Optimal Data Analysis
Vol. 9 (June 23, 2020), 190-194

2155-0182/10/$3.00

Assessing Reproducibility of Novometric Bootstrap Confidence Interval Analysis Using Multiple Seed Numbers (*Invited*)

Nathaniel J. Rhodes, Pharm.D., M.Sc.
Chicago College of Pharmacy, and the Pharmacometrics
Center of Excellence, Midwestern University

I study the role of the random seed number in affecting the reliability of a statistical finding, which in turn determines upper and lower bounds of statistical confidence in expected gain or loss yielded from associated decision-making. Simulation research reveals that obtaining a bootstrap solution consistent with an effect (e.g., the lower exact, discrete 2.5th percentile bound of the *model* bootstrap exceeds the upper exact, discrete 97.5th percentile bound of the *chance* bootstrap)—*more than once* using *any two* independent seeds—supports the hypothesis that the effect is not attributable to random chance because it was replicated vis-à-vis an independent seed. Based upon this finding I suggest consistent use of a primary seed and confirmation using a secondary seed when undertaking model qualification via simulation. This procedure reduces doubt about the veracity of borderline statistically significant findings, and eliminates false positive results identified by replication failure.

The *NOVOboot()* function, now available in the ODA package for R, generates exact, discrete novometric confidence intervals (CIs) for *model* and *chance* using 50% replacement bootstrap analyses involving 25,000 iterations each. Initial simulation analysis evaluated a relatively weak effect, an effect of moderate strength, and a relatively strong effect. For the weak effect, overlapping CIs for model and chance motivated the rejection of the alternative hypothesis. However, for moderate (slight CI overlap) and strong (no overlap) effects, the alternative hypothesis was not rejected.[1] Initial research used a random seed value of 1234, however it is unknown if the use of an arbitrary seed value can significantly influence findings in novometric CI simulation. The present study thus investigates the effect of using alternative seeds on bootstrap-generated exact discrete novometric CIs, by assessing the extent of overlap in chance-adjusted classification accuracy (ESS) of exact, discrete 95% confidence intervals (CIs) for model *vs*. chance.

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

I consider three scenarios having ESS varying from weak to strong.[1-3] In each, observations and predictions were resampled at 50% with replacement ("model"). Observations' class data were then scrambled and resampled at 50% with replacement. Resampling was conducted for model and chance to obtain *n*= 25,000 bootstrap replicates each. ESS overlap of model LB (2.5th percentile) and chance UB (97.5th percentile) was assessed. I used the function *NOVOboot()* which is available[4] in the ODA package for R: the code for the analysis conducted herein is given in the *Appendix*.

In the original research three analyses were run using a single random seed value of 1234: one analysis for a relatively weak effect, another for a moderate effect, and a third for a relatively strong effect (corresponding ESS values=24, 48, 72).[2] In this study I retain the initial findings obtained using the seed value 1234, and compare the results with findings of re-analyses obtained using the seed value 4321.

To evaluate robustness of these results I generated 25,000 bootstrap re-samples for both model and chance (scrambled class data) for 1000 different seed numbers: the code is given in the Appendix at the end of this article.

Example 1: Relatively Weak ESS (24%)

I first present an example of a model having a relatively weak effect, ESS = 24%. Table 1 is the confusion matrix obtained using the model to classify the total sample.

Table 1: Training Confusion Matrix, Relatively Weak Model

	Predicted Class	
Actual Class	class = 0	class = 1
class = 0	62	38
class = 1	38	62

Figure 1A summarizes the findings of analysis using the original seed value of 1234, and Figure 1B summarizes the analysis findings using the new seed value of 4321.

Figure 1A. Distribution of Model vs. Chance Weak ESS from 25,000 bootstrap samples (seed = 1234)

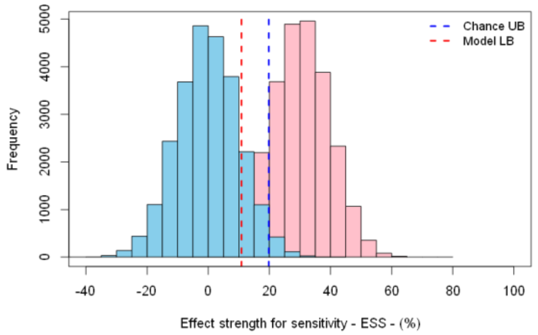
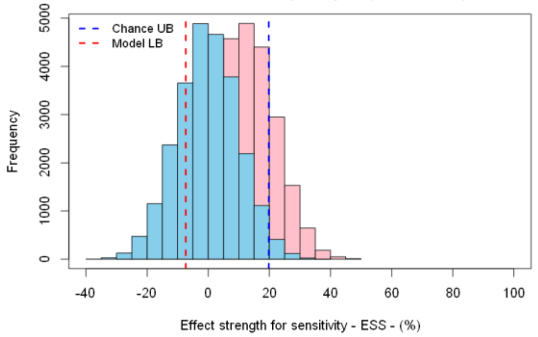


