# Outline of a Simulation Approach to a *reductio ad absurdum* for an Outcomes Imputation Procedure of Kessler et al

David C. Norris, MD
David Norris Consulting, LLC
david@dnc-llc.com

August 28, 2015

**Abstract**

A simulation approach is motivated and developed, by which the outcomes imputation procedure of Kessler et al (*JAMA*. 2014;311(9):937-947) may be criticized. To provide a concrete basis for understanding this otherwise opaque imputation procedure, we first develop a causal model that appears to be at least consistent with the procedure, if not deducible from it. This causal model is then employed as a data-generating process (DGP) for simulating the underlying data, and thence the Kessler et al analysis itself. With the published analysis itself being simulable, we are able to demonstrate a pattern of forensic re-analysis that can be used to abstract several layers of arbitrariness from the published effect estimates in the Kessler et al paper.

## 1 Introduction

In a March 2014 publication in *JAMA*, Kessler et al reported that, at follow-up 10 to 15 years later, boys from households receiving housing vouchers in the Moving to Opportunity Demonstration experienced 12-month prevalence of posttraumatic stress disorder (PTSD) at several times the rate of boys from control households. Because the DSM-IV PTSD criteria were not fully operationalized in the MTO Final Youth Survey questionnaire, the PTSD 'outcomes' analyzed in this paper were in fact *imputed* using a logistic regression model estimated against the National Comorbidity Survey Replication(NCSR), which operationalized these criteria more completely.[1]

There are many levels on which this imputation procedure can be criticized:

1. The opaqueness of this *outcomes imputation* to a reader of the research report is problematic, especially since the report described in some detail the comparatively workaday matter of imputing missing *covariates*.

2. The choice of model form for the logistic regression appears to be of the desultory kind that would be appropriate for imputing partially missing covariates, not a completely missing outcome. The original analysis did not address how this imputation model was selected, whether the possibility of interaction terms or nonlinearity were explored, nor address the question of overfitting through bootstrap validation.

3. The NCSR population on which the imputation model was estimated comprises adults in the general population (TODO: verify this!), and generalizability to the MTO adolescent population would seem questionable.

[1] Although the imputation of partially missing *covariates* was described in some detail in the *JAMA*article itself, this *outcomes imputation* procedure was not. It came to this author's attention through a footnote on page 38 of this document, and confirmed through personal communication with Dr. Kessler.

4. The decision to analyze a *singly* imputed PTSD outcome introduces a superfluous, information-destroying and noise-generating transformation of the predictive model's real-valued outputs–logit probabilities on the continuous interval $(-\infty, \infty)$–to pseudorandom 'outcomes' in the set *no*, *yes*. In order to defend the conclusions about these imputed outcomes as conclusions about *genuine PTSD*, one would have to defend the outputs of the predictive model as *genuine probabilities*. But if these were real probabilities, then they could be analyzed directly without interposing a noisy and information-destroying pseudorandom number generation (RNG) step. The reported confidence intervals around the reported effects estimates are in fact a manifestation strictly of this RNG step, acting therefore almost like a decoy that distracts attention from the the substantive sources of uncertainty in the reported findings, which remain therefore as opaque as the fact that a completely missing outcome was imputed.

The core challenge addressed in this paper is to show how a focused criticism *of a strictly statistical kind* may be developed around points 4 and 2 above, avoiding the admixture of the questions of scientific judgement implicit in points 1 and 3. To accomplish this, we develop a causal model under which the imputation procedure might be thought *scientifically* appropriate. This model then 'motivates' the imputation procedure, and also provides a data-generating process (DGP) suitable for simulation experiments that address the statistical properties of the procedure in isolation from the larger scientific questions.

Of course, having thus clarified what was actually done in the original analysis, and perhaps showing what the motivation for it might have been, a productive conversation about the larger scientific questions can ensue.

## 2 A model to motivate the imputation procedure

Let us suppose that there are two studies, called 'MTO' and 'NCSR', that they describe two largely identical populations, and that they administer questionnaires intended to diagnose a disease $Y$. The NCSR study administers a more comprehensive questionnaire that includes two sets of questions, $S_1$ and $S_2$, whereas for reasons of cost and burden on participants, the MTO study administers only the $S_1$ questions. Suppose that the questions $S_1 \cup S_2$ suffice to deliver a highly accurate diagnosis $Y$ in the NCSR study–perhaps because $Y$ is *defined* as a deterministic function of the responses to $S_1 \cup S_2$. Whereas NCSR is an observational study, MTO is an social experiment in which the treatment is a voucher, $V$. It is desired to know how $V$ affected $Y$ in the MTO study, but because the $S_2$ questions were not asked, the $Y$ outcome is effectively missing. Might there be a way to use the NCSR data to impute this missing outcome?

