R: Regulatory Compliance and Validation Issues
A Guidance Document for the Use of R in Regulated Clinical Trial Environments

October 18, 2021

Contents

Acknowledgements

The contributions of Marc Schwartz, Frank Harrell, Jr., Anthony Rossini and Ian Francis, who wrote and updated the initial versions of this document, are gratefully acknowledged.

1 Introduction

The far-reaching domain of clinical trials for pharmaceuticals and medical devices ranges from initial research and discovery to post-regulatory approval surveillance. These studies and clinical trials are conducted by manufacturers, academic and commercial research organizations and individual clinical investigators.

Activities surrounding human clinical trials must follow regulations specified by governmental, quasi-governmental, and harmonization-oriented agencies. These regulations are put in place to protect current and future participants with respect to safety and privacy as well as to drive honest decision making based on the study results by preserving scientific integrity. The particular practice, interpretation, or implementation of these regulations is driven by the characterization of the intended population and by the intended use for the devices or pharmaceuticals. The spectrum of guidelines apply to myriad aspects of these studies, including clinical practices, manufacturing standards, and decision-making guidance.

As these documents are not prescriptive, the entities engaged in these activities will interpret the guidance provided with some level of variation and will impose their own internal operational requirements. These requirements will be based upon prior experience, the nature of the organization, internal business practices and external audits of processes and documentation.

Key guidance documents are put forth by two principal regulatory entities. The United States Food and Drug Administration (hereafter referred to as the FDA) and the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals in Human Use (hereafter referred to as the ICH). Similar governmental and regulatory bodies in the international community, such as EMEA (European Medicines Agency) and PMDA (the Japanese Pharmaceuticals and Medical Devices Agency) oversee activities within their respective domains, but are heavily influenced by the standards promulgated by the FDA and ICH. Thus, the content of this document is largely influenced by the regulatory guidance provided under the imprimatur of these two bodies.

The use of statistical software for the analysis and presentation of data collected in the course of these regulated activities is itself regulated, also to varying levels. There are several documents that are relevant to this particular domain.

First, applicable documents collectively referred to as GxP:

- 21 CFR Part 11 - Electronic Records; Electronic Signatures
- Guidance for Industry: Part 11, Electronic Records; Electronic Signatures - Scope and Application
- 21 CFR Part 58 - Good Laboratory Practice (GLP)
- 21 CFR Part 312 - Good Clinical Practice (GCP)
- 21 CFR Part 210 - Current Good Manufacturing Practice (cGMP)
- ICH E6 - Good Clinical Practice Consolidated Guideline

Second, principal software guidance documents:

- Guidance for Industry - Computerized Systems Used in Clinical Investigations (2007)
- General Principles of Software Validation; Final Guidance for Industry and FDA Staff (2002)

Third, principal statistical guideline documents:

- ICH E9 - Statistical Principles for Clinical Trials
- Guidance for Industry and FDA Staff - Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (2010)

Finally, in May of 2015, the FDA published a "Statistical Software Clarifying Statement", which contained the following text:

"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.

As noted in the FDA guidance, E9 Statistical Principles for Clinical Trials (available at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm), "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."

It should be noted that the above list, related to the use of statistical software for the analysis and presentation of data is not complete. It should also be noted that there is a level of overlap amongst these documents, such that specific topics may be covered in more than one of them. It is left for the reader to determine whether or not other such documents and/or regulations are applicable within specific domains. Additional information may be found at the FDA Web Site.

This document will address specific areas within the GxP domain. It is intended to provide a reasonable consensus position on the part of the R Foundation for Statistical Computing (hereafter referred to as the R Foundation) relative to the use of R within these regulated environments and to provide a common foundation for end users to meet their own internal standard operating procedures, documentation requirements and regulatory obligations.

The R Foundation for Statistical Computing makes no warranties, expressed or implied, in this document.

The R Foundation For Statistical Computing

2 The Scope of this Document

It is important to clarify that this document is SOLELY applicable to R software that is part of the official R distribution, as formally released by the R Foundation. This software is commonly referred to as "Base R plus Recommended Packages" and is released in both source code and binary executable forms under the Free Software Foundation's GNU Public License (hereafter referred to as the GPL).

