Reproducible research and the substantive context

Niels Keiding

Department of Biostatistics

University of Copenhagen

Validation in Statistics and Machine Learning

Weierstrass Institute for Applied Analysis and Stochastics

Berlin 6 October 2010

Inductive and deductive biostatistics

- Reproducibility Is Good
- What do we want to reproduce?
  - Mathematical derivations
  - Algorithms
    - Donoho
      - Imagination?
  - Choice of statistical design and analysis
  - Correct calculation and analysis
    - JAMA
- Conclusion: reproducible research and the substantive context

The archetypical American biostatistics project

Method-driven

Find method that can be *generalized*.

Generalize it, run simulation (Monte Carlo) study to confirm that it works.

Illustrate briefly on data from the literature gathered to study some long forgotten problem.

Publish! (Journal editors apparently like these papers)

The (European?) puritan attitude

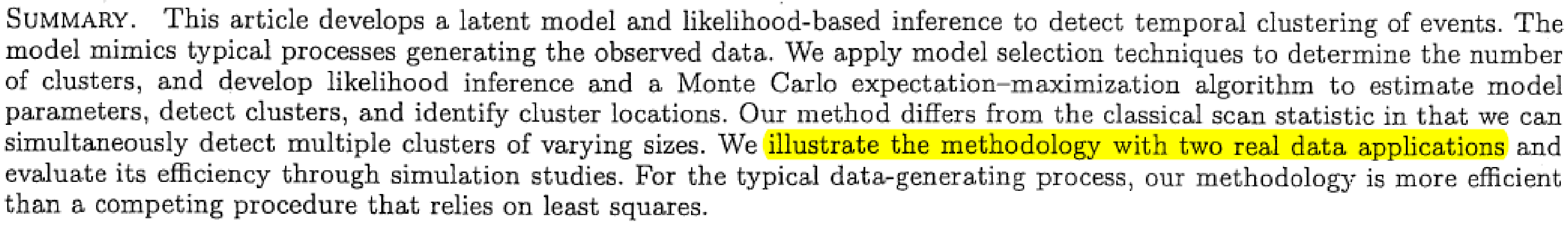
Problem-driven

Biostatistical methodology research is more interesting (and fun) when driven by a **problem** from out there.

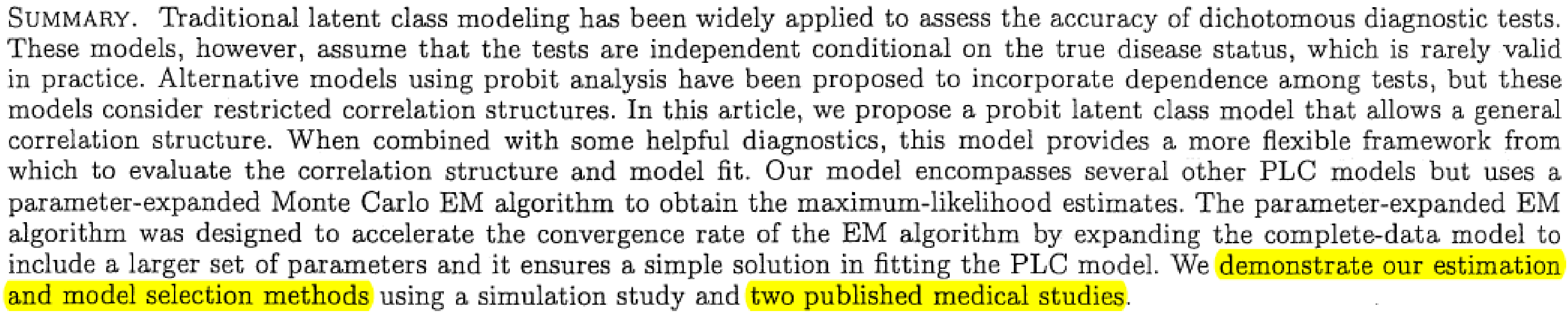
This requires a wide net of contacts with real problems, engagement and time in discussions with the subject-matter people. These will almost always co-author the publications.