Let us suppose that, in addition to the variables mentioned above, the MTO and NCSR studies also share observations on baseline characteristics $X$ of the study participants. Then, from the perspective of the MTO investigators, the relevant variables might be thought to be connected causally as in Figure 2.

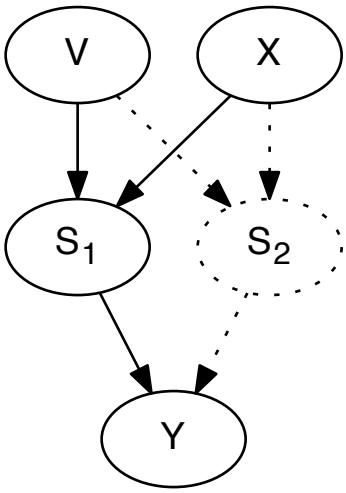


Figure 1: A view of the inferential problem for the MTO investigator. The outcome of interest is determined by $S_1, S_2$, but $S_2$ is unmeasured.

Under these circumstances, it might be hoped that the deterministic relationship $Y = f(S_1, S_2)$ could be replaced in the MTO study by a *statistical* relationship extracted from the NCSR study by the regression of $Y$ on $S_1$ in the latter. Given the availability of additional measures $X$ shared between the two studies, it might further be supposed that these should be included on the right-hand side of the regression. Thus, the following procedure might suggest itself:

(A) Estimate a logistic regression with the known NCSR $Y$ on the left-hand side, and regressors $S_1, X$ on the right-hand side

(B) Assume that the statistical relationship hereby discovered in NCSR will apply also in MTO

(C) Use the model to estimate the probability $\hat{P}_i$ of $Y_i = 1$ for each MTO study participant $i$

(D) In order to avoid modeling probabilities directly, use the estimated $\hat{P}_i$ to generate a pseudorandom $\hat{Y}_i$ for each MTO participant

(E) Regress $\hat{Y}$ on $V$ and other baseline characteristics, as if $\hat{Y}$ were a measured outcome

(F) Identify the estimated $V$ coefficient with $V$'s causal effect on the real, but unknown $Y$

(G) Present the confidence interval (CI) around this logistic regression coefficient as a full measure of the uncertainty about the inferential outcome of this procedure

Expressed in this way, this procedure appears vulnerable on numerous counts. What we wish to accomplish here, however, is to single out the *purely statistical* flaws in the procedure, as manifested in: failing to account for model selection and overfitting in step (A); introducing pseudorandom noise in step (D); presenting the CIs from the regression as a full account of the uncertainty in the effects estimates—i.e., step (G).

## 3 Simulation-based examination of imputation errors

We simulate the above model, as follows:

```
# Simulate 1400 MTO boys and 5000 NCSR males
N.mto <- 1400
N.ncsr <- 5000
N <- N.mto + N.ncsr
# Suppose S1, S2 and X are independent standard normal variates
df <- data.frame(study = c(rep("NCSR", 5000), rep("MTO", 1400)), S1 = rnorm(N),
    S2 = rnorm(N), X = rnorm(N))
# Generate Y deterministically as a function of S1 and S2, namely as
# the upper 6% tail of (S1+S2)
df <- upData(df, Y = S1 + S2 > qnorm(0.94, sd = sqrt(2)))

## Input object size:  180656 bytes;  4 variables
## Added variable  Y
## New object size:  206376 bytes;  5 variables

# Split the joint NCSR+MTO data into separate data sets, dropping the
# now-unnecessary 'study' variable.
ncsr <- subset(df, study == "NCSR", select = -study)
mto <- subset(df, study == "MTO", select = -study)
# Drop the unmeasured S2 data from 'mto' and rename Y
mto <- upData(mto, drop = "S2", rename = list(Y = "Y.actual"))
```

```
## Input object size:  45776 bytes;  4 variables
## Renamed variable  Y  to Y.actual
## Dropped variable  S2
## New object size:  34496 bytes;  3 variables

# Estimate an 'outcome imputation model' in the NCSR data
dd <- datadist(ncsr)
oim <- lrm(Y ~ S1 + X, data = ncsr)
# Use the data to estimate P.hat in MTO, then singly impute Y.imp
mto$P.hat <- predict(oim, newdata = mto, type = "fitted.ind")
mto$Y.imp <- runif(N.mto) < mto$P.hat
```