As of this writing, "Base R" includes the following packages:

- base
- compiler
- datasets
- graphics
- grDevices
- grid
- methods
- parallel
- splines
- stats
- stats4
- tcltk
- tools
- utils

and the "Recommended Packages" includes the following packages:

- boot
- class
- cluster
- codetools
- foreign
- KernSmooth
- lattice
- MASS
- Matrix
- mgcv

- nlme

- nnet

- rpart

- spatial

- survival

This document is NOT in any fashion, applicable to other R-related software and add-on packages made available via other parties, such as users or even members of the R Development Core Team, who may, from time to time, make their software available via the Comprehensive R Archive Network (CRAN) or other software distribution repositories and vehicles.

It is important to note that there is a significant obligation on the part of the end-user's organization to define, create, implement and enforce R installation, validation and utilization related Standard Operating Procedures (SOPs) within the end-user's environment. These SOPs should define appropriate and reasonable quality control processes to manage end-user related risk within the applicable operating framework. The details and content of any such SOPs are beyond the scope of this document.

This document is not intended to be prescriptive, does not render a legal opinion and does not confer or impart any binding or other legal obligation. It should be utilized by the reader and his or her organization as one component in the process of making informed decisions as to how best to meet relevant obligations within their own professional working environment.

The R Foundation for Statistical Computing makes no warranties, expressed or implied, in this document.

The R Foundation For Statistical Computing

3 The R Foundation For Statistical Computing

The R Foundation is a not-for-profit organization working in the public interest. It was founded by the members of the R Development Core Team in order to:

- Provide support for the R project and other innovations in statistical computing. We believe that R has become a mature and valuable tool and we would like to ensure its continued development and the development of future innovations in software for statistical and computational research.

- Provide a reference point for individuals, institutions or commercial enterprises that want to support or interact with the R development community.

- Hold and administer the copyright of R software and documentation.

R is an official part of the Free Software Foundation's GNU project, and the R Foundation has similar goals to other open source software foundations, such as the Apache Foundation and the GNOME Foundation.

Among the goals of the R Foundation are the support of continued development of R, the exploration of new methodology, teaching and training for statistical computing and the organization of meetings and conferences with a statistical computing orientation.

The R Foundation is seated in Vienna, Austria and currently hosted by the Vienna University of Technology. It is a registered association under Austrian law and active worldwide. The R Foundation can be contacted at:

The R Foundation for Statistical Computing
c/o Institute for Statistics and Mathematics
Wirtschaftsuniversität Wien
Welthandelsplatz 1
1020 Vienna, Austria

Tel: (+43 1) 31336 4754
Fax: (+43 1) 31336 904754
Email: R-foundation-board@R-project.org

The R Foundation Statutes are available from the Foundation's web site:

http://www.r-project.org/foundation/

4   What is R?

Introduction to R

R is a language and environment for statistical computing and graphics. It is a GNU project and is similar to the S language and environment that was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and his colleagues[2]. R can be considered as a distinct implementation of S, developed separately from the original implementation at Bell Laboratories. Although there are some important differences between these two implementations, much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is readily extensible. The S language is often the vehicle of choice for research in statistical methodology and R provides an open source route to participation in that activity.

One of R's strengths is the ease with which well designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- an effective data handling and storage facility,

- a suite of operators for calculations on arrays, in particular matrices,

- a large, coherent, integrated collection of intermediate tools for data analysis,

- graphical facilities for data analysis and display either on-screen or on hardcopy, and

- a well developed, simple and effective programming language that includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term "environment" is intended to characterize R as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R, like S, is designed around a true programming language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of the S language, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are a number of

---

[2] See references [Becker et al.(1988)Becker, Chambers, and Wilks], [Chambers and Hastie(1992)] and [Chambers(1998)]

packages listed previously supplied with the R distribution and many more, covering a very wide range of modern statistics, are available through the CRAN family of Internet sites.

R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

In addition, as R is open source, the availability of R's source code provides for superior and thorough documentation of R's functionality and designed behavior and is open to inspection by all users.

5 Qualification and Validation of Systems for 21 CFR Part 11 Compliance

5.1 Part 11: Electronic Records, Electronic Signatures

21 CFR Part 11 was issued by the FDA to provide regulatory requirements for processes and controls that must be applied to electronic *records* and electronic signatures. The FDA released Guidance for Industry: Part 11, Electronic Records; Electronic Signatures – Scope and Application to clarify their position with respect to the scope and application of the regulation.

