The pwrEWAS User's Guide

Stefan Graw, Devin C. Koestler

27 October 2020

Abstract

pwrEWAS is a user-friendly tool to estimate power in EWAS as a function of sample and effect size for two-group comparisons of DNAm (e.g., case vs control, exposed vs non-exposed, etc.). Detailed description of in-/outputs, instructions and an example, as well as interpretations of the example results are provided in the following vignette.

Package

pwrEWAS 1.4.0

Contents

Introduction

When designing an epigenome-wide association study (EWAS) to investigate the relationship between DNA methylation (DNAm) and some exposure(s) or phenotype(s), it is critically important to assess the sample size needed to detect a hypothesized difference with adequate statistical power. However, the complex and nuanced nature of DNAm data makes direct assessment of statistical power challenging. To circumvent these challenges and to address the outstanding need for a user-friendly interface for EWAS power evaluation, we have developed pwrEWAS. The current implementation of pwrEWAS accommodates power estimation for two-group comparisons of DNAm (e.g. case vs control, exposed vs non-exposed, etc.), where methylation assessment is carried out using the Illumina Human Methylation BeadChip technology. Power is calculated using a semi-parametric simulation-based approach in which DNAm data is randomly generated from beta-distributions using CpG-specific means and variances estimated from one of several different existing DNAm data sets, chosen to cover the most common tissue-types used in EWAS. In addition to specifying the tissue type to be used for DNAm profiling, users are required to specify the sample size, number of differentially methylated CpGs, effect size(s), target false discovery rate (FDR) and the number of simulated data sets, and have the option of selecting from several different statistical methods to perform differential methylation analyses. pwrEWAS reports the marginal power, marginal type I error rate, marginal FDR, and false discovery cost (FDC). The R-Shiny web inter-face allows for easy input of user-defined parameters and includes an advanced settings button that offers additional options pertaining to data generation and computation.

Installation

pwrEWAS can be installed with the following R code:

```
if (!requireNamespace("BiocManager"))
    install.packages("BiocManager")
BiocManager::install("pwrEWAS")
```

Usage

To execute the main pwrEWAS function the following to codes can be used. pwrEWAS allows the user to specify the effect size in one of two ways, by either providing a target maximal difference in methylation ("targetDelta"), or by providing the standard deviation of the simulated differnces ("deltaSD"). Only one of both arguments can be provided. If "targetDelta" is specified, pwrEWAS will automatically identify a standard deviation to simulate differences in methylation, such that the 99.99th percentile of the absolute value of simulated differences falls within a range around the targeted maximal difference in DNAm (see paper for additional details). If "deltaSD" is specified, pwrEWAS will simulate differences in methylation using the provided standard deviation (additional information provided in paper).

```
library("pwrEWAS")
```

```
# providing the targeted maximal difference in DNAm
results_targetDelta <- pwrEWAS(minTotSampleSize = 10,
    maxTotSampleSize = 50,
    SampleSizeSteps = 10,
    NcntPer = 0.5,
    targetDelta = c(0.2, 0.5),
    J = 100,
    targetDmCpGs = 10,
    tissueType = "Adult (PBMC)",
    detectionLimit = 0.01,
    DMmethod = "limma",
    FDRcritVal = 0.05,
    core = 4,
    sims = 50)

# providing the targeted maximal difference in DNAm
results_deltaSD <- pwrEWAS(minTotSampleSize = 10,
    maxTotSampleSize = 50,
    SampleSizeSteps = 10,
    NcntPer = 0.5,
    deltaSD = c(0.02, 0.05),
    J = 100,
    targetDmCpGs = 10,
    tissueType = "Adult (PBMC)",
    detectionLimit = 0.01,
    DMmethod = "limma",
    FDRcritVal = 0.05,
    core = 4,
    sims = 50)
```

Input parameter

The following table provides a description of the input arguments:

Parameter	Description
minTotSampleSize	Lowest total sample sizes to be considered
maxTotSampleSize	Highest total sample sizes to be considered
SampleSizeSteps	Steps with which total sample size increases from minTotSampleSize to maxTotSampleSize
NcntPer	Rate by which the total sample size is split into groups (0.5 corresponds to a balanced study; rate for group 2 is equal to 1 rate of group 1)
targetDelta	Standard deviations of the simulated differences is automatically determined such that the 99%til of the simulated differences are within a range around the provided values
deltaSD	Differences in methylation will be simulated using provided standard deviation
J	Number of CpG site that will simulated and tested (increasing Number of CpGs tested will require increasing RAM (memory))
targetDmCpGs	Target number of CpGs simulated with meaningful differences (differences greater than detection limit)
tissueType	Heterogeneity of different tissue types can have effects on the results. Please select your tissue type of interest or one you believe is the closest
detectionLimit	Limit to detect changes in methylation. Simulated differences below the detection limit will not be consider as meaningful differentially methylated CpGs
DMmethod	Method used to perform differential methylation analysis
FDRcritVal	Critical value to control the False Discovery Rate (FDR) using the Benjamini and Hochberg method
core	Number of cores used to run multiple threads. Ideally, the number of different total samples sizes multiplied by the number of effect sizes should be a multiple (m) of the number of cores (#sampleSizes * #effectSizes = m * #threads). An increasing number of threads will require an increasing amount of RAM (memory)
sims	Number of repeated simulation/simulated data sets under the same conditions for consistent results

Output parameter

Running pwrEWAS will result in an object with the following four attributes: meanPower, powerArray, deltaArray, and metric. The first attribute ''meanPower''is a 2D matrix with empirically estimated marginal mean power for sample sizes and target Δ_βs (averaged over simulated data sets). The second attribute''powerArray''provides the full set of empirically estimated marginal power for sample sizes and target Δ_βs for each simulated data sets in a 3D matrix. The third attribute''deltaArray''contains a 3D matrix with simulated Δ_βs for sample sizes, target Δ_β, and simulated data sets. The last attribute''metric" contains 2D matrices with the marginal type I error rate (marTypeI), power in the classical sense (classicalPower), actual FDR (FDR), False Discovery Cost (FDC), and probabilities of identifying at least one true positive in table format, where sample sizes are shown as rows and effect sizes are columns. Examples results can be found in the example section.

Runtime

In general, the computational complexity of pwrEWAS depends on four major components: (1) assumed number and magnitude of sample size(s), (2) number of target Δ_β's (effect sizes), (3) number of CpGs tested, and (4) number of simulated data sets. To enhance the computational efficiency, pwrEWAS allows users to process simulations in parallel. While (1) and (2) are usually dictated by the study to be conducted, (3) and (4) can be modified to either increase the precision of power estimates (increased run time) or reduce the computational burden (decreased precision of estimates). The following table provides the run time of pwrEWAS for different combinations of sample sizes and effect sizes. In all scenarios presented the number of tested CpGs was assumed to be 100,000, number of simulated data sets was 50, and the method to perform the differential methylation analysis as limma. A total of 6 clusters/threads were used.

Total sample size Effect size Δ_β	0.1	0.1, 0.2	0.1, 0.3, 0.5
10	2min 21sec	3min 11sec	3min 50sec
100	6min 22sec	7min 39sec	8min 33sec
500	24min 43sec	27min 36sec	29min 22sec
10-100 (increments of 10)	9min 40sec	16min 34sec	23min 44sec
300-500 (increments of 100)	27min 58sec	30min 01sec	52min 00sec

Example

Running pwrEWAS

Running pwrEWAS by providing target maximal difference in methylation or by providing standard deviation of difference in methylation:

```
library(pwrEWAS)
set.seed(1234)
results_targetDelta <- pwrEWAS(minTotSampleSize = 20,
    maxTotSampleSize = 260,
    SampleSizeSteps = 40,
    NcntPer = 0.5,
    targetDelta = c(0.02, 0.10, 0.15, 0.20),
    J = 100000,
    targetDmCpGs = 2500,
    tissueType = "Blood adult",
    detectionLimit = 0.01,
    DMmethod = "limma",
    FDRcritVal = 0.05,
    core = 4,
    sims = 50)

results_deltaSD <- pwrEWAS(minTotSampleSize = 20,
    maxTotSampleSize = 260,
    SampleSizeSteps = 40,
```

```
        NcntPer = 0.5,
        deltaSD = c(0.00390625, 0.02734375, 0.0390625, 0.052734375),
        J = 100000,
        targetDmCpGs = 2500,
        tissueType = "Blood adult",
        detectionLimit = 0.01,
        DMmethod = "limma",
        FDRcritVal = 0.05,
        core = 4,
        sims = 50)
```

