Uncertainties using Genomic Information for Evidence-Based Decisions*

Pasky Pascual

U.S. Environmental Protection Agency
Washington, DC, USA

Abstract. For the first time, technology exists to monitor the biological state of an organism at multiple levels. It is now possible to detect which genes are activated or deactivated when exposed to a chemical compound; to measure how these changes in gene expression cause the concentrations of cell metabolites to increase or decrease; to record whether these changes influence the over-all health of the organism. By integrating all this information, it may be possible not only to explain how a person's genetic make-up might enhance her susceptibility to disease, but also to anticipate how drug therapy might affect that individual in a particularized manner.

But two related uncertainties obscure the path forward in using these advances to make regulatory decisions. These uncertainties relate to the unsettled notion of the term "evidence" — both from a scientific and legal perspective. From a scientific perspective, as models based on genomic information are developed using multiple datasets and multiple studies, the weight of scientific evidence will need to be established not only on long established protocols involving p-values, but will increasingly depend on still evolving Bayesian measures of evidentiary value. From a legal perspective, new legislation for the Food and Drug Administration has only recently made it possible to consider information beyond randomized, clinical trials when evaluating drug safety. More generally, regulatory agencies are mandated to issue laws based on a "rational basis," which courts have construed to mean that a rule must be based, at least partially, on the scientific evidence. It is far from certain how judges will evaluate the use of genomic information if and when these rules are challenged in court.

Keywords: genome, Bayesian model, scientific evidence, evidence-based decisions, regulatory decisions, systems biology, meta-analysis

In 2000, in an event announcing that one of biology's long-standing challenges — the sequencing of the human genome — had finally been scaled, then US President Bill Clinton issued a bold prognostication: "It will revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases" [3]. A

---

decade later, while most biologists agree that mapping the human genome has revolutionized science, some also admit that it has increased — not diminished — the complexity of biological science by orders of magnitude. The complexity arises not because more genes have been discovered than had been previously anticipated. Indeed, before the Human Genome Project, biologists estimated that the genome might contain about 100,000 genes. The current estimate is that the human genome contains just a fraction of that — about 21,000 [7]. Rather, the challenge of interpreting genomic information lies in understanding the network of events through which genes are regulated.

The traditional model through which genes were thought to be expressed — that in response to environmental signals, one gene codes for one protein that may metabolize one or a few cellular functions — is insufficient to describe the full panoply of cellular behavior. The problem is that metabolic pathways, the series of cell-mediated chemical reactions necessary to maintain life, rarely proceed in a linear fashion. If a gene that triggers a series of reactions is deactivated, there still may be multiple other genes to ensure that the reactions continue to occur. In the words of Cell Biologist Tony Pawson, "When we started out, the idea was that signalling pathways were fairly simple and linear. Now, we appreciate that the signalling information in cells is organized through networks of information rather than simple discrete pathways. It's infinitely more complex" [7].