The guidance can be summarized as follows:

- If there are *predicate rules that require records* to be maintained, and these *records* are managed electronically, then Part 11 controls apply to those *records*.

- If there are *predicate rules that require a signature* to be applied, and this signature is applied electronically or digitally, then Part 11 controls apply.

A predicate rule is an underlying regulation that describes the regulatory requirements for business processes, *records* and signatures. Examples of predicate regulations are:

- 21 CFR Part 58: Good laboratory practice for nonclinical laboratory studies
- 21 CFR Part 211: Current Good Manufacturing Practice For Finished Pharmaceuticals
- 21 CFR Part 312: Investigational New Drug Application
- 21 CFR Part 314: Applications for FDA approval to market a new drug
- 21 CFR Part 601: Licensing
- 21 CFR Part 606: Current good manufacturing practice for blood and blood components
- 21 CFR Part 820: Quality system regulation
- 21 CFR Part 814: Premarket approval of medical devices

Therefore, an assessment of Part 11 applicability should be performed when determining the requirement to validate a system. However, even if an organization determines that a system is not Part 11 relevant, a system may still be GxP relevant and still warrant validation (e.g. 211.68(b) and 820.70(i)). This is driven by the need for basic data integrity of GxP critical data (prior to an electronic *record* being "created").

Note that the revised definition of Part 11 relevant *records* states (emphasis by the authors):

"Records submitted to FDA, under predicate rules in electronic format [are Part 11 records]. However, a record that is *not itself submitted*, but is used in *generating a submission*, is not a part 11 record unless it is otherwise required to be maintained under a predicate rule and it is maintained in electronic format."

Therefore, it is not mandated that 21 CFR Part 11 is appropriate to data analysis software systems that are not primarily intended for storage and transmission of electronic medical *records*. It remains the responsibility of an individual organization however to define the applicability of Part 11 and validation to their systems.

For readers who agree that Part 11 does not apply to these types of systems, this document still serves the purpose of providing a high degree of confidence that R can comply with these and other validation regulations.

5.2 Validation

The term "validation" is interpreted in different ways in different fields. The FDA clearly defined the term in guidance[3] and it should be noted that validation is more than a verification or testing exercise.

Validation is defined by the FDA as: "Establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes."[4]

The FDA explains that validation encompasses the overall program and is designed to assure quality and consistency for a process/product throughout its lifecycle. In contrast, verification is an activity performed during and/or between phases of the overall lifecycle. Software testing is one form of verification.

Qualification can be seen as a phase of verification and/or testing within an overall validation program.

The purpose of this document is to demonstrate that R, when used in a qualified fashion, can support the appropriate regulatory requirements for validated systems, thus ensuring that resulting electronic *records* are "trustworthy, reliable and generally equivalent to paper records."

It is crucial to note that many validation requirements, as described in the following pages, may be met by the operational characteristics of software systems (i.e. operating systems and database applications) and other technologies or processes outside of R itself, where R will be used as a component in an overall data management, analysis and presentation process.

---

[3] General Principles of Software Validation; Final Guidance for Industry and FDA Staff
[4] Glossary Of Computerized System and Software Development Terminology

6 Software Development Life Cycle (SDLC)

6.1 Operational Overview

The development, release and maintenance of R is, broadly, a collaborative process involving the R Development Core Team (hereafter referred to as R Core). Members of R Core represent multiple statistical disciplines and are based at academic, not-for-profit and industry-affiliated institutions on multiple continents.

Most communications amongst the members of R Core take place electronically via e-mail and similar means. A non-public e-mail list (r-core) provides a common forum for discussions amongst the members of R Core. An archive of the list is available to facilitate R Core in documenting and reviewing these discussions, as they pertain to development decisions and related issues.

R Core does meet, collectively and/or in smaller groups, with a level of frequency dictated by multiple factors, including taking advantage of regularly scheduled conferences where members of R Core may already be in attendance. Such conferences include those that are specific to statistical computing and R itself (http://www.r-project.org/conferences.html). These routine communications and meetings ensure that the collaborative efforts are appropriately coordinated and prioritized as ongoing development takes place.

Reasonable software development and testing methodologies are employed by R Core in order to maximize the accuracy, reliability and consistency of R's performance. While some aspects of R's development are handled collaboratively, others are handled by members of the team with specific interests and expertise in focused areas.