The first question one might ask is, how accurate was the single imputation?

```
s <- summary(cbind(`Y.imp=Yes` = Y.imp, `Y.imp=No` = !Y.imp) ~ Y.actual,
    method = "response", data = mto)
attr(s, "ylabel") <- "Accuracy of imputation"
latex(s, file = "", label = "tbl:2x2", where = "htbp")
```

Table 1: Accuracy of imputation N=1400

| | N | Y.imp=Yes | Y.imp=No |
|---|---|---|---|
| **Y.actual** | | | |
| No | 1319 | 0.05 | 0.95 |
| Yes | 81 | 0.23 | 0.77 |
| **Overall** | | | |
| | 1400 | 0.06 | 0.94 |

Thus, we see that 3 out of every 4 true diagnoses are missed by the imputation. Clearly, this kind of tabulation would be impossible in a forensic reproduction of the analysis of Kessler et al, since the *Y.actual* is unobserved. However, if the estimated probabilities of Kessler et al are taken at face value, then a *reductio ad absurdum* argument becomes readily available, such that the *expectation* of the error rate among imputed 'yes' diagnoses is at least recoverable:

```
# Assuming the estimated probabilities 'P.hat' are correct, the
# expected number of true 'yes' diagnoses is given straightforwardly by
# this calculation:
expected.N.really.yes <- with(mto, sum(P.hat))
# Assuming that the pseudorandom number generators used are reasonably
# good, our expectation of the number of imputed 'yes' diagnoses is the
# same:
expected.N.imputed.yes <- expected.N.really.yes
# The number of imputed 'yes' diagnoses that may be expected to be
# incorrect is:
expected.N.incorrect.yes <- with(mto, sum(P.hat * (1 - P.hat)))
# The *fraction* of imputed 'yes' diagnoses expected to be wrong is:
expected.N.incorrect.yes/expected.N.imputed.yes

## [1] 0.68867

# This result can be computed more directly via:
with(mto, mean((1 - P.hat) * P.hat)/mean(P.hat))

## [1] 0.68867
```

The close agreement with the corresponding figure from Table 1 should be noted. A similar calculation is immediately available from a reproduction of the Kessler et al analysis, provided that the probabilities corresponding to $\hat{P}_i$ are set to zero for those individuals $i$ who do not meet criteria for PTSD.

# 4 Simulation-based examination of effect estimates

To estimate the voucher effect in our simulated MTO data set, we require a treatment variable $V$. We simulate a binary treatment[2] consistent with a 'null hypothesis' of *no effect*:

```
mto$V <- rbinom(N.mto, 1, p = 0.5)
dd <- datadist(mto)
fit <- lrm(Y.imp ~ V + X, data = mto)
latex(summary(fit), file = "", label = "tbl:effects", where = "htbp")
```

Table 2: Effects Response : Y.imp

| | Low | High | Δ | Effect | S.E. | Lower 0.95 | Upper 0.95 |
|---|---|---|---|---|---|---|---|
| V | 0.00000 | 1.00000 | 1.0000 | 0.135250 | 0.21967 | -0.29529 | 0.56579 |
| *Odds Ratio* | 0.00000 | 1.00000 | 1.0000 | 1.144800 | | 0.74431 | 1.76080 |
| X | -0.67551 | 0.69811 | 1.3736 | 0.045404 | 0.15207 | -0.25264 | 0.34345 |
| *Odds Ratio* | -0.67551 | 0.69811 | 1.3736 | 1.046500 | | 0.77675 | 1.40980 |