If pwrEWAS is executed with providing target maximal difference, first τ will be determined. The beginning and finish of this process will be printed with time stamps (see below for an example). If the standard deviation of difference is provided, this step will be skipped. \ Next, pwrEWAS will run the simulations to empirically estimate power. pwrEWAS will indicate when the simulations are started. To monitor the process pwrEWAS will display a process bar. pwrEWAS will print a statement including a time stamps once finished (see below for an example).

```
## [2019-02-12 18:40:23] Finding tau...done [2019-02-12 18:42:53]
## [1] "The following taus were chosen: 0.00390625, 0.02734375, 0.0390625, 0.052734375"
## [2019-02-12 18:42:53] Running simulation
## |======================================================================| 100%
## [2019-02-12 18:42:53] Running simulation ... done [2019-02-12 19:27:03]
```

Outputs

Running pwrEWAS will result in an object, that stores the following four attributes:

```
attributes(results_targetDelta)
## $names
## [1] "meanPower"  "powerArray" "metric"     "deltaArray"
## $names
## [1] "meanPower" "powerArray" "deltaArray" "metric"
```

meanPower

The primary results will be provided in the attribute ''meanPower''. It is essentially a summary of the attribute''powerArray''. meanPower will be provide a 7x4 table with the average power by total sample size as rows (here 20-260 patients with increments of 40) and by target Δ_β, if''targetDelta''was provided, or''$SD(\Delta_\beta)$'', if deltaSD was provided, as columns (here targetDelta was provided as: 0.02, 0.1, 0.15, 0.2):

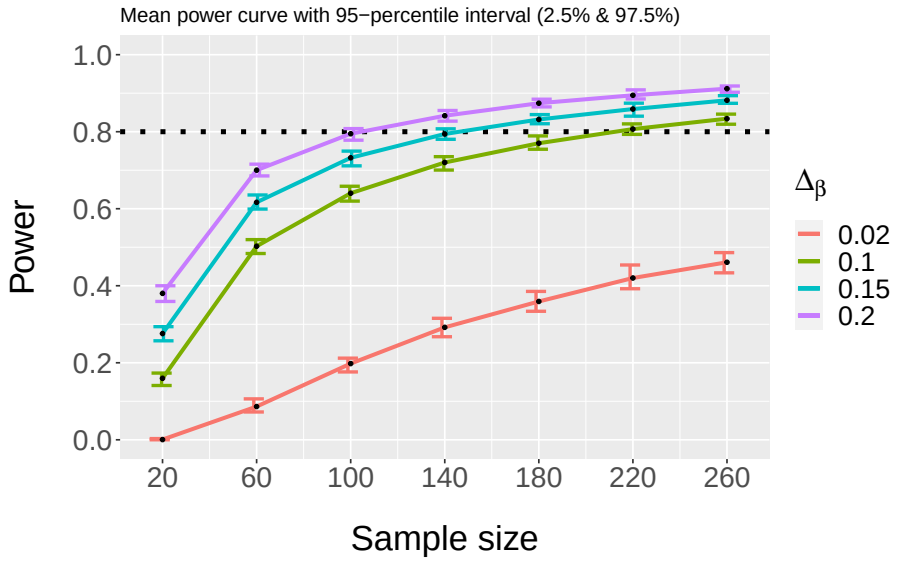
```
dim(results_targetDelta$meanPower)
## [1] 7 4
print(results_targetDelta$meanPower)
##             0.02       0.1      0.15       0.2
## 20  0.0005415101 0.1596165 0.2758319 0.3801848
## 60  0.0863276853 0.5026172 0.6166725 0.7001472
```

```
## 100 0.1978966524 0.6402466 0.7322670 0.7947203
## 140 0.2919669218 0.7201027 0.7940429 0.8414375
## 180 0.3592038789 0.7700964 0.8317636 0.8739818
## 220 0.4201022535 0.8068096 0.8588536 0.8945975
## 260 0.4609956067 0.8338529 0.8816222 0.9117306
```