Importantly, as R is released under the terms of the GPL, all of the source code underlying R, whether it be in R, C or FORTRAN, is available for peer review by all members of the R user community. Thus, all of the functionality embodied within R is subject to continuous critique and improvement relative to its accuracy, reliability and consistency.

The size of the R user community (difficult to define precisely, because there are no sales transactions, but conservatively estimated as being in the tens of thousands, with some independent estimates in the hundreds of thousands), provides for extensive review of source code and testing in "real world" settings outside the confines of the formalized testing performed by R Core. This is a key distinction, related to product quality, between R and similar software that is only available to end users in a binary, executable format. In conjunction with detailed documentation and references provided to end users, the size of the R user community, all having full access to the source code, enables a superior ability to anticipate and verify R's performance and the results produced by R.

Additional documentation regarding the activities of R Core as they pertain to development, goals and related activities, including coding guidelines, are available for review:

- R Developer Page (http://developer.r-project.org/)
- R Internals – A Guide to the Internal Structures of R and Coding Standards for the R Core Team (http://cran.r-project.org/doc/manuals/R-ints.html)

6.2 Source Code Management

All of R's source code is managed in a source code version control repository based on Subversion. The R Subversion Repository is access controlled, such that only members of R Core have write access to the

source code tree. Various security, access control and archival procedures are in place to provide reasonable protection and to maintain the integrity of the hosting server and the source code management system.

Separate source code branches for version control are maintained by R Core. The current Release Branch and the ongoing Development Version are kept in separate branches to facilitate non-conflicting source code management. The Release Branch is designed for bug fixes and allows only minor feature enhancements. Major features are introduced in the Development Version, from which a new Release Branch is made prior to the next x.y.0 release.

Daily logs of code changes are maintained within the Subversion repository and reflect all aspects of code changes made by R Core. These logs are available for public review as http://developer.r-project.org/R_svnlog_YYYY, where 'YYYY' is a placeholder for a four-digit year specification (e.g. 2017).

In addition, a "NEWS" file is actively maintained by R Core to make it easier for users to track changes made to past, present and future versions of R. The current version of this file is available for public viewing at https://svn.r-project.org/R/trunk/doc/NEWS.Rd. This file is also included in all source code and binary executable versions of R to enable end users to review and gain insight into the ongoing changes to R via the `news()` function.

The typical format of the NEWS file contains detailed, version-specific information on:

- Significant User-Visible Changes
- New Features
- Graphics Devices
- Installation
- Package Installation
- Utilities
- Deprecated and Defunct
- C-Level Facilities
- Bug Fixes

The entire list (and any additions) may or may not be present for each R version as appropriate.

Further, older versions of the NEWS file are available as https://svn.r-project.org/R/trunk/doc/NEWS.3, https://svn.r-project.org/R/trunk/doc/NEWS.2, https://svn.r-project.org/R/trunk/doc/NEWS.1 and https://svn.r-project.org/R/trunk/doc/NEWS.0. These files enable R users to gain insight into the full history of R's ongoing development, back to version 0.50, which was released in 1997.

6.3   Testing and Validation

Within the R Core development related documents, as identified in the aforementioned references (see Section 6.1), guidelines are provided relative to modifications to source code, regression tests, validation tests and similar issues. These guidelines are in place to maximize code quality and to facilitate ongoing code validation during development and during the "run-up" to each version release.

A set of validation tests are maintained and upgraded by R Core to enable the testing of source code against known data and known results. Any errors noted during this testing are resolved prior to release.

The tests are located in the "tests" sub-directory of the extracted source code tarball. A README file is also available in that directory to describe the procedures to run the tests and various options related to selecting all tests or only a subset of the tests to run. The source code and expected results for these tests are available for review and use in other applications as may be appropriate.

These tests are also available to end users and/or system administrators and can be run as part of their installation process to provide further documentation and objective evidence as to the accuracy, reliability and consistency of their installation of R.

As with any statistical software, the user should take care to consider the appropriateness of any R software, and the statistical methods implemented in the software, to the intended application. The potential exists in any statistical software for the lack of consistency and reliability in results due to the inappropriate application of statistical methodologies. Reasonable judgment in this regard should be rendered by users with appropriate expertise.