Publication rate smaller.

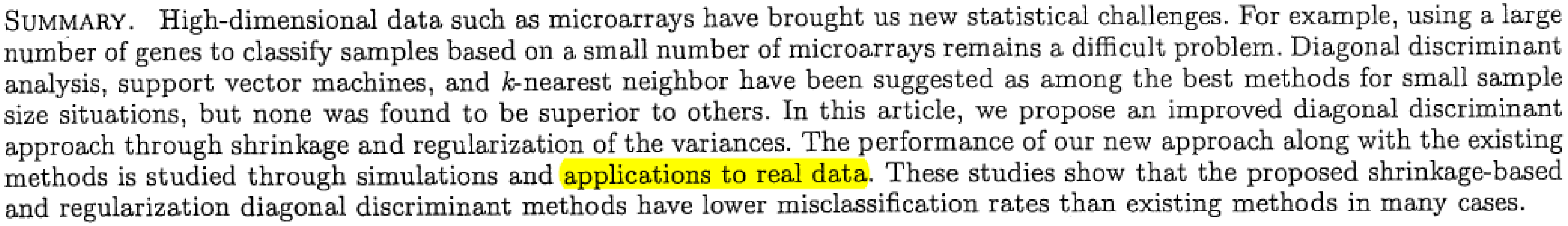
… real data example …



SUMMARY. This article develops a latent model and likelihood-based inference to detect temporal clustering of events. The model mimics typical processes generating the observed data. We apply model selection techniques to determine the number of clusters, and develop likelihood inference and a Monte Carlo expectation–maximization algorithm to estimate model parameters, detect clusters, and identify cluster locations. Our method differs from the classical scan statistic in that we can simultaneously detect multiple clusters of varying sizes. We illustrate the methodology with two real data applications and evaluate its efficiency through simulation studies. For the typical data-generating process, our methodology is more efficient than a competing procedure that relies on least squares.



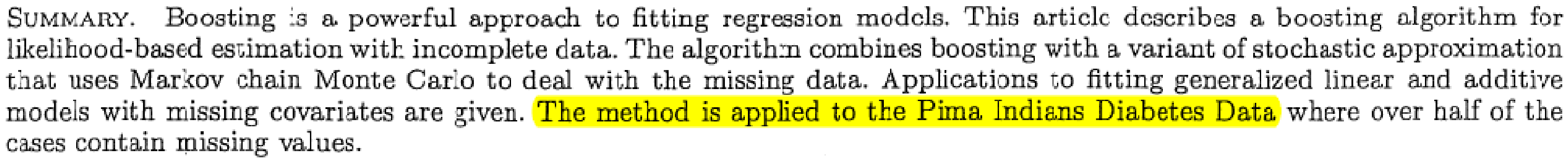
SUMMARY. Traditional latent class modeling has been widely applied to assess the accuracy of dichotomous diagnostic tests. These models, however, assume that the tests are independent conditional on the true disease status, which is rarely valid in practice. Alternative models using probit analysis have been proposed to incorporate dependence among tests, but these models consider restricted correlation structures. In this article, we propose a probit latent class model that allows a general correlation structure. When combined with some helpful diagnostics, this model provides a more flexible framework from which to evaluate the correlation structure and model fit. Our model encompasses several other PLC models but uses a parameter-expanded Monte Carlo EM algorithm to obtain the maximum-likelihood estimates. The parameter-expanded EM algorithm was designed to accelerate the convergence rate of the EM algorithm by expanding the complete-data model to include a larger set of parameters and it ensures a simple solution in fitting the PLC model. We demonstrate our estimation and model selection methods using a simulation study and two published medical studies.



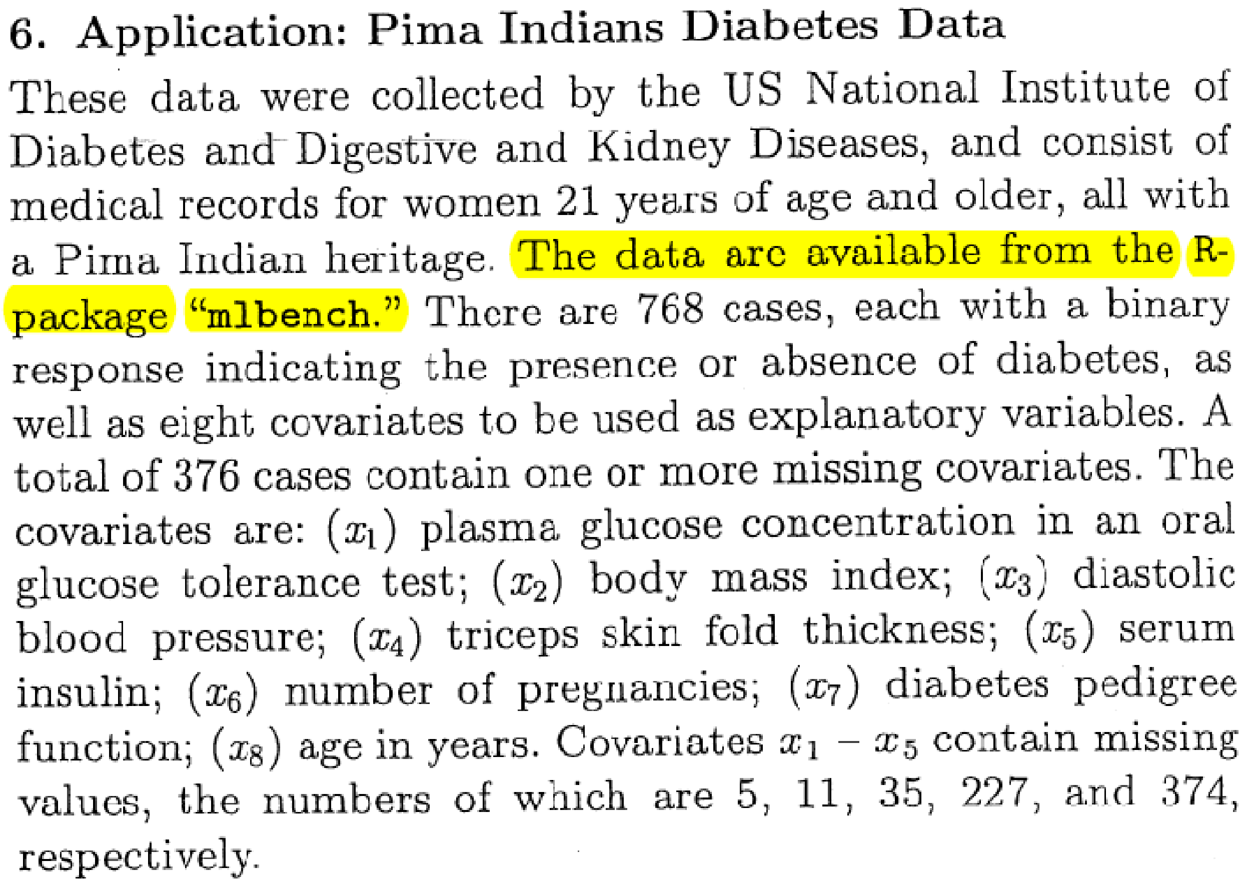
SUMMARY. High-dimensional data such as microarrays have brought us new statistical challenges. For example, using a large number of genes to classify samples based on a small number of microarrays remains a difficult problem. Diagonal discriminant analysis, support vector machines, and k-nearest neighbor have been suggested as among the best methods for small sample size situations, but none was found to be superior to others. In this article, we propose an improved diagonal discriminant approach through shrinkage and regularization of the variances. The performance of our new approach along with the existing methods is studied through simulations and applications to real data. These studies show that the proposed shrinkage-based and regularization diagonal discriminant methods have lower misclassification rates than existing methods in many cases.

The Pima Indians Diabetes data

Biometrics December 2009



SUMMARY. Boosting is a powerful approach to fitting regression models. This article describes a boosting algorithm for likelihood-based estimation with incomplete data. The algorithm combines boosting with a variant of stochastic approximation that uses Markov chain Monte Carlo to deal with the missing data. Applications to fitting generalized linear and additive models with missing covariates are given. The method is applied to the Pima Indians Diabetes Data where over half of the cases contain missing values.



6. Application: Pima Indians Diabetes Data

These data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases, and consist of medical records for women 21 years of age and older, all with a Pima Indian heritage. The data are available from the R-package "mlbench." There are 768 cases, each with a binary response indicating the presence or absence of diabetes, as well as eight covariates to be used as explanatory variables. A total of 376 cases contain one or more missing covariates. The covariates are: (x_1) plasma glucose concentration in an oral glucose tolerance test; (x_2) body mass index; (x_3) diastolic blood pressure; (x_4) triceps skin fold thickness; (x_5) serum insulin; (x_6) number of pregnancies; (x_7) diabetes pedigree function; (x_8) age in years. Covariates $x_1 - x_5$ contain missing values, the numbers of which are 5, 11, 35, 227, and 374, respectively.

Details

The data set `PimaIndiansDiabetes2` contains a corrected version of the original data set. While the UCI repository index claims that there are no missing values, closer inspection of the data shows several physical impossibilities, e.g., blood pressure or body mass index of 0. In `PimaIndiansDiabetes2`, all zero values of `glucose`, `pressure`, `triceps`, `insulin` and `mass` have been set to `NA`, see also Wahba et al (1995) and Ripley (1996).

Source

- Original owners: National Institute of Diabetes and Digestive and Kidney Diseases
- Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)

These data have been taken from the UCI Repository Of Machine Learning Databases at

- `ftp://ftp.ics.uci.edu/pub/machine-learning-databases`
- `http://www.ics.uci.edu/~mlearn/MLRepository.html`

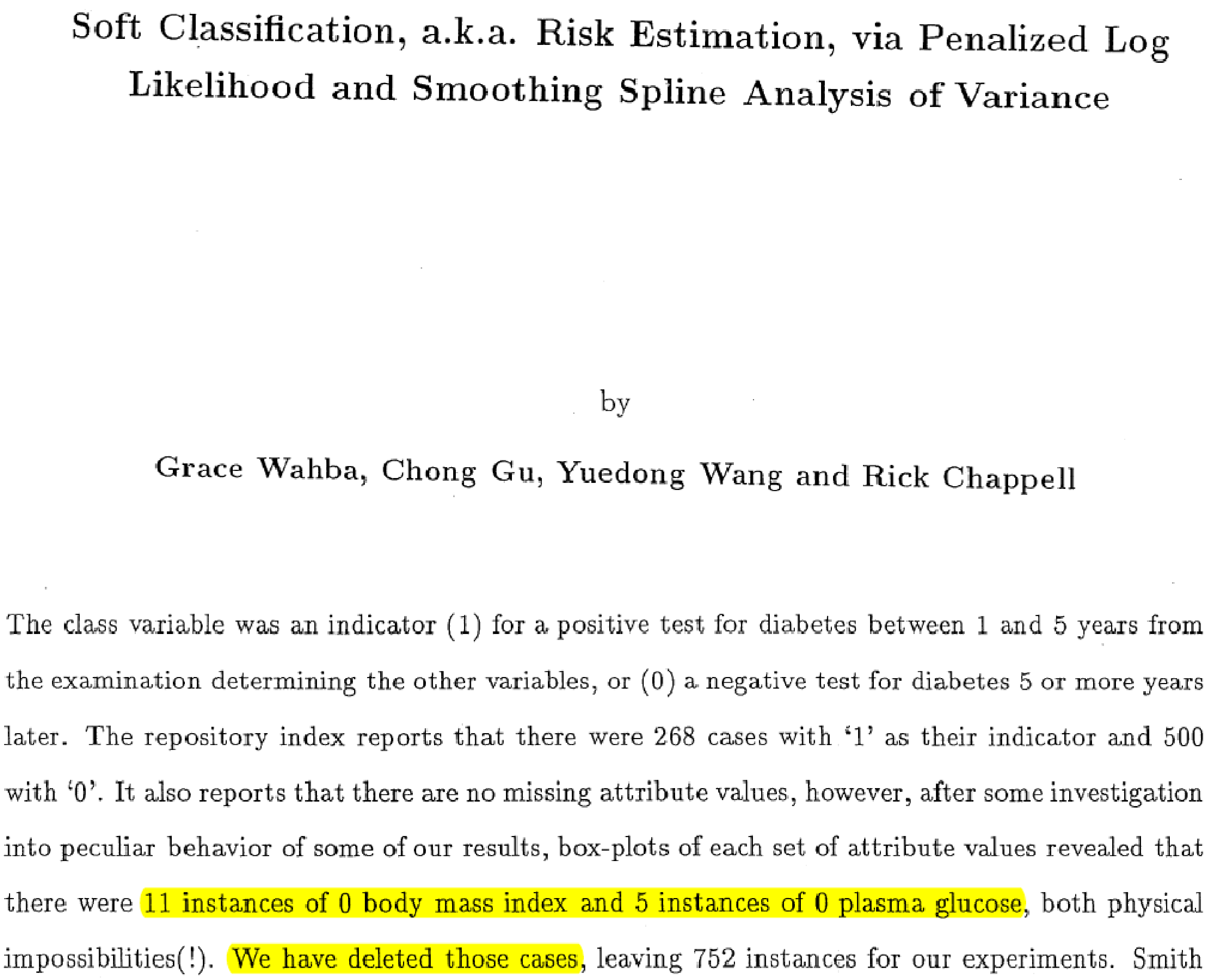
and were converted to R format by Friedrich Leisch.

