SOFTWARE TOOL ARTICLE

REVISED R package “QRISK3”: an unofficial research purposed implementation of ClinRisk’s QRISK3 algorithm into R [version 3; peer review: 1 approved, 1 approved with reservations, 1 not approved]

Yan Li ¹, Matthew Sperrin¹, Tjeerd van Staa¹⁻³

¹Health e-Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

²Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

³Alan Turing Institute, British Library, London, UK

v3 First published: 23 Dec 2019, 8:2139
<https://doi.org/10.12688/f1000research.21679.1>
 Second version: 28 Feb 2020, 8:2139
<https://doi.org/10.12688/f1000research.21679.2>
 Latest published: 22 May 2020, 8:2139
<https://doi.org/10.12688/f1000research.21679.3>

Abstract

Cardiovascular disease has been the leading cause of death for decades. Risk prediction models are used to identify high risk patients; the most common model used in the UK is ClinRisk’s QRISK3. In this paper we describe the implementation of the QRISK3 algorithm into an R package. The package was successfully validated by the open sourced QRISK3 algorithm and QRISK3 SAS program. We provide detailed examples of the use of the package, including assigning QRISK3 scores for a large cohort of patients. This R package could help the research community to improve future risk prediction models based on a currently used risk prediction model. The package is available from CRAN: <https://cran.r-project.org/web/packages/QRISK3/index.html>.

Keywords

CVD, risk prediction model, R, QRISK3

This article is included in the **RPackage** gateway.

Open Peer Review

Approval Status

	1	2	3
version 3 (revision) 22 May 2020		view	view
version 2 (revision) 28 Feb 2020	view	view	
version 1 23 Dec 2019	view		

- Mohieddin Jafari** , University of Helsinki, Helsinki, Finland
Ali Amiryousefi, University of Helsinki, Helsinki, Finland
- Angela K. Birnbaum** , University of Minnesota, Minneapolis, USA
Ashwin Karanam , University of Minnesota, Minneapolis, USA
- Fadratul Hafinaz Hassan** , School of

Computer Sciences, Universiti Sains Malaysia,
Pulau Pinang, Malaysia

Any reports and responses or comments on the
article can be found at the end of the article.

Corresponding author: Tjeerd van Staa (tjeerd.vanstaa@manchester.ac.uk)

Author roles: **Li Y:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation; **Sperrin M:** Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **van Staa T:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the China Scholarship Council.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Li Y *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Li Y, Sperrin M and van Staa T. **R package “QRISK3”: an unofficial research purposed implementation of ClinRisk’s QRISK3 algorithm into R [version 3; peer review: 1 approved, 1 approved with reservations, 1 not approved]**

F1000Research 2020, 8:2139 <https://doi.org/10.12688/f1000research.21679.3>

First published: 23 Dec 2019, 8:2139 <https://doi.org/10.12688/f1000research.21679.1>

REVISED Amendments from Version 2

We appreciate the time and effort that reviewers spent on this paper. We edited the paper according to the comments and provided more references to clarify technical details.

Any further responses from the reviewers can be found at the end of the article

Introduction

Cardiovascular disease (CVD) was responsible for 17.9 million deaths in 2016, which represents 31% of overall global deaths, and over 75% of these deaths happened in low/middle-income countries¹. People who are at high risk of CVD need to be identified and treated early¹. Risk prediction models that use risk factors to calculate the probability of patients developing diseases are often used to identify high risk patients². QRISK3 is the most popular risk prediction model for CVD developed in the UK. It calculates risk of patients developing CVD in the next 10 years and has been incorporated into the electronic health records (EHRs) system in the UK in order to detect high risk CVD patients and help clinicians make treatment decisions^{3,4}. NICE guidelines recommend clinicians to consider prescribing statins to patients with a risk over 10% identified from QRISK3⁵. QRISK3 was developed from historical patients' EHR data using Cox proportional hazard model⁶ and has been well validated at population level corresponding to discrimination and calibration^{3,4,7}.

The implementation of QRISK3 into R would not only benefit researchers to improve future risk prediction but also enable them to use QRISK3 scores to identify patients at certain risk levels, e.g. for clinical trial recruitment. There is also scope to improve these risk predictions; it has been found that QRISK3 has uncertainty on individual risk prediction^{7,8} due to unmeasured heterogeneity between practices, which was not captured. A follow-up study suggests that QRISK3 may need to include additional causal risk factors as this uncertainty on individual risk prediction was not related to data quality and variation of association between disease and outcome⁹. The current QRISK3 can only be accessed through an online web calculator or specialised commercial software¹⁰ and its original algorithm was written by C, which is a low level programming language appealing to software engineering rather than data science¹¹. R is the most popular statistical programming language in the data science field due to its great advantage as free and open-source, with fast computing and a well-supported community¹². This paper explains the incorporation of the QRISK3 algorithm into R for ease of research concerning QRISK3 and how the package was developed and validated. The package aims to help researchers to improve risk prediction models and better detect high risk CVD patients.

Methods

Extraction of the QRISK3 algorithm

The original QRISK3 algorithm was written in C by ClinRisk under a GNU Lesser General Public License¹¹. Their previously published QRISK3 paper was used to understand the original algorithm and the associations between variables used in the original algorithm and risk factors of QRISK3³.

Development and validation of the QRISK3 R package

The QRISK3 algorithm was written in both R (3.4.2) and SAS (9.4)¹³ independently, in order to mimic double programming, with a plan to use the SAS implementation to validate the R package. An additional C program, which could directly call the original QRISK3 algorithm to calculate risk, was written for validation. Two validation datasets (QRISK3_2017_test and QRISK3_2019_test) were then created and included in the R package. Dataset QRISK3_2017_test was created by manually recording the calculated QRISK3 risk score from the original QRISK3 algorithm for a group of simulated patients. The simulated patient groups were generated by changing each risk factor sequentially covering the changes of all QRISK3 risk factors. For example, patient 1 in QRISK3_2017_test does not have any positive CVD risk factors, patient 2 is similar to patient 1 except he has atrial fibrillation, patient ID 3 is similar to patient 2 except he is on atypical antipsychotic medication rather than atrial fibrillation and so on until all the change of CVD predictors are covered. Therefore, each patient is similar to the previous patient except the change of one CVD predictor. QRISK3_2019_test was the version recorded using the original QRISK3 algorithm with different value changes for each risk factor. Risk scores of the same simulated patient groups (QRISK3_2017_test and QRISK3_2019_test) was compared among different versions of QRISK3, including QRISK3 R package, QRISK3 SAS program and QRISK3 C function for validation. The R package was created using R CMD tool¹⁴ with several useful online tutorials¹⁵⁻¹⁸.

