Open Source Software in Pharmaceutical Research

Gregory R. Warnes[1 2 3] James A. Rogers[2] Max Kuhn[2]

[1] Center for Biodefense Immune Modeling, University of Rochester
[2] Department of Biostatistics and Computational Biology, University of Rochester
[3] REvolution Computing, Inc.
[4] Research Statistics, Pfizer, Inc.

Joint Statistical Meetings, Seattle, WA, Aug 6-9, 2006

Abstract

Open-Source statistical software is being used with increasing frequency for the analysis of pharmaceutical data, particularly in support of “omics” technologies within discovery. While it is relatively straightforward to employ open-source tools for basic research, software used in any regulatory context must meet more rigorous requirements for documentation, training, software life-cycle management, and technical support.

We will focus on R, a full-featured open-source statistical software package. We’ll briefly outline the benefits it provides, as seen from the perspective of a discovery statistician, show some example areas in which it may be used, and then discuss the documentation, training, and support required for this class of use.

Next we will discuss what is needed for organizations to be comfortable with employing open-source statistical software for regulatory use within clinical, safety, or manufacturing. We will then talk about how well or

Outline

1. Introduction
2. Requirements
3. What is R?
4. Status of R
5. Moving Forward
6. News Flash!
7. More Information

Introduction

Open-Source statistical software is being used with increasing frequency for the analysis of pharmaceutical data, particularly in support of “omics” technologies within discovery. While it is relatively straightforward to employ open-source tools for basic research, software used in any regulatory context must meet more rigorous requirements for documentation, training, software life-cycle management, and technical support.

Requirements

Software used in mission critical and regulated contexts must exhibit 7 key attributes:

1. Functional
2. Verifiable
3. Repeatable
4. Documentable
5. Auditable
6. Stable
7. Supported

Requirements: Details (I)

Functional Performs the required tasks

Verifiable Demonstrate that computer output is correct, or at least consistent..

Repeatable Given the same data, the same results can be obtained, potentially much later in time.

Documentable Documentation is available or can easily be generated for the entire software life-cycle: Specification, Design, Development. Testing, Deployment, Change Management

Requirements: Details (I)

Functional Performs the required tasks

Verifiable Demonstrate that computer output is correct, or at least consistent..

Repeatable Given the same data, the same results can be obtained, potentially much later in time.

Documentable Documentation is available or can easily be generated for the entire software life-cycle: Specification, Design, Development. Testing, Deployment, Change Management

Requirements: Details (I)

Functional Performs the required tasks

Verifiable Demonstrate that computer output is correct, or at least consistent..

Repeatable Given the same data, the same results can be obtained, potentially much later in time.

Documentable Documentation is available or can easily be generated for the entire software life-cycle: Specification, Design, Development. Testing, Deployment, Change Management

Requirements: Details (I)

Functional Performs the required tasks

Verifiable Demonstrate that computer output is correct, or at least consistent..

Repeatable Given the same data, the same results can be obtained, potentially much later in time.

Documentable Documentation is available or can easily be generated for the entire software life-cycle: Specification, Design, Development. Testing, Deployment, Change Management

Requirements: Details (II)

Auditable	Track everything done to data and the system
Stable	Doesn't change too fast, so that there is enough time to develop required documentation
Supported	Guaranteed (by $$) availability of external expense for installation, problem resolution, bug fixes, feature development, training, application development, consulting

Requirements: Details (II)

Auditable Track everything done to data and the system

Stable Doesn’t change too fast, so that there is enough time to develop required documentation

Supported Guaranteed (by $$) availability of external expense for installation, problem resolution, bug fixes, feature development, training, application development, consulting

Requirements: Details (II)

Auditable	Track everything done to data and the system
Stable	Doesn't change too fast, so that there is enough time to develop required documentation
Supported	Guaranteed (by $$) availability of external expense for installation, problem resolution, bug fixes, feature development, training, application development, consulting

What is R?

- System for statistical computing and graphics
- Language is very similar to the S-Plus
- Full featured support for statistical and graphical techniques:
    - linear and nonlinear modeling,
    - classical statistical tests,
    - time-series analysis,
    - classification,
    - clustering
    - ...
- Highly extensible with good development tools
- *Huge* library of user-contributed add-on packages: $>$ 850 !
- Source code is freely available

What is R?