The first point we would like to make is that the standard errors reported in connection with this effect estimate merely describe the *noise* introduced by the RNG required for the imputation. That is, these standard errors (and the associated confidence intervals) are *purely artifactual*. To make this demonstration, we employ a loop that reifies the frequentist conceit of 'multiple parallel universes' in which different random seeds are chosen for the imputation:

```
NU <- 500  # number of frequentist 'Universes' to simulate
beta <- numeric(NU)  # vector to hold NU effect estimates ..
serr <- numeric(NU)  # ... and standard errors
for (U in 1:NU) {
    # 'U' for 'Universe' It is of course unnecessary to set the RNG seed
    # explicitly within this kind of loop; we do this solely as an aid to
    # the exposition.
    set.seed(U)
    ## Re-impute the Y.imp values (note: the estimated P.hat remain fixed
    ## within this loop).
    mto$Y.imp <- runif(N.mto) < mto$P.hat
    ## Estimate the effects model
    dd <- datadist(mto)
    fit <- lrm(Y.imp ~ V + X, data = mto)
    ## Extract the treatment effect coefficient and its standard error
    beta[U] <- fit$coefficients["V"]
    serr[U] <- sqrt(fit$var["V", "V"])
}
```

[2]The housing voucher treatment in MTO was actually more complex, involving 2 distinct types of housing voucher. As with the rest of our developments here, we abstract away from such details.

```r
# Now demonstrate that the estimated betas have nearly the normal
# distribution implied by the reported standard errors:
estimates <- data.frame(series = "effect estimate", Beta = beta)
normality <- data.frame(series = "asymptotic theory", Beta = rnorm(50000,
    mean = mean(beta), sd = serr))
demo <- rbind(estimates, normality)
demo$series <- as.factor(demo$series)
densityplot(~Beta, groups = series, data = demo, rug = FALSE, auto.key = TRUE,
    plot.points = FALSE)
```