Implementation

The QRISK3 package can be directly installed from CRAN¹⁹ using "install(QRISK3)" or GitHub respiratory²⁰ with "install_github("YanLiUK/QRISK3")". The package contains one function (QRISK3_2017) to calculate the risk of patients developing CVD in the next 10 years using the QRISK3 algorithm¹¹ and the two datasets for testing.

Variables used by the QRISK3 package were summarised and compared to the original algorithm in [Table 1](#). All variables have the same definition as the QRISK3 paper³, most of variables were coded into numeric variables similar to the original algorithm. The coding of ethnicity and smoking was different from the original algorithm (written in C), as the C index starts from 0 but R's index starts from 1.

Table 1. Description of QRISK3 variables.

Parameters in QRISK3 R package	Meaning of variables	Variables in original algorithm
age	Specify the age of the patient in year (e.g. 64 years-old)	age
atrial_fibrillation	Atrial fibrillation? (0: No, 1: Yes)	b_AF
atypical_antipsy	On atypical antipsychotic medication? (0: No, 1: Yes)	b_atypicalantipsy
regular_steroid_tablets	On regular steroid tablets? (0: No, 1: Yes)	b_corticosteroids
erectile_disfunction	A diagnosis of or treatment for erectile dysfunction? (0: No, 1: Yes)	b_impotence2 (only for men)
migraine	Do patients have migraines? (0: No, 1: Yes)	b_migraine
rheumatoid_arthritis	Rheumatoid arthritis? (0: No, 1: Yes)	b_ra
chronic_kidney_disease	Chronic kidney disease (stage 3, 4 or 5)? (0: No, 1: Yes)	b_renal
severe_mental_illness	Severe mental illness? (0: No, 1: Yes)	b_semi
systemic_lupus_erythematosus	Systemic lupus erythematosus (SLE)? (0: No, 1: Yes)	b_sle
blood_pressure_treatment	On blood pressure treatment? (0: No, 1: Yes)	b_treatedhyp
diabetes1	Diabetes status: type 1? (0: No, 1: Yes)	b_type1
diabetes2	Diabetes status: type 2? (0: No, 1: Yes)	b_type2
weight (kg)	Weight	Not available
height (cm)	Height	Not available
weight (m) / (height (cm) / 100) ²	Body mass index (BMI)	bmi
ethnicity	1 White or not stated 2 Indian 3 Pakistani 4 Bangladeshi 5 Other Asian 6 Black Caribbean 7 Black African 8 Chinese 9 Other ethnic group	ethrisk: 0, --not stated 1, --white 2, --inidan 3, --Pakistani 4,--Bangladeshi 5,--Other Asian 6,--Black Caribbean 7,--Black African 8,--Chinese 9--Other ethnic group
heart_attack_relative	Angina or heart attack in a 1st degree relative < 60? (0: No, 1: Yes)	fh_cvd
cholesterol_HDL_ratio	Total cholesterol/HDL ratio? (range from 1 to 11, e.g. 4)	rati
systolic_blood_pressure	Systolic blood pressure (mmHg, e.g. 180 mmHg)	sbp
std_systolic_blood_pressure	Standard deviation of at least two most recent systolic blood pressure readings(mmHg)	sbps5
smoke	1 non-smoker 2 ex-smoker 3 light smoker (less than 10) 4 moderate smoker (10 to 19) 5 heavy smoker (20 or over)	smoke_cat: 0 non-smoker 1 ex-smoker 2 light smoker (less than 10) 3 moderate smoker (10 to 19) 4 heavy smoker (20 or over)
townsend	Townsend deprivation scores	town

Validation

The two datasets `QRISK3_2017_test` and `QRISK3_2019_test` were used for validation. Risk scores calculated from this QRISK3 package, the original algorithm and the SAS version on the same group of patients were exactly the same. Applying this QRISK3 package to a big CPRD cohorts²¹ with 3.6 million patients⁷ showed a good discrimination (C statistic: 0.85) and calibration²², similar to the original QRISK3³.

Usage and features

A patient cohort with anonymous patient identifiers and CVD risk factors should first be extracted and coded similarly to QRISK3 by the user. Missing values in the dataset should be handled (e.g. multiple imputation) before using this package²³. Column names of CVD risk factors (e.g. "age") should then be specified correctly to the `QRISK3_2017` function. The function returns calculated risk scores through a dataset with three columns, including patient identifier, calculated QRISK3 score and calculated QRISK3 score with one digit. It also reminds users to double check whether the definition of their variables was the same as the definition of QRISK3. The package also automatically detects whether all variables were coded as numeric and whether age of patients was ranged between 25 and 84, if not an error message returns (explained in [Table 2](#)).

Table 2. Description of error message in the QRISK3 R package.

Error message	Conditions	Explanation
"Variables including XXX, XXX must be coded as numeric (0/1) variable."	When at least one of variables in dataset are not numeric	QRISK3 algorithm needs numeric variable (0/1) to calculate risk
"Age of patients must be between 25 and 84."	When at least one patient in the dataset has age below 25 or above 84	QRISK3 algorithm was developed from a population with age between 25 and 84
"Variables including XXX, XXX has missing values."	When at least one of variables in dataset has missing value	Missing values must be handled before using this QRISK3 algorithm

Workflow

1. Set path and read data from CSV file

```
dataPath <- "yourPath"
dataName <- "yourDataName.csv"

setwd(dataPath)
myData <- read.csv(dataName, check.names=FALSE)
```

