Assessing Microsoft R Open for implementation in a production environment in the clinical industry

ClinBAY

Christos Stylianou, Ph.D

Jan 2021

ClinBAY Ltd
182 Agias Fylaxeos street
Office 101, Kofteros Business center
3083 Limassol, Cyprus

Table of Contents

Introduction

R is a complete environment for data scientists, and one of the most popular if not the most popular statistics programming language. Despite in its popularity, the adoption of R in the clinical industry has been slower, due to the increased complexity in the regulatory submission requirments. However, due to it being a free software and its openness, it has been gaining momentum and increased popularity within the clinical industry as well, with efforts being initiated by multiple groups, with a range of different scopes, such as generation of validation packages, clinical trial reporting libraries, example code etc. Some good examples are the R/PHARMA conference and the R Validation Hub, a cross-industry initiative whose mission is to enable the use of R by the Bio-Pharmaceutical Industry in a regulatory setting.

Microsoft R Open is an enhanced R Distribution, created from Microsoft. Like R, Microsoft R Open is open source and free to use. However, it is advertised as having the following benefits which could be important for use in a clinical trials environment:

1. Multi-threaded math libraries that bring parallel computations to R: Which could improve performance of both simulations and computations on large datasets.
2. A high-performance default CRAN repository that provide a consistent and static set of packages to all Microsoft R Open users: Which could assist in the IQ/OQ/PQ testing required in our industry.
3. The checkpoint package that makes it easy to share R code and replicate results using specific R package versions: Which could help for consistency when sponsors work with CRO's that could have different versions of packages installed, and even within companies when updates are made.

In this white paper, we attempted to assess benefit “1”, the improved performance of the Microsoft R Open 4.0.2 in comparison to our already validated R 4.0.2. Benefit “2”, due to the current list of packages as of the writing date, in the authors opinion provides little benefit for clinical data analysis, with the sole exception of “genomic data” usually out of scope of the main analysis of clinical trials. Benefit “3”, the checkpoint package is also available on CRAN and can be added to existing validated environments.

Methodology

We provide time comparisons for execution of identical code between Microsoft R Open 4.0.2 and the R 4.0.2 from CRAN. The assessments were done in a Windows 10 64bit environment, which is what our production environment is using. Some basic tests were done in a test Linux environment, with similar results, but due to the complexity of switching between versions in a Linux server a formal comparison was out of scope of this White Paper.

Microsoft R Open is packaged and used along with the Intel Math Kernel Library (MKL) for Windows. The MKL library provides the BLAS and LAPACK library functions used by R, and is supposed to improve the performance of the R code, especially where many vector/matrix operations are used. MKL's default behavior is to use as many parallel threads as there are available cores. Theorically, no change to the R scripts are required to benefit from this multi-thread performance improvement.

The following simulations were performed and compared in the two environments:

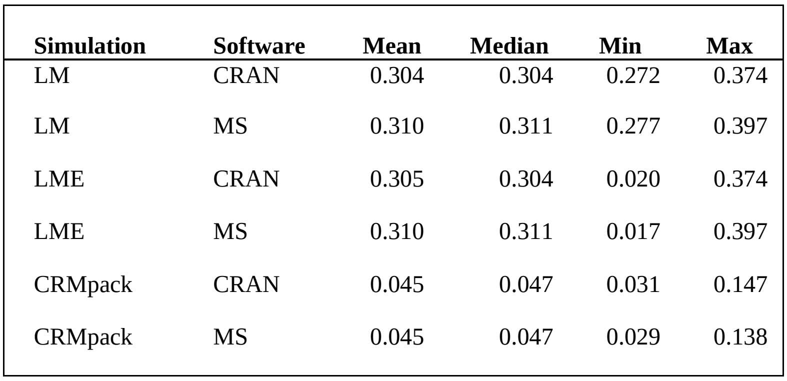
1. Linear model on a large 1 million records dataset.
2. Linear mixed model on a dataset with 13 patients, with 10 timepoints each using the lme4 package.
3. Interaction with external software (Jags), using CRMpack on 21 patients.

Furthermore, simulations were also run using the doParallel package, to assess impact of having multiple packages competing for threads.

Results

Results from the simulation are summarized in Table 1 and visualized in Appendix Figures 1-3, Microsoft Open R is presented as “MS” and the R from CRAN as “CRAN”. Despite expectations of increased performance from Microsoft Open R, both for LM (simulation 1) and LME (simulation 2), both the mean and the median performance of the Microsoft Open R was worse than CRAN R.

Running the simulations with the doParallel package, reduced for LM the total simulation time of 304sec to 96sec for CRAN R, but for the Microsoft Open R total simulation time was reduced from 310sec to 252sec. For LME, the total simulation time of 305sec was reduced to 7sec for CRAN, and for the Microsoft Open R the 310ms were reduced to 7sec.



Simulation	Software	Mean	Median	Min	Max
LM	CRAN	0.304	0.304	0.272	0.374
LM	MS	0.310	0.311	0.277	0.397
LME	CRAN	0.305	0.304	0.020	0.374
LME	MS	0.310	0.311	0.017	0.397
CRMpack	CRAN	0.045	0.047	0.031	0.147
CRMpack	MS	0.045	0.047	0.029	0.138

Table 1: Simulation summary results in seconds

Conclusion

This assessment, investigated potential benefits, of switching from the CRAN maintained by the R foundation of statistical computing, to Microsoft Open R. Despite the advertised gains in the utilization of the multicore architecture, such gains were not observed in real processing time in the scenarios tested.

