It's easy to make mistakes in computational models... and hard to catch them.

By Kristin Sainani, PhD

In 2006, a paper in *Nature Medicine* suggested a novel and potentially revolutionary method for predicting patient responses to cancer therapies using gene signatures. The finding piqued the interest of oncologists at The University of Texas MD Anderson Cancer Center, who sought help from two statisticians, **Keith A. Baggerly, PhD**, and **Kevin Coombes, PhD**, to recreate the approach.

Baggerly and Coombes, both professors of bioinformatics and computational biology, unexpectedly uncovered multiple errors with the data: off-by-one indexing errors, label reversals, inconsistencies, and duplications. The consequence: the original results were not reproducible and the approach was ultimately discredited.

The story received unusual public attention (more details follow). But it is by no means an isolated case. Errors in biomedical computing are surprisingly common. Strictly speaking, every biomedical model contains error in the sense that it is an imperfect representation of the truth. But more troubling are the errors that are avoidable—such as misadventures in Excel, glitches in the software, bad assumptions, and typos. As datasets and models become increasingly complex, errors of this type become both harder to avoid and harder to detect.

"When you've got a complicated model with a bunch of stuff in it, it's hard to tell when it's wrong," says **James Bassingthwaighte, MD, PhD**, professor of bioengineering at the University of Washington. The point of a complex model is to predict behavior beyond the limits of intuition; but, in this realm, our intuition for spotting errors also becomes unreliable.

Plus, the current publication system wasn't designed to catch errors buried within high-dimensional data or intricate models. Reviewers and editors rarely have direct access to datasets or code; and when they do, they don't have time to check every step of the authors' analyses. "There's no way that your typical reviewer can catch some of these problems, unless the journal editor is willing to give you a year and a half to review a paper," Baggerly says. Furthermore, some reviewers may lack an understanding of the biology while others lack an understanding of the computation. Thus, it's inevitable that errors—both inconsequential and serious—will slip through.

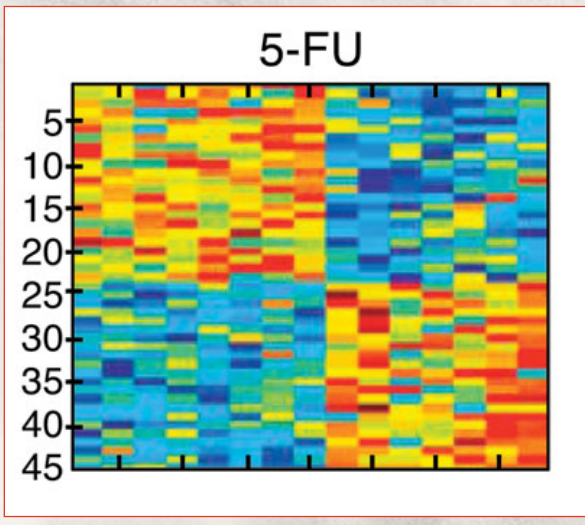
Errors are a touchy and uncomfortable subject. Many researchers avoid the topic for fear of stirring up controversy, making enemies, or casting a shadow of doubt over the field. But keeping silent threatens both the integrity and long-term credibility of biomedical computing. Thus, the best way to address errors is head-on, with the attitude that errors are opportunities for learning rather than for embarrassment. This article reviews examples where researchers have boldly identified errors—in the data, software, methodology, or paper—as well as the lessons that can be gleaned from these errors.

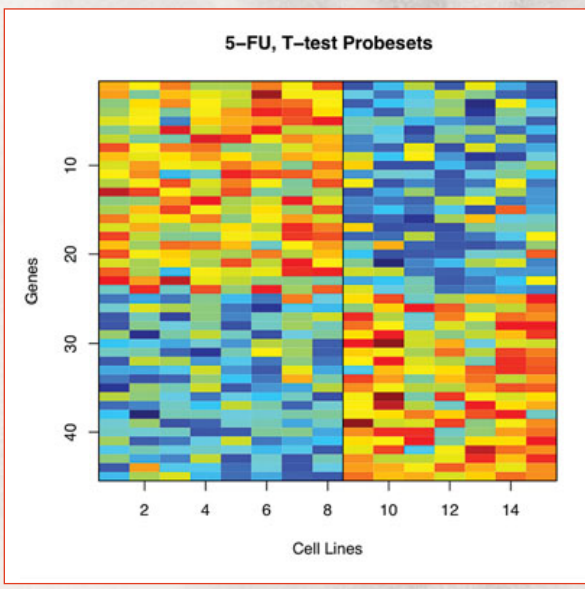
Errors In the Data

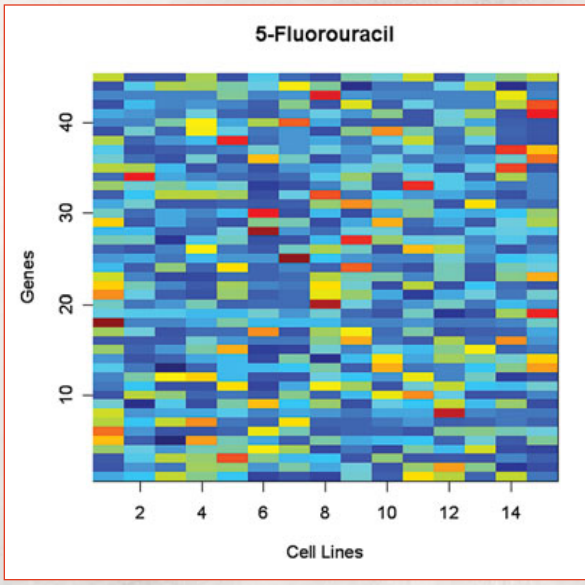
The first opportunity to introduce error is within the data—either in datasets that will be fed into statistical models; or in parameter values or biochemical structures that will be fed into simulations.

