Reproducible Research

Frank E Harrell Jr

Department of Biostatistics
Vanderbilt University School of Medicine

Fourth Annual Bayesian Biostatistics Conference
University of Texas MD Anderson Cancer Center

Houston 26 Jan 2011

Non-reproducible Research

- Misunderstanding statistics
- "Investigator" moving the target
- Lack of a blinded analytic plan
- Tweaking instrumentation / removing "outliers"
- Pre-statistician "normalization" of data and background subtraction
- Poorly studied high-dimensional feature selection

canstockphoto.com

Non-reproducible Research, *continued*

- Programming errors
- Lack of documentation
- Failing to script multiple-step procedures
  - using spreadsheets and other interactive approaches for data manipulation
- Copying and pasting results into manuscripts
- Insufficient detail in scientific articles
- No audit trail

General Importance of Sound Methodology

- Hackam and Redelmeier [2006]: Translation of research evidence from animals to humans
- Screened articles having preventive or therapeutic intervention in in vivo animal model, $>$ 500 citations
- 76 "positive" studies identified
- Median 14 years for potential translation
- 37 judged to have good methodological quality (flat over time)
- 28 of 76 replicated in human randomized trials; 34 remain untested
- $\uparrow$ 10% methodology score $\uparrow$ odds of replication $\times$ 1.28 (0.95 CL 0.97–1.69)
- Dose-response demonstrations: $\uparrow$ odds $\times$ 3.3 (1.1–10.1)

Note: The article misinterpreted *P*-values

BMJ 1994;308:283 (Published 29 January 1994)

Editorial

The scandal of poor medical research

D G Altman

We need less research, better research, and research done for the right reasons

What should we think about a doctor who uses the wrong treatment, either wilfully or through ignorance, or who uses the right treatment wrongly (such as by giving the wrong dose of a drug)? Most people would agree that such behaviour was unprofessional, arguably unethical, and certainly unacceptable.

What, then, should we think about researchers who use the wrong techniques (either wilfully or in ignorance), use the right techniques wrongly, misinterpret their results, report their results selectively, cite the literature selectively, and draw unjustified conclusions? We should be appalled. Yet numerous studies of the medical literature, in both general and specialist journals, have shown that all of the above phenomena are common.[1 2 3 4 5 6 7] This is surely a scandal.

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

BY JONAH LEHRER

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in

ity is that the scientific community can correct for these flaws.

But now all sorts of well-established, multiply confirmed findings have started to look increasingly uncertain. It's as if our facts were losing their truth: claims that have been enshrined in textbooks are suddenly unprovable. This phenomenon

New Yorker Dec 13, 2010

The Truth Wears Off

- Prescribe drugs while they still work
- Verbal overshadowing: decreases in performance when subjects asked to put perceptions into words
- Rhine and ESP: "the student's extra-sensory perception ability has gone through a marked decline"
- Regression to the mean
- Floating definitions of X or Y: association between physical symmetry and mating behavior; acupuncture

The Truth Wears Off, continued

- Selective reporting and publication bias
- Journals seek confirming rather than conflicting data
- Damage caused by hypothesis tests and cutoffs
- Ioannidis: $\frac{1}{3}$ of articles in *Nature* never get **cited**, let alone replicated
- Biologic and lab variability
- Weak coupling ratio exhibited by decaying neutrons fell by 10 SDs from 1969–2001

"The *decline effect* is actually a decline of *illusion*"

What's Gone Wrong with Omics & Biomarkers?

- Subramanian and Simon [2010]: Gene expression-based prognostic signatures in lung cancer: Ready for clinical use?
- NSCLC gene expression studies 2002–2009, $n \geq 50$
- 16 studies found
- Scored on appropriateness of protocol, stat validation, medical utility
- Average quality score: 3.1 of 7 points
- No study showed prediction improvement over known risk factors; many failed to validate
- Most studies did not even consider factors in guidelines
  - Completeness of resection only considered in 7
  - Similar for tumor size
  - Some only adjusted for age and sex

Clinical Epidemiology Researchers Need Training in Clinical Epidemiology

Biases might pose a special challenge for laboratory researchers who are used to biological reasoning and the tightly controlled conditions of experimental research. Such researchers unwittingly become non-experimental observational epidemiologists when they apply molecular assays in studies of diagnosis and prognosis, for which the experimental method is not available and for which biological reasoning might have limited usefulness.

