Reply to 'Poor trial design leaves gene therapy death a mystery'

To the editor:

We were heartened to see that *Nature Medicine* chose to cover the National Institutes of Health (NIH) Recombinant DNA Advisory Committee (RAC)'s review of the death of Jolee Mohr, which occurred while she was enrolled in a gene therapy trial[1], but we are concerned that the article did not reflect the panel's discussion and conclusions.

The RAC committee members were mindful of the dangers of drawing definitive conclusions on the basis of limited information and did not declare a cause of death in the case. Still, the data they were presented with clearly showed that the amount of gene therapy product outside the knee—where it was injected and where it was designed to remain—was negligible. Instead, the panel was presented with extensive evidence that systemic antiarthritis immunosuppressants, which Mohr was also taking, have been definitively linked to opportunistic infections such as the histoplasmosis that contributed to her death.

Yet the article, in quoting a witness who was neither a member of the blue-ribbon NIH committee nor an immunologist or rheumatologist, gave the impression that the gene therapy Mohr received remains a prime candidate for a cause of death, a supposition that was not borne out by the evidence shown at the meeting.

We are also concerned that the headline, "Poor trial design leaves gene therapy death a mystery," suggests that the phase 1/2 trial of our therapy for rheumatoid arthritis was not designed according to current good practices because patients were permitted to use other arthritis drugs. Yet continuance of maintenance therapy has been a mainstay of trial design in this or any field. This protocol underwent a most rigorous screening process and was evaluated not only by leading rheumatologists and the institutions that performed the research, but also by the NIH itself. Indeed, the assertion that the trial design was flawed was never made by any member of the RAC during the meeting, so the claim that this was an official outcome of the meeting is troublesome.

Mohr's death was a tragedy. We at Targeted Genetics have made every effort to assist the doctors, pathologists, researchers and regulatory bodies who are seeking to better understand why she died. We respect the conclusions of the NIH's expert panel, which reviewed all the available information and cast no aspersions on the gene therapy product, and we hope that their measured assessment will receive more prominent mention in the future.

H Stewart Parker

President and CEO, Targeted Genetics Corporation, 1100 Olive Way, Ste. 100, Seattle, Washington 98101, USA.
e-mail: Stewart.Parker@targen.com

COMPETING INTERESTS STATEMENT

The author declares competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemedicine/.

1. Wadman, M. *Nat. Med.* **13**, 1124 (2007).

Microarrays: retracing steps

To the editor:

Recently, Potti *et al.*[1] published an article in *Nature Medicine* reporting an approach predicting whether a tumor will respond to chemotherapy. Using publicly available data, they derived signatures from microarray profiles of the NCI-60 human cancer cell lines with known *in vitro* sensitivity or resistance to a particular drug. They used these profiles to predict *in vivo* chemotherapeutic response to seven different drugs. In order to help investigators at our institution use similar approaches, we tried to reproduce their results. We used the same published data and additional information generously supplied by the authors regarding methods, lists of cell lines called sensitive or resistant, and the software used to perform their analysis.

We report here our inability to reproduce their findings. Details of our methods and results are described in the supplementary information (**Supplementary Reports 0–9**) and are summarized here.

1. We cannot reproduce their selection of cell lines. The most sensitive and resistant lines should be used to focus on drug effects. However, the GI_{50} (the concentration needed to reduce the growth of treated cells to half that of untreated cells) concentrations for their sensitive and resistant lines overlap (**Supplementary Report 3**). Our analyses used both their cell lines and ones we selected independently.

2. The lists of genes initially reported in the supplementary information on the *Nature Medicine* website[1] are wrong because of an 'off-by-one' indexing error (**Supplementary Report 9**): for example, probe set 1881_at was reported instead of probe set 1882_at. These lists were revised but are still incorrect.

3. Using their software and lists of cell lines, we reproduced their published heatmaps for six out of seven drugs. (We could not reproduce the heatmap for cytoxan.) However, after correcting for the off-by-one error, we matched the reported gene lists exactly for only three out of seven drugs (**Supplementary Report 9**). The other lists contain outliers.

4. For docetaxel, their software yields only 31 of their 50 reported genes. Of the remaining 19 (**Supplementary Report 9**), Chang *et al.*[2] name 14 as useful discriminators in the paper that described the test set used by Potti *et al.*[1]. We do not know how these 19 can be obtained from the training data, and we suspect that they were included by mistake. The model may more easily predict test classes with these genes.

5. Their software does not maintain the independence of training and test sets, and the test data alter the model. Specifically, their software uses 'metagenes': weighted combinations of individual genes. Weights are assigned through a singular value decomposition (SVD). Their software applies SVD to the training and test data simultaneously, yielding different weights than when SVD is applied only to the training data (**Supplementary Report 9**). Even using this more extensive model, however, we could not reproduce the reported results.