powerArray

The attribute ''powerArray''should primarily be used to create a power plot but can also be used to investigate the power results for the individual simulations. pwrEWAS includes a function''pwrEWAS_powerPlot''that will create a power plot, where power (y-axis) is shown as a function of sample sizes (x-axis) for different effect sizes (color coded). For each sample size, the mean power as well as the 95%tile interval (2.5% and 97.5%) is shown.''sd''should be set to''FALSE''if''targetDelta''was specified in pwrEWAS, and''TRUE''if''deltaSD" was specified in pwrEWAS.

```
dim(results_targetDelta$powerArray) # simulations x sample sizes x effect sizes
## [1] 50  7  4
pwrEWAS_powerPlot(results_targetDelta$powerArray, sd = FALSE)
```



deltaArray

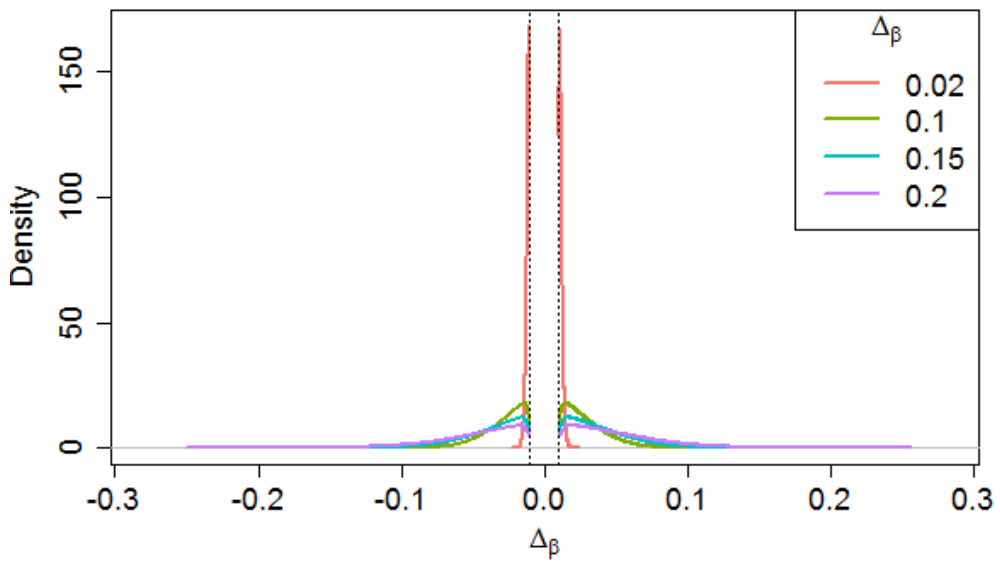
The third attribute ''deltaArray''contains the simulated differences in mean DNAm. Each Δ_β is drawn from a truncated normal, where either the standard devation is provided (''deltaSD'') or automatically determined based on the user-specified target Δ_β (''targetDelta'') and the expected number of differentially methylated CpGs (''targetDmCpGs"). To automatically determined the standard devation, it is adjusted stepwise until the 99.99th percentile of the absolute value of simulated Δ_βs falls within a range around the targeted maximal difference in DNAm (see paper for additional details). The maximal value of Δ_β can exceed the user-specified target Δ_β, but about 99.99% of simulated differences will be below user-specified target Δ_β (as seen below):

```
# maximum value of simulated differences by target value
lapply(results_targetDelta$deltaArray, max)
## $`0.02`
## [1] 0.02095302
##
## $`0.1`
## [1] 0.1265494
##
## $`0.15`
## [1] 0.2045638
##
## $`0.2`
## [1] 0.2458416

# percentage of simulated differences to be within the target range
mean(results _ targetDelta$deltaArray[[1]] < 0.02)
## [1] 0.9999999
mean(results _ targetDelta$deltaArray[[2]] < 0.10)
## [1] 0.9998882
mean(results _ targetDelta$deltaArray[[3]] < 0.15)
## [1] 0.9999386
mean(results _ targetDelta$deltaArray[[4]] < 0.20)
## [1] 0.9999539
```