Ransohoff [2005]

Over-interpretation of Clinical Applicability in Molecular Diagnostic Research

Authors of 108 studies who were laboratory scientists were 19-fold more likely to over-interpret the clinical utility of molecular diagnostic tests compared with clinic-based authors.

Lumbreras et al. [2009]

Natural History of New Fields

Each new field has a rapid exponential growth of its literature over 5–8 years ("new field phase"), followed by an "established field" phase when growth rates are more modest, and then an "over-maturity" phase, where the rates of growth are similar to the growth of the scientific literature at large or even smaller. There is a parallel in the spread of an infectious epidemic that emerges rapidly and gets established when a large number of scientists (and articles) are infected with these concepts. Then momentum decreases, although many scientists remain infected and continue to work on this field. New omics infections continuously arise in the scientific community.

Ionnidis [2010]

Biomarker Discoveries

Izvestia (News)	Pravda (Truth)
Big Effects	Validated Effects

Strong Inference

16 October 1964, Volume 146, Number 3642

SCIENCE

Strong Inference

Certain systematic methods of scientific thinking may produce much more rapid progress than others.

John R. Platt

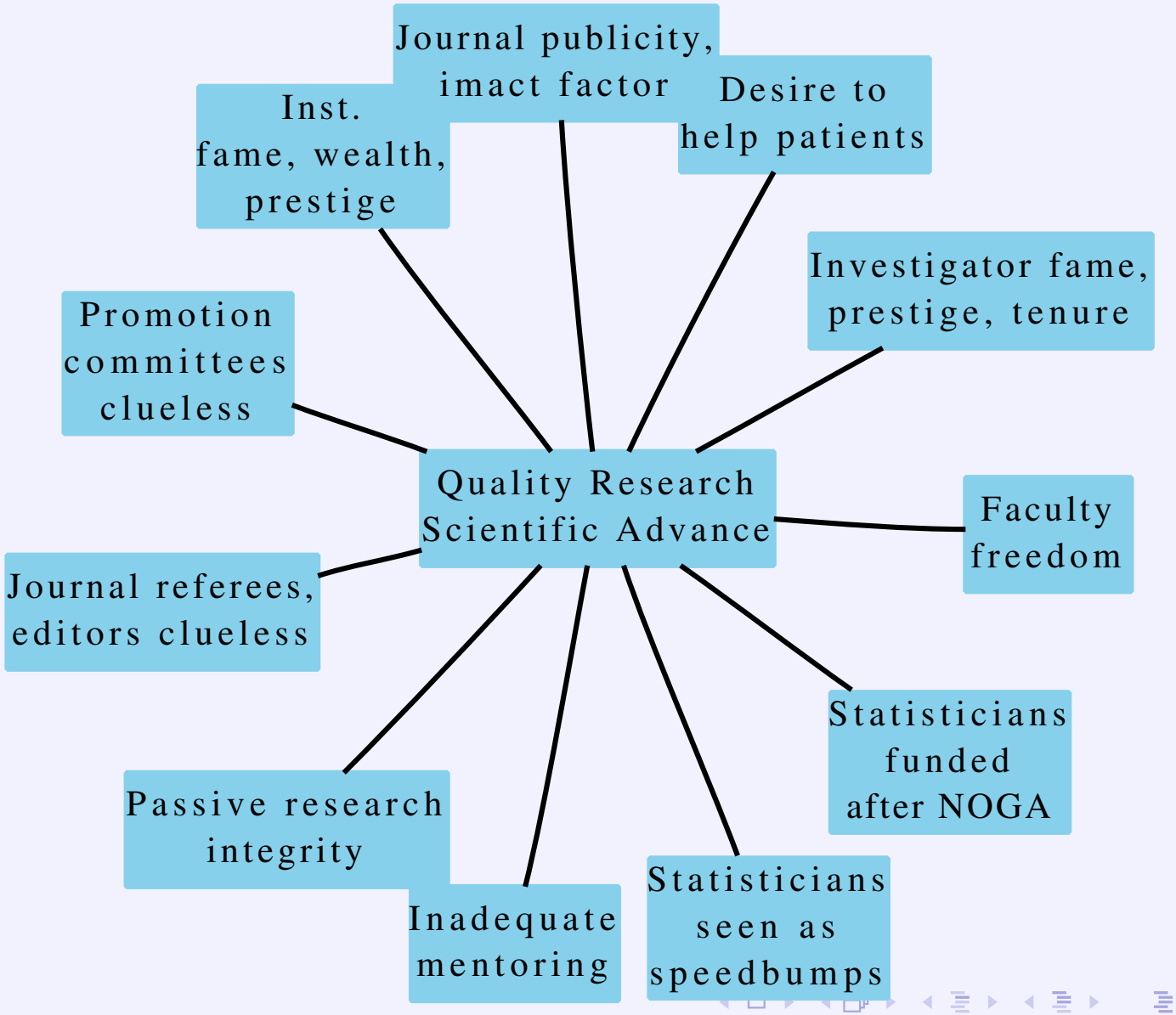
"nature" or the experimental outcome chooses—to go to the right branch or the left; at the next fork, to go left or right; and so on. There are similar branch points in a "conditional computer program," where the next move depends on the result of the last calculation. And there is a "conditional inductive tree" or "logical tree" of this kind written out in detail in many first-year chemistry books, in the table of steps for qualitative analysis of an unknown sample, where the student

