brackets are environment variables and values derived at output production.

3

| TitleFootAOM |          |         |           |                                                                                 |
|--------------|----------|---------|-----------|---------------------------------------------------------------------------------|
| titleFootID  | type     | ordinal | alignment | text                                                                            |
| dm_t01_t1    | Title    | 1       | Center    | {OutputType}{OutputNumber}: Summary of Demographic and Baseline Characteristics |
| dm_t01_fl    | Footnote | 1       | Left      | [1] Age is imputed when full date of birth is not provided.                     |

## ColumnGroupAOM and ColumnAOM

ColumnGroupAOM and ColumnAOM define the columns in the output. These can be stub columns (the leftmost columns that identify the rows) and result columns (the rightmost columns that contain the actual results of the analyses).

| ColumnGroupAOM |                        | ColumnAOM |         |        |        |                 |        |           |                   |       |           |       |       |
|----------------|------------------------|-----------|---------|--------|--------|-----------------|--------|-----------|-------------------|-------|-----------|-------|-------|
| columnGroupID  | columnGroupLabel       | columnID  | ordinal | type   | subSet | variables       | pageBy | header    | proportionalWidth | sort  | breakLine | print | group |
| 4              | Demography Table stubs | dm_t01_c1 | 1       | Stub   |        | groupLabel      | FALSE  |           | 20                | FALSE | TRUE      | TRUE  | TRUE  |
|                |                        | dm_t01_c2 | 2       | Stub   |        | sex * Statistic | FALSE  |           | 20                | TRUE  | FALSE     | TRUE  | FALSE |
| 5              | Treatment results      | trt_r1    | 17      | Result | trta   |                 | FALSE  | trta_full | 10                | FALSE | FALSE     | TRUE  | FALSE |
|                |                        | trt_r2    | 18      | Result | trtb   |                 | FALSE  | trta_full | 10                | FALSE | FALSE     | TRUE  | FALSE |
|                |                        | trt_r3    | 19      | Result | trttot |                 | FALSE  | trta_full | 10                | FALSE | FALSE     | TRUE  | FALSE |

## SectionAOM

SectionAOM defines the different sections in the output, what analyses to include and how to format them using FormatGroupAOM.

| SectionAOM |              |                |          |                     |                               |              |      |
|------------|--------------|----------------|----------|---------------------|-------------------------------|--------------|------|
| sectionID  | sectionLabel | ordinal        | analyses | groupLabel          | resultFormat                  | suppressZero |      |
| 6          | dm_t01_sex   | DM sex counts  | 1        | dm_sex_any * dm_sex | Sex                           | freqpercent  | TRUE |
| 7          | dm_t01_aage1 | DM AAGE1 stats | 2        | dm_aage1            | Age (YEARS) at First Dose [1] | sumstat      | TRUE |

## FormatGroupAOM and FormatAOM

FormatGroupAOM and FormatAOM define formatting rules for results. Values like +1 for decimals mean that the statistic should be printed with one decimal more than the input precision.

| FormatGroupAOM |                      | FormatAOM  |            |             |            |                      |          |
|----------------|----------------------|------------|------------|-------------|------------|----------------------|----------|
| formatGroupID  | formatGroupLabel     | formatID   | formatCode | formatLabel | subSet     | format               | decimals |
| 8              | Count and percentage | count      | N          | n           | count      |                      | 0        |
|                |                      | percent0   | Percent    | %           | percent0   | 0                    |          |
|                |                      | percent1   | Percent    | %           | percent1   | <1%                  |          |
|                |                      | percent100 | Percent    | %           | percent100 | {N}.0 ({Percent}.0%) |          |
|                |                      | percent99  | Percent    | %           | percent99  | >99%                 |          |
|                |                      | percent    | Percent    | %           | percent    | {N}.0 ({Percent}.0%) |          |
| 9              | Summary statistics   | n          | N          | n           |            |                      | 0        |
|                |                      | mean       | Mean       | Mean        |            |                      | +1       |
|                |                      | sd         | SD         | SD          |            |                      | +2       |
|                |                      | median     | Median     | Median      |            |                      | +1       |
|                |                      | min        | Min        | Min.        |            |                      | +0       |
|                |                      | max        | Max        | Max.        |            |                      | +0       |

## ANALYSIS RESULTS METADATA

The AOM contains the definition of analyses and is spread across multiple sets of metadata. In addition to the examples below, the model also contains RoleARM, which defines additional roles for variables in the analysis.