The entire R source code tree is available to end users (either via the Subversion repository or via source code archive files, known as "tarballs", that are automatically created with daily updates). Additional testing is solicited from the user community during so-called "Alpha", "Beta" and "Release Candidate" testing cycles. Progressively stronger restrictions are imposed on modifications to the source code during the testing cycles to minimize the risk of unexpected side effects. This provides further opportunities to identify and resolve issues that may have been missed during the development process, such as "boundary" issues that may represent unusual or atypical circumstances, including unique operating system and/or hardware configurations.

Feedback from the community is facilitated by the use of the r-devel e-mail list [5] and via the R Bug Tracking System [6]. This open and public process enables a wider array of code testing and further increases the likelihood of resolving issues prior to the release of a stable version of R.

6.4 Release Cycles

Once the in-development version of R has been approved for release by R Core's designated Release Manager, a public announcement is made via the R e-mail lists to the user community.

Source code archive files ("tarballs") are made available via the CRAN mirror infrastructure.

Pre-built executable binary install files follow and are made available for common operating system and CPU architectures. These include Linux, Windows and MacOS platforms.

R's major release cycles are generally predictable. Prior to R version 3.0.0, x.y.0 releases occurred on or about April 1 and October 1 of each calendar year. Effective with R version 3.0.0, x.y.0 releases will occur on or about April 1 of each calendar year.

x.y.z versions, so called patch releases are made available when required in order to fix issues discovered in the current release. Generally one or two of these releases have been made before the semi-annual release of the next x.y.0 version and this number may increase with the introduction of the annual release cycle.

Additional instructions regarding the utilization of R source code, installation requirements, compilation and

---

[5] https://stat.ethz.ch/mailman/listinfo/r-devel
[6] http://bugs.r-project.org/

platform and operating system related issues are extensively documented in the R Installation and Administration Manual, which is available with source code and binary executables and online at http://cran.r-project.org/manuals.html.

6.5 Availability of Current and Historical Archive Versions

Current and historical versions of R are available in source code tarballs from the main CRAN server (http://cran.r-project.org/src/base/) and its worldwide mirrors (http://cran.r-project.org/mirrors.html).

Pre-built executable binary install files for current versions are made available for common operating system and CPU architectures. These include Linux, Windows and MacOS platforms and are available from the main CRAN server binary tree (http://cran.r-project.org/bin/) and from its worldwide mirrors as referenced above.

6.6 Maintenance, Support and Retirement

Each Released Version of R is actively supported by R Core with respect to bug reporting, fixes and patches. Patched versions are made available, generally as source code only, to end users to facilitate their installation of these. Binary executable installation files for the patched Release Versions are made available at the discretion of the individual maintainers of the platform specific versions.

Source code tarballs of the daily incremental patched versions of each current R release are made available to the community via an FTP server (ftp://ftp.stat.math.ethz.ch/Software/R/) for download to enable R users to update their systems between formal releases as their local needs may dictate.

In addition, users with Subversion clients can download the latest copy of the source code tree at any time, via a direct connection to the Subversion server.

As each version of R is released, there are a variety of support resources that are made available to the community of end users.

Extensive documentation is provided by R Core and is available both within the source code and binary executable versions of R as well as online in HTML and PDF formats at http://cran.r-project.org/manuals.html.

Function-specific help is also available within R including, where appropriate, extensive references to algorithms and methods to facilitate the user's comprehension of R's functionality and expected behavior.

R FAQs (Frequently Asked Questions) are also available to facilitate answers to commonly asked end user questions. These are available at:

- The Main R FAQ (http://cran.r-project.org/doc/FAQ/R-FAQ.html)
- R FAQ for Windows (http://cran.r-project.org/bin/windows/base/rw-FAQ.html)
- R FAQ for MacOS (http://cran.r-project.org/bin/macosx/RMacOSX-FAQ.html)

R's Bug Reporting system, available online at http://bugs.r-project.org/, facilitates end user reporting of bugs identified during the course of use. In addition, an internal R function, `bug.report()`, is available to enable end users to generate and send bug reports directly from an interactive R session.

An extensive set of public e-mail lists exist. These are the primary vehicle for interactive support and communications between R Core and the user community. There are two primary lists, called r-devel and r-help.

The former list is principally for issues surrounding R's development and lower level coding issues that are more technical.

The latter list, which is the primary end user support forum, is an active discussion on various R coding or usage issues and related concerns.