- System for statistical computing and graphics
- Language is very similar to the S-Plus
- Full featured support for statistical and graphical techniques:
  - linear and nonlinear modeling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering
  - ...
- Highly extensible with good development tools
- *Huge* library of user-contributed add-on packages: $>$ 850 !
- Source code is freely available

What is R?

- System for statistical computing and graphics
- Language is very similar to the S-Plus
- Full featured support for statistical and graphical techniques:
    - linear and nonlinear modeling,
    - classical statistical tests,
    - time-series analysis,
    - classification,
    - clustering
    - ...
- Highly extensible with good development tools
- *Huge* library of user-contributed add-on packages: $>$ 850 !
- Source code is freely available

What is R?

- System for statistical computing and graphics
- Language is very similar to the S-Plus
- Full featured support for statistical and graphical techniques:
    - linear and nonlinear modeling,
    - classical statistical tests,
    - time-series analysis,
    - classification,
    - clustering
    - ...
- Highly extensible with good development tools
- *Huge* library of user-contributed add-on packages: $>$ 850 !
- Source code is freely available

What is R?

- System for statistical computing and graphics
- Language is very similar to the S-Plus
- Full featured support for statistical and graphical techniques:
    - linear and nonlinear modeling,
    - classical statistical tests,
    - time-series analysis,
    - classification,
    - clustering
    - ...
- Highly extensible with good development tools
- *Huge* library of user-contributed add-on packages: $>$ 850 !
- Source code is freely available

What is R?

- System for statistical computing and graphics
- Language is very similar to the S-Plus
- Full featured support for statistical and graphical techniques:
  - linear and nonlinear modeling,
  - classical statistical tests,
  - time-series analysis,
  - classification,
  - clustering
  - ...
- Highly extensible with good development tools
- *Huge* library of user-contributed add-on packages: $>$ 850 !
- Source code is freely available

Status of R (I)

Functional +++ This is R's strength. Largely provided by the $>$ 850 user-supplied add-on packages. R currently provides more functionality than any other statistical software system and is growing rapidly.

Verifiable — Most of the functionality of R comes from user-developed add-on packages ($>$ 850!), but there is currently no formal mechanism for evaluating the level of quality of these packages (e.g.: development, test, production, peer reviewed, validated) or documentation that they accomplish the required tasks.

Repeatable — Currently, add on packages do not display version information when loaded, making it difficult to know what versions were utilized for a given analysis, and thus impossible to reliably replicated.

Status of R (I)

Functional +++ This is R's strength. Largely provided by the $>$ 850 user-supplied add-on packages. R currently provides more functionality than any other statistical software system and is growing rapidly.

Verifiable — Most of the functionality of R comes from user-developed add-on packages ($>$ 850!), but there is currently no formal mechanism for evaluating the level of quality of these packages (e.g.: development, test, production, peer reviewed, validated) or documentation that they accomplish the required tasks.

Repeatable — Currently, add on packages do not display version information when loaded, making it difficult to know what versions were utilized for a given analysis, and thus impossible to reliably replicated.

Status of R (I)

Functional +++ This is R's strength. Largely provided by the $>$ 850 user-supplied add-on packages. R currently provides more functionality than any other statistical software system and is growing rapidly.

Verifiable — Most of the functionality of R comes from user-developed add-on packages ($>$ 850!), but there is currently no formal mechanism for evaluating the level of quality of these packages (e.g.: development, test, production, peer reviewed, validated) or documentation that they accomplish the required tasks.

Repeatable — Currently, add on packages do not display version information when loaded, making it difficult to know what versions were utilized for a given analysis, and thus impossible to reliably replicated.

Status of R (II)

Documentable — While the R core team has a well defined and managed process for design, development, testing, release, and change management, no formal documentation of this process appears to exists (aside from the specifications of the language itself). No centrally defined or managed process appears to exist for add-on packages.

Auditable — R has no built-in no audit log, either for data analysis steps or for changes to the system (e.g.: package updates, patches)

Status of R (II)

Documentable — While the R core team has a well defined and managed process for design, development, testing, release, and change management, no formal documentation of this process appears to exists (aside from the specifications of the language itself). No centrally defined or managed process appears to exist for add-on packages.

Auditable — R has no built-in no audit log, either for data analysis steps or for changes to the system (e.g.: package updates, patches)

Status of R (III)

Stable — The R core team releases minor (major.minor.patch) versions twice a year. Since bug fixes are currently applied only to the latest released version of the system, it is difficult to properly support embedded and validated systems where one may need to resolve bugs in R, but must constrain the R version to remain constant for long periods due to the burden of documentation and testing that must be performed.