Investigation of the CPU suggested that only the LM package utilized the multiple threads. The CPU was used at 100% by Open R, compared to 35% usage by R from CRAN, for Simulation 1. For Simulations 2 (LME) and 3 (CRMpack), CPU usage was identical between both versions of R. This was not unexpected, as the author expected that some packages would utilize the improved/threaded functions, but others would not.

In the authors opinion, the 100% CPU usage when Linear Models were executed, also negatively impacted the performance when these models were executed using the doParallel package; leading to approximately a 20% decrease in time for Open R using this package, compared to approximately 70% using the CRAN R. This was not unexpected, as two packages would be competing for the available threads, and was actually the reason the “doParallel” investigation was done, even though it is acknowledged that at the time the Microsoft Open R was expected to outperform CRAN R without the “doParallel” command.

It is unclear to the author why the expected gain in performance was not observed when the linear models (Simulation 1) were executed, alternate packages may have had an increase in performance instead. However, the results have led the author to decide that for now, the implementation of Microsoft Open R in the ClinBAY production enviromets will not be recommended.

Key Takeaways

- Multi-threading could improve run times of large datasets, but in the scenarios tested Microsoft Open R did not improve performance.
- For simulations the performance of packages like “doParallel” are unlikely to be matched, even if performance of individual processes is improved using multi-threading. Simulations should therefore still be run in parallel, rather than relying on multithreading across simulations rather than using multi-threading to improve individual functions. Should mathematical functions that enable multi-threading be used alongside “doParallel”, tests need to be performed to assess the potential impact of having two resources competing for the threads.
- The checkpoint package will be investigated further in its utility as it could guarantee reproducibility between companies and along time.

ClinBAY is a company that provides biometrics solutions to decision makers. Feel free to contact us (info@clinbay.com) if you are interested in our services or products.

.

Appendix

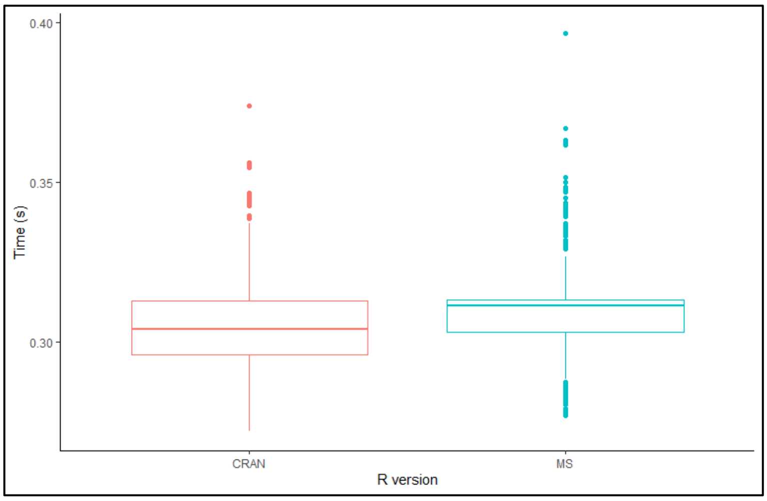


Figure 1: Comparison of LM (Simulation 1)

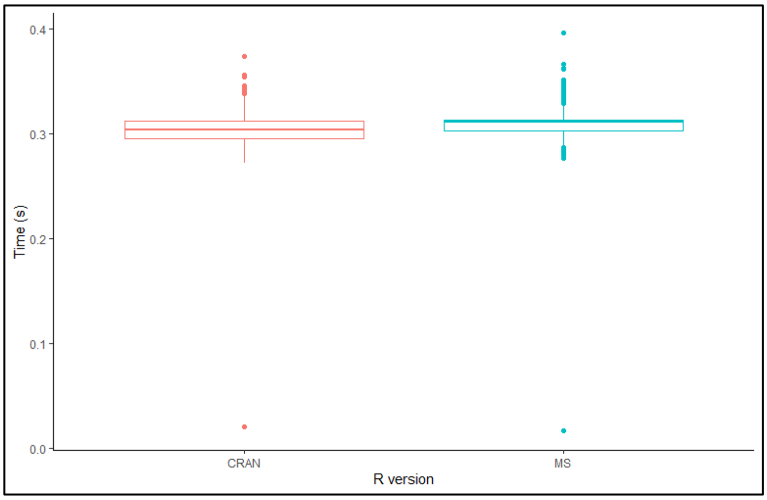


Figure 2: Comparison of LM (Simulation 2)

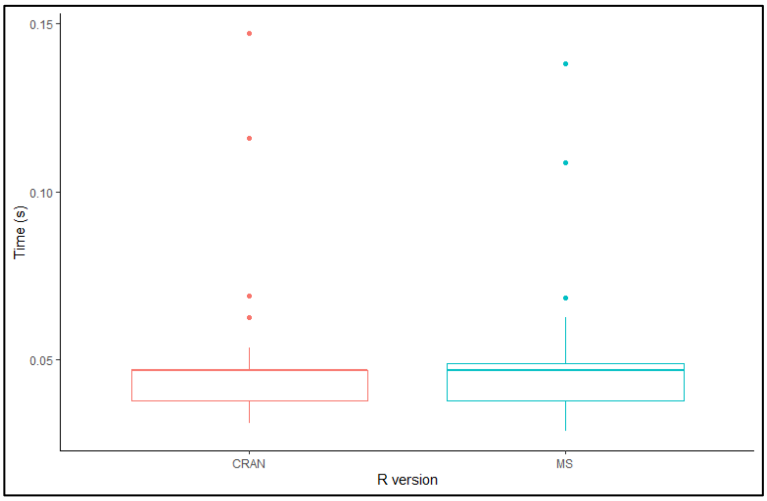


Figure 3: Comparison of CRMPack (Simulation 3)