Figure 1B. Distribution of Model vs. Chance Weak ESS from 25,000 bootstrap samples (seed = 4321)



Findings demonstrate that, for model and for chance, the exact, discrete CIs overlapped: i.e., the dashed red (model 2.5th percentile) and the dashed blue (chance 97.5th percentile) lines cross relative to their respective distributions (light blue and rose, respectively). Numerical simulation results are summarized in Table 2.

Table 2: Model *vs*. Chance Bootstrap Distribution for Random Seed Values of 1234 *vs*. 4321, for Weak, Moderate and Strong ESS

ESS (%)	Seed (#)	Model 2.5th CI	Chance 97.5th CI	Exact discrete 95% CI:
24	1234	10.87	19.81	Overlap
24	4321	-7.34	19.81	Overlap
48	1234	32.77	19.80	No overlap
48	4321	28.48	19.81	No overlap
72	1234	62.10	19.63	No overlap
72	4321	48.00	19.95	No overlap

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

As seen in Table 2, classification of the overlap between the model LB and chance UB exact discrete 95% CIs was not different within each seed evaluated. Use of a secondary seed here supported the primary analysis.

Example 2: Moderate ESS (48%)

I next considered a model which achieved a moderate level of ESS (48%). Table 3 gives the confusion matrix obtained by using the model to classify the total sample.

Table 3: Training Confusion Matrix, Moderate Strength Model

	Predicted Class	
Actual Class	class = 0	class = 1
class = 0	74	26
class = 1	26	74

Figure 2A summarizes the findings of analysis using the original seed value of 1234, and Figure 2B summarizes the analysis findings using the new seed value of 4321.

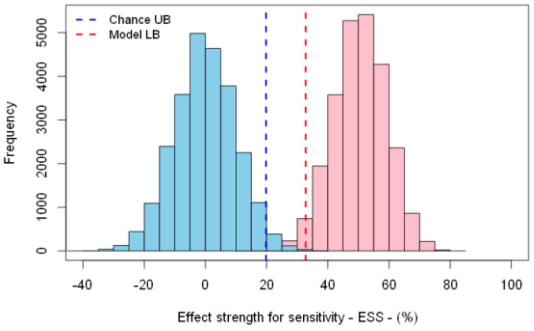


Figure 2A. Distribution of Model vs. Chance Moderate ESS from 25,000 bootstrap samples (seed = 1234)

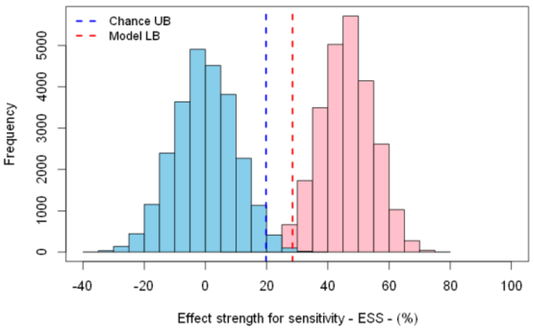


Figure 2B. Distribution of Model vs. Chance Moderate ESS from 25,000 bootstrap samples (seed = 4321)

As seen, for model (2.5% LB) and for chance (97.5% UB), exact, discrete CIs did not significantly overlap, and the red dashed line is on the same side as the rose histogram. Use of a secondary seed supported the primary analysis.

Example 3: Relatively Strong ESS (72%)

Finally, I considered a model with a relatively strong effect as measured by ESS (72%). Table 4 is the confusion matrix obtained by using the model to classify the total sample.

Table 4: Training Confusion Matrix, Relatively Strong Model

	Predicted Class	
Actual Class	class = 0	class = 1
class = 0	86	14
class = 1	14	86

Figure 3A summarizes the findings of analysis using the original seed value of 1234, and Figure 3B summarizes analysis findings using the new seed value of 4321.