Additional e-mail lists focus on specific special interest areas that range from database interfaces to robust statistics and financial modeling.

More information on these e-mail lists is available at http://www.r-project.org/mail.html.

Extensive search facilities, accessible at http://www.r-project.org/search.html, are also available to search the list archives, enabling users to perform keyword-based searches of prior discussions and the online documentation. An internal R function, `RSiteSearch()`, is also available to facilitate such searches during an interactive R session.

The R Journal (ISSN 2073-4859), formerly R News, a peer-reviewed newsletter is available electronically as a periodical from http://journal.r-project.org/. The R Journal provides general information on R and R Core and user contributed articles in specific domains of interest.

A large set of published books, several by members of R Core, are available to support the use of R, both generally and within subject-matter-specific domains. A periodically updated but partial list of these books is available at http://www.r-project.org/doc/bib/R-books.html.

The x.y.0 releases are maintained via a series of x.y.z patch releases. At a new x.y.0 version of R, the prior version is retired from formal support. R Core's efforts are then focused on the new Release (and the ongoing Development) version. No further development, bug fixes or patches are made available for the retired versions. Thus there is always only one current version of R. However, the SVN repository will allow older release branches to be reopened, should the need arise.

6.7   Qualified Personnel

As noted in Section 6.1, members of R Core represent multiple statistical disciplines and are based at academic, not-for-profit and industry-affiliated institutions on multiple continents.

All members of R Core hold Ph.D. and/or Master's degrees (all but one have Ph.D.s) from accredited academic institutions and have published extensively in peer reviewed journals. Several have written books on statistical computing technologies and applications. The members of R Core constitute a widely recognized, international team of experts on statistical computing and software development.

Institutions at which the members of R Core currently hold or have previously held appointments include:

- University of Wisconsin – Madison
- Bell Laboratories
- Copenhagen Business School (Inst. of Finance)
- Fred Hutchinson Cancer Research Center

- Harvard Medical School

- Wirtschaftsuniversität Wien

- University of Auckland

- Czech Technical University in Prague

- Genentech

- Ludwig-Maximilians-Universität – München

- Technische Universität Dortmund

- University of Washington

- Eidgenössische Technische Hochschule Zürich

- Friedrich-Alexander-Universität Erlangen-Nürnberg

- Centre International de Recherche sur le Cancer, Lyon

- Oxford University

- Indian Statistical Institute, Delhi Centre

- University of California – Davis

- University of Iowa

- AT&T Research Labs

6.8   Physical and Logical Security

The R Foundation maintains its key servers within the brick and mortar infrastructure of university-supported computing facilities. In accordance with defined security policies, only personnel with authorized access may enter.

User names and passwords are required by all R Core members to gain access to computing systems for R Foundation-related activities. User accounts are limited in access based upon standard security policies and functional requirements.

Network access is controlled via the use of typical hardware and software controls, including the use of firewalls, security policies and related mechanisms.

6.9   Disaster Recovery

As a result of having R Foundation servers within the confines of university-hosted computing facilities, disaster recovery plans for R Foundation computing systems are in sync with those of the host facilities.

In addition, the worldwide network of CRAN mirrors provides for an alternative means of accessing key components of R, should primary servers be temporarily unavailable.

7 21 CFR Part 11 Compliance Functionality

7.1 Overview

Within the regulated domain, R is intended to be utilized as a component within a larger data management framework, with respect to data acquisition, validation and related source electronic *records* tasks. R's design and development are focused on reporting, by enabling leading edge statistical analysis and presentation, rather than on data management tasks as illustrated by transaction/data processing and related functionality.

To that end, the following sections discuss important components of the 21 CFR Part 11 Regulation, provides the R Foundation's interpretation of each, and discuss how R and/or other enabling technologies, within an overall data management framework, can meet the guidance interpretations.

Note that sections 11.10(a) and (i), pertaining to system validation and qualified personnel, respectively, have already been covered previously in Section 6 and do not appear below.

In the following sections, the term *record* means an electronic record that is interpreted to fall within the remit of Part 11 as defined in FDA Guidance for Industry Part 11, Electronic Records; Electronic Signatures – Scope and Application (2003).

R is not intended to create, maintain, modify or delete Part 11 relevant *records* but to perform calculations and draw graphics.