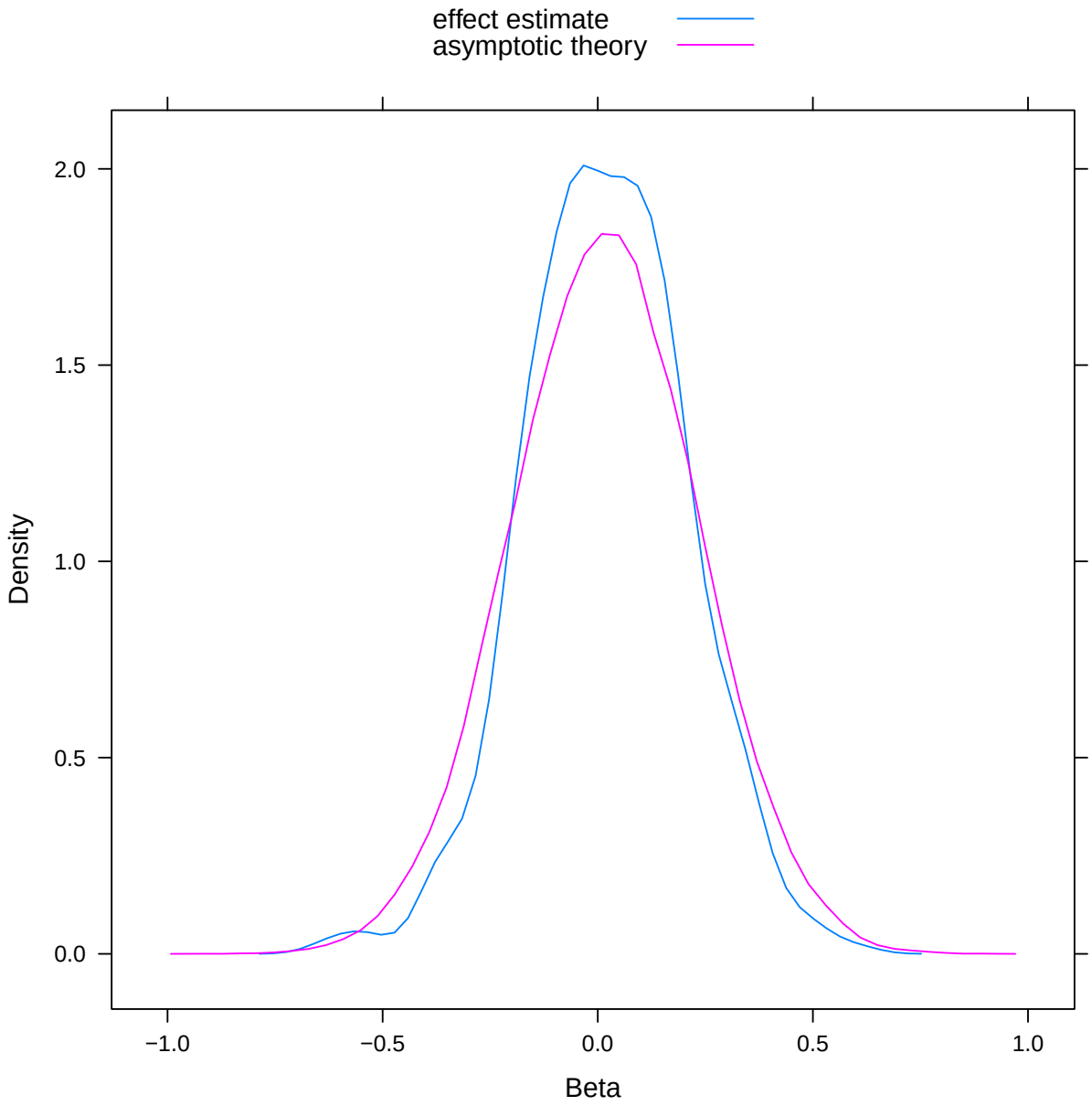


Figure 2: The distribution of 500 voucher effect estimates with different random seeds used for PTSD diagnosis imputation, compared with a normal density with standard deviation equal to the average standard error across these fits. The close alignment of these two densities demonstrates that the standard errors reported in these procedures describes *only* the noise introduced by the random number generation employed during the outcomes imputation. That is, the reported standard errors of the effects estimates derived by the outcomes imputation procedure of Kessler et al are *strictly artifactual*.

# 5 Abstracting a higher order of arbitrariness

The simulation-based analysis of the previous section abstracted away only the most proximate level of arbitrariness immanent in the effect estimate reported in Table 2, namely the arbitrariness introduced by pseudorandom number generation during the outcomes imputation step. (As was shown above, in fact, this arbitrariness is precisely what the standard errors from the effects regression accounts for.) But this is by no means the only arbitrariness present in the estimated effect; a further level of arbitrariness derives from *overfitting* of the imputation model. In this section, Efron's bootstrap is used to abstract away this source of arbitrariness.

To illustrate the concept that two nested 'layers of arbitrariness' are being abstracted, we retain the `for` loop from the code above, but wrap it inside an outer loop that implements the bootstrap.

```
B <- 100  # number of bootstrap samples
NU <- 10  # number of frequentist 'Universes' to simulate per bootstrap re-estimation
betaM <- matrix(numeric(B * NU), nrow = B, ncol = NU)  # one matrix to hold B*NU effect estimates..
serrM <- betaM  # ... and another for the standard errors
for (i in 1:B) {
    ## In each bootstrap iteration, re-estimate the imputation model against
    ## the original data *resampled with replacement* ...
    ncsr.resamp <- ncsr[sample(nrow(ncsr), replace = TRUE), ]
    dd <- datadist(ncsr.resamp)
    oim <- lrm(Y ~ S1 + X, data = ncsr.resamp)
    ## .. then use the model to estimate P.hat in MTO.
    mto$P.hat <- predict(oim, newdata = mto, type = "fitted.ind")
    ## With these estimated probabilities in hand, we can proceed as before:
    ## 'U' for 'Universe' It is of course unnecessary to set the RNG seed
    ## explicitly within this kind of loop; we do this solely as an aid to
    ## the exposition.
    for (U in 1:NU) {
        set.seed(U + (i - 1) * U)  # seed runs through 1..B*NU
        ## Re-impute the Y.imp values (note: the estimated P.hat remain fixed
        ## within this loop).
        mto$Y.imp <- runif(N.mto) < mto$P.hat
        ## Estimate the effects model
        dd <- datadist(mto)
        fit <- lrm(Y.imp ~ V + X, data = mto)
        ## Extract the treatment effect coefficient and its standard error
        betaM[i, U] <- fit$coefficients["V"]
        serrM[i, U] <- sqrt(fit$var["V", "V"])
    }
}
```

```
# Now demonstrate that the estimated betas have nearly the normal
# distribution implied by the reported standard errors:
estimates <- data.frame(series = "effect estimate", Beta = beta)
normality <- data.frame(series = "asymptotic theory", Beta = rnorm(50000,
    mean = mean(beta), sd = serr))
bootstrap <- data.frame(series = "bootstrapped", Beta = as.vector(betaM))
demo <- rbind(demo, bootstrap)
demo$series <- as.factor(demo$series)
densityplot(~Beta, groups = series, data = demo, rug = FALSE, auto.key = TRUE,
    plot.points = FALSE)
```

