

# yspec: an R package to create and deploy data set specification objects in a modeling and simulation workflow

R/Pharma 2020

Kyle Baron, PharmD, PhD

2020-10-14



# Motivation

Data set documentation is ...

1. really important
2. takes a lot of work and time

Need to create: `define.pdf` (analysis data sets)

Want to track:

- ▶ column name and order
- ▶ data source
- ▶ data type
- ▶ scope
  - ▶ range for continuous data
  - ▶ levels for discrete data + decode
- ▶ other attributes - convenience

# Motivation (2)

- ▶ programmatic handling of data attributes
  - ▶ avoid MS Word
- ▶ specify data set attributes once (only once)
  - ▶ re-use in different contexts
- ▶ easy data entry process
  - ▶ human readable format

# How yspec works

- ▶ Use yaml markup language to declare data column names and attributes
- ▶ Read yaml code from file into an object (list) in R
- ▶ Validate data contents
  - ▶ enforce standards and requirements (column names  $\leq 8$  characters)
- ▶ Extract column info
  - ▶ maximize data reuse
- ▶ Template `define.pdf` document
  - ▶ options for customizing format for different use contexts

# Get started

```
spec <- ys_load("nmdata.yml")
```

## Meta data for nmdata data set

```
SETUP__:  
  description: "Example PopPK analysis data set"  
  sponsor: "Tru Valu Pharmaceuticals"  
  projectnumber: "TVP11020F"  
  use_internal_db: true  
  character_last: true  
  comment_col: "C"
```

## Column data (same file as meta data)

```
C:  
  short: "comment character"  
  values: {comment: "C", non-comment: "."}  
NUM:  
USUBJID:  
  short: "unique subject identifier"  
  type: "character"  
TIME:  
  unit: "hour"  
SEQ:  
  short: "data type"  
  values: {observation: 0, dose: 1}  
WT:  
  range: [40, 140]
```

# Access

```
spec <- ys_load("nmdata.yml")
```

```
spec$WT
```

```
.  name  value  
.  col   WT  
.  type  numeric  
.  short subject weight at baseline  
.  unit  kg  
.  range 40 to 140
```

## Access (2)

```
spec$RF
```

```
.  name  value  
.  col   RF  
.  type  character  
.  short renal function stage  
.  value 0 : normal function  
.         1 : mild impairment  
.         2 : moderate impairment  
.         3 : severe impairment
```

Make factor columns in the data set from values + decode (labels)

```
data <- ys_add_factors(data, spec)
```



## Access (3)

spec is a list that we can work on like any other list

```
spec$TIME$unit
```

```
. [1] "hour"
```

```
labels <- map(spec, ys_get_label)
```

```
names(spec)
```

```
summary(spec)
```

```
as.list(spec)
```

```
spec[c("WT", "ALB", "ACTARM")]
```

```
ys_select(spec, WT, ALB, ACTARM)
```

# Create define document

1. `define.pdf` for regulatory submission
2. `mydataset.pdf` for working use

```
ys_document(  
  spec,  
  title = "Data Definitions",  
  author = "Kyle Baron",  
  sponsor = "Tru Valu Pharmaceuticals",  
  type = "regulatory"  
)
```

# Data Definitions

Kyle T. Baron, Pharm.D., Ph.D.

2020-10-05

## 1 Datasets

Description	Location
Example PopPK analysis data set	nmdata.xpt
Example AE analysis data set	ae.xpt

### 1.1 Example PopPK analysis data set (nmdata.xpt)

VARIABLE	LABEL	TYPE	CODES
C	comment character	character	C = comment, . = non-comment
NUM	record number	numeric	
ID	subject identifier	numeric	
TIME	time after first dose (unit: hour)	numeric	
SEQ	data type	numeric	0 = observation, 1 = dose
CMT	compartment number	numeric	
EVID	event ID	numeric	values: 0, 1
AMT	dose amount (unit: mg)	numeric	
DV	dependent variable	numeric	
AGE	age (unit: years)	numeric	
WT	subject weight at baseline (unit: kg)	numeric	

# Thank you

- ▶ R/Pharma Organizers
- ▶ MetrumRG
  - ▶ Bill Knebel and Marc Gastonguay
  - ▶ M&S scientists at MetrumRG