7.2 11.10(b) The ability to generate accurate and complete copies of records in both human readable and electronic form suitable for inspection, review, and copying

The R Foundation understands this item to mean that any *records* created or maintained in the system must be accurate and complete. These *records* must be available in both human readable and electronic form.

R is not intended to create, maintain, modify or delete Part 11 relevant *records* but to perform calculations and draw graphics. Where R's use may be interpreted as creating *records*, however any such *records* (for example data objects such as vectors, matrices, lists and data frames, and graphics, plots and images) are available to be output in various industry-standard formats. Because R provides for the routine generation of these outputs as standard features, the output is available in both machine- and human-readable formats.

Using these industry-standard formats, the output is available to be read by other products that also utilize these same industry standards and these *records* are therefore readable independent of the use of R.

In conjunction with local policies regarding *record* access control, retention and archival, R meets the FDA requirements for the inspection, review and copying of *records* as defined above.

7.3 11.10(c) Protection of records to enable their accurate and ready retrieval throughout the records retention period

The R Foundation understands this item to mean that all *records* created or maintained in R must be stored in a manner that enables accurate and ready retrieval.

R is not intended to create, maintain, modify or delete Part 11 relevant *records* but to perform calculations and draw graphics.

Records created by R will, therefore, reside within and be managed by a separate host system.

The host system is required to provide for compliance with this part using local policies regarding the retention and archival of such *records* and the mechanisms and access controls in place.

7.4   11.10(d) Limiting system access to authorized individuals

The R Foundation understands this item to mean that access to the computer system that creates, maintains or modifies a *record* is limited to only authorized individuals.

R is an application that runs within the hosting computer environment, which must provide user access controls at hardware and/or operating system levels.

The requirement for this section is typically met via system level functionality and is based on user roles, object level security and related security policies.

Approved users must be supplied unique user account identifiers and passwords, which are required to gain access to the hosting system and thus to R. Upon connection to the hosting system, further access and functional restrictions will be in place to limit the activities in which the user may engage.

These limitations can also limit user access to objects, such as data files and programs, to further constrain the user's activities.

7.5   11.10(e) Use of secure, computer-generated, time-stamped audit trails to independently record the data and time of operator entries and actions that create, modify, or delete electronic records. Record changes shall not obscure previously recorded information. Such audit trail documentation shall be retained for a period at least as long as that required for the subject electronic records and shall be available for agency review and copying

The R Foundation understands this item to mean that the creation, modification or deletion of *records* must have an associated audit trail describing who, when and why an action was performed. Additionally, any such audit trail will be also considered an electronic *record* within the scope of Part 11.

R is not intended to create, maintain, modify or delete Part 11 relevant *records* but to perform calculations and draw graphics.

Where R's use may be interpreted as creating *records*, however, R can support audit trail creation within the *record*.

R includes `date(), Sys.time(), Sys.Date() and Sys.timezone()` functions which enable users to include date and time stamps on report, graphical and other output, thus enabling the use of this information in the tracking of user sessions.

Records created by R necessarily reside within and are managed by a separate host system. Therefore, after *record* creation, any subsequent changes to the *record* must have an audit history imposed by the

host system. This may be implemented technically via system-level logging as a component of the hosting computer system in which R operates.

For session-based logging focusing on data analysis, the organization using R would need to provide extensions using R or other tools to facilitate the generation of a session-based audit trail that meets the local implementation requirements of the organization's quality assurance group. The security and integrity of this log would be ensured through the use of the hosting system's user and object-based security models.

7.6 11.10(f) Use of operational system checks to enforce permitted sequencing of steps and events, as appropriate

The R Foundation understands this item to mean that effective user technology, processes, and interfaces must be in place to reduce errors made by an operator to the extent that system errors can be minimized.

R was designed with an architecture, technology, process and interface that provide for operational system checks for software function or features. Components in R provide for error checking mechanisms to preclude certain actions, which when combined with computer system level functionality, can limit certain user operations.

In conjunction with code reviews and validation conducted by R Core and community peer review (as described elsewhere in this document), these features provide for the use of R in a production environment.

These capabilities are similar to those of any statistical software application and are consistent with the implementation of good analytical practice.

Appropriate coding techniques that implement good and defensive programming style are documented and described in many books, including Software for Data Analysis (Chambers)[7].