### MainARM

MainARM defines analyses. Besides analysis concepts this is the only place with direct references to input data sets and variables. It refers to other MainARM records when using another analysis as denominator. It also links to duplications that need to be applied prior to the analysis.

| MainARM |            |                              |            |              |        |             |        |                      |                  |                 |
|---------|------------|------------------------------|------------|--------------|--------|-------------|--------|----------------------|------------------|-----------------|
|         | analysisID | analysisLabel                | denomID    | duplications | domain | analysisSet | subSet | method               | analysisVariable | byVariables     |
| 11      | as_saf     | Analysis Set - Safety counts |            | trta_tot     | ADSL   | saf         |        | countdistinctsubject |                  | trta_full       |
| 12      | dm_sex_any | DM any sex counts            |            | trta_tot     | ADSL   | saf         | sex    | countdistinctsubject |                  | trta_full       |
| 13      | dm_sex     | DM sex counts                | dm_sex_any | trta_tot     | ADSL   | saf         | sex    | countdistinctsubject |                  | trta_full * sex |
| 14      | dm_aage1   | DM AAGE1 stats               |            | trta_tot     | ADSL   | saf         |        | summarystat          | AAGE1            | trta_full       |

## DuplicationARM and DuplicationItemARM

DuplicationARM and DuplicationItemARM define duplications that need to be applied prior to the analysis. In this case only duplication to create the total column is defined.

| DuplicationARM |                  |                          | DuplicationItemARM |                    |                          |
|----------------|------------------|--------------------------|--------------------|--------------------|--------------------------|
| duplicationID  | duplicationLabel | variableGroupID          | duplicationItemID  | inputCodeListItems | duplicatedCodeListItemID |
| 10             | trta_tot         | Actual Treatment (total) | trttot             | trta * trtb        | trttot                   |

## ANALYSIS RESULTS DATA

### MainARD

MainARD contains the actual results of the analyses in a normalized structure. It is not meant to be human-readable nor is it meant to be consumed without the corresponding ARM that defines the analyses.

| MainARD |            |             |            |            |            |           |             |
|---------|------------|-------------|------------|------------|------------|-----------|-------------|
|         | analysisID | qualifierID | role       | variable   | valueNum   | valueCode | valueDecode |
| 11      | as_saf     | 1           | ByVariable | TRTAN:TRTA | 1          | TRTA      | Treatment A |
|         | as_saf     | 1.01        | Statistic  |            | xxx        | N         |             |
|         | ...        | ...         | ...        | ...        | ...        | ...       | ...         |
| 12      | dm_sex_any | 1           | ByVariable | TRTAN:TRTA | 1          | TRTA      | Treatment A |
|         | dm_sex_any | 1.01        | Statistic  |            | xxx        | N         |             |
|         | ...        | ...         | ...        | ...        | ...        | ...       | ...         |
| 13      | dm_sex     | 1           | ByVariable | TRTAN:TRTA | 1          | TRTA      | Treatment A |
|         | dm_sex     | 1           | ByVariable | SEXN:SEX   | 1          | F         | Female      |
|         | dm_sex     | 1.01        | Statistic  |            | xxx        | N         |             |
|         | dm_sex     | 1.02        | Statistic  |            | xxx        | Denom     |             |
|         | dm_sex     | 1.03        | Statistic  |            | xx.xxxxxxx | Percent   |             |
|         | ...        | ...         | ...        | ...        | ...        | ...       | ...         |
|         | ...        | ...         | ...        | ...        | ...        | ...       | ...         |
| 14      | dm_age     | 1           | ByVariable | TRTAN:TRTA | 1          | TRTA      | Treatment A |
|         | dm_age     | 1.01        | Statistic  | AAGEI      | xxx        | N         |             |
|         | dm_age     | 1.02        | Statistic  | AAGEI      | xx.xxxxxxx | Mean      |             |
|         | dm_age     | 1.03        | Statistic  | AAGEI      | xx.xxxxxxx | SD        |             |
|         | dm_age     | 1.04        | Statistic  | AAGEI      | xx         | Min       |             |
|         | dm_age     | 1.05        | Statistic  | AAGEI      | xx.x       | Median    |             |
|         | dm_age     | 1.06        | Statistic  | AAGEI      | xx         | Max       |             |
| ...     | ...        | ...         | ...        | ...        | ...        | ...       |             |

## ANALYSIS CONCEPTS

Analysis Concepts define metadata that is reused across ARM and AOM. Besides MainARM this is the only place where direct references to input data sets and variables are included.

### VariableGroup

VariableGroup links variables and their codelist. In addition, it defines whether at the time of analysis a full matrix should be created for that variable based on the possible values and not the values present in the data.

| VariableGroup   |                         |             |              |            |            |
|-----------------|-------------------------|-------------|--------------|------------|------------|
| variableGroupID | variableGroupLabel      | variableNum | variableCode | codeListID | fullMatrix |
| trta_full       | Actual Treatment (full) | TRTAN       | TRTA         | trt        | TRUE       |
| sex             | Sex                     | SEXN        | SEX          | sex        | TRUE       |

### CodeList and CodeListItem

CodeList and CodeListItem define codelists. It explicitly links the triplicate of numeric, code and decode and includes values not present in the input data (e.g. aggregate values like total treatment).

| CodeList   |               | CodeListItem   |          |           |             |
|------------|---------------|----------------|----------|-----------|-------------|
| codeListID | codeListLabel | codeListItemID | valueNum | valueCode | valueDecode |
| trt        | Treatment     | trta           | 1        | TRTA      | Treatment A |
|            |               | trtb           | 2        | TRTB      | Treatment B |
|            |               | trttot         | 9        | TRTTOT    | Total       |
| sex        | Sex           | sexm           | 1        | F         | Female      |
|            |               | sexf           | 2        | M         | Male        |