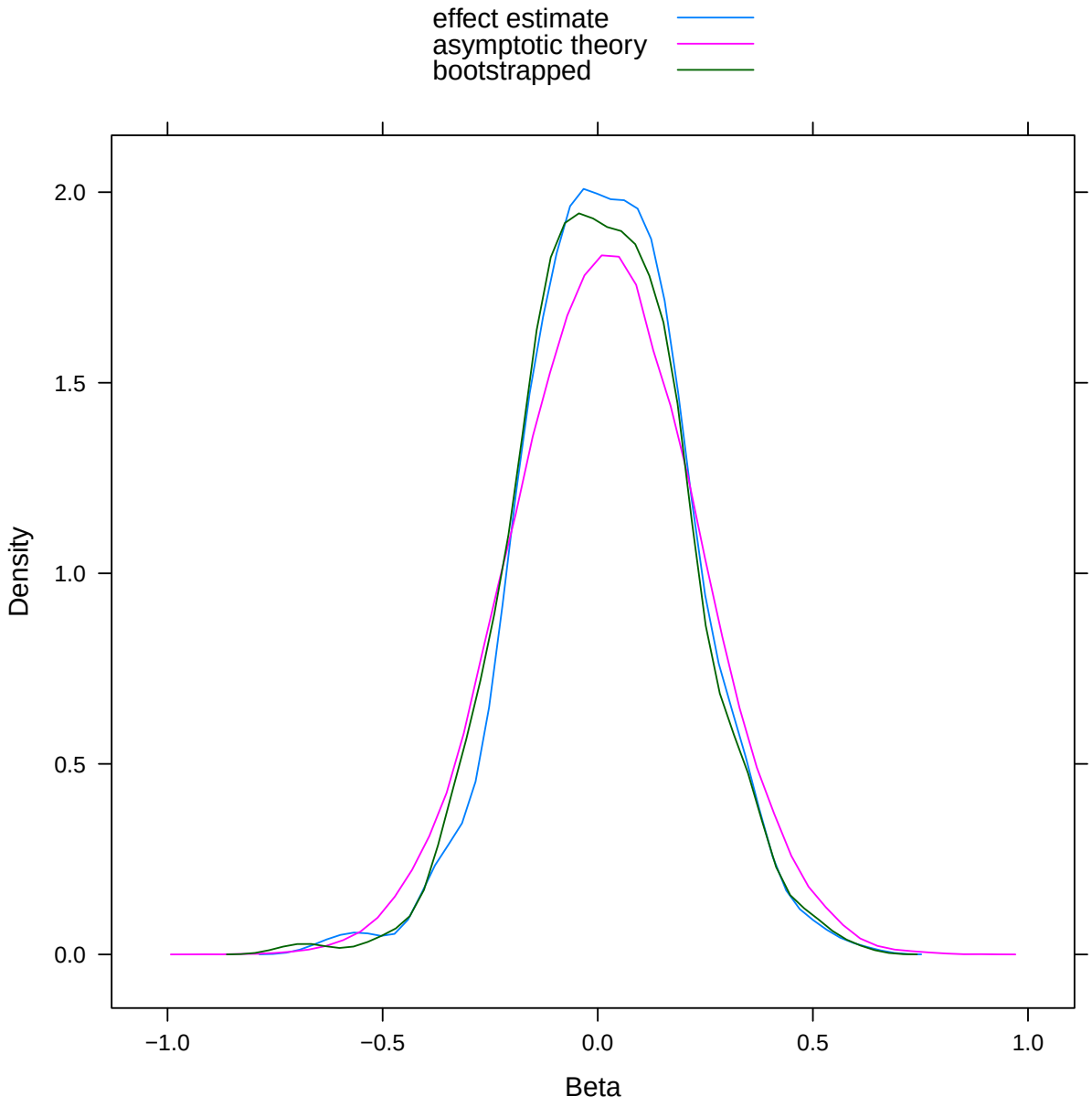


Figure 3: The distribution of 1000 voucher effect estimates with nested simulation loops performing 10 re-imputations within each of 100 bootstrapped re-estimations of the imputation model.

# A Imputation model employed by Kessler et al

TODO: Format the imputation model, with coefficients as provided by Ronald Kessler.