7.7 11.10(g) Use of authority checks to ensure that only authorized individuals can use the system, electronically sign a record, access the operation or computer system input or output device, alter a record, or perform the operation at hand

The R Foundation understands this item to mean that the system must provide for authority checks to allow users to perform system operations, such as applying electronic signatures, access to input and output devices, the ability to alter a *record* and perform functions.

Authority checks (such as user name/password controls) must be implemented within the host system, as described in section 11.10(d). This provides for controlled access for authorized users to the R application.

Within R there are no controls to enable/disable user access to particular subsets of functionality. Any restrictive functionality must be implemented within the host system according to relevant business processes and documented operating procedures as defined by the R user's organization.

---

[7] See reference [Chambers(2008)]

7.8   11.10(h) Use of device (e.g., terminal) checks to determine, as appropriate, the validity of the source of data input or operational instruction

The R Foundation understands that these checks are warranted where only certain devices have been selected as legitimate sources of data input or commands. The device checks would be used to determine if the data or command source was authorized. If R is used as a primary-source data management and data entry system, such checks would need to be implemented.

R supports the host environment in providing these capabilities as discussed previously, notably in sections 11.10(d) and 11.10(f).

R is not intended to create, maintain, modify or delete Part 11 relevant *records* but to perform calculations and draw graphics.

7.9   11.10(j) The establishment of, and adherence to, written policies that hold individuals accountable and responsible for actions initiated under their electronic signatures, in order to deter record and signature falsification

The R Foundation understands that individuals must understand their responsibility and accountability when performing actions using their electronic signatures. This must be communicated with documented policies.

R is not intended to create *records* but to perform calculations and draw graphics. Following from this, it is not intended to allow for signature of *records*.

Where R's use may be interpreted as creating *records*, however, a predicate rule requirement to sign these *records* must first be established, since not every *record* requires a signature.

In the event that an electronic signature must be applied to the *record*, this must be achieved using a third-party system. R does not provide functionality for the application of electronic signatures.

7.10   11.10(k) Use of appropriate controls over systems documentation

21 CFR Part 11.10(k) indicates that these controls must include:

- Adequate controls over the distribution of, access to, and use of documentation for system operation and maintenance
- Revision and change control procedures to maintain an audit trail that documents time-sequenced development and modification of systems documentation

The R Foundation understands this item to mean that there must be control over who can access and change system documentation and also that there exists revision and change control in place for system documentation.

All releases of R include documentation covering installation, administration, programming and related user guides. R documentation is created once per Release Version; thus these documents are uniquely identifiable and associated with a specific release of the software.

This documentation is published and maintained by R Core as part of the Software Development Life Cycle (see Section 6) using the Subversion version-control system. This documentation is controlled in the same manner as R source code.

This documentation is provided to R users in both printed and electronic formats.

The maintenance and distribution of this documentation at the R user site is the sole responsibility of the user site and should be handled in accordance with their training and other standard operational procedures.

7.11 Section 11.30 Controls for Open Systems - the system shall employ procedures and controls designed to ensure the authenticity, integrity and as appropriate the confidentiality of electronic records from the point of their creation to the point of their receipt. Additional measures such as document encryption and use of appropriate digital signature standards to ensure, as necessary under the circumstances record authenticity, integrity and confidentiality

R supports the host environment (see previous discussion, particularly section 11.10(d)) that provides these capabilities.

It is the sole responsibility of the R user to ensure that the appropriate safeguards are implemented for a particular hosting system.

8 Bibliography

References

[Becker et al.(1988)Becker, Chambers, and Wilks] R. A. Becker, J. M. Chambers, and A. R. Wilks. *The New S Language*. Chapman & Hall, London, 1988. ISBN 0-534-09193-8.

[Chambers(1998)] J. M. Chambers. *Programming with Data*. Springer, New York, 1998. URL http://cm.bell-labs.com/cm/ms/departments/sia/Sbook/. ISBN 0-387-98503-4.

[Chambers(2008)] J. M. Chambers. *Software for Data Analysis: Programming with R*. Springer, 2008. URL http://stat.stanford.edu/~jmc4/Rbook/. ISBN 978-0-387-75935-7.

[Chambers and Hastie(1992)] J. M. Chambers and T. J. Hastie. *Statistical Models in S*. Chapman & Hall, London, 1992. ISBN 0-534-16764-0.