2. See the data structure and other information

```
#See data structure
str(myData)
```

```
## 'data.frame':    48 obs. of  26 variables:
## $ QRISK_C_algorithm_score : num  17.2 36 21.6 24.1 17.2 19.1 20.9 22.3 19.3 23.5 ...
## $ age                     : int   64 64 64 64 64 64 64 64 64 64 ...
## $ gender                  : num   1 1 1 1 1 1 1 1 1 1 ...
## $ b_AF                    : int   0 1 0 0 0 0 0 0 0 0 ...
## $ b_atypicalantipsy       : int   0 0 1 0 0 0 0 0 0 0 ...
## $ b_corticosteroids       : int   0 0 0 1 0 0 0 0 0 0 ...
## $ b_impotence2           : int   0 0 0 0 1 0 0 0 0 0 ...
```

```
## $ b_migraine      : int  0 0 0 0 0 1 0 0 0 0 ...
## $ b_ra            : int  0 0 0 0 0 0 1 0 0 0 ...
## $ b_renal         : int  0 0 0 0 0 0 0 1 0 0 ...
## $ b_semi         : int  0 0 0 0 0 0 0 0 1 0 ...
## $ b_sle           : int  0 0 0 0 0 0 0 0 0 1 ...
## $ b_treatedhyp    : int  0 0 0 0 0 0 0 0 0 0 ...
## $ b_type1         : int  0 0 0 0 0 0 0 0 0 0 ...
## $ b_type2         : int  0 0 0 0 0 0 0 0 0 0 ...
## $ weight          : int  70 70 70 70 70 70 70 70 70 70 ...
## $ height          : int  180 180 180 180 180 180 180 180 180 180 ...
## $ ethrisk         : int  2 2 2 2 2 2 2 2 2 2 ...
## $ fh_cvd          : int  0 0 0 0 0 0 0 0 0 0 ...
## $ rati            : int  4 4 4 4 4 4 4 4 4 4 ...
## $ sbp             : int  180 180 180 180 180 180 180 180 180 180 ...
## $ sbps5           : int  20 20 20 20 20 20 20 20 20 20 ...
## $ smoke_cat       : int  1 1 1 1 1 1 1 1 1 1 ...
## $ surv            : int  10 10 10 10 10 10 10 10 10 10 ...
## $ town            : int  0 0 0 0 0 0 0 0 0 0 ...
## $ ID              : int  1 2 3 4 5 6 7 8 9 10 ...

#See missing value
# summary(myData)

#If there is any missing value
#please use methods (e.g. multiple imputation) to impute missing value

#Once there is no missing value
#Get all variable names in your data
# colnames(myData)

#Use help of this package to map your variable to QRISK3 variables
# ?QRISK3_2017
```

3. Call the QRISK3 function to calculate risk score

```
test_all_rst <- QRISK3_2017(data= myData, patid="ID", gender="gender", age="age",
atrial_fibrillation="b_AF", atypical_antipsy="b_atypicalantipsy",
regular_steroid_tablets="b_corticosteroids", erectile_disfunction="b_impotence2",
migraine="b_migraine", rheumatoid_arthritis="b_ra",
chronic_kidney_disease="b_renal", severe_mental_illness="b_semi",
systemic_lupus_erythematosus="b_sle",
blood_pressure_treatment="b_treatedhyp", diabetes1="b_type1",
diabetes2="b_type2", weight="weight", height="height",
ethniciy="ethrisk", heart_attack_relative="fh_cvd",
cholesterol_HDL_ratio="rati", systolic_blood_pressure="sbp",
std_systolic_blood_pressure="sbps5", smoke="smoke_cat", townsend="town")

##
## This R package was based on open-sourced original QRISK3-2017 algorithm.
## <https://qrisk.org/three/src.php> Copyright 2017 ClinRisk Ltd.
##
## The risk score calculated from this R package can only be used for research purpose.
```

```
##
## Please refer to QRISK3 website for more information
## <https://qrisk.org/three/index.php>
##
## Important: Please double check whether your variables are coded the same
## as the QRISK3 calculator
##
## Height should have unit as (cm)
## Weight should have unit as (kg)
##
## Ethnicity should be coded as:
##
## Ethnicity_category Ethnicity
## 1 White or not stated      1
## 2 Indian                  2
## 3 Pakistani               3
## 4 Bangladeshi             4
## 5 Other Asian             5
## 6 Black Caribbean         6
##
## Smoke should be coded as:
##
## Smoke_category Smoke
## 1 non-smoker              1
## 2 ex-smoker               2
## 3 light smoker (less than 10) 3
## 4 moderate smoker (10 to 19) 4
## 5 heavy smoker (20 or over)  5
##
## The head of result in all patients is:
##
## ID QRISK3_2017 QRISK3_2017_1digit
## 1 1 17.22985 17.2
## 2 2 17.89260 17.9
## 3 3 36.02081 36.0
## 4 4 21.60346 21.6
## 5 5 24.06195 24.1
## 6 6 17.22985 17.2
```

Use case

Users first need to structure their data file similar to the provided test dataset (e.g. QRISK3_2019_test) which contains information of patients' identifier and QRISK3 risk factors and mimics QRISK3's training cohort³. The structured data file (statistical analysis dataset) should be set out so that each row (observation) represents one individual patient and each column represents one QRISK3 predictor. The exact definition of all QRISK3 predictors can be found from Box 1 of the original QRISK3 paper³. Variables used by QRISK3 can be extracted from EHR databases, such as CPRD²⁴ or QResearch²⁵. Code lists (Read code) for the outcome variable (CVD) can be obtained from the supplementary materials of the QRISK3 paper³. Code lists for variables included in QRISK2 can be extracted from a previous study²⁶. Code lists for other variables including anxiety, alcohol abuse, atypical anti-psychotic medication, erectile dysfunction, HIV/AIDS, left ventricular hypertrophy, migraine and systemic lupus erythematosus could be found from CPRD²⁷ or clinical codes website²⁸. All CVD risk factors should be coded as numeric, binary variables should be coded as 0 or 1, categorical variables such as smoking status should be coded as the same as this package. Any differences between users' variables and QRISK3 predictors (e.g. different criteria to define smoking status) should be mentioned in users' final report. Once the analysis dataset was extracted, it is recommended to compare the distribution of users' analysis dataset to QResearch's cohort using their baseline table^{3,8}. Missing values should be imputed with multiple

imputation²⁹. Finally, users should follow the above workflow and carefully match their variable names to pre-defined QRISK3 predictors to calculate risk score. The function will return a dataset with patient identifier, calculated score and calculated score with one digit.

Discussion

This R package successfully implements the QRISK3 algorithm into R, which allows researchers to calculate CVD risk of patients in the next 10 years. The R package was validated by the original algorithm and a SAS version. This is also the first R implementation of the QRISK3 algorithm at the date of writing.