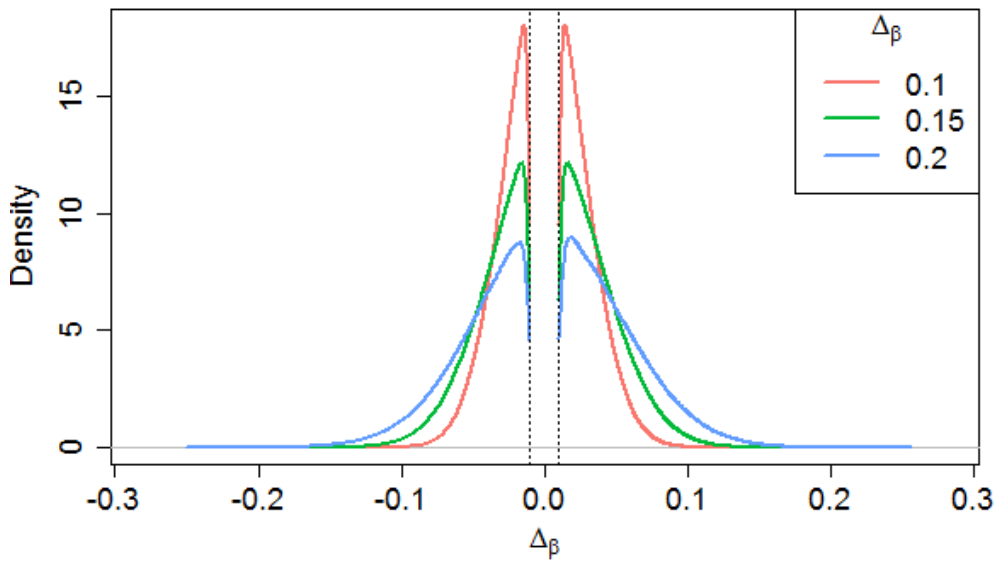
To get a better understanding of how the differences in mean DNAm are distributed, pwrEWAS provides a density plot, where the distribution of simulated differences in mean DNAm is plotted by target differences in DNAm (Δ_β). The color theme matches the colors of the power plot. Simulated differences within the detection limit around zero are removed, as they are here not defined as meaningful differences. ''sd''should be set to''FALSE''if''targetDelta''was specified in pwrEWAS, and''TRUE''if''deltaSD'' was specified in pwrEWAS.

```
pwrEWAS_deltaDensity(results_targetDelta$deltaArray, detectionLimit = 0.01, sd = FALSE)
```



In the figure above, the densities are very compress, because the first effect size is clearly different from the other three. The following code will provide the figure after removing the first effect size:

```
temp <- results_targetDelta$deltaArray
temp[[1]] <- NULL
pwrEWAS_deltaDensity(temp, detectionLimit = 0.01, sd = FALSE)
```


Density
0
5
10
15
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Δβ
Δβ
0.1
0.15
0.2

metric

The fourth attribute ''metric''contains tables on marginal type I error rate (''marTypeI''), power in the classical sense (classicalPower), actual FDR (FDR), False Discovery Cost (FDC, see paper for additional details), and probabilities of identifying at least one true positive, for each sample size and effect size combination:

```
results_targetDelta$metric
## $marTypeI
##             0.02          0.1         0.15          0.2
## 20  0.0000000000 0.0001927742 0.0003533407 0.0004575820
## 60  0.0003435644 0.0006813394 0.0008155174 0.0009199059
## 100 0.0011254329 0.0008494543 0.0009544978 0.0010784126
## 140 0.0023155015 0.0009987010 0.0011007120 0.0011301504
## 180 0.0031646165 0.0010936404 0.0011537869 0.0011709668
## 220 0.0043017549 0.0011711588 0.0011688609 0.0011961111
## 260 0.0050251766 0.0011825572 0.0012256528 0.0012703139
##
## $classicalPower
##          0.02       0.1      0.15       0.2
## 20  0.0000230 0.1140913 0.2188748 0.3211014
## 60  0.0072840 0.3665948 0.4969722 0.5948422
## 100 0.0243978 0.4749589 0.5952528 0.6816387
## 140 0.0447472 0.5386504 0.6500822 0.7252384
## 180 0.0650878 0.5859314 0.6852212 0.7575621
## 220 0.0848528 0.6181004 0.7131985 0.7789187
## 260 0.1031464 0.6445875 0.7373262 0.7980121
##
## $FDR
##            0.02        0.1       0.15        0.2
## 20  0.000000000 0.04402833 0.04704810 0.04431709
## 60  0.004517889 0.04837274 0.04781140 0.04793149
## 100 0.004418029 0.04660417 0.04675174 0.04899762
## 140 0.004953809 0.04824695 0.04925629 0.04831297
## 180 0.004662894 0.04856488 0.04900589 0.04792516
## 220 0.004844068 0.04925471 0.04774217 0.04763118
## 260 0.004668246 0.04777146 0.04839480 0.04928052
##
## $FDC
##           0.02        0.1       0.15        0.2
## 20  0.00000000 0.04638179 0.04959881 0.04652390
## 60  0.03043364 0.05207772 0.05098151 0.05083022
## 100 0.04361648 0.05082257 0.05032126 0.05243835
## 140 0.06076271 0.05345012 0.05358616 0.05197136
## 180 0.06718904 0.05446132 0.05373615 0.05186793
## 220 0.07820442 0.05588333 0.05261712 0.05173305
## 260 0.08348369 0.05445357 0.05369060 0.05387600
##
## $probTP
##     0.02 0.1 0.15 0.2
## 20   0.4   1    1   1
## 60   1.0   1    1   1
```

```
## 100  1.0   1    1   1
## 140  1.0   1    1   1
## 180  1.0   1    1   1
## 220  1.0   1    1   1
## 260  1.0   1    1   1
```