1   "Hairballs" as a Metaphor for Systems Biology

To do full justice to this complexity, Lander [11] suggests that the double helix — that icon of 20th century biology — should be replaced by the hairball as a metaphor for genomic science (see Fig. 1). A ubiquitous visualization tool for genomic data, the hairball consists of "nodes" (representing genes, proteins, or metabolites) and "edges" (which represent the associations among the nodes). A particular node may be the focus of a researcher's entire program. In 1977 for example, Andrew Schally, Roger Guillemin, and Rosalyn Sussman Yalow shared the Nobel Prize in Medicine for their investigations into a biologically significant "node" showing a connection between the nervous and endocrine systems [24]. Their work demonstrated that hormones secreted by an organism's hypothalamus could trigger the release of other hormones from its pituitary and gonadal glands. Elucidating biology's nodes — such as this so-called hypothalamus-pituitary-gonadal axis — is necessary to understand how an organism operates. But to the systems biologist intent on using genomic information in a quantitative way, the focal point of understanding is the hairball, i.e. the computational, systems-oriented model of how nodes relate to and function within a broader network of other nodes. And so, for example, Basu [2] in research funded by the US Environmental Protection Agency (EPA) proposes modeling how environmental toxicants disrupt fish reproduction and ultimately diminish fish populations by way of perturbations to the hypothalamus-pituitary-gonadal axis. That is, they will model how knowledge about the nodes describing the hypothalamus,
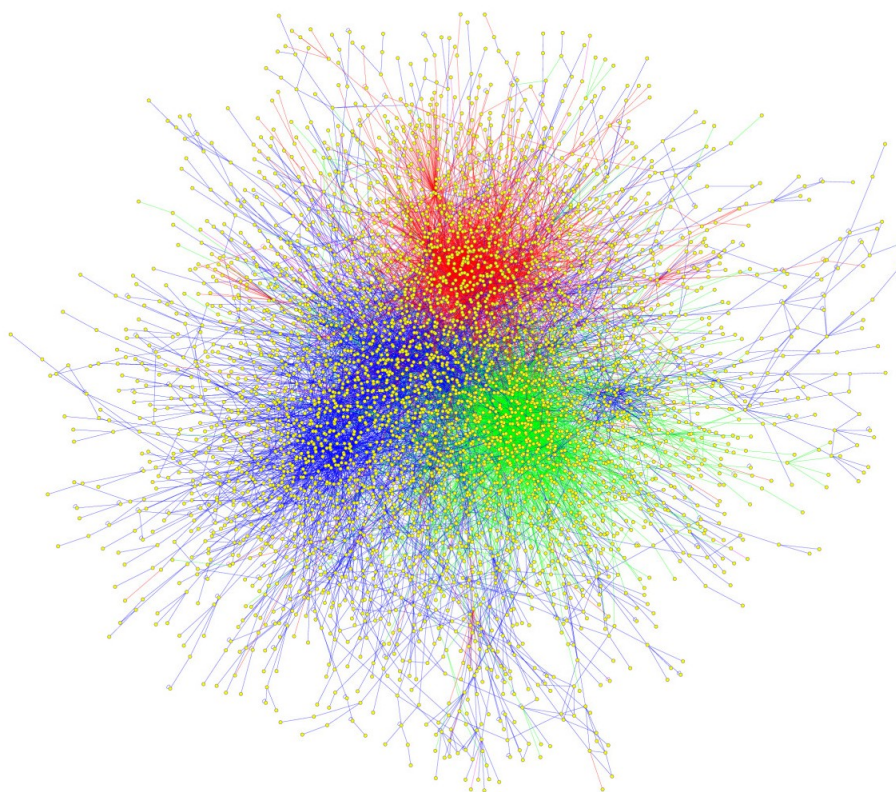
Fig. 1. The "hairball" of systems biology, consisting of "nodes" (representing genes, proteins, or metabolites) and "edges" (which represent the associations among the nodes). Original figure in color provided by Nicolas Simonis and Marc Vidal (see [5]).

pituitary, and gonadal glands interact within a hairball to ultimately impact a resource protected by a regulatory agency's statutory mandate.

A report issued under the aegis of the National Academy of Sciences, *Toxicity Testing in the 21st Century: a Vision and a Strategy* [16], laid out a path for explaining the etiology of environmental disease by using the tools of genomic science. In that report, the Academy proposed that toxicity testing should become less reliant on whole animal tests and eventually rely instead on systems-oriented, computational models, which can be used to screen large numbers of chemicals, based on information from in vitro assays and in vivo biomarkers. Technology exists to monitor the biological state of an organism at multiple levels. It is now possible to detect which genes are activated or deactivated when exposed to a chemical compound; to measure how these changes in gene expression cause the concentrations of cell metabolites to increase or decrease; and to record whether these changes influence the over-all health of the organism. By integrating all this information, it may be possible not only to explain how a person's genetic

make-up might enhance her susceptibility to disease, but to also anticipate how drug therapy might affect that individual in a particularized manner. One of the scientific leaders of the human genome project put it this way: "All biological science works by collecting the complexity and recognizing it is part of a limited repertoire of events. What's exciting about the genome is it's gotten us the big picture and allowed us to see the simplicity" [4].