Though QRISK3 was already published and released from the online website, it is time consuming for researchers to calculate QRISK3 risk score, as the online calculator cannot be used as a service to obtain QRISK3 scores for a large cohort, this package bridges this gap by helping researchers to apply QRISK3 model to their own cohort.

Although it is easy to use this R function to calculate a risk score, researchers should carefully check whether their variables are coded the same as the original QRISK3 cohort, otherwise the calculated score might not be the correct risk of the patient in the cohort. For example, a patient who is a smoker and has the smoking variable coded as “1” would conflict with the definition of the QRISK3 algorithm (“smoking” equals 1 in this R package means non-smoker). Since QRISK is updated annually every spring, researchers who are interested in the latest work should refer to their website¹⁰.

In conclusion, we developed this R package to allow researchers to obtain QRISK3 scores for large cohorts. This tool could help researchers to improve risk prediction modelling based on a currently used risk prediction model.

Data availability

Underlying data

Original QRISK3 algorithm: <https://qrisk.org/three/src.php>

Software availability

Package available from CRAN: <https://cran.r-project.org/web/packages/QRISK3/index.html>

Source code available from: <https://github.com/YanLiUK/QRISK3>

Archived source code as at time of publication: <https://doi.org/10.5281/zenodo.3570682>²⁷

License: GPL-3

C source code, SAS version and QRISK3_2017_test and QRISK3_2019_test datasets used for validation available from: https://github.com/YanLiUK/QRISK3_valid

Archived C code, SAS version and test datasets as at time of publication: <https://doi.org/10.5281/zenodo.3571304>²⁸

License: GPL-3

References

1. **Cardiovascular diseases (CVDs)**. Accessed November 16, 2019.
[Reference Source](#)
2. Grant SW, Collins GS, Nashef SAM: **Statistical Primer: developing and validating a risk prediction model**. *Eur J Cardiothorac Surg*. 2018; 54(2): 203–208.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Hippisley-Cox J, Coupland C, Brindle P: **Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study**. *BMJ*. 2017; 357: j2099.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Collins GS, Altman DG: **An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study**. *BMJ*. 2010; 340: c2442.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. **CVD risk assessment and management - NICE CKS**. Accessed November 16, 2019.
[Reference Source](#)
6. Cox DR: **Regression Models and Life-Tables**. 1972; Accessed February 21, 2019; 34(2): 187–202.
[Publisher Full Text](#)
7. Li Y, Sperrin M, Belmonte M, *et al.*: **Do population-level risk**

- prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep.* 2019; 9(1): 11222.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Pate A, Emsley R, Ashcroft D, *et al.*: The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med.* 2019; 17(1): 134.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 9. Li Y, Sperrin M, Martin GP, *et al.*: Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *Int J Med Inform.* 2020; 133: 104033.
[PubMed Abstract](#) | [Publisher Full Text](#)
 10. QRISK3. Accessed November 16, 2019.
[Reference Source](#)
 11. <https://qrisk.org/three/src.php>. Accessed November 16, 2019.
 12. R: The R Project for Statistical Computing. Accessed April 28, 2019.
[Reference Source](#)
 13. SAS® 9.4 Statements: Reference, Fifth Edition. Accessed August 20, 2017.
[Reference Source](#)
 14. R Installation and Administration. Accessed November 17, 2019.
[Reference Source](#)
 15. Submitting your first package to CRAN, my experience | R-bloggers. Accessed November 17, 2019.
[Reference Source](#)
 16. Writing an R package from scratch | Not So Standard Deviations. Accessed November 17, 2019.
[Reference Source](#)
 17. R package primer. Accessed November 17, 2019.
[Reference Source](#)
 18. Collins D, Lee J, Bobrovitz N, *et al.*: whoishRisk – an R package to calculate WHO/ISH cardiovascular risk scores for all epidemiological subregions of the world [version 2; peer review: 3 approved]. *F1000Res.* 2016; 5: 2522.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 19. CRAN - Package QRISK3. Accessed December 8, 2019.
[Reference Source](#)
 20. YanLiUK/QRISK3: A QRISK3 R package implements QRISK3 algorithm into R. Accessed December 12, 2019.
[Reference Source](#)
 21. Herrett E, Gallagher AM, Bhaskaran K, *et al.*: Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol.* 2015; 44(3): 827–836.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 22. Steyerberg EW, Vickers AJ, Cook NR, *et al.*: Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010; 21(1): 128–138.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 23. Steyerberg EW: Clinical Prediction Models : A Practical Approach to Development, Validation, and Updating. Springer; 2009.
[Publisher Full Text](#)
 24. Clinical Practice Research Datalink - CPRD. Accessed August 20, 2017.
[Reference Source](#)
 25. Home - QResearch. Accessed December 8, 2019.
[Reference Source](#)
 26. van Staa TP, Gulliford M, Ng ES, *et al.*: Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One.* 2014; 9(10): e106455.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 27. CPRD @ Cambridge - Code Lists - Primary Care Unit. Accessed November 18, 2019.
[Reference Source](#)
 28. ClinicalCodes Repository. Accessed November 18, 2019.
[Reference Source](#)
 29. Sterne JA, White IR, Carlin JB, *et al.*: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* 2009; 338: b2393.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 3

Reviewer Report 07 January 2021

<https://doi.org/10.5256/f1000research.26578.r75384>

© 2021 Hassan F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Fadratul Hafinaz Hassan

School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia

QRISK3 algorithm is openly available and already embedded in SAS. However, the authors didn't clearly state what is the rationale behind embedding the algorithm QRISK3 algorithm in another statistical software such as R.

The conclusion doesn't strongly show the significant performance of QRISK3 algorithm in R. Also, there is not a clear comparison performance between QRISK3 in SAS and QRISK3 in R.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Optimization and Machine Learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Jan 2021

Yan Li, the University of Manchester, Manchester, UK

Many thanks to reviewer's interest in our work, we would like to address the comments as below.

To our knowledge, QRISK3 was not already embedded or available in SAS before this R package (if this is not the case, we would like to ask for a reference for any pre-released QRISK3-SAS version). We did provide QRISK3 SAS version (https://github.com/YanLiUK/QRISK3_valid/blob/master/SAS/QRISK3_valid.sas) along with this R package as a procedure to mimic double-programming to verify the correctness of R version of QRISK3 (The SAS program can be used to calculate QRISK3 score too with minor correct mapping of variables).

