Reproducibility of Dressman JCO 2007

VJ Carey

July 20, 2010

In the light of recent high-level challenges to reproducibility of microarray studies (Ioannidis 2009 and others) 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.

In the 2007 paper that is the subject of BC's correspondence, Dressman et al. (JCO 25(5):517-25) studied 119 ovarian cancers with respect to responsiveness to platinum therapy and presented analyses showing association of Src and E2F3 pathway activation with survival time in platinum non-responsive tumors. BC discovered a number of clerical errors in the published archive associated with the 2007 paper, corrected the errors to the best of their ability, extracted information to group arrays into chronologically ordered batches, and presented graphics and analyses indicating that expression distributions of some genes, and survival time distributions, varied systematically across batches. In analyses of the pathway activation:survival associations that adjust for batch effects, pathway activation was no longer statistically significant. One way of reading BC's contribution is as an identification of a published association that is based, at least in part, on uncontrolled confounding. It happens; the title given to the correspondence is consistent with this gentle reading: "Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer".

The response by DPN is vehement but has two main substantive components. First, they argue that confounding could not be present because RMA preprocessing was followed by sparse factor regression corrections that would eliminate artifacts like batch effects. Second, they argue that BC have no basis challenging reproducibility of the original paper because BC do not 'follow the methods outlined in [the] original report."

The first response indicates excessive trust in complex statistical modeling to remove confounding effects in the analysis of observational data. Although there are indications that sparse factor regression can ameliorate batch effects, there is no reason to believe that it can eliminate them completely, and BC provide evidence, using Dressman's published quantifications, which incorporate the benefits of sparse factor regression, that batch effects remain. See Figure 1(a) below, which shows clear structural variation in both location and spread of RPS11, an element of the Src pathway signature, across batches. The expression measures are those used by Dressman et al., and thus are denoted RMA+SFR (RMA followed by sparse factor regression).

The second response by DPN is flatly incorrect. BC do follow the methods outlined in the original report with one minor exception. Because coefficients used to score tumors for pathway activation have not been published, not by the original authors of the cell-line based method for decisionmaking on pathway activation (Bild et al., 2006), and not by Dressman et al., BC use Bild's published data to derive their own coefficients. The derivation seems straightforward enough, and in fact allows rather close reproduction of Dressman's Figure 2B.

I was able to approximately reproduce Dressman's original Figure 2B myself, independently of code disseminated by BC in the supplement to their letter – see Figure 1(b) below. This figure is an excellent sanity check – it shows that one of the associations published in the original data is approximately recoverable using published archival data. BC had a similar display

```
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)