Baggerly and Coombes' investigation is a case in point of errors in high-throughput data. The *Nature Medicine* paper claimed that gene signatures built using publicly available cancer cell lines could predict patient responses to specific chemotherapy drugs. It was a striking claim—and was named one of the "Top 6 Genetics Stories of 2006" by *Discover* magazine.

But after months of work, including multiple rounds of emailing with the authors from Duke University, Baggerly and Coombes concluded that the data were riddled with bookkeeping errors. For example, in one instance, the authors accidentally shifted data cells in Excel, causing all the gene labels to be off by one; and, in another instance,

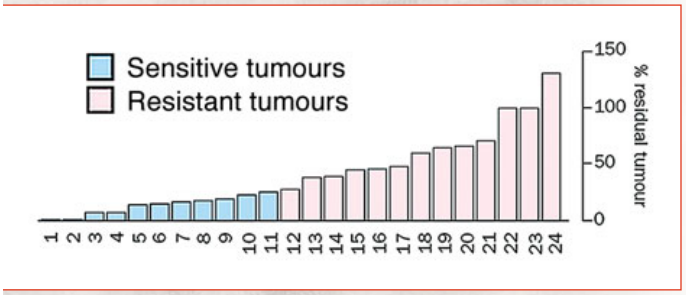


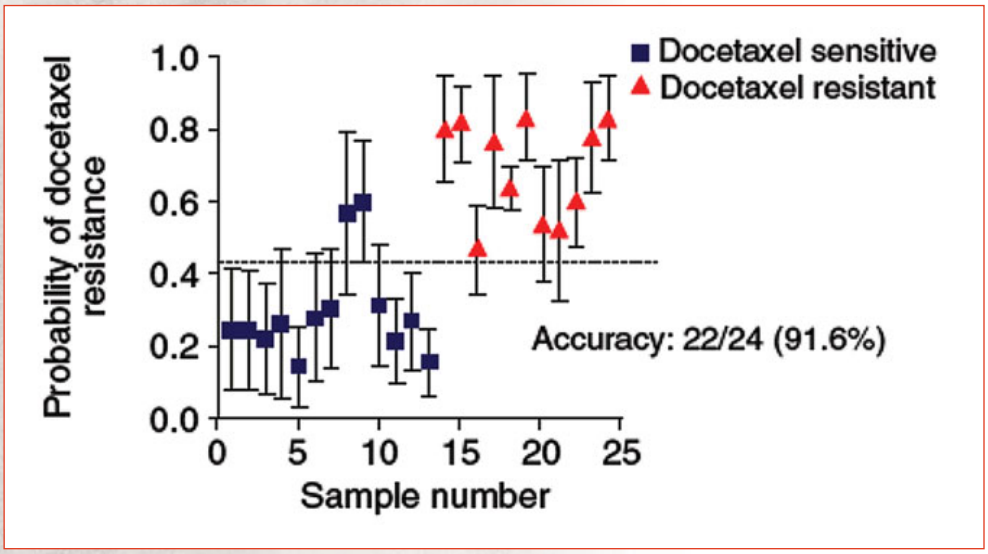




Disappearing Act. ***Potti and Nevins identified cell lines that were sensitive and resistant to the chemotherapy drug 5-fluorouracil (5-FU). Panel 1 shows expression levels of 45 genes (y-axis) that best separate these two groups (cell lines are on the x-axis). Baggerly and Coombes were able to generate the identical heatmap (panel 2). However, the list of 45 genes that Potti and Nevins said were present in the signature—which they claimed made biological sense—was completely wrong, due to an off-by-one indexing error. When Baggerly and Coombes produced the expression profile for these biologically plausible genes, there was no separation between resistant and sensitive cells (panel 3). Panel 1 reproduced by permission from Macmillan Publishers Ltd from Figure 2a in Potti A et al. Genomic signatures to guide the use of chemotherapeutics.*** **Nature Medicine** ***2006; 12: 1294-1300. Panels 2 and 3 courtesy of: Keith Baggerly, MD Anderson Cancer Center.***

the authors reversed the labeling of drug-sensitive and drug-resistant cells. Once these errors were corrected, the impressive predictions disappeared.





