Reproducibility of Dressman 2007

VJ Carey

July 20, 2010

The dispute between Baggerly and Coombes (BC) and Dressman, Potti and Nevins (DPN) in J Clin Oncology, 2008; 26(7):1186-1187, is of broad interest. Because general reproducibility of genome-scale data analysis has been questioned at high levels (cite Ioannidis and others), specific debates on difficulties of reproduction of analytic findings should be examined carefully. Study of these debates can reveal patterns of analysis and interpretation that should be avoided to reduce risks of conflicts over reproducibility, or promoted to increase ease with which new findings can be prioritized for application and extension.

The response of DPN to BC is harsh and dismissive. DPN write that "[t]o 'reproduce' means to repeat, following the methods outlined in an original report", and then argue that BC's claims are erroneous and inconsistent. The debate is difficult to follow in detail, in part because much of the work of BC is devoted to correcting clerical errors in the archive for the original paper, in part because the primary documentation of BC's findings is spread over 7 substantial supplementary documents and hundreds of supporting data files, and in part because BC's investigations of non-reproducibility are interlarded with investigations of confounding. Confounding is a common problem for statistical analysis of observational data, and epidemiologists frequently hedge their conclusions with admissions that unmeasured and uncorrected confounding could be responsible for some component of observed associations. BC's primary claims are a) good faith attempts to reproduce findings such as Dressman et al.'s Figures 2B and 2C do not succeed on the basis of the published data, even after admitted clerical errors are corrected, and b) any findings associating pathway activation with survival that can be teased out of the public archive become nonsignificant when confounding adjustments are made.

The response by DPN has three main components. First, they admit various clerical errors but assert that these errors are connected only with the creation of the public archive and no not affect the paper's inferences. Second, they argue that post-RMA processing using sparse factor regression would eliminate any artifacts due to batch and thus eliminate confounding. Third, they argue that BC did not use the same methods of analysis as the original paper and therefore cannot produce any judgment on the reproducibility of Dressman's analyses.

The second and third components of response are of serious concern. The claim that sparse factor regression can *completely* eliminate batch-related artifacts seems unsupportable, although DPN would be completely justified to argue that they did something reasonable – possibly something very effective – to reduce effects of array batch in their analysis. Residual confounding could still be present and should be acknowledged. Additional analyses "adjusting for batch" would be called for. This makes the third component of DPN's response extremely disconcerting. The fact is that BC did use the "same methods of analysis" as Dressman et al. – BC implicitly used the sparse factor regression adjustments to RMA when using "Dressman's quantifications" in their supplementary analyses. Using these quantifications, BC are able to show that considerable numbers of genes have distributions that vary significantly across array batches, despite the corrections introduced by sparse factor regression. There are passages in BC's analyses, for example p 16 of ovca6.pdf, that suggest that sparse factor regression does tend to diminish batch effects (XLS patterns show less clumping by batch than CEL patterns), supporting DPN's hopes for reduced confounding. However, the confounding persists. The only analysts who tested for impact of batch on the pathway activation:survival association using Dressman's data were BC.

In the course of preparing an invited chapter on reproducible research discipline in a forthcoming volume on cancer bioinformatics, I decided to try to get to the bottom of the BC-DPN conflict in as independent and concise a manner as possible. The key targets of my investigation were Figures 2B and 2C of the Dressman paper. The necessary ingredients are 1) transcript profiles and survival times for patients analyzed in those figures and 2) coefficients for the tumor scoring procedure that determines activity of Src and E2F3 for each sample.

The archive at `http://data.cgt.duke.edu/platinum.php` includes the 'corrected RMA' transcript profiles, clinical data including survival time and authors' determination of clinical responsiveness or nonresponsiveness to platinum therapy, and a large collection of raw CEL files related to the study. I assume that 'corrected RMA' refers to application of 'sparse factor regression' to remove artifacts such as batch effects from the RMA transformations. Coefficients for the tumor scoring procedure are not available, to the best of my knowledge, but the cell-line transcript profiles of Bild et al. can be obtained from GSE3151 and the singular value decomposition can be used to obtain coefficients facilitating sample scoring.

Figure 1(a) shows the association between Src dysregulation and survival among nonresponders using Dressman et al.'s 'corrected RMA' with the given sample identifiers. Figure 1(b) shows the same association after sample identifiers are redefined using the method of BC to map Dressman's quantifications to the original CEL files which are assumed to be accurately labeled to correspond to records in the clinical data. Figure 1(b) is a closer approximation to Dressman's original Figure 2B, and shows that BC are correct when they say that the labels in the 'corrected RMA' archive need to be revised. The mild discrepancies between Figure 1(b) and Dressman's original 2B might be explained through differences in tumor scoring coefficients, or through differences in

exact numbers of patients/events available – only 116 of 119 transcript profiles could be reliably mapped to CEL files for relabeling. In any case, Figure 1(b) suggests that we can use the public archive, with some adjustments, to technically reproduce original findings to a reasonable approximation.