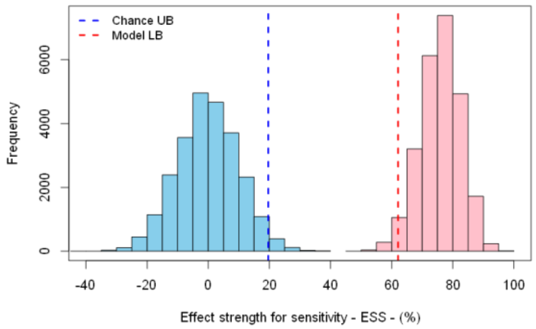


Figure 3A. Distribution of Model vs. Chance Weak ESS from 25,000 bootstrap samples (seed = 1234)

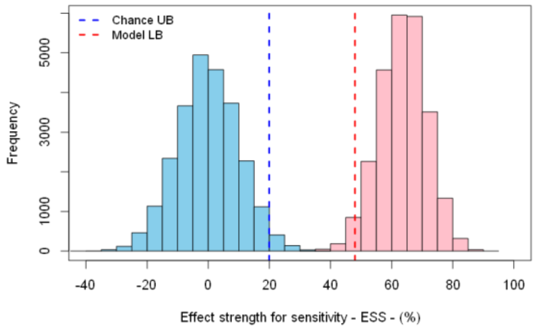


Figure 3B. Distribution of Model vs. Chance Weak ESS from 25,000 bootstrap samples (seed = 4321)

As seen, findings demonstrate that for model and for chance, the exact discrete CIs did not overlap at all using seed 1234. Overlap was observed at the extremes with seed 4321 (Model minimum=30%, *vs*. Chance maximum=37.8%), but the exact discrete 95% CI remained separate (Table 2). Again, the use of a secondary seed supported the primary analysis.

As a final confirmation I incrementally iterated through integer seed numbers 1 to 1000, and found that for a strong ESS (i.e., ≥72%) the exact discrete model and chance 95% CIs didn't overlap in any of the 1000 seeds evaluated.

In summary, in this paper I evaluated the reproducibly of a bootstrap methodology used to obtain exact discrete 95% CIs for novometric analysis using two independent seed numbers. For moderate and strong effects the results were reproducible, and the decisions regarding the overlap (or lack thereof) between exact discrete 95% CIs remained the same. Researchers who utilize seed numbers to enhance reproducibility may consider the approach of using at least two independent seeds to reflect a confirmatory step in model diagnostic workflow. The use of two or more independent seeds may be particularly salient, in terms of inhibiting false positives and thereby increasing prospective validity, for data analysis involving marginal effects that degrade in cross-generalizability analysis.

References

[1] Rhodes NJ, Yarnold PR (2020). Generating novometric confidence intervals in R: Bootstrap analyses to compare model and chance ESS. *Optimal Data Analysis*, *9*, 172-177.

[2] Yarnold PR, Soltysik RC (2005). *Optimal data analysis: Guidebook with software for Windows*. APA Books.

[3] Rhodes NJ (2020). Statistical power analysis in ODA, CTA and novometrics (Invited). *Optimal Data Analysis*, *9*, 21-25.

[4] Rhodes NJ, Yarnold PR (2020). *ODA: a package and R-interface for the MegaODA software suite.* Version: 1.0.1.3.

Author Notes

No conflict of interest was reported.