## 2   Three Enabling Technologies for Genomic Information

Rusyn and Daston [20] highlight three interconnected, technological breakthroughs that have been accelerating developments in genomic science: continuing progress in computational power; advances in quickly and efficiently producing data streams with high information content; and novel biostatistical methods that take advantage of the previous two breakthroughs. More than 45 years ago, Intel's former Chief Executive Officer, Gordon Moore, first reported the observation that has come to be popularly referred to as "Moore's Law": the number of transistors that can be placed on an integrated circuit doubles every 18 months, even as the cost of producing these transistors has diminished over time [14]. In turn, the rapid increase in the cost-effectiveness of computing power has fueled the speed and economic efficiencies with which the genome can be sequenced. The National Institutes of Health's National Human Genome Research Institute has tracked data on the costs of sequencing a human-sized genome during the ten years since the genome was first mapped [15]. These costs have tracked and, since 2007, even exceeded the progress of Moore's Law (see Fig. 2). Similarly, computing power and the use of robotics have made it possible to test thousands of chemicals in plates containing hundreds of wells in order to evaluate a biological response — binding to a receptor site in a cell; producing a particular enzyme; transcribing a gene. These so-called "high-throughput technologies" have generated considerable data about an organism's reaction to chemical exposure.

By itself, this profusion of biological information would be nothing more than unrelated terabytes of data. Complemented with the appropriate analytical methods, the data can yield important insights into the human response to synthetic chemicals. The modeling objective for the systems biologist is the usual one for any modeler, which is to solve for (using the standard regression model):

$$Y = X\beta + \epsilon, \tag{1}$$

where Y is the n-vector of the categorical biological response in which the modeler is interested; X is the $[n \times p]$-matrix of predictors; β is the p-vector of parameters relating biological response to the predictors; and ϵ is the n-vector error term.

For modelers in genomic science, the high dimension of genomic information raises several challenges. Because high-throughput technologies can monitor for multiple biological and chemical attributes simultaneously, these modelers typically confront a situation in which the $[n \times p]$-matrix of predictors, X, is "short
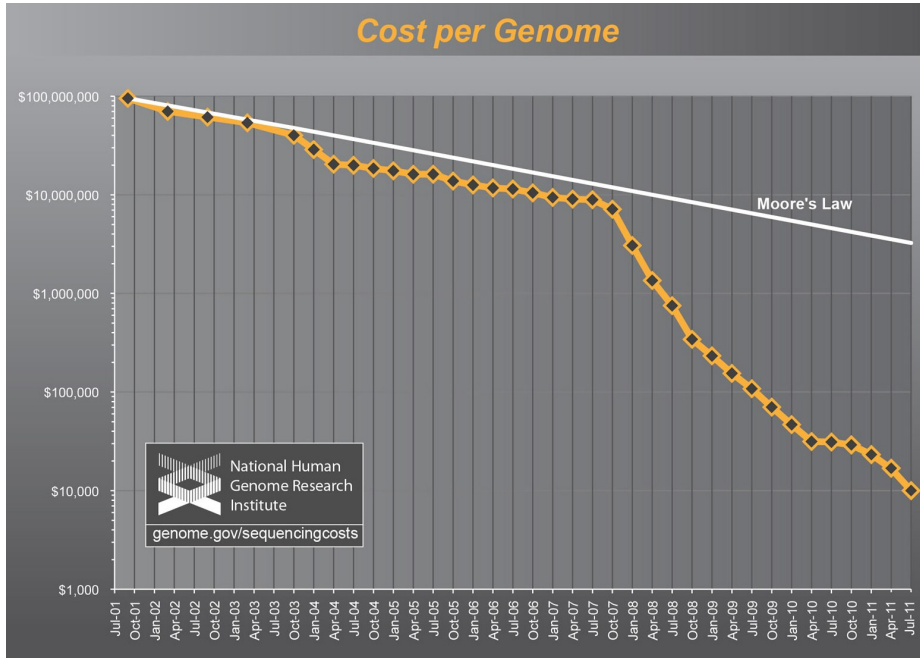
Fig. 2. Costs of sequencing the human genome. Figure from `http://genome.gov/SequencingCosts`.

and wide," i.e. the number of predictors far exceed the sample size, $p >> n$ [25]. At the same time, the biological attributes monitored by high-throughput technologies may often be co-regulated by the same genes or may be involved in metabolic pathways that are correlated [10].

Fortunately for the modeler, the basic tenets of biology suggest that assuming an underlying structure can approximate biological data is not only analytically convenient, but also reasonable, plausible, and empirical. Natural selection, the key mechanism through which evolution selects biological traits that enable survival, imposes constraints on an organism's physical attributes. This is evidenced most clearly by cellular pathways that are conserved over long timescales and among widely disparate organisms [12]. The National Academy's report, *Toxicity Testing in the 21st Century*, defines a "toxicity pathway" as a cellular response that, when sufficiently perturbed, is expected to result in an adverse health effect [16]. Implicit in this definition is the notion that an organism's response to a toxic compound is the result of perturbation away from a stable, homeostatic system of cellular behavior that has evolved over time.