## Precision

Precision defines the input precision of continuous input variables. For BDS data this uses information from WhereClause to store precision per parameter.

| Precision         |       |
|-------------------|-------|
| domain            | ADSL  |
| variable          | AAGE1 |
| preferredDecimals | 0     |

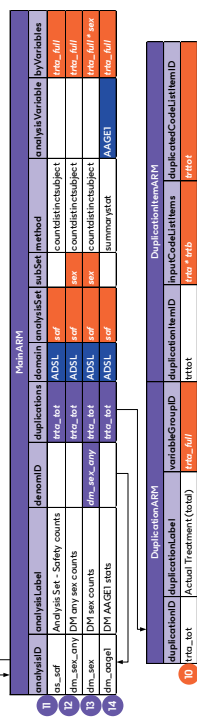
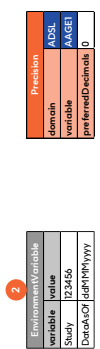
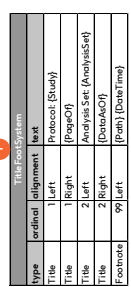
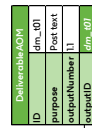
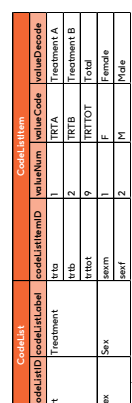
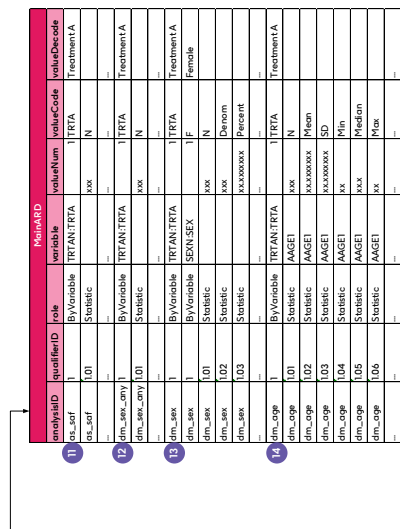
## WhereClause and RangeCheck

WhereClause and RangeCheck define how to subset the data. This includes analysis set definitions. One or more range checks are combined into a where clause using AND as separator.

| WhereClause   |                  |                 | RangeCheck   |                 |                  |
|---------------|------------------|-----------------|--------------|-----------------|------------------|
| whereClauseID | whereClauseLabel | whereClauseRole | rangeCheckID | rangeCheckLabel | rangeCheck       |
| saf           | Safety           | AnalysisSet     | saf          | Safety          | SAFFL EQ "Y"     |
| sex           | Sex              | SubSet          | sex          | Sex             | SEX IN ("M","F") |
| trta          | Treatment A      | SubSet          | trta         | Treatment A     | TRTAN EQ 1       |
| trtb          | Treatment B      | SubSet          | trtb         | Treatment B     | TRTAN EQ 2       |
| trttot        | Total            | SubSet          | trttot       | Total           | TRTAN EQ 9       |
| count         | No denominator   | SubSet          | count        | No denominator  | Denom EQ null    |
| percent0      | Count 0          | SubSet          | percent0     | Count 0         | N EQ 0           |
| percentl      | Percent lt 1     | SubSet          | percentltl   | Percent lt 1    | Percent LT 1     |
| percent100    | Percent 100      | SubSet          | percent100   | Percent 100     | Percent EQ 100   |
| percent99     | Percent gt 99    | SubSet          | percentgt99  | Percent gt 99   | Percent GT 99    |
|               |                  |                 | percentlt100 | Percent lt 100  | Percent LT 100   |
| percent       | Percent 1-99     | SubSet          | percentge1   | Percent ge 1    | Percent GE 1     |
|               |                  |                 | percentle99  | Percent le 99   | Percent LE 99    |

## EVERYTHING COMBINED

How it all fits together is shown on the next page (see the published poster as well).

[illegible]

| VariableGroup   |                         |             |              |            |            |
|-----------------|-------------------------|-------------|--------------|------------|------------|
| variableGroupID | variableGroupLabel      | variableNum | variableCode | codeListID | fullMatrix |
| tra_full        | Actual Treatment (full) | TREAT       | TREA         | tra        | TRUE       |
| sex             | Sex                     | SEXN        | SEX          | sex        | TRUE       |

## **ACKNOWLEDGMENTS**

This metadata model is designed by the author, but heavily relies on previous and ongoing input by and discussions with the following people (in alphabetical order): Alexey Kuznetsov, Dean Grundy, Fabian Noelle, Grzegorz Lach, Jeremy Teoh, Roza Antonyan, Stephen Law, and Warwick Benger.

## **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Edwin van Stein

Company: GSK

Email: [edwin.j.stein@gsk.com](mailto:edwin.j.stein@gsk.com)

LinkedIn: [www.linkedin.com/in/ejvanstein](https://www.linkedin.com/in/ejvanstein)

Brand and product names are trademarks of their respective companies.