Oops! Mixed-up Labels. *Potti and Nevins claimed that the signatures they derived from cell lines could predict patient responses to the chemotherapy agent docetaxel. They used clinical data on women with breast cancer published previously in the* **Lancet** *(Chang* **et al.***). However, when Baggerly and Coombes retrieved the original paper, they noticed that Chang et al. (top panel) had 11 docetaxel-sensitive cases (blue) and 13 docetaxel-resistant cases (red). But Potti and Nevins (lower panel) claimed to have data on 13 sensitive cases (blue squares) and 11 resistant case (red triangles). Potti and Nevins had gotten the labels reversed. Thus, if their signatures were predictive, they would be giving docetaxel to the women most likely to be resistant. Top panel reproduced from: Figure 2a from: Chang JC* **et al.** *Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer.* **Lancet** *2003; 362:362-9. Lower panel reproduced by permission from Macmillan Publishers Ltd from Figure 1d in Potti A* **et al.** *Genomic signatures to guide the use of chemotherapeutics.* **Nature Medicine** *2006; 12: 1294-1300.*

Baggerly and Coombes reported their findings in a 2007 letter to the editor of *Nature Medicine*.

Meanwhile, the Duke team, led by Anil Potti and Joseph Nevins, continued to publish similar results for different drugs and cancers—all with fatal errors. Baggerly and Coombes continued to follow the case, occasionally writing letters to the editors (several of which were rejected). Then, in 2009, they learned that multiple clinical trials of the approach were underway at Duke.

"At which point, we're going: 'They're using stuff that's this screwed up to choose what treatment patients get?' This is bad," Baggerly says. "So that's pretty much when we started shifting it from a mild stink to a loud stink." They published a detailed paper laying out their criticisms in the *Annals of Applied Statistics*. (They initially sent a draft to the editor of a prominent biological journal, but were told that the story was "too negative," Baggerly says.) Their criticisms still didn't gain traction, however—Duke suspended clinical trials in October of 2009, only to restart them a few months later.

The turning point finally came in July of 2010. *The Cancer Letter* (a publication known for controversy) revealed that Potti had lied on his resume, including about being a Rhodes Scholar. This sensational revelation finally thrust the case into the limelight and incited action on several fronts. The clinical trials were permanently halted; several papers were retracted (including the one in *Nature Medicine*); and Potti resigned from Duke.

"It took us over three years to get some of the problems paid attention to. During that time, about 110 patients were treated," Baggerly reflects.

Unfortunately, these types of simple data errors are "far from uncommon," he says. In 2002, Coombes' team won the Competitive Analysis of Microarray Data (CAMDA) contest because they were the only team to discover and fix a major (and unintentional) screw up in the competition data. A one-cell deletion in Excel followed by an inappropriate shifting of values scrambled the annotations for about half the samples. Baggerly says he has also caught such errors in work at MD Anderson.

As a result, he and others at MD Anderson have implemented procedures to systematically avoid or catch this type of error. For example, they require reports to be written in *Sweave*, a function within the statistical programming language R which integrates code and data with a written report that is automatically updated whenever code or data change, and can easily be rerun and checked.

Baggerly and Coombes have also become champions of the "reproducible research" cause. "There are quite a few things in the literature where we look and, quite honestly, we can't tell if there's a mistake. And we can't tell because there's not enough detail for us to check," Baggerly says. Complex analyses involve so many tiny decision points that without a complete record, it is painstaking to retrace the authors' steps. Baggerly and Coombes spent about 1500 hours recreating Potti and Nevins' analyses. In their own papers, Baggerly and Coombes provide supplements containing detailed transcripts of their analyses.

Data errors can also occur in simulations. Though modelers don't have to worry about the integrity of data spreadsheets, they do have to worry about the parameter values that populate their models. They typically cull these numbers from the literature. But rather than citing the original source of the data, they cite the most recent use. For example, Smith *et al.* measure a rate constant in guinea pig cells; Jones *et al.* use this number in their simulation; the next team cites Jones *et al.* as the source of the parameter rather than Smith *et al.*; and the chain continues, much like the children's game of "telephone"—where a phrase whispered from child to child becomes distorted and comical.

"It's amazing how often this happens," says **Daniel A. Beard, PhD**, professor of physiology at the Medical College of Wisconsin. "This is the rule and not the exception."

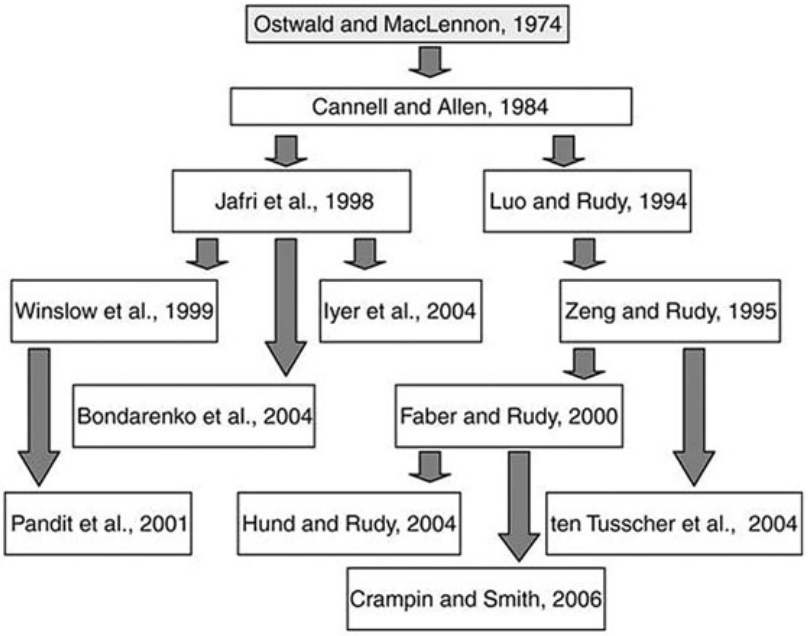
Thus, researchers unwittingly use parameter values that are decades old, from the wrong species, or from inappropriate experimental conditions. They may even propagate explicit errors, Bassingthwaighte notes. "For example, there was one model that introduced a rate constant for hexokinase that was wrong. It went through about eight generations of models continually wrong."