Bayesian methods are particularly well-suited for generating models of genomic information. The Bayesian approach is grounded in the view that because intractable uncertainties obscure any model's objective truth, one can only express the degree to which one believes in a model's truthfulness. If one can

assume that these models conform to probability distributions and to certain axioms, then any initial, hypothesized model can accommodate emergent evidence according to the following relationship:

$$\text{Posterior model} \sim \text{Likelihood} \times \text{Prior model.}$$

Hierarchical Bayesian modeling, based on the notion that computational functions and probabilistic relationships can capture the underlying structure of data organized into discrete levels, conform to information about biological pathways that occur across multiple scales of biological information — from gene to cell to tissue to organ to the whole organism [13]. Additionally, several public databases are available that store data on genomic information, such as the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [23]. Using hierarchical Bayes, the modeler can merge the datasets available at these repositories in order to improve statistical power, while accounting for sources of variability inherent in the experimental protocols used to generate each dataset [1].

3   Example: a Multinomial Probit Model for Genomic Information

As an example of how Bayesian methods can be used to develop computational models of biological information, Sha et al. [21] investigated the use of gene expression data in predicting rheumatoid arthritis, an autoimmune disease characterized by chronic inflammation and destruction of cartilage and bone in the joints. To glean useful insights from their data, the researchers used a multinomial probit (MNP) model. Like the more familiar multinomial logit model, the MNP is used to estimate how categorical, unordered response variables might be functionally related to explanatory variables. The MNP model is more appropriate in modeling genomic information because, unlike the multinomial logit, it allows for the possibility that the categories of response variables are not independent. The MNP model allows for dependence among these categories by estimating the variance-covariance matrix that quantifies any co-variability among them [27]. While this approach had long-standing theoretical appeal, applications of the MNP model were restricted by the computational complexities in fitting them. However, recent advances now implement a Markov Chain Monte Carlo (MCMC) method in order to estimate the MNP posterior model by taking random walks through the given data set [8].

In an MNP model, the response variable, Y_i, is modeled in terms of a latent variable $W_i = (W_{i1}, \ldots, W_{i,p-1})$, where

$$W_i = X_i \beta + \epsilon_i \quad \epsilon_i \sim \mathrm{N}(0, \Sigma), \quad \text{for } i = 1, \ldots, n, \tag{2}$$

and Σ is a $p - 1 \times p - 1$ variance-covariance matrix. The response variable, Y_i, is then modeled using the latent variable W_i, as

$$Y_i(W_i) = \begin{cases} 0 \text{ if } \max(W_i) < 0 \\ j \text{ if } \max(W_i) = W_{ij} > 0 \end{cases} \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, p - 1, \tag{3}$$

where $\max(W_i)$ is the largest element of the vector W_i and Y_i equal to 0 corresponds to an arbitrarily chosen base category.

In Sha et al.'s study [21], patients afflicted with rheumatoid arthritis were differentiated by whether they were in early or late stages of the disease, as measured by erythrocyte sedimentation, the rate of red blood cell sedimentation that is commonly used as an indicator of inflammation. As well, gene expression data for major functional categories were derived for these patients. Applying their MNP model, the investigators noted that genes regulating two sets of biological pathways were associated with patients afflicted with the late stages of rheumatoid arthritis — those regulating aspects of the cytoskeleton (i.e. the system of filaments that provide cells with their structure and shape) and those influencing cytokines (i.e., molecules that participate in regulating immune responses and inflammatory reactions).

It bears highlighting that in applying MCMC to estimate the MNP model, the quantification of uncertainty pervades the entire model estimation process. That is, the objective of model fitting is not merely to estimate the model parameters, but rather to estimate the entire probability distributions underlying the system being modeled.

Baragatti [1] extended this basic, single-level model to a hierarchical model with a categorical, binary response variable. Her study focused on the estrogen receptor status of a patient, a clinically measured indicator of breast cancer. The data were drawn from three different datasets and therefore, a hierarchical model of fixed and random effects were used. The former corresponded to gene expression measurements, while the latter corresponded to the variability introduced by using the different datasets.