QRISK3 was only initially provided with C code on the official website without clear definition of variables and outputs. Researchers who do not know C language (which as mentioned in the paper is a low-level programming language not specialised for data analysis comparing to R) and unfamiliar with risk prediction models would find it hard to use QRISK3 in their study. R is currently one of the top languages with well-established support of data structure and useful statistical tool packages for data analysis, this is reason we provide this R version of QRISK3 to bridge this gap.

Since in the package we provided both of R version and SAS version of QRISK3 algorithm, it really depends on users to decide which one better suit their project. QRISK3 algorithm itself is rather a simple calculation of a score from a polynomial function, so we find it less interest to compare the performance between SAS version and R version especially when they return the same results very quickly. The general difference between SAS and R is that R would load the data into Random Access Memory (RAM) in calculation which means for medium size/small size of data (as long as the size of the data fits into the RAM that R can access), R would calculate the score in one run with very high speed. SAS load data into hard disk drive (HDD) but it can cope with large data size in one run (i.e., when the sample size is too large for RAM). However, when the sample size is larger than the accessible size of RAM, R users can cope with this issue by calculating risk score proportionally from overall dataset (e.g., run 10 times and each time for 1/10 number of patients). The performance also depends on what data structure software they were using in SAS or R and how they programmed. Overall, we think to compare performance of R version of SAS version is beyond the scope of this paper, as users with knowledge of both languages could easily calculate them both very quickly and would end up with the same calculated score.

We again are thankful for the time and effort that the reviewer has spent on this paper.

Competing Interests: No competing interests to be disclosed

Reviewer Report 17 August 2020

<https://doi.org/10.5256/f1000research.26578.r63810>

© 2020 Birnbaum A et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Angela K. Birnbaum

Department of Experimental and Clinical Pharmacology, College of Pharmacy, University of Minnesota, Minneapolis, MN, USA

Ashwin Karanam

University of Minnesota, Minneapolis, MN, USA

Although the authors addressed our previous comments we believe there are still some outstanding issues that were not adequately addressed.

- The reply to the comment concerning development and validation of the QRISK3 R package states that "There are no sequential addition of predictors in two datasets." However, the use of a sequential design is stated in the paper: "The simulated patient groups were generated by changing each risk factor sequentially covering the changes of all QRISK3 risk factors." The readers would still benefit from an explanation of the reasoning behind the use of the created dataset.
- For the Validated section: The comments still need addressed. The authors should include more information on how this was done. This could just consist of more detail so the reader can have a general idea without reading other papers. Some pertinent questions are: how are the statistic values defined? how was the statistic reported performed? what do the reported statistics indicate? how were the statistics compared with your study and the literature study? Is there a p-value associated with the comparison? Additionally, include the definition for CPRD. One of your new sentences now has a grammar issue. See below. Applying this QRISK3 package to a big DEFINE HERE (CPRD) cohorts with 3.6 million patients showed a good discrimination (C statistic: 0.85) and calibration similar to the original QRISK3.
- Usage and features concerning missing values. There are various ways to handle missing data. Is multiple imputation the only option with this package? Usually there are multiple options and imputation may not be the best option in all cases. A reference can help, but it would need to cover more than one method or that method should be one that has been chosen to be the only acceptable one for use in the package. Simply stating that standard methods should be employed would be an improvement.
- Discussion: This still needs to be fixed. How does your package help researchers? Maybe state that the package will allow for a consistent and convenient method of calculating

QRISK for large cohort databases. The revised sentence is confusing. The original comment simply asked to remove "to better understand" as the package only calculates a number and does not provide an understanding of methodology or underlying pathology. We still believe a larger package with more than one function is needed for a published R package. For this package, at this level the vignette associated with the package should suffice rather than a publication.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Pharmacometric modeling, neuropharmacology, clinical pharmacology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Version 2

Reviewer Report 30 April 2020

<https://doi.org/10.5256/f1000research.24939.r61993>

© 2020 Birnbaum A et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Angela K. Birnbaum

Department of Experimental and Clinical Pharmacology, College of Pharmacy, University of Minnesota, Minneapolis, MN, USA

Ashwin Karanam

University of Minnesota, Minneapolis, MN, USA

This paper describes an R package for calculating QRISK3 scores in large datasets. The introduction reads well and provides justification for the paper although the package only describes one function. The methods section needs more detail and adjustments. Some of the items can be fixed by improving grammar and readability although others need clarification or details added. Please find our particular comments below.

1. Development and validation of the QRISK3 R package: The description of the simulation used to create the QRISK3_2017_test and QRISK3_2019_test datasets are unclear. It would be beneficial to explain the reasoning behind creating a dataset with sequential addition of risk factors. Clinically, it would be very unlikely to encounter such a dataset. A more translatable dataset would have been patients with randomly sampled risk factors rather than a sequential addition. Additionally, it could be beneficial for the authors to show a summary of the test datasets (number of individuals, demographic information etc.). The last sentence “with several useful online tutorials^{15–18}” should be deleted as these are not relevant to the manuscript.
2. Table 1: The ratio concerning cholesterol should read “Total cholesterol/HDL ration?” in order to clarify the input data for users.
3. For the Validation section CPRD needs to be defined. A description of the CPRD dataset along with the methodology and statistical terminology (e.g. discrimination, calibration, C-statistic etc.) should be clearly described in the methods. Additionally, the study that includes 3.6 million patients that was used is not described and there is no assurance of what the data were, how it was formatted, or how it performed. The authors emphasize the importance of how fields are entered and presented into the R package, therefore, a description of the validation set showing any issues and to elaborate on the results with the R package are needed in order to determine how well the package is performing. The last sentence needs to be adjusted as it is confusing and needs to read better in order to determine what the authors are trying to convey.
4. Usage and features: The authors state that missing values should be handled, but do not elaborate as to how they should be handled. What if a value is truly missing? If data handling is an issue with QRISK3 then that should be presented and how it will affect the QRISK3 score being calculated should be explained. A reference could be added to indicate how these should be handled.
5. Use case: The instructions for creation of a dataset is unclear. Does the user just need to structure their data file similar to the test dataset as opposed to creating a statistical analysis dataset?
6. Discussion, 2nd paragraph: Please rewrite the last half of the second paragraph. The reason for creation of this package was presented in the Introduction and the description in this

section is not useful. The last portion of the second sentence needs to be deleted or rewritten.

7. Discussion, 3rd paragraph: The second to last sentence is confusing. It could be rewritten from "...who is a smoker is coded as "1" in the variable....." to "...who is a smoker and has the smoking variable coded as "1" would be in conflict.....".

8. Discussion: Please delete the words "to better understand" as this package only calculates a number and does not enable exploration of understanding.

Overall observation: Although the package implements the QRISK3 algorithm in an R package, the ultimate goal as stated by the authors is for investigators to readily use the algorithm for large cohorts. Therefore, one would require working knowledge of R to use the package. One way to increase the ease of use would be to create a Shiny App using the algorithm written in R to make the program more accessible to clinicians and investigators. Additionally, there are no functions to create data/results visualizations. A package should be somewhat standalone in the sense that it provides all functions from data exploration to modeling. A single function usually does not merit a research article.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Pharmacometric modeling, neuropharmacology, clinical pharmacology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 18 May 2020

Yan Li, the University of Manchester, Manchester, UK

Reviewer comments:

This paper describes an R package for calculating QRISK3 scores in large datasets. The introduction reads well and provides justification for the paper although the package only describes one function. The methods section needs more detail and adjustments. Some of the items can be fixed by improving grammar and readability although others need clarification or details added. Please find our particular comments below.

- **Reply:** We appreciate the time and effort that reviewers spent on this paper. We thank for those helpful comments and improved the paper.

Reviewer comments:

Development and validation of the QRISK3 R package: The description of the simulation used to create the QRISK3_2017_test and QRISK3_2019_test datasets are unclear. It would be beneficial to explain the reasoning behind creating a dataset with sequential addition of risk factors. Clinically, it would be very unlikely to encounter such a dataset. A more translatable dataset would have been patients with randomly sampled risk factors rather than a sequential addition. Additionally, it could be beneficial for the authors to show a summary of the test datasets (number of individuals, demographic information etc.). The last sentence “with several useful online tutorials^{15–18}” should be deleted as these are not relevant to the manuscript.

- **Reply:** There are no sequential addition of predictors in two datasets. Each patient only has one positive CVD predictor in the test datasets and all the patients covered change of all predictors. In this case dataset with randomly sampled risk factors is equivalent to datasets which covered change of all the predictors for package testing purpose, as each observation of the data with randomly sampled risk factors is a combination of each individual predictor, i.e. for each individual predictor, the package calculates the same score as original algorithm then it would calculate the same score for patients with a combination of individual predictors. Also because the more positive predictors a patient have a higher risk would be expected, it would be more difficult to verify whether the implementation is correct when a patient with many predictors has a very high risk such as 99.9% in R version and 99.8% in original version of QRISK3, and data with randomly sampled risk factors cannot directly provide insights of what predictors were implemented incorrectly (e.g. patients with 10 predictors result different score, which predictors among these 10 predictors contribute to this?). Therefore, comparing low risk patients (i.e. with only one predictor each and cover the change of all predictors) is preferred. Furthermore, we and ClinRisk (in the terms of agreement) encourage users with more hesitation to verify this R package with their own simulated datasets using the provided original QRISK3 algorithm and R package as a part of their data quality control process.

The demographic information of testing data was not shown as it was mainly used for testing whether the implemented QRISK3 calculates the same score as the original algorithm, which only serves for package testing and example. The structure information of data shows it has 48 patients records and covered changes of all QRISK3 predictors.

We did refer to contents of those online tutorials to create this package, i.e. three of them guided us how to publish the package and one guided how to organize package paper. As

reference aims to acknowledge the contribution of other writers and researchers in our work ¹, these are relevant to this paper.

Reviewer comments:

Table 1: The ratio concerning cholesterol should read "Total cholesterol/HDL ration?" in order to clarify the input data for users.

- **Reply:** Thanks. corrected as suggested.

Reviewer comments:

For the Validation section CPRD needs to be defined. A description of the CPRD dataset along with the methodology and statistical terminology (e.g. discrimination, calibration, C-statistic etc.) should be clearly described in the methods. Additionally, the study that includes 3.6 million patients that was used is not described and there is no assurance of what the data were, how it was formatted, or how it performed. The authors emphasize the importance of how fields are entered and presented into the R package, therefore, a description of the validation set showing any issues and to elaborate on the results with the R package are needed in order to determine how well the package is performing. The last sentence needs to be adjusted as it is confusing and needs to read better in order to determine what the authors are trying to convey.

- **Reply:** We added references to clarify these. Description of CPRD could be found here ². Explanation of model performance measurements including discrimination and calibration were detailed explained here ³. The study includes 3.6 million patients can be found here ⁴. The main validation of this R package is to calculate the same risk score as original algorithm, and the test group which covered change of all the predictors has shown this. Applying this R package to a larger cohort is an addition of verification, as incorrect implementation would result very poor model performance. We have rephrased the last sentence according to the suggestion.

In manuscript

"Applying this QRISK3 package to a big CPRD cohorts ² with 3.6 million patients ⁴ showed a good discrimination (C statistic: 0.85) and calibration ³ similar to the original QRISK3 ⁵."

Reviewer comments:

Usage and features: The authors state that missing values should be handled, but do not elaborate as to how they should be handled. What if a value is truly missing? If data handling is an issue with QRISK3 then that should be presented and how it will affect the QRISK3 score being calculated should be explained. A reference could be added to indicate how these should be handled.

- **Reply:** There are no truly missing risk factors for the GP, as they can enter additional data when reviewing QRISK; for researchers, dealing with missing value with methods such as multiple imputation can be found in chapter 7 of the reference ⁶. We have added this reference in manuscript.

In manuscript

"Missing values in the dataset should be handled (e.g. multiple imputation) before using this package ⁶."

Reviewer comments:

Use case: The instructions for creation of a dataset is unclear. Does the user just need to structure their data file similar to the test dataset as opposed to creating a statistical analysis dataset?

- **Reply:** Yes, user need to structure their data file to a data frame similar to the test dataset. We rephrased sentences in main manuscript to better clarify this.