A 2007 paper in the *Journal of Experimental Biology* formally explored this phenomenon. Authored by **Nicolas Smith, PhD**, professor of biomedical engineering at Kings College London, his then graduate student Steven Niederer, and Beard and Bassingthwaighte, the paper showed how a parameter for calcium binding affinity from 1974 worked its way through five generations of models, reappearing (with increasingly recent citations) in models in the 1980s, 1990s, and 2000s.

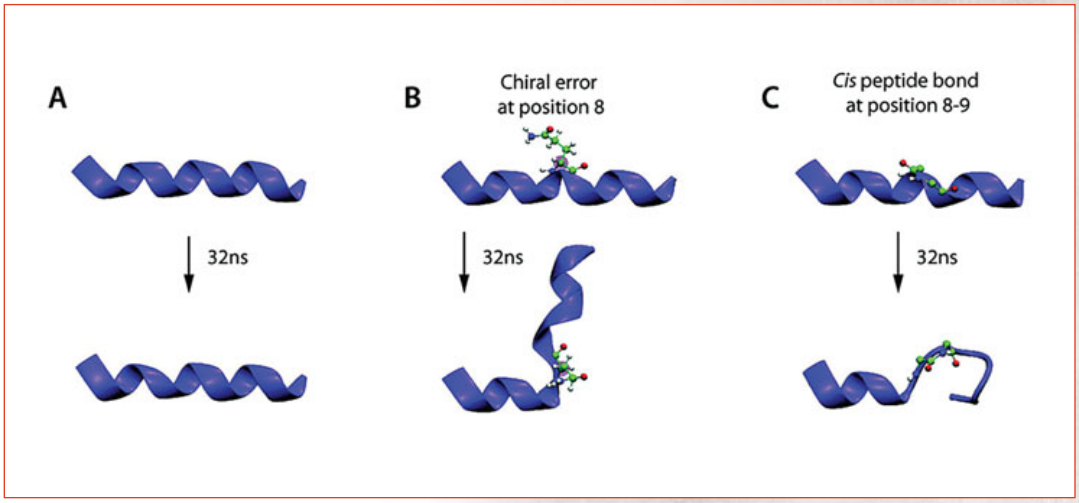
In a 2009 paper in *Experimental Physiology*, Smith and Niederer traced back the entire genealogical history of two state-of-the art models of human heart cells. The model parameters turned out to be decades old and derived from a host of different animals and temperatures. "This is fundamental; almost any physiologist would tell you that the differences in function between species and temperatures are profound," Smith says. And the problem is widespread. "We could pretty much have gone through every existing cardiac model and I assure you that for 90 percent of them we would have found the same thing."

Despite the significance, it was difficult to get the paper published, Smith says. "Parameter sensitivity is the kind of thing nobody wants to talk about." But, he says, the paper has prompted discussions at recent meetings and suggested some easy fixes going forward. For example, authors should include a supplemental table that gives each parameter's original citation and details about how the parameter value was derived. They can also use sensitivity analyses—which show how sensitive the model is to changes in particular parameters—to gauge the potential impact of uncertain values. In the future, markup languages like CellML and SBML may make it easier to directly link model parameters with experimental data held within dedicated databases, he adds.

For molecular dynamics simulations, errors may also occur in the structural data. For example, these structures may have subtle stereo-chemical errors, such as incorrect chirality, where the shape used is a mirror image of the real shape, says **Eduard Schreiner, PhD**, a postdoc at the University of Illinois at Urbana-Champaign. "If you feed that into a simulation, the error will persist because the force field also supports this form." This can dramatically



Citation Tree. *The binding affinity of calcium to calsequestrin was measured in a 1974 experimental study. This parameter value was then propagated, with increasingly recent citations, through five generations of heart cell models. Reprinted with permission from Figure 4B from: Smith NP* et al., *Computational biology of cardiac myocytes: proposed standards for the physiome,* J Exp Biol. *May;210(Pt 9):1576-83 (2007).*



Stereochemical Errors. *This figure shows the impact of two different stereochemical errors on a molecular dynamics simulation involving an α-helix. (A) shows a stereochemically correct helix; (B) shows a helix with a chirality error; and (C) shows a helix with an erroneous* cis *bond. The stereochemical errors introduce incorrect turns and coils into the helix. Figure 3 from: Schreiner* et al., *Stereochemical errors and their implications for molecular dynamics simulations,* BMC Bioinformatics *12:190:1471-2105 (2011).*

impact the results, for example causing helices to kink or unwind when they shouldn't, he showed in a 2011 paper in *BMC Bioinformatics*.

"I think these errors are more common than one thinks," Schreiner says. "It's just people don't think about it and they don't check." His team has written open source software to identify and fix stereo-chemical errors before and during simulations.

Errors In the Software

Widely used—and well-trusted—software packages may contain bugs or inherent limitations. This can lead to widespread errors within a community, especially when users are unaware of the technical details of their tools.

For example, a problem in the statistical package S-Plus led to statistical errors in 37 papers on air pollution and health. Several groups of researchers used a particular statistical model to test the association between daily changes in air pollution and morbidity and mortality. However, the defaults in S-Plus were not set correctly for this type of data. When the data were re-analyzed correctly for a 2003 EPA report, the links between air pollution and morbidity/mortality were greatly diminished (by nearly 50 percent overall). This led to changes in the default parameters in S-Plus and prompted EPA officials to warn that "widespread use does not guarantee that a software or algorithm has no drawbacks."

The molecular dynamics community is often heralded as a paradigm of excellence for building and disseminating standardized software tools and force fields (the energy functions used for molecular dynamics simulations), such as AMBER, CHARMM, and GROMACS. However, even these tools have imperfections.

For example, a 2008 paper in the *Journal of Chemical Theory and Computation* identified a bug in AMBER and GROMACS. Unless the user specified otherwise, these programs always used the same seed in their random number generators, causing streams of "random" numbers to repeat. This can create artificial patterns. "If you run a bunch of simulations with the same seed, you can actually see periodic behavior emerging in the system [where there may actually be none]," says study co-author **Peter Freddolino, PhD**, a postdoc in the Lewis-Sigler Institute for Integrative Genomics at Princeton University. Though the error has been corrected in both programs, many published simulations may have been affected and "it shows how sensitive these simulations can be," he says.

The force fields have inherent limitations as well. For example, certain force fields work better in certain situations. "From my personal experience, the CHARMM force field is good for proteins but bad for nucleic acids," Schreiner says. CHARMM tends to make RNA too floppy, he says.

"These models all have different propensities," Freddolino agrees. For example, in a 2010 paper, he used 16 different models—combinations of four force fields and four water models—to simulate the same crystal lattice. "In some of them, the crystallographic lattice totally melted, which is really unfortunate because we know this should not be happening," he says. AMBER did the best job of maintaining the integrity of the lattice; and CHARMM also did well with certain corrections, he found.

Errors in the Methodology

The specific steps that people take to analyze data or build models provide abundant opportunities for errors, due to the many assumptions, subtleties, and choices involved.

Misuse of statistics in the medical literature is a well-known problem. For example, some researchers have recently called into question the statistics used in a series of high-profile papers that claim to show that traits such as obesity, happiness, loneliness, and divorce can spread "contagiously" through social networks. In a 2007 paper in the *New England Journal of Medicine*, the authors showed that (not surprisingly)

***Melting Lattice.** Different force fields give different results for simulations of the same crystal lattice. The top cartoon depicts two interacting monomers after simulation with a "good" force field for this application; the simulated structures line-up nicely with the original structure (pictured as a transparent outline). The bottom cartoon depicts the same two monomers after simulation with a "bad" force field for this application; the simulated structures line up poorly with the original structure. Courtesy of: Peter Freddolino, Princeton University.*

participants in a large community study who became obese were more likely to have friends and relatives who became obese. More interestingly, they used statistical models to (purportedly) show that this clustering wasn't just due to shared environment, shared genes, or self-selection (the tendency for similar people to seek each other out).

But their statistical arguments and models contain fatal flaws, says **Russell Lyons, PhD**, professor of mathematics at Indiana University. The models aren't appropriate and actually contradict their conclusions, Lyons says. Even if the models had been sound, they wouldn't adequately account for shared environment and selection, Lyons and others have argued. For example, other researchers used similar models to show that acne, height, and headaches are "contagious" among teenagers.

"No one is arguing that people don't influence other people. Everyone agrees that it could be. But the issue is whether they've actually added any knowledge," Lyons says. "And they haven't."

Lyons had difficulty publishing his critique. The *New England Journal of Medicine* rejected his paper without explanation. "Clearly they just weren't interested," Lyons says. "But they should have been." The paper was published in a relatively new and obscure journal called *Statistics, Politics, and Policy* in 2011.

The example points to broader issues with publication, peer review, and statistical training, Lyons says. For example, the authors of the contagion papers did not adequately write out their models, making it difficult for others to check them. Authors need to make their code and data available, he says. Also, biomedical researchers need better training in statistics and computation, so that errors don't slip through peer review, he says.

Peer review does have problems, agrees **Richard Simon, D.Sc.**, chief of the Biometric Research Branch of the Division of Cancer Treatment and Diagnosis at the National Cancer Institute. "You see papers that are just really wrong, and you wonder how did these things ever get published?" Simon says. The problem is that journal editors don't always know enough to select the right reviewers, he says.

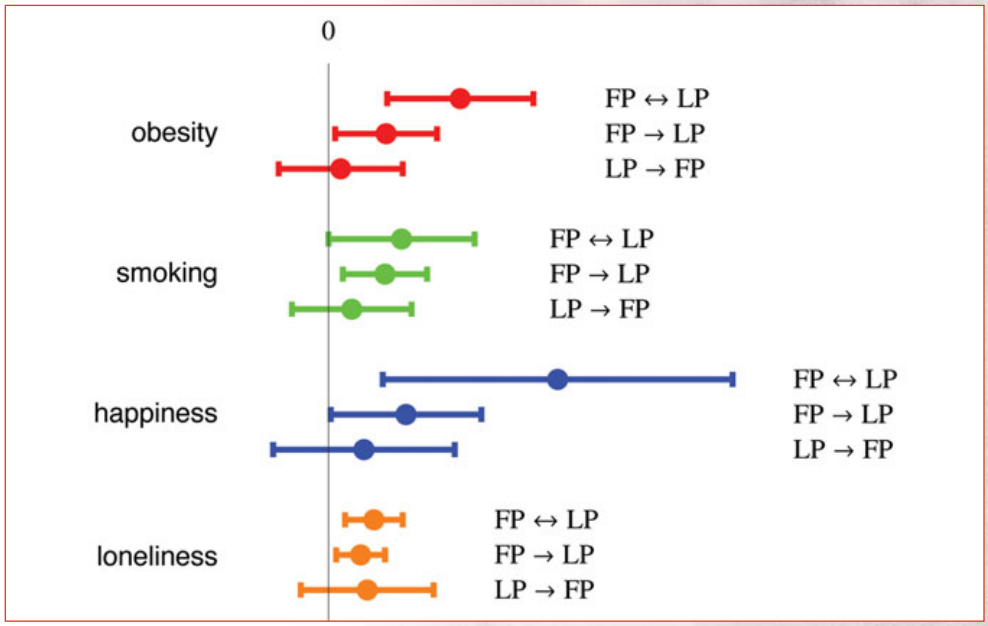
In a 2007 review of 42 microarray studies for cancer prognosis (many published in high-impact journals), Simon's team found that half had basic statistical flaws in the analysis. For example, several papers had incorrectly implemented a statistical technique called cross-validation. With high-dimensional data, it's easy to spot patterns that are actually just random noise. Therefore, it's essential that the data used to fit the model ("training" data) differ from those used to evaluate the model ("test" data). In cross-validation, researchers divide their data into temporary training and test sets and then fit and test the model. They repeat this process many times and then calculate the average model fit over all iterations.

The problem is that people often implement this wrong, Simon says. Rather than redoing their gene selection during each loop of training/test partition, they do this selection only once and then "cross-validate" how much weight each gene is given in the signature as opposed to which genes belong in the signature. "It's not intuitive that it would make that big of a difference, but it makes an enormous difference," Simon says. And it wasn't just novices who got this wrong—those who had the most computationally intensive and fancy algorithms were actually more likely to make this error, probably because they wanted to avoid rerunning the algorithm each time, Simon says.

Since their paper was published, people have slowly begun to catch onto this error, Simon says. But they still make other mistakes. For example, even when authors validate or cross-validate their results correctly, they still publish the "resubstitution" statistics—statistics garnered from fitting and testing the model on exactly the same data. "The biased results are so impressive, even though they are so biased, that they want to give them," he says. In a simulation, his team showed that gene signatures culled from completely random data always appear to have impressive prognostic ability when the resubstitution statistics are used.

These errors negatively impact the credibility of genomics, because highly publicized findings turn out to be much less exciting than initially pronounced. "A lot of the population gets turned off to all of genomics. They think it's all garbage, which it's not," he says.

Modeling studies are similarly fraught with opportunities for errors and shaky assumptions. Molecular dynamics simulations start with a "horrendous set of assumptions," Freddolino says. For



Statistical Fallacy. *In the social contagion papers, the authors claim to have found a directional effect: a friend increases your risk of obesity (or smoking, happiness, or loneliness) the most when the friendship is mutual (top); less when you name him/her as a friend but not vice versa (middle); and the least when he/she names you as a friend but not vice versa (bottom). The authors argue that this is evidence of transmission rather than shared environment. However, the differences in effects between the three types of friendships are not statistically significant, as can be seen from the overlapping confidence intervals. FP=focal participant; LP=linked participant. Figure 1 from: Lyons, Russell (2011) "The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis,"* **Statistics, Politics, and Policy***: Vol. 2: Iss. 1, Article 2.*

example, researchers simulate molecules in a small box of salt water, and allow the box to interact with itself on all sides to avoid boundary effects. This creates an artificial spatial periodicity, and may lead compact structures to be over-stabilized if the box is too small, he says. "So, you need to think very carefully about all of the decisions that you make when you are setting up your system." It's also important to be upfront with readers, especially experimentalists, about the assumptions and potential pitfalls of the model, Freddolino says. "The really impressive part, in some ways, is that molecular dynamics gets so much right despite all these assumptions," he adds.

Errors can also be introduced in the algorithms used to run simulations. Researchers commonly divide space or time into discrete chunks to make simulations computationally tractable—but making these chunks too big can cause the model to be imprecise and unstable. For example, in biomechanical simulations, researchers divide objects into a grid, or "mesh," of repeated polygons (e.g., squares, triangles, or cubes). With many published papers, one can tell with a "quick look" that "there's no way that the mesh was adequate for the problem," says **Jeffrey A. Weiss, PhD**, associate professor of bioengineering at the University of Utah. Analysts should perform mesh convergence studies—where they progressively decrease the size of the mesh until the results change by only a negligible amount, Weiss says.

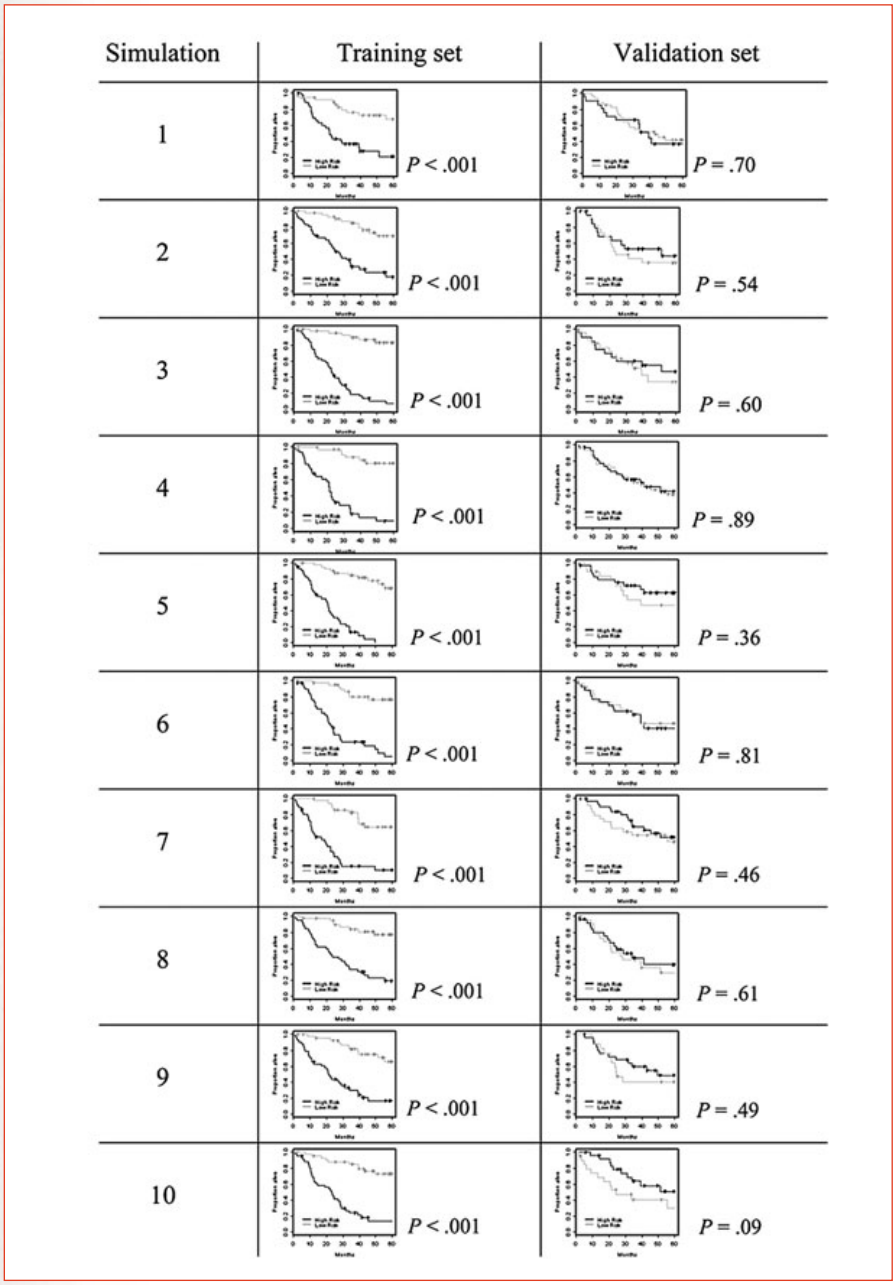
Similarly, with molecular dynamics simulations, researchers break the simulation into discrete time steps. The standard is to make these time steps two femtoseconds, Schreiner says. "But one finds simulations where people go to even larger time steps. The force fields were never designed for that."

Another issue for molecular dynamics simulations is model convergence. If researchers don't run the simulation for long enough, they may reach a conformation that is stable within a short timeframe but not at physiologic time scales, says **Scott C. Schmidler, PhD**, associate professor of statistical science and computer science at Duke University. "I see people making claims about simulations that I know have not been run long enough to make those claims reliably,"

Schmidler's team has developed an automated procedure to diagnose model convergence. The algorithm samples the conformational space by running many simulations in parallel, starting from different initial conditions; then it uses statistics to predict whether the model has converged (or when it's likely to converge). They plan to make their software freely available.

Molecular modelers at least face a relatively well-defined set of potential errors, Beard says. In contrast, he likens physiology modeling to the "Wild West." The problems and approaches are so vast that it's been difficult to develop infrastructure and standards, he says. "I bet there are many errors that we don't even know about."

For example, Smith's team invited 20 international teams to simulate electrical signaling in a simple cube of heart tissue (with known behavior). Of the 12 teams who eventually contributed a solution, three of them "got horribly different results" initially. They were able to use the fact that they got different results to figure out what was wrong and fix their



So Impressive, But So Biased. *This figure shows results from a simulation study in which prognostic gene signatures were created from randomly generated gene expression data. The resulting model does an excellent job of predicting survival in the training set (the data which were used to fit the model), but has no predictive value when applied to the validation set. Figure 2 from: Subramanian J and Simon R. Gene Expression–Based Prognostic Signatures in Lung Cancer: Ready for Clinical Use?* **J Natl Cancer Inst** *2010;102(7): 464-474, by permission of Oxford University Press.*

code, Smith says. "But until that point they'd had no idea [their approach was wrong]. And they'd published lots of work up to that point. We promised not to say who was who in that process."

In a similar example (with a known solution), the FDA asked several modeling teams to simulate blood flow through a Left Ventricular Assist Device (LVAD), an implantable device that helps the heart pump blood. They also asked teams to rate themselves as beginners, intermediates, or experts. The self-rated "experts" actually did the worst. "Some of the solutions that were submitted by people who were claiming to be advanced users were just ridiculous," Smith says. "That makes me worried." These types of benchmark studies can help uncover errors and suggest best practices. Standards for physiome modeling are emerging, though there's still a long way to go, Beard says.

In all fields of biomedical computing, the methods used for "validating" the model are often flawed. "There's so much confusion in the field about what constitutes proper validation," Simon says. For example, it is not appropriate to "validate" a prognostic gene signature against biological data; it needs to be validated for its intended use, as a decision tool for physicians.

Weiss says the same confusion exists for modeling studies. "I have seen many, many studies published with either no validation or the validation that was done was just wholly inadequate," he says. Investigators have such a poor understanding of validation that they often think they have done a good job of it when they haven't, he says. Many papers say that "the model has been validated;" but validation is not an on/off switch, he says.

There's actually a rich literature on the validation of computational models, but it exists outside of biomedical computing, within traditional engineering fields such as computational mechanics, Weiss says. Researchers need to look beyond PubMed to find these papers and standards, he says.

Errors In the Paper

Published papers frequently have typos, omissions, and otherwise poor documentation of methods. These errors make it impossible to figure out exactly what was done or to reproduce the results.

Researchers who curate models for repositories—such as the Physiome Project or the CellML model repository—are especially attuned to these types of errors. "Currently there are a few hundred models in the CellML repository. I think there are maybe five or six of those that didn't have errors in the original publication," says **David Nickerson, PhD**, a research fellow at the University of Auckland. The errors include missing equations and parameter values, typos in equations, or ambiguities about which equations were used for which analyses, he says.

"Ninety-nine percent of models are not reproducible," agrees Bassingthwaighte, who leads the Physiome project. "Name any element and it will be wrong somewhere," he says. "There's usually a process of iterating with the authors to get things right."

Beard recalls a paper he published with a 40-page supplement. "I couldn't tell you with any confidence that there are not typos in that supplement. I know that I would never want to sit down with that supplement and try to reproduce that model," he says. This is why authors need to make their code available and to use markup languages such as CellML and SBML, so that everyone is on the same page, Beard says.

The same publication errors occur in high-throughput studies. In a 2009 paper in *Nature Genetics*, teams of analysts tried to reproduce a table or figure from 18 microarray studies published in the journal from 2005 to 2006. Even though data were available in theory for all 18 studies, only two studies were fully reproducible.

Baggerly recalls that one of Potti and Nevins' papers involved 59 patients, but gave a link to a dataset with 153 patients. "So technically they fulfilled the letter of the law, by saying 'this is where our data came from.' But they didn't tell us which 59 were chosen," he says. "And we are ornery and geeky, but attempting all 153 choose 59 combinations, that's beyond even our tolerance."

Common Lesson

A common lesson that emerges from these examples is the need for changes in the publication system. Many journals now encourage authors to make code and data available, but compliance is still spotty at best; and "availability" currently doesn't guarantee usability.

To address these problems, the journal *Biostatistics* named the first-ever "editor of reproducibility," **Roger Peng, PhD**, associate professor of biostatistics at The Johns Hopkins Bloomberg School of Public Health. Once a paper is accepted, authors may request a "reproducibility review," in which Peng does the hard work of going through the code and data to make sure that the paper's tables and figures are completely reproducible. If so, the paper is marked with "R" for reproducible (as well as "D" for data available and "C" for code available.)

Reproducibility doesn't mean the paper is right. But reproducibility has to be met before one can begin to ferret out errors in the data, software, or methods, Peng says.

Many envision more sweeping changes to the publication system. In this vision, published papers will consist of a human-readable summary linked to a complete implementation of the model in standard formats. "In that way, the journal article can ignore the underlying details and just present what the study was about," Nickerson says. "But then, from there, you can relatively straightforwardly link back to the underlying encoding."

Implementing this vision will not be straightforward, however. Journals will need new tools and infrastructure. Editors will have to figure out how to handle proprietary data and algorithms with commercial potential. Researchers will have to be convinced of the benefits of making their work—and, inevitably, their errors—more transparent. This may require researchers to embrace a more positive view of errors, where they see them as teaching moments for the community rather than as individual failures.

Imagine if Potti and Nevins had viewed errors in this way—they may have heeded the early warnings of Baggerly and Coombes, giving that story a much happier ending.

Identifying and fixing errors is a fundamental part of the scientific process, says **Ron Dror, PhD**, a scientist at D. E. Shaw Research. "It's almost taken for granted that if you read the literature on any hot scientific topic from 10 years ago, you'll discover a lot of papers where it turns out that some of the conclusions were incorrect, often because a given set of data can be interpreted in multiple ways."

He adds: "I don't want to make excuses for errors, but I feel like some of that is the nature of science." □