Appendix

The code for the bootstrap resampling procedure using NOVOboot within the *ODA* package for R, and for evaluation of the impact of seed number on the conclusion reached by analysis via novometric analysis.

```
#Enter data for primary confidence interval analysis
library(ODA)
library(epitools)
#Table Weak ESS 24%
data <-
matrix(c(62,38,38,62),ncol=2,nrow=2,dimnames=list(c(0,1)
,c(0,1)))
data.raw <- expand.table(data)
data.tab <-
list(table(cbind(data.raw[1],data.raw[2]),dnn=c("v1","x")))
oda.list.1 <- list()
oda.list.1[[1]] <- do.call("list",data.tab)
oda.list.1[[1]]

#Table Moderate ESS 48%
data <-
matrix(c(74,26,26,74),ncol=2,nrow=2,dimnames=list(c(0,1)
,c(0,1)))
data.raw <- expand.table(data)
data.tab <-
list(table(cbind(data.raw[1],data.raw[2]),dnn=c("v1","x")))
oda.list.2 <- list()
oda.list.2[[1]] <- do.call("list",data.tab)
oda.list.2[[1]]

#Table Strong ESS 72%
data <-
matrix(c(86,14,14,86),ncol=2,nrow=2,dimnames=list(c(0,1)
,c(0,1)))
data.raw <- expand.table(data)
data.tab <-
list(table(cbind(data.raw[1],data.raw[2]),dnn=c("v1","x")))
oda.list.3 <- list()
oda.list.3[[1]] <- do.call("list",data.tab)
oda.list.3[[1]]

#Run the bootstrap analyses and plot the CI distributions
## Weak ESS = 24%
NOVOboot(data=oda.list.1,run=1,predictor=1,outcome=1,s
eed=1234, nboot=25000)
```

```
hist(novo.boot.1$ess.model,xlim=c(-40,100),main="Figure
1A. Distribution of Model vs. Chance Weak ESS\n from
25,000 bootstrap samples (seed =
1234)",col="pink",xlab="Effect strength for sensitivity -
ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.025),
quantile(novo.boot.1$ess.chance,probs=0.975)),col=c("red
", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topright", c("Chance UB", "Model LB"),
col=c("blue", "red"), lwd=2, lty=2, cex=0.9, bty = "n")

NOVOboot(data=oda.list.1,run=1,predictor=1,outcome=1,s
eed=4321, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-40,100),main="Figure
1B. Distribution of Model vs. Chance Weak ESS\n from
25,000 bootstrap samples (seed =
4321)",col="pink",xlab="Effect strength for sensitivity -
ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.025),
quantile(novo.boot.1$ess.chance,probs=0.975)),col=c("red
", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"), col=c("blue",
"red"), lwd=2, lty=2, cex=0.9, bty = "n")

## Moderate ESS = 48%
NOVOboot(data=oda.list.2,run=1,predictor=1,outcome=1,s
eed=1234, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-40,100),main="Figure
2A. Distribution of Model vs. Chance Moderate ESS\n
from 25,000 bootstrap samples (seed =
1234)",col="pink",xlab="Effect strength for sensitivity -
ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.025),
quantile(novo.boot.1$ess.chance,probs=0.975)),col=c("red
", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"), col=c("blue",
"red"), lwd=2, lty=2, cex=0.9, bty = "n")

NOVOboot(data=oda.list.2,run=1,predictor=1,outcome=1,s
eed=4321, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-40,100),main="Figure
2B. Distribution of Model vs. Chance Moderate ESS\n
from 25,000 bootstrap samples (seed =
4321)",col="pink",xlab="Effect strength for sensitivity -
ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.025),
quantile(novo.boot.1$ess.chance,probs=0.975)),col=c("red
", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"), col=c("blue",
"red"), lwd=2, lty=2, cex=0.9, bty = "n")

## Strong ESS = 72%
NOVOboot(data=oda.list.3,run=1,predictor=1,outcome=1,s
eed=1234, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-40,100),main="Figure
3A. Distribution of Model vs. Chance Weak ESS\n from
25,000 bootstrap samples (seed =
1234)",col="pink",xlab="Effect strength for sensitivity -
ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.025),
quantile(novo.boot.1$ess.chance,probs=0.975)),col=c("red
", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"), col=c("blue",
"red"), lwd=2, lty=2, cex=0.9, bty = "n")

NOVOboot(data=oda.list.3,run=1,predictor=1,outcome=1,s
eed=4321, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-40,100),main="Figure
3B. Distribution of Model vs. Chance Weak ESS\n from
25,000 bootstrap samples (seed =
4321)",col="pink",xlab="Effect strength for sensitivity -
ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.025),
quantile(novo.boot.1$ess.chance,probs=0.975)),col=c("red
", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"), col=c("blue",
"red"), lwd=2, lty=2, cex=0.9, bty = "n")

## Iterating through 1000 seed numbers to evaluate
stability of results for a strong ESS of 72% ##
boot <- list()
for(j in seq(1,1000,1)){
  boot[[j]] <-
NOVOboot(data=oda.list.3,run=1,predictor=1,outcome=1,s
eed=j,nboot=25000)
}

boot2 <- do.call(cbind, boot)
boot2 <- as.data.frame(boot2)
boot3 <- list()
for(i in seq(1,length(boot2),2)){
  boot3[[i]] <- cbind(boot2[[i]][2], boot2[[i+1]][6])
}

bootlist <- list()
for(i in seq(1,length(boot2),2)){
  bootlist[[i]] <- ifelse(boot2[[i]][2] > boot2[[i+1]][8],1,0)
}
mean(as.numeric(unlist(bootlist)),na.rm=T)
```