Interpretation

To detect differences up to 10%, 15% and 20% in CpG-specific methylation across 2,500 CpGs with at least 80% power, we would need about 220, 180 and 140 total subjects, respectively. As expected, 80% power was not achieved for a difference in DNAm up to 2% for the selected total sample size range. However, it can be observed that the probability of detecting at least one CpG out of the 2500 differentially methylated CpGs is about 40% for 20 total patients and virtually 100% for 60 and more total patients.

SessionInfo

```
toLatex(sessionInfo())
```

- R version 4.0.3 (2020-10-10), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Running under: `Windows Server 2012 R2 x64 (build 9600)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: BiocStyle 2.18.0, foreach 1.5.1, pwrEWAS 1.4.0, pwrEWAS.data 1.3.0, shinyBS 0.61
- Loaded via a namespace (and not attached): AnnotationDbi 1.52.0, AnnotationHub 2.22.0, Biobase 2.50.0, BiocFileCache 1.14.0, BiocGenerics 0.36.0, BiocManager 1.30.10, BiocVersion 3.12.0, DBI 1.1.0, ExperimentHub 1.16.0, IRanges 2.24.0, Matrix 1.2-18, R6 2.4.1, RSQLite 2.2.1, Rcpp 1.0.5, S4Vectors 0.28.0, XML 3.99-0.5, abind 1.4-5, annotate 1.68.0, assertthat 0.2.1, bit 4.0.4, bit64 4.0.5, blob 1.2.1, bookdown 0.21, codetools 0.2-16, colorspace 1.4-1, compiler 4.0.3, crayon 1.3.4, curl 4.3, dbplyr 1.4.4, digest 0.6.27, doParallel 1.0.16, doSNOW 1.0.19, dplyr 1.0.2, ellipsis 0.3.1, evaluate 0.14, farver 2.0.3, fastmap 1.0.1, genefilter 1.72.0, generics 0.0.2, ggplot2 3.3.2, glue 1.4.2, grid 4.0.3, gtable 0.3.0, htmltools 0.5.0, httpuv 1.5.4, httr 1.4.2, interactiveDisplayBase 1.28.0, iterators 1.0.13, knitr 1.30, later 1.1.0.1, lattice 0.20-41, lifecycle 0.2.0, limma 3.46.0, magrittr 1.5, memoise 1.1.0, mime 0.9, munsell 0.5.0, parallel 4.0.3, pillar 1.4.6, pkgconfig 2.0.3, promises 1.1.1, purrr 0.3.4, rappdirs 0.3.1, rlang 0.4.8, rmarkdown 2.5, scales 1.1.1, shiny 1.5.0, snow 0.4-3, splines 4.0.3, stats4 4.0.3, stringi 1.5.3, stringr 1.4.0, survival 3.2-7, tibble 3.0.4, tidyselect 1.1.0, tools 4.0.3, truncnorm 1.0-8, vctrs 0.3.4, xfun 0.18, xtable 1.8-4, yaml 2.2.1