In manuscript

"Users first need to structure their data file similar to the provided test dataset (e.g. QRISK3_2019_test) which contains information of patients' identifier and QRISK3 risk factors and mimic QRISK3's training cohort ⁵. The structured data file (statistical analysis dataset) would be each row (observation) represents one individual patient and each column represents one of QRISK3 predictors."

Reviewer comments:

Discussion, 2nd paragraph: Please rewrite the last half of the second paragraph. The reason for creation of this package was presented in the Introduction and the description in this section is not useful. The last portion of the second sentence needs to be deleted or rewritten.

- **Reply:** Thanks. We have rewritten this paragraph.

In manuscript

"Though QRISK3 was already published and released from the online website, it is time consuming for researchers to calculate QRISK3 risk score, as the online calculator cannot be used as a service to obtain QRISK3 scores for a large cohort, this package bridges this gap by helping researchers to apply QRISK3 model to their own cohort."

Reviewer comments:

Discussion, 3rd paragraph: The second to last sentence is confusing. It could be rewritten from "...who is a smoker is coded as "1" in the variable....." to "...who is a smoker and has the smoking variable coded as "1" would be in conflict.....".

- **Reply:** Thanks. We have rephrased this sentence.

In manuscript

"For example, a patient who is a smoker and has the smoking variable coded as "1" would conflict with the definition of the QRISK3 algorithm ("smoking" equals 1 in this R package means non-smoker)."

Reviewer comments:

Discussion: Please delete the words "to better understand" as this package only calculates a

number and does not enable exploration of understanding.

- **Reply:** Okay, we deleted and rephrased this.

In manuscript

"In conclusion, we developed this R package to allow researchers to obtain QRISK3 scores for large cohorts. This tool could help researchers to improve risk prediction modelling based on a currently used risk prediction model."

Reviewer comments:

Overall observation: Although the package implements the QRISK3 algorithm in an R package, the ultimate goal as stated by the authors is for investigators to readily use the algorithm for large cohorts. Therefore, one would require working knowledge of R to use the package. One way to increase the ease of use would be to create a Shiny App using the algorithm written in R to make the program more accessible to clinicians and investigators. Additionally, there are no functions to create data/results visualizations. A package should be somewhat standalone in the sense that it provides all functions from data exploration to modelling. A single function usually does not merit a research article.

- **Reply:** The Shiny app works for clinicians and investigators was already made by the QRISK3 developer and could be found in <https://qrisk.org/three/>, it was also integrated with electronic health records system such as EMIS ⁷.

The main function returns a dataset with calculated QRISK score. As suggested before, the visualisation is purely depending on what research question was asked. Also, the web version of QRISK3 did not provide visualisation as well ⁸. For a continuous variable like calculated QRISK risk score, a summary function in R (i.e.

"summary(test_all_rst\$QRISK3_2017)") would provide basic statistics such as mean and standard deviation, and with a histogram function (i.e. "hist(test_all_rst\$QRISK3_2017)") would plot histogram of the calculated risk score and

"plot(density(test_all_rst\$QRISK3_2017))" to plot the distribution of the calculated risk score.

We understand the desire of collecting all functions into one place, but not a single R package from CRAN ⁹ can provide all such functions in one place. This is because either these functions were already provided by other packages or massive number of functions in one package could mask those truly helpful functions. However, CRAN task review may help, as it integrates all sort of R packages within a relevant topic. "CRAN Task View: Missing Data ¹⁰" describes packages to deal with missing data. "CRAN Task View: Survival Analysis ¹¹" provides all sorts of R packages which supports model development in survival analysis. Other generic R programming skill could be found in this free online tutorial ¹².

Overall, we believe the merit of this work is to provide a tool along with references to save researchers from re-implementing and re-validating QRISK3 into R. This tool could help data scientists to compare model performance and predicted risk of their own risk prediction model to a currently used risk prediction model, help other researchers to apply QRISK3 to a large cohort and help clinical trials which requires identify patients at certain risk levels.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

References

1. Why is Referencing Important? | UNSW Current Students. <https://student.unsw.edu.au/why-referencing-important>. Accessed May 12, 2020.
2. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098
3. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
4. Li Y, Sperrin M, Belmonte M, Pate A, Ashcroft DM, van Staa TP. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep*. 2019;9(1):11222. doi:10.1038/s41598-019-47712-5
5. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357(3):j2099. doi:https://doi.org/10.1136/bmj.j2099
6. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer; 2009.
7. EMIS Web - QRISK2®-2015 calculator. https://www.emisnow.com/community?id=kb_article_view&sys_kb_id=19c88ac4dbe100d09aa4641d0b96197. Accessed May 13, 2020.
8. QRISK3. <https://qrisk.org/three/index.php>. Accessed December 2, 2019.
9. CRAN Task Views. <https://cran.r-project.org/web/views/>. Accessed May 13, 2020.
10. Julie Josse NTN (r-miss-tastic team). CRAN Task View: Missing Data. April 2020.

11. Arthur Allignol AL. CRAN Task View: Survival Analysis. April 2020.
12. R for Data Science. <https://r4ds.had.co.nz/>. Accessed May 13, 2020.

Competing Interests: There are no conflicts of interest to disclose

Reviewer Report 16 March 2020

<https://doi.org/10.5256/f1000research.24939.r60685>

© 2020 Jafari M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mohieddin Jafari

University of Helsinki, Helsinki, Finland

Ali Amiryousefi

University of Helsinki, Helsinki, Finland

We still think that an R package should be more comprehensive and larger. It would be better to include more novel functions and datasets to create a package. We can say that in this situation a Shiny App would be a better choice.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biologist

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 12 February 2020

<https://doi.org/10.5256/f1000research.23899.r59681>

© 2020 Jafari M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mohieddin Jafari

University of Helsinki, Helsinki, Finland

Ali Amiryousefi

University of Helsinki, Helsinki, Finland

The authors provided an R package for the QRISK3 algorithm to predict the risk of cardiovascular disease. My main comment is related to the size of this package, which has only one simple function with two simple datasets. I am not sure that this tiny package can be presented as an F1000Research article. There are not any new findings represented in this article compared to the author's previous works. I suggest enriching this package with more visualization and data manipulation functions and datasets. Besides, I do not think it needs to highlight validation of R performance with SAS or C. This is part of your quality control, not an outcome. Also, I have some minor comments:

1. Why didn't you use the R 3.6.2 instead of 3.4.2?
2. Why you did not make the categorical variables as ordered factor R object instead of a simple numeric object? (in this case, you do not need to think about the conflict of the meaning of "1" and "0").
3. The range of patient ages is not consistent between text and table.
4. "Height", "Weight" and "Weight/Height" in table 1 should be lowercase the same as the package.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biologist

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 12 Feb 2020

Yan Li, the University of Manchester, Manchester, UK

We appreciate reviewers' time and effort in review this package paper, we would like to make following reply to these comments.

Reviewer's comments:

The authors provided an R package for the QRISK3 algorithm to predict the risk of cardiovascular disease. My main comment is related to the size of this package, which has only one simple function with two simple datasets. I am not sure that this tiny package can be presented as an F1000Research article. There are not any new findings represented in this article compared to the author's previous works.

Reply: This implementation of QRISK3 algorithm was needed due to lack of metadata, explanations of variables and accessible source code. There is a need of this R package as researchers cannot easily estimate risks based on the QRISK3 algorithm¹. This is also the first implementation of QRISK3 into R (validated by SAS and C version) and was approved by CRAN. By the date of 12-02-2020 (67 days since the first release), this R package has been downloaded by 1343 times and rank 42 among total 66 packages which were published in the same day (data available from "cranlogs").

We believe the novelty of this package and this paper is:

QRISK3 is different from other model such as Framingham model (one other popular American CVD risk prediction model). For Framingham model, the model formula was simple and clearly presented in its original paper, and user could easily implement it into R. While for QRISK3, this is different and difficult. The main challenge implementing QRISK into R is the original algorithm written in low level language C and there is no document to

explain the meaning of these variables in original algorithm. Users also need to validate their implemented version is the same as the original version of QRISK3, this requires users to have programming knowledge of both R and C while R is the dominated language in data science. Therefore, researchers don't have table 1 of this paper if they start to implement QRISK3 from the original algorithm. This creates challenges that it is hard to acquire a similar statistic model formula as presented in Framingham, and it is hard to identify which variable represents what risk factors especially considering they were written in C. For example, how to know variable such as "b_impotence2", "ethrisk" and "fh_cvd" represents "erectile disfunction", "ethnicity" and "relative of CVD". QRISK was used to predict 10 years risk, why there are 11 values in a vector which represents baseline risk from the original algorithm, and which variable represents this baseline risk? There is a long distance between input value of predictors to output a correct predicted risk. This package solved all the challenges mentioned above and bridges the gaps.

This R package acts like a bridge which enables researchers to easily access and use QRISK3 model in R like using Framingham model.

Without this R package and this paper described how this R package was developed and validated, researchers may spend months to re-implement QRISK into R and they also need to spend a lot of effort on validating whether their implemented version of QRISK3 does the similar thing as the original algorithm. Overall, our package and this paper solved these, and saved time and energy for researchers who only require calculated QRISK3 score as part of their new research.

Reviewer's comments:

I suggest enriching this package with more visualization and data manipulation functions and datasets. Besides, I do not think it needs to highlight validation of R performance with SAS or C. This is part of your quality control, not an outcome. Also, I have some minor comments:

Reply: We appreciate reviewers' suggestion, but this package was a simple implementation of QRISK3 algorithm. We decide to make this package as simple as we can as:

1. It would be easier for user to quickly obtain QRISK3 score. Rather than searching a list of redundant functions, user could instantly test and use this package.
2. There are mature R packages which support reviewer suggested functions. Due to the nature of this R package (i.e. only write with base R language without dependent packages), it is well compatible to all the other R packages. For visualization, users could easily use R package like "ggplot2" with this R package. For data manipulation, R package like "dplyr" is well compatible to our package.
3. The main function of this R package is to provide QRISK3 calculated risk score. What visualization and datasets user might require were purely dependent on their research question. In the current version, we have provided a dataset with patient identifier and calculated QRISK score, which could be easily used by researchers to draw their own visualization.

Though there are only two datasets present in this package, it contains all the key information of what the datasets would look like and what variables were needed. They contain enough information for user to understand and use this package easily. Overall, simplistic is a part of design of this R package and there are well established R packages

which could be used with this simple R package. We do not find any necessary reasons to add more functions in this R package, as the aim of this R package is to help users to obtain a QRISK3 score from large dataset quickly and current version fulfilled this purpose. We are happy to add more functions or tools if there are more specific requirements from users in future.

We also believe it is essential to show user how this R package was validated with the original algorithm, as this R package aims to provide a correct QRISK3 predicted score. Users who decide to use our package may wish to replicate our validation process.

Reviewer's comments:

Why didn't you use the R 3.6.2 instead of 3.4.2?

Reply: R 3.4.2 was used to develop the prototype of this R package in 2017. However, all functions in this package used R base language, so it can be used in any R version including 3.6.2 (tested).

Reviewer's comments:

Why you did not make the categorical variables as ordered factor R object instead of a simple numeric object? (in this case, you do not need to think about the conflict of the meaning of "1" and "0").

Reply: This was due to how QRISK3 score was calculated. It was calculated from a mathematic formula where we need to calculate a linear predictor which requires numeric value of predictors. We coded all variables into numeric rather than some are numeric, and some are factor is for the ease of users (follow the simplistic principle). In this case, users would know all variables need to be numeric rather than to remember which variable should be a factor especially when there are over 20 predictors in the model. We also provided an automate function to help user double check which variables were not coded into numeric. However, we do realize that we may never prevent all unforeseen mistake, so we highlight this in descriptions of both of paper and package.

Reviewer's comments:

The range of patient ages is not consistent between text and table.

Reply: thanks. Corrected in the next version.

Reviewer's comments:

"Height", "Weight" and "Weight/Height" in table 1 should be lowercase the same as the package.

Reply: thanks. Updated.

Other reviewer's comments:

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

References

1. I need SAS, or SPSS or Stata or R code for :Two cardiac risk scores: QRISK and SCORE . May you share?

https://www.researchgate.net/post/I_need_SAS_or_SPSS_or_Stata_or_R_code_for_Two_cardiac_risk_scores_Q
Accessed February 12, 2020.

Competing Interests: No competing interests should be disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research