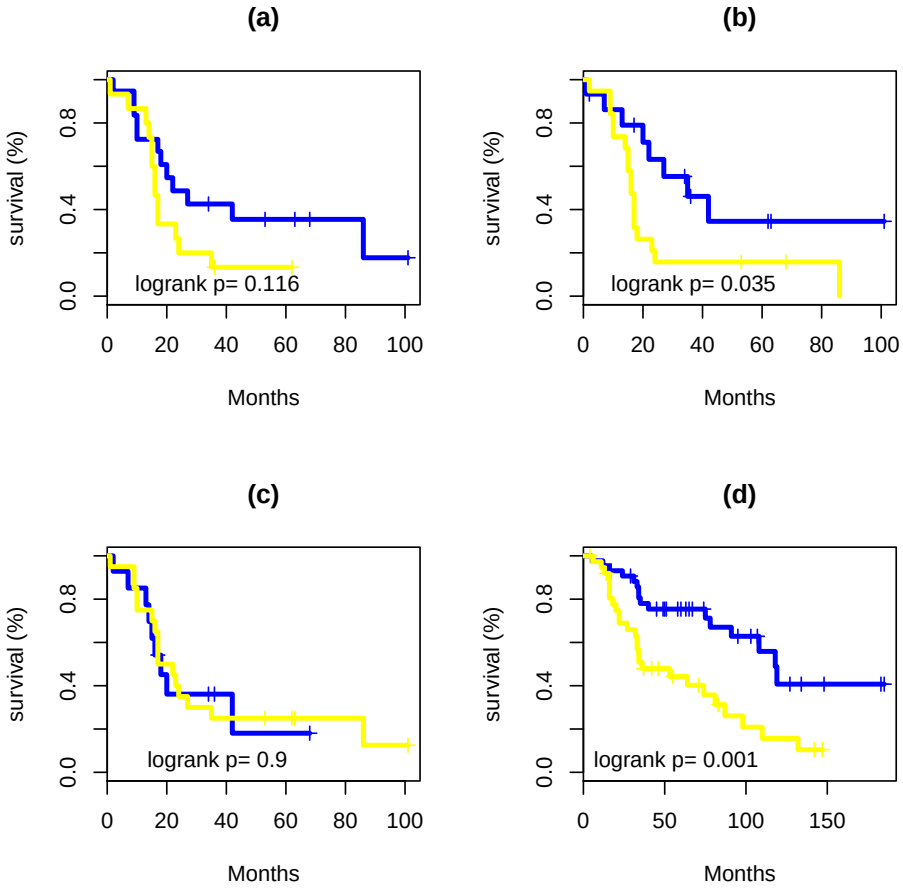
Figures 1(c) and 1(d) are more troubling. Figure 1(c) reproduces the methods for Figure 1(b) in application to E2F3 activation. Among patients labeled as nonresponders by Dressman in the public archive, there is no association between pathway activation and survival. Figure 1(d) shows that there is an association, but among the responders. These findings are congruent with those of BC, who explored more variations on the data sources and could not recover the finding of the original Figure 2C.

We now have two basic problems. First, ignoring concerns about confounding, and using only (and all) quantifications provided by Dressman, and thereby 'following their methods', we cannot reproduce Figure 2C. Second, the significance of the association indicated in my Figure 1(b) is lost when a simple allowance for batch effects is made in the test for different survival distributions by pathway activation.

```
Call:
survreg(formula = Surv(Survival, dead) ~ sdys + poly(chron(rundate),
    2), subset = CR == 0)
                            Value Std. Error      z        p
(Intercept)                 3.767      0.298 12.657 1.02e-36
sdys                       -0.292      0.402 -0.726 4.68e-01
poly(chron(rundate), 2)1    5.805      1.929  3.009 2.62e-03
poly(chron(rundate), 2)2 -1.567      1.711 -0.916 3.60e-01
Log(scale)                 -0.250      0.168 -1.489 1.36e-01

Scale= 0.779

Weibull distribution
Loglik(model)= -109   Loglik(intercept only)= -115.2
        Chisq= 12.32 on 3 degrees of freedom, p= 0.0064
Number of Newton-Raphson Iterations: 6
n= 34
```


(a)
survival (%)
0.0
0.4
0.8
logrank p= 0.116
0
20
40
60
80
100
Months
(b)
survival (%)
0.0
0.4
0.8
logrank p= 0.035
0
20
40
60
80
100
Months
(c)
survival (%)
0.0
0.4
0.8
logrank p= 0.9
0
20
40
60
80
100
Months
(d)
survival (%)
0.0
0.4
0.8
logrank p= 0.001
0
50
100
150
Months