References

Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Brian D. Ripley (1996), Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge.

Grace Whaba, Chong Gu, Yuedong Wang, and Richard Chappell (1995), Soft Classification a.k.a. Risk Estimation via Penalized Log Likelihood and Smoothing Spline Analysis of Variance, in D.



Soft Classification, a.k.a. Risk Estimation, via Penalized Log Likelihood and Smoothing Spline Analysis of Variance

by

Grace Wahba, Chong Gu, Yuedong Wang and Rick Chappell

The class variable was an indicator (1) for a positive test for diabetes between 1 and 5 years from the examination determining the other variables, or (0) a negative test for diabetes 5 or more years later. The repository index reports that there were 268 cases with '1' as their indicator and 500 with '0'. It also reports that there are no missing attribute values, however, after some investigation into peculiar behavior of some of our results, box-plots of each set of attribute values revealed that there were 11 instances of 0 body mass index and 5 instances of 0 plasma glucose, both physical impossibilities(!). We have deleted those cases, leaving 752 instances for our experiments. Smith

Reproducibility Is Good

Truism: the opposite statement is impossible

What do we want to be able to reproduce?

Are mathematical derivations correct?

Do algorithms do what they are claimed to do?

Is the choice of statistical design and analysis relevant?

Are the actual statistical calculations and conclusions correct?

Mathematical derivations

Documented on paper in main text or web appendix.

Referees spend most of their time here.

Algorithms

Usually little effort spent in the reviewing process to check validity of algorithms

Donoho’s Invitation

D.L. Donoho (2010). An invitation to reproducible computational research. *Biostatistics* **11**, 385-388.

An article about computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result

Biostatistics proposed framework

One has to plan a single script, in the R language, that generates all the figures and tables used in one’s paper

Donoho's reasons why we would want to work reproducibly

We as scientists

(a) Improved work and work habits
(b) Improved teamwork
(c) Greater impact
  (i) Less inadvertent competition
  (ii) More acknowledgement

We as scientists and educators

(d) Greater continuity and cumulative impact

We as taxpayers

(e) Stewardship of public goods
(f) Public access to public goods

The role of imagination

Science	*Confirmatory clinical trials* *Donoho’s scripts?*
Imagination essential	Imagination forbidden Pre-specified hypotheses
Statisticians should help generate new knowledge	Statisticians should take care that rules are obeyed
Clever plots extracting essential features after the fact	Neyman-Pearson hypothesis testing theory

Is choice of statistical design and analysis relevant?