6. The interaction between point 4 (accidentally including genes from Chang *et al.*[2] whose expression levels separate the test set responders from

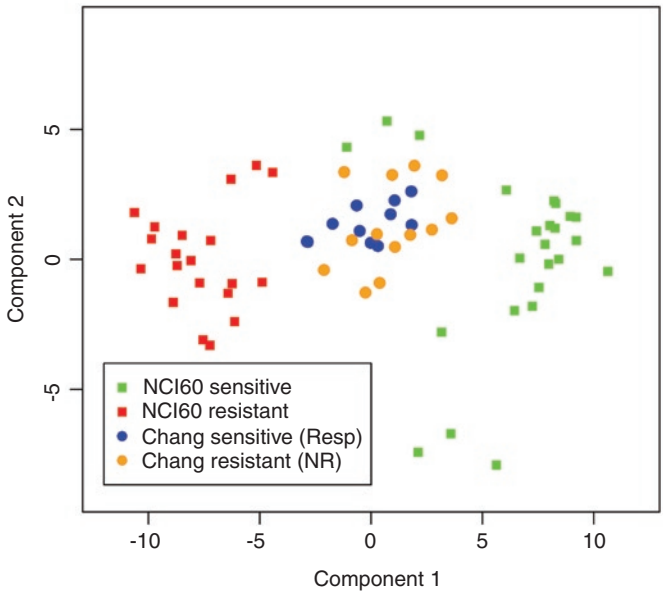

Figure 1 Plot of the first two principal components from the NCI-60 training set for docetaxel, into which the validation set from Chang *et al.*[2] has been projected. The first principal component completely separates sensitive from resistant cell lines. Test samples from breast cancer patients treated with docetaxel project into the center of the space, with responders (Resp) and nonresponders (NR) overlapping.

nonresponders) and point 5 (combining training and testing data when choosing gene weights) can produce 'better than chance' predictions in the wrong direction. This appears to have happened with "another, independent dataset of samples cultured from adriamycin-treated individuals (GEO accession numbers GSE650 and GSE651)"[1]. These GEO datasets, from Holleman *et al.*[3], include samples from pediatric patients with acute lymphocytic leukemia. There are 28 samples that are resistant to adriamycin (daunorubicin) and 94 that are sensitive. Figure 2c of Potti *et al.*[1] shows 99 resistant and 23 sensitive samples, suggesting that most labels are reversed. If the labels are reversed, the model suggests administering the drug only to the patients it would not benefit.

7. When we apply the same methods but maintain the separation of training and test sets, predictions are poor (**Fig. 1** and **Supplementary Report 7**). Simulations show that the results are no better than those obtained with randomly selected cell lines (**Supplementary Report 8**).

We do not believe that any of the errors we found were intentional. We believe that the paper demonstrates a breakdown that results from the complexity of many bioinformatics analyses. This complexity requires extensive double-checking and documentation to ensure both data validity and analysis reproducibility. We believe that this situation may be improved by an approach that allows a complete, auditable trail of data handling and statistical analysis. We use Sweave[4,5], a package that allows analysts to combine source code (in R)[6] and documentation (in LaTeX)[7] in the same file. Our Sweave files are available at (http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo/). Running them reproduces our results and generates figures, tables and a complete PDF manuscript.

The idea of using the NCI-60 cell lines to predict patient response to chemotherapy is exciting. Our analysis, however, suggests that it did not work here.

Kevin R Coombes, Jing Wang & Keith A Baggerly

Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA.
e-mail:kcoombes@mdanderson.org

Note: Supplementary information is available on the Nature Medicine website.

1. Potti, A. *et al. Nat. Med.* **12**, 1294–1300 (2006).
2. Chang, J.C. *et al. Lancet* **362**, 362–369 (2003).
3. Holleman, A. *et al. N. Engl. J. Med.* **351**, 533–542 (2004).
4. Leisch, F. & Rossini, A.J. *Chance* **16**, 46–50 (2003).
5. Gentleman, R. *Stat. Appl. Genet. Mol. Biol.* **4**, Article 2 (2005).
6. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2006).
7. Lamport, L. *LaTeX: A Document Preparation System* (Addison Wesley, Boston, 1994).

Potti *et al.* reply:

We appreciate the interest that Coombes *et al.*[1] (pages **aaa–bbb**) have shown in our reported results and agree that an algorithm that tracks the results of the various steps would be useful in such complex analyses. Unfortunately, they have not followed our methods in several crucial contexts and have made unjustified conclusions in others, and as a result their interpretation of our process is flawed.

Coombes *et al.*[1] raise three main issues. First, they cannot reproduce our methods for cell selection. This process involves using not only GI_{50} (the concentration needed to reduce the growth of treated cells to half that of untreated cells) concentrations but also LC_{50} (the concentration that kills 50% of treated cells) and TGI (the concentration required to completely halt the growth of treated cells) concentrations for each drug, as well as raw data available at the National Cancer Institute website in cases in which the –log concentrations are truncated. We have provided details describing these steps on our web page (http://data.cgt.duke.edu/NatureMedicine.php). Because Coombes *et al.*[1] did not follow these methods precisely and excluded cell lines and experiments with truncated –log concentrations, they have made assumptions inconsistent with our procedures.

Second, they point to inaccuracies in the gene lists we reported. As they note, software problems resulted in an off-by-one error in the matching of probe IDs with gene names. Additional inaccuracies resulted from errors made when we assembled the gene lists. We have corrected these errors, and accurate gene lists were posted on the *Nature Medicine* website on 10 October. We regret any inconvenience this may have caused for other investigators but emphasize that these errors in no way influence the primary results of our study, as the models are defined by the training set, not by gene lists.

Third, they suggest that our method of including both training and test data in the generation of metagenes (principal components) is flawed. We feel this approach is entirely appropriate, as it does not include any information regarding the actual patient response and thus does not influence the generation of the signature with respect to predicting patient outcome. The aim of generating metagenes from test and validation data is to accommodate differences among the characteristics of the data from cancer cell lines and human tumors, and is similar to the use of methods of 'standardization' that are intended to correct for intrinsic differences in data, including batch effects, before analysis. Indeed, we find that the predictions are equally robust if the data are first standardized and the predictions are then carried out on independent validation cohorts with metagenes generated from only the training data (A.P. and J.N., unpublished data). Additionally, there was no accidental inclusion of genes from the validation data distinguishing responders from non-responders and this