Strong (Inductive) Inference, *continued*

- Devise alternative hypotheses
- Devise an experiment with alternative possible outcomes each of which will exclude a hypothesis
- Carry out the experiment
- Repeat
- Regular, explicit use of alternative hypotheses & sharp exclusions $\rightarrow$ rapid & powerful progress
- "Our conclusions . . . might be invalid if . . . (i) . . . (ii) . . . (iii) . . . We shall describe experiments which eliminate these alternatives."
- Rushton: "A theory which cannot be mortally endangered cannot be alive."

Platt [1964]

System Malfunctions



System Cost of Investigating Research Malpractice

Published online 17 August 2010 | Nature | doi:10.1038/news.2010.414

News

High price to pay for misconduct investigations

A single investigation into research malpractice cost US$525,000.

Eugenie Samuel Reich

Investigations into research misconduct cost US institutions more than US$110 million per year, estimates a study published this week. But experts contacted by *Nature* question whether calculating the cost of investigation is the right way to measure the impact of research misconduct.

The research, published in *PLoS Medicine*[1], is based on the costs of a single recent case of research misconduct at the Roswell Park Cancer Institute in Buffalo, New York. In the case, a senior scientist was accused of fabricating data in at least one grant application, and an internal investigation reached a conclusion of research misconduct. As the work was partly funded by the US Department of Health and Human Services, the matter was referred to the department's Office of Research Integrity (ORI), which has

A study has attempted to put a figure on the cost of miscoduct investigations

Pre-Specified Analytic Plans

- Long the norm in multi-center RCTs
- Needs to be so in **all** fields of research using data to draw inferences (Rubin [2007])
- Front-load planning with investigator
  - too many temptations later once see results (e.g., $P = 0.0501$)
- SAP is signed, dated, filed
- Pre-specification of reasons for exceptions, with exceptions documented (when, why, what)
- Becoming a policy in VU Biostatistics

Goals of Reproducible Analysis/Reporting

- Be able to reproduce your own results
- Allow others to reproduce your results

> *Time turns each one of us into another person, and by making effort to communicate with strangers, we help ourselves to communicate with our future selves. (Schwab and Claerbout)*

- Reproduce an entire report, manuscript, dissertation, book with a single system command when changes occur in:
  - operating system, stat software, graphics engines, source data, derived variables, analysis, interpretation
- Save time
- Provide the ultimate documentation of work done for a paper

History

- Donald Knuth found his own programming to be sub-optimal
- Reasons for programming attack not documented in code; code hard to read
- Invented **literate programming** in 1984
  - mix code with documentation in same file
  - "pretty printing" customized to each, using TEX
  - not covered here: a new way of programming
- Knuth invented the `noweb` system for combining two types of information in one file
  - *weaving* to separate non-program code
  - *tangling* to separate program code