Can only be assessed in dialogue between substantive researchers and statisticians.

Very hard to reproduce.

Availability of naked datasets may well be counterproductive.

Are the actual statistical calculations and conclusions correct?

Availability of *‘final dataset’* may allow others to check that calculations have been correct. And illustrate results of other approaches.

But where is the documentation of all the choices made before arriving at the final dataset?

To reproduce those requires contact with substantive researchers.

The statistician needs to understand how data were generated and selected in order to produce relevant analyses.

Reanalysis: old form

Some journals requested “affirmation from independent statistician” that statistical treatment OK.

Why should we waste our time doing these?

We are busy enough already, and to do it for the money … !

The Journal of the American Medical Association (JAMA) initiative

Reporting Conflicts of Interest, Financial Aspects of Research, and Role of Sponsors in Funded Studies

Phil B. Fontanarosa, MD, MBA; Annette Flanagin, RN, MA; Catherine D. DeAngelis, MD, MPH

JAMA. 2005;294:110-111.

Requirements for Reporting Industry-Sponsored Studies. For all reports (regardless of funding source) containing original data, at least 1 author (eg, the principal investigator) must indicate that she or he “had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.” For industry-sponsored studies, this statement must be provided by an investigator (preferably the principal investigator) who is not employed by any commercial funding source. Moreover, in industry-sponsored studies, it is strongly preferred that data collection and data management are conducted independently of the study sponsor and with additional monitoring and oversight, such as under the auspices of an independent data and safety monitoring committee.

In addition, industry-sponsored studies in which the data analysis has been conducted only by statisticians employed by the company sponsoring the research will not be accepted for publication in *JAMA*. This does not mean that the names of industry-employed statisticians, epidemiologists, or others involved with the data management or analyses should be removed from the manuscript reporting these studies; their roles as authors or nonauthor contributors should be clearly identified. However, for these studies, an additional independent analysis of the data must be conducted by statisticians at an academic institution, such as a medical school, academic medical center, or government research institute.

For these analyses, the entire raw data set should be given to the independent biostatistician, along with the study protocol and the prespecified plan for data analysis. The independent biostatistician should verify the appropriateness of the analytic plan and conduct an independent analysis of the raw data. The results of these analyses should be reported in the manuscript. The independent statistician should clearly describe his or her involvement in conducting the analyses, and provide written confirmation of the data analysis. Details if this independent statistical analysis, as well as the name and academic institution of the independent statistician and whether compensation or funding was received for conducting the analyses, must be reported and will be included in the published article. We recognize that this requirement for an independent statistical analysis of industry-sponsored studies entails additional effort, time, and cost, but in our view, this additional verification of the data and the analyses, as well as an additional layer of institutional oversight for these studies, are essential.

Statistical primary analysis vs. the proposed reanalysis

Primary: Statistician part of subject matter discussion, trial design, choice of endpoints, judgments in inevitable cases of doubt even under the most elaborate protocol.

Reanalysis: Check whether the protocol-specified SAS program actually produces the results that the company statistician claims when applied to the company-supplied data file

Reanalysis: the JAMA proposal

Incentives for academic institutions?

None, except cash.

No intellectual stimulus, problematic to waste scarce resources e.g. senior people’s time

Funding – no incentive for third parties, so funding must come from author or journal

Integrity of academic statistician funded by industry??

Possible implementation of JAMA proposal

Establish non-profit reanalysis centres attached to respected biostatistical university departments.

Centres are jointly sponsored by all relevant companies

Create governing board of professors of biostatistics.

Centres can compete by applying to this board to get contract for doing reanalyses. Thus break the direct cash flow from company to statistician.

Conclusions: Reproducible research and the substantive context

There is much more to research than statistical analysis, and there is much more to the statistical analysis than correctness of calculations.

Reproducible research is not only about repeating calculations on the same data but requires deep insight into the substantive context.

Existing proposals for reproducible research in statistics do not take the substantive context sufficiently seriously.