Supported — While there is an increasingly large pool of statisticians and statistical consulting groups that have R expertise, no organization formally supports R at this time.

Status of R (III)

Stable — The R core team releases minor (major.minor.patch) versions twice a year. Since bug fixes are currently applied only to the latest released version of the system, it is difficult to properly support embedded and validated systems where one may need to resolve bugs in R, but must constrain the R version to remain constant for long periods due to the burden of documentation and testing that must be performed.

Supported — While there is an increasingly large pool of statisticians and statistical consulting groups that have R expertise, no organization formally supports R at this time.

Moving Forward (I)

Functional Already a strength. Continue!

Verifiable RForge proposal

1. Develop a SourceForge-like system for contributed packages;
2. Support package status categories, including clear standards
    - development,
    - testing,
    - production, or
    - peer-reviewed/validated.

Repeatable Display versions of packages on load

Moving Forward (I)

Functional Already a strength. Continue!

Verifiable RFORGE proposal

1. Develop a SourceForge-like system for contributed packages:
2. Support package status categories, including clear standards
   - development,
   - testing,
   - production, or
   - peer-reviewed/validated.

Repeatable Display versions of packages on load

Moving Forward (I)

Functional Already a strength. Continue!

Verifiable RFORGE proposal

1. Develop a SourceForge-like system for contributed packages:
2. Support package status categories, including clear standards
   - development,
   - testing,
   - production, or
   - peer-reviewed/validated.

Repeatable Display versions of packages on load

Moving Forward (I)

Functional Already a strength. Continue!

Verifiable RFORGE proposal

1. Develop a SourceForge-like system for contributed packages:
2. Support package status categories, including clear standards
    - development,
    - testing,
    - production, or
    - peer-reviewed/validated.

Repeatable Display versions of packages on load

Moving Forward (I)

Functional Already a strength. Continue!

Verifiable RFORGE proposal

1. Develop a SourceForge-like system for contributed packages:
2. Support package status categories, including clear standards
   - development,
   - testing,
   - production, or
   - peer-reviewed/validated.

Repeatable Display versions of packages on load

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

Moving Forward (II)

Documentable

1. Formally document the development process used for R
2. Provide tools to perform and document this process for add-on packages
3. Develop validation templates for use by organizations
4. Encourage commercial vendors to support R and to provide additional validation effort and associated documentation.

Auditable Add an audit-log facility

Stable Establish a system for back-porting bug fixes to previous versions.

Supported Encourage commercial vendors to formally support R.

News Flash!: RPRO from *REvolution Computing*

2006-08-01 **New Haven, CT**: *REvolution Computing* announces the immediate availability of RPRO, an enterprise-strength statistical computing environment providing the strengths of the open source R statistical software system from the R-Project coupled with the enterprise-level support and high-performance computing expertise of *REvolution Computing*.

Additions to R:

- Technical Support
- Simple Installation and Maintenance
- Performance Tuning
- Documentation and Training
- Validation Materials
- Consulting and Services

One Century Tower • 265 Church Street, Suite 1006
New Haven, CT 06510 • 203.777.7442
www.revolution-computing.com

News Flash!: NetWorkSpaces from *REvolution Computing*

2006-08-01 **New Haven, CT**: *REvolution Computing* announces the immediate availability of NetWorkSpaces for RPro (NWS). NWS enables calculations to be automatically distributed across multiple processors in clusters. Distributing the data and/or work across multiple processors permits a dramatic decrease in time to completion of large computational tasks or permits a dramatic increase in those calculations size, length or complexity. NWS fully supports Microsoft Windows Compute Cluster Server 2003 (CCS), which provides a security enhanced and affordable high performance computing solution.

Microsoft Windows
Compute Cluster Server 2003

REvolution
computing

www.microsoft.com/hpc
Sales:
hpinfo@microsoft.com

One Century Tower • 265 Church Street, Suite 1006
New Haven, CT 06510 • 203.777.7442
www.revolution-computing.com

Contact Information

- Personal:

  Email greg@warnes.net
  Web http://www.warnes.net/Research

- University of Rochester:

  Email warnes@bst.rochester.edu
  Web http://www.urmc.rochester.edu/smd/biostat

- REvolution Computing:

  Email greg@revolution-computing.com
  Web http://www.revolution-computing.com