4 Weight of Evidence and Meta-analysis

In their review of highly cited studies that have used in vitro and in vivo biological information in order to predict disease risk, Ioannidis and Panagiotou [9] suggest that the associations uncovered in these studies generally tend to be exaggerated, when compared to larger studies and subsequent meta-analysis. The authors attribute these false positives and spurious results partially to publication bias, i.e., the original researchers report only the data which indicate statistically significant results. To guard against misleading results, Zeggini and Ioannidis [26] propose the greater use of meta-analysis in genomic studies, for which a Bayesian framework provides an intuitive framework.

Once again, fixed and random effects models serve as a useful approach. When using these models in meta-analysis, one assumes that a common, fixed
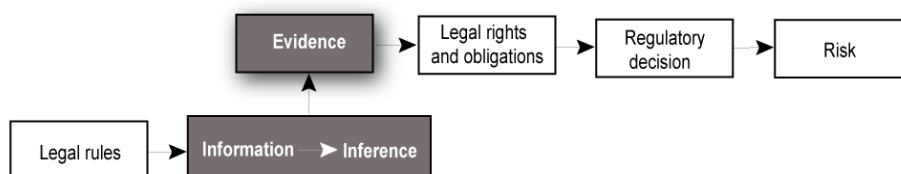
Fig. 3. Evidence is often the precursor to obligations, rights and responsibilities established by law, necessitating a decision that influences risk. But just as evidence leads to legal rights and obligations that constrain regulatory decisions, so too does the law constrain how evidence is established, sometimes in ways that are inconsistent with scientific best practices.

effect underlies every single study in the meta-analysis; i.e., if each study was infinitely large, there would be no heterogeneity between studies. However, no study is infinitely large and therefore one must assume that individual studies exert random effects. These random effects have some mean value and some measurement of variability stemming from between-study differences.

By meta-analyzing genomic studies, one increases sample size as well as the variation in genomic data, thereby enhancing the power to detect true associations. The weight of evidence for genomic information accumulates over time. Previous, individual studies form the prior belief. With each additional study, estimates are updated to form the posterior belief in a way that takes into account all available evidence.

5   Linking Genomic Information to Regulatory Decisions with Evidence

While high-throughput technologies, the proliferation of genomic information, and evolving analytical techniques will continue to spur the scientific community's understanding of the cellular basis for disease, an important issue that remains uncertain is how these advances will be used to make regulatory decisions. The issue arises because, as a matter of administrative law, governmental agencies must issue regulations that have a "rational basis," which the courts have taken to mean that a regulation, among other things, must be based on the scientific evidence [17]. Some threshold of evidentiary burden must be satisfied before the evidence triggers legal obligations, rights or responsibilities that thereby necessitate a decision influencing risk (see Fig. 3). But just as scientific evidence constrains regulatory decisions, so too does the law constrain the way that evidence is established, sometimes in ways that are inconsistent with best scientific practices.

An example of how the law on scientific evidence can hamper the use of science for regulatory decisions is provided by the controversy surrounding the Food and Drug Administration's approval of the painkilling drug, Vioxx. Vioxx

works by suppressing enzymes regulating the body's production of compounds associated with inflammation. Unfortunately, these compounds also play a role in maintaining the cardiovascular system [6]. Even before FDA's approval of Vioxx, there had been evidence indicating that inhibiting these enzymes may elevate blood pressure, may thicken artery walls, and may increase blood clots — all of which affect the risk of heart disease [19]. Two complications obfuscated the drug's risks. First, each single piece of evidence of risk — taken alone — did not dispositively evince a hazard [6]. Second, current research indicates that the response to Vioxx within a population is subject to genetic variation [22].

But in 1962, Congress mandated that FDA must assess whether a drug was effective for its intended use based on "substantial evidence" from "adequate and well-controlled investigations." The agency interpreted this statute to mean that a regulatory decision on drug effectiveness must be based on *randomized, replicated, controlled, clinical trials (RCTs). Controlled* means a clinical trial is designed so that ideally, treatment and control groups are identical in every way but one, which is in the levels of treatment being tested. Ergo, any variability among these groups is attributed solely to the treatment. In reality, it is difficult to eliminate extraneous sources of variability. Therefore, one *randomizes* how treatments are assigned to the various groups so that, ideally, the effects of any extraneous sources of variability cancel out. Finally, