



# Open-source software for regulatory submissions and regulatory environments.

Paul Schuette Ph.D.

Deputy Director, Division of Analytics and Informatics  
Office of Biostatistics, Office of Translational Sciences  
Center for Drug Evaluation and Research  
Food and Drug Administration

# Disclaimer



This presentation reflects the views of the author and should not be construed to represent the FDA's views or policies.

# Statistical Software Clarifying Statement



<https://www.fda.gov/media/109552/download>

FDA does not require use of any specific software for statistical analyses, and statistical software is not explicitly discussed in Title 21 of the Code of Federal Regulations [e.g., in 21CFR part 11]. However, the software package(s) used for statistical analyses should be fully documented in the submission, including version and build identification.

# Statistical Software Continued

As noted in the FDA guidance, *E9 Statistical Principles for Clinical Trials*, “The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available.”

Sponsors are encouraged to consult with FDA review teams and especially with FDA statisticians regarding the choice and suitability of statistical software packages at an early stage in the product development process.

May 6, 2015



# Which programs should be submitted?

“4.1.2.10 Software Programs ([Study Data Technical Conformance Guide](#)) (SDTCG)

Sponsors should provide the source code used to create all ADaM datasets, tables, and figures associated with primary and secondary efficacy analyses. Sponsors should submit source code in single byte ASCII text format. Files with MS Windows executable extensions (.cmd, .com, and .exe) should NOT be submitted. **For a list of acceptable file extensions, refer to the document entitled Specifications for File Format Types Using eCTD Specifications.**

Furthermore, sponsors should submit the source code used to generate additional information included in Section 14 CLINICAL STUDIES of the Prescribing Information,<sup>32</sup> if applicable. **The specific software utilized (version and operating system) should be specified in the ADRG.”**

# Which file formats are accepted ?



- A text file does not have to be a .txt file.
- For CDER, as of 2021, .c, .cpp, .m, .mat, .rmd, .py, .jl are accepted in M5, while .sas and .r are acceptable in M3-M5.

<https://www.fda.gov/media/85816/download>, *Specifications for File Format Types Using eCTD Specifications*

# Caveats from the SDTCG

## 3.3.5 Special Characters: Variables and Datasets

“Variable names, as well as variable and dataset labels should include American Standard Code for Information Interchange (ASCII) text codes only. Variable values are the most broadly compatible with software and operating systems when they are restricted to ASCII text codes (printable values below 128). Use UTF-8 for extending character sets; however, the use of extended mappings is not recommended. Transcoding errors, variable length errors, and lack of software support for multi byte UTF-8 encodings can result in incorrect character display and variable value truncations. Ensure that LBSTRESC and controlled terminology extensions in LBTEST do not contain byte values 160-191 as some character mappings in that range may interfere with agency processes.”

# R Programming Language



- Based on S, which was originally developed at Bell Labs, by Chambers et al, 1976
- R, Ihaka and Gentleman, 1993
- Comprehensive R Archive Network (CRAN) for R packages, GitHub, Bioconductor, etc.
- RStudio is a widely used IDE, (2011)
- Governance: R Core Team, R Foundation





# R Consortium Efforts

- R Consortium (<https://www.r-consortium.org/>)
  - Supports the R Foundation, members include pharma, tech companies, and ASA
- R Consortium Working Groups
  - R Validation Hub
  - R Repositories
  - R Tables for Regulatory Submissions
  - R Submissions

# R Submissions WG

## Pilots:

- Pilot 1: R based submission of tables, graphs and analyses for CDER using ADaM datasets. (completed)
- Pilot 2: CDER submission with an interactive Shiny component (completed)
- Pilot 3: R based CDER submission, derivation of ADaM datasets from SDTM. (pending)
- Pilot 4: R based CDER submission with a container or WebAssembly (Wasm) component (in development)

<https://github.com/RConsortium/submissions-wg>

Data: [CDISC Pilot Submission](#)

# Demographics Table, Pilot 1

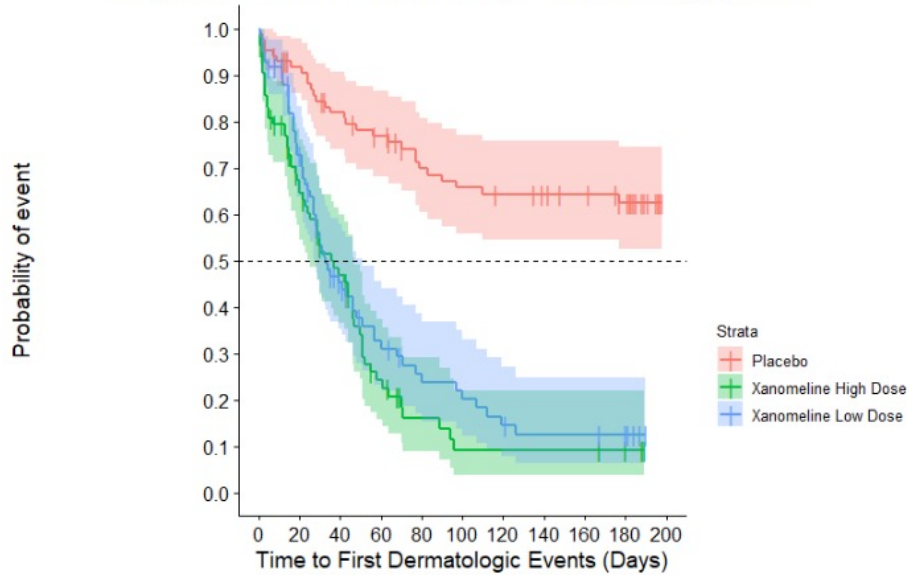
**Table 14-2.01 Summary of Demographic and Baseline Characteristics**

	Placebo N=86	Xanomeline Low Dose N=84	Xanomeline High Dose N=84
<b>Age</b>			
Mean (SD)	75.21 (8.59)	75.67 (8.29)	74.38 (7.89)
Median	76.00	77.50	76.00
Min, Max	52.0, 89.0	51.0, 88.0	56.0, 88.0
<b>Pooled Age Group 1</b>			
<65	14 (16)	8 (10)	11 (13)
65-80	42 (49)	47 (56)	55 (65)
>80	30 (35)	29 (35)	18 (21)
<b>Race</b>			
White	78 (91)	78 (93)	74 (88)
Black or African American	8 (9)	6 (7)	9 (11)
American Indian or Alaska Native	0 (0)	0 (0)	1 (1)
<b>Baseline Height (cm)</b>			
Mean (SD)	162.57 (11.52)	163.43 (10.42)	165.82 (10.13)
Median	162.60	162.60	165.10
Min, Max	137.2, 185.4	135.9, 195.6	146.1, 190.5
<b>Baseline Weight (kg)</b>			
N	86	83	84
Mean (SD)	62.76 (12.77)	67.28 (14.12)	70.00 (14.65)
Median	60.55	64.90	69.20
Min, Max	34.0, 86.2	45.4, 106.1	41.7, 108.0
Missing	0	1	0
<b>Baseline BMI (kg/m<sup>2</sup>)</b>			
N	86	83	84
Mean (SD)	23.64 (3.67)	25.06 (4.27)	25.35 (4.16)
Median	23.40	24.30	24.80
Min, Max	15.1, 33.3	17.7, 40.1	13.7, 34.5
Missing	0	1	0
<b>MMSE Total</b>			
Mean (SD)	18.05 (4.27)	17.87 (4.22)	18.51 (4.16)
Median	19.50	18.00	20.00
Min, Max	10.0, 23.0	10.0, 24.0	10.0, 24.0

Source: adsl.xpt  
2021-12-01 13:26:33

# Pilot 1, KM Plot

KM plot for Time to First Dermatologic Event: Safety population

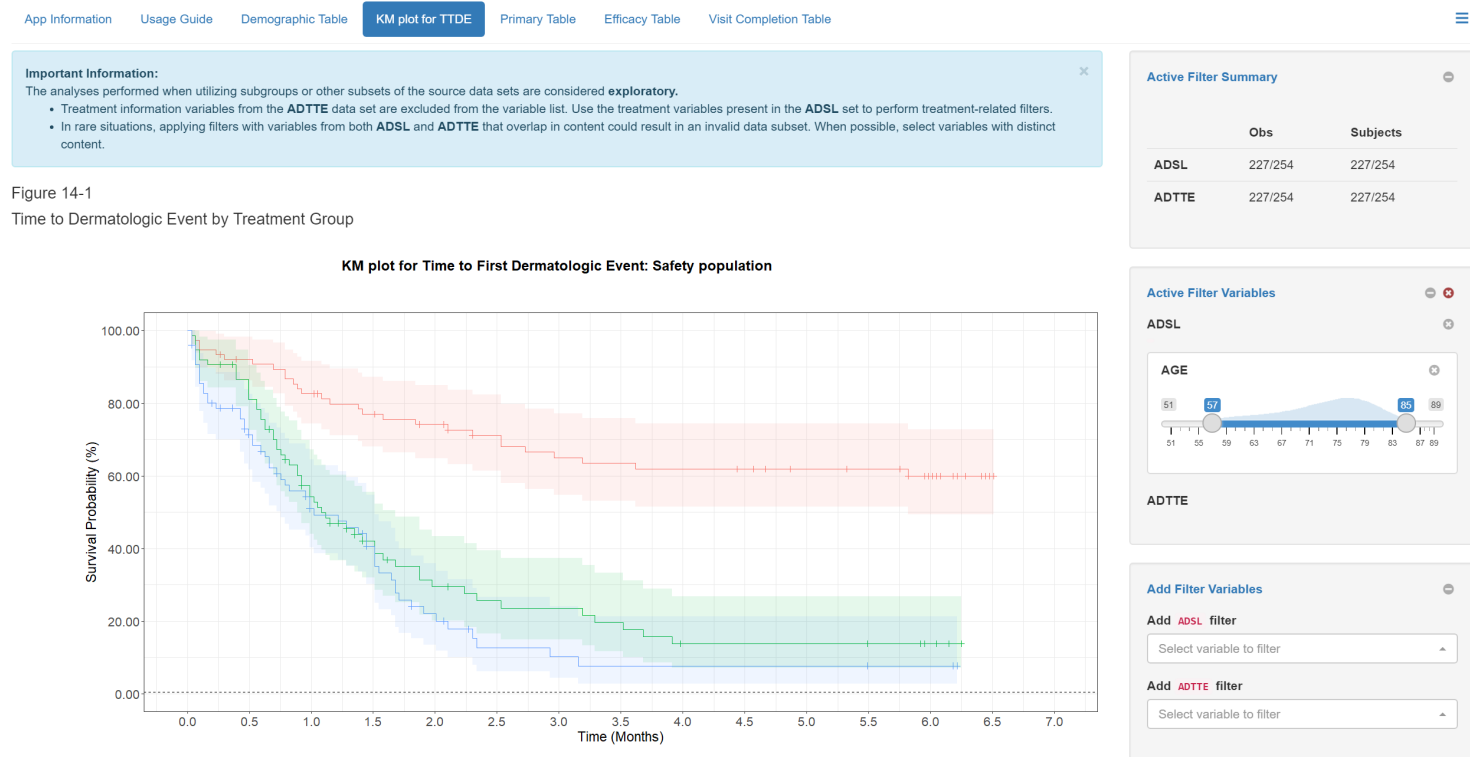


At risk

Placebo	86	75	65	59	50	47	45	42	40	35	0
Xanomeline High Dose	84	48	31	14	7	4	4	4	4	3	0
Xanomeline Low Dose	84	58	31	20	14	12	8	6	6	5	0

# Pilot 2, Interactive KM Plot

## Pilot 2 Shiny Application





# Other Groups Supporting R

- R in Pharma conference, Oct 24-26, 2023, <https://rinpharma.com/>
- R in Medicine conference, June
- Pharmaverse, <https://pharmaverse.org/>
- useR! The R User Conference, July 8-11, 2024
- R Foundation

# What about Python?

- Natural Language Processing (NLP)
- Text modelling, OCR, machine learning
- Data Science applications
- Less focus on statistics and visualization
- Some commercial packages use Python
- Shiny apps for Python in development

# Challenges with Open-Source



- Rate of change
- “Who are you going to call?”
- Support by IT (open-source isn’t always free)
- Plethora of Packages
  - Which ones can be trusted?
  - Dependencies and version control issues





# Reflections & Observations

- CDER has yet to experience a completely R based submission
- Hybrid submissions and hybrid workflows are reported
- Recent graduates tend to have more experience with open-source tools than proprietary alternatives.
- Recommend that sponsors reach out to review division(s)  
Consider the perspective of FDA staff with your submissions

# Reflections & Observations 2

- Open-source tools are popular for AI/ML.
- Digital Health Technologies (DHTs), while not specifically open-source, have the potential to be a disruptive technology.
- “With great power comes great responsibility” (Stan Lee, 1962). Open-source tools and methods don’t necessarily promote good statistical practices. Avoid p-hacking and cherry picking.



## Contact Info:

[Paul.Schuetter@fda.hhs.gov](mailto:Paul.Schuetter@fda.hhs.gov)



**U.S. FOOD & DRUG**  
ADMINISTRATION