`http://www.ctan.org/tex-archive/help/LitProg-FAQ`

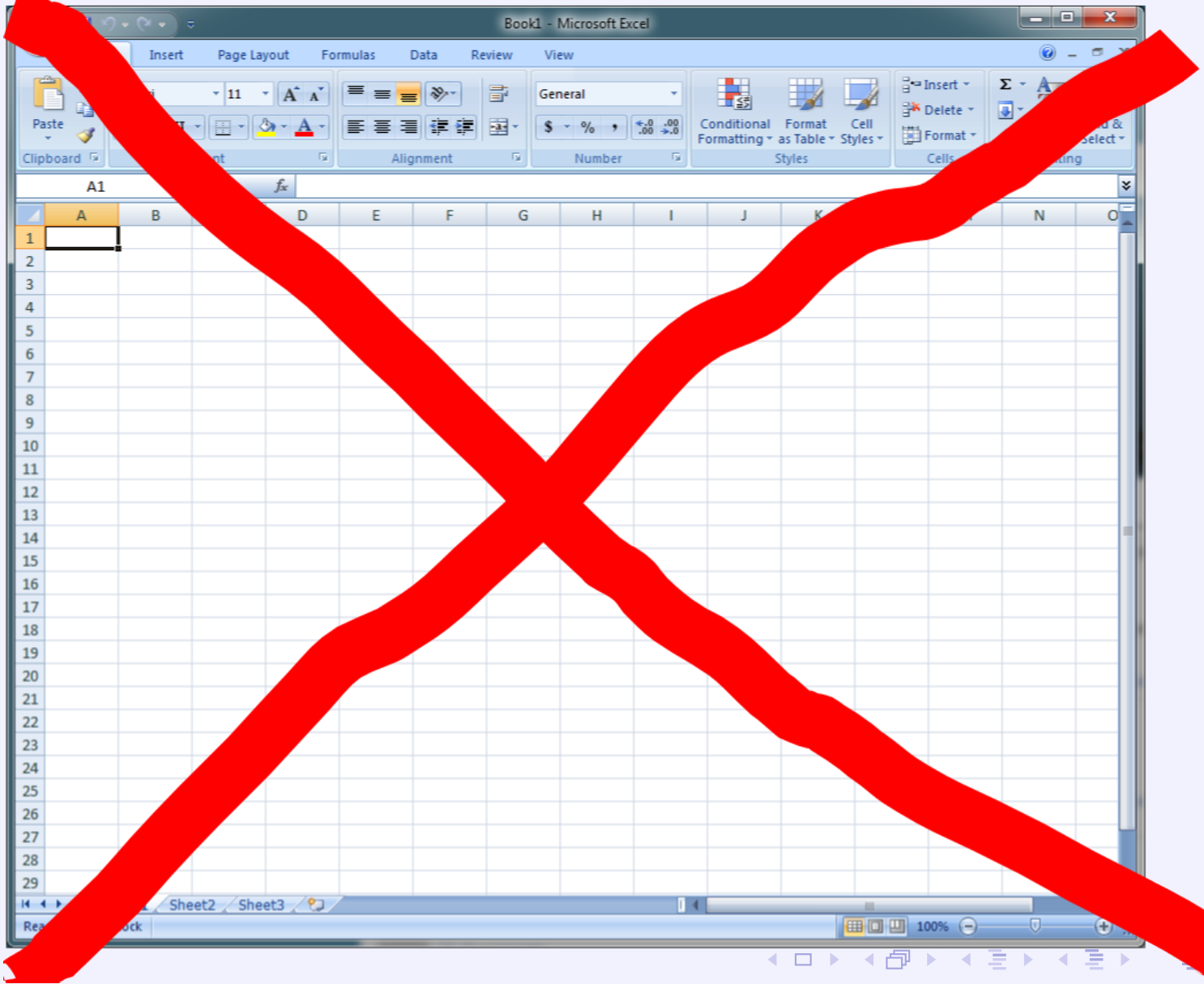
History, *continued*

- Leslie Lamport made TEX easier to use with a comprehensive macro package LATEXin 1986
- Allows the writer to concern herself with structures of ideas, not typesetting
- LATEX is easily modifiable by users: new macros, variables, *if-then* structures, executing system commands (Perl, etc.), drawing commands, etc.
- S system created by Chambers, Becker, Wilks of Bell Labs, 1976
- R created by Ihaka and Gentleman in 1993, grew partly as a response to non-availability of S-Plus on Linux and Mac
- Friedrich Leisch developed `Sweave` in 2002

Sweave Approach

- Sweave is a function in the R tools package
- Uses noweb and an sweave style in LaTeX
- *Insertions* are a major component
  - R printout after code chunk producing the output; plain tables
  - single pdf or postscript graphic after chunk, generates LaTeX includegraphics command
  - direct insertion of LaTeX code produced by R functions
  - computed values inserted outside of code chunks
- Major advantages over Microsoft Word: composition time, batch mode, easily maintained scripts, beauty
- Sweave produces self-documenting reports with nice graphics, to be given to clients
  - showing code demonstrates you are not doing "push-button" research

A Bad Alternative to Sweave



What Do Methodologists Offer?

Biostatisticians and clinical epidemiologists play important roles in

- assessing the needed information content for a given problem complexity
- minimizing bias
- maximizing reproducibility

For more information see:

- `ctspedia.org`
- `reproducibleresearch.net`
- `groups.google.com/group/reproducible-research`

Some Random Thoughts

Kelvin's curse: The unthinking and inappropriate worship of quantifiable information in medicine

Feinstein [1977]

. . . monetization of intellectual property appears to be a powerful force favoring methodological limitations and an excessive reductionism and fragmentation of biologic knowledge

Porta et al. [2007]

There is nothing wrong with cancer research that a little less money wouldn't cure.

Nathan Mantel, NCI

References

A. R. Feinstein. *Clinical Biostatistics*, chapter 16, pages 229–242. C. V. Mosby Co., St. Louis, MO, 1977.

D. G. Hackam and D. A. Redelmeier. Translation of research evidence from animals to humans. *JAMA*, 296: 1731–1732, 2006.

J. P. A. Ionnidis. Expectations, validity, and reality in omics. *J Clin Epi*, 63:945–949, 2010.

B. Lumbreras, L. A. Parker, M. Porta, M. Pollan, J. P. Ioannidis, and I. Hernandez-Aguado. Overinterpretation of clinical applicability in molecular diagnostic research. *Clinical Chemistry*, 55: 786–94, 2009.

J. R. Platt. Strong inference. *Science*, 146(3642):347–353, 1964.

M. Porta, I. Hernández-Aguado, B. Lumbreras, and M. Crous-Bou. "omics" research, monetization of intellectual property and fragmentation of knowledge: can clinical epidemiology strengthen integrative research? *J Clin Epi*, 60:1220–1225, 2007.

D. F. Ransohoff. Bias as a threat to validity of cancer molecular-marker research. *Nat Rev*, 5:142–149, 2005.

D. B. Rubin. The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized studies. *Stat Med*, 26:20–36, 2007.

J. Subramanian and R. Simon. Gene expression-based prognostic signatures in lung cancer: Ready for clinical use? *J Nat Cancer Inst*, 102:464–474, 2010.

Reproducible Research

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville TN

Much of research that uses data analysis is not reproducible. This can be for a variety of reasons, the most major one being poor design and poor science. Other causes include tweaking of instrumentation, the use of poorly studied high-dimensional feature selection algorithms, programming errors, lack of adequate documentation of what was done, too much copy and paste of results into manuscripts, and the use of spreadsheets and other interactive data manipulation and analysis tools that do not provide a usable audit trail of how results were obtained. Even when a research journal allows the authors the "luxury" of having space to describe their methods, such text can never be specific enough for readers to exactly reproduce what was done. All too often, the authors themselves are not able to reproduce their own results. Being able to reproduce an entire report or manuscript by issuing a single operating system command when any element of the data change, the statistical computing system is updated, graphics engines are improved, or the approach to analysis is improved, is also a major time saver.

It has been said that the analysis code provides the ultimate documentation of the "what, when, and how" for data analyses. Eminent computer scientist Donald

Knuth invented literate programming in 1984 to provide programmers with the ability to mix code with documentation in the same file, with "pretty printing" customized to each. Lamport's LaTeX, an offshoot of Knuth's TeX typesetting system, became a prime tool for printing beautiful program documentation and manuals. When Friedrich Leisch developed `Sweave` in 2002, Knuth's literate programming model exploded onto the statistical computing scene with a highly functional and easy to use coding standard using R and LaTeX and for which the Emacs text editor has special dual editing modes using ESS. This approach has now been extended to other computing systems and to word processors. Using R with LaTeX to construct reproducible statistical reports remains the most flexible approach and yields the most beautiful reports, while using only free software. One of the advantages of this platform is that there are many high-level R functions for producing LaTeX markup code directly, and the output of these functions are easily directly to the LaTeX output stream created by `Sweave`.

See `ctspedia.org`, `reproducibleresearch.net` and `biostat.mc.vanderbilt.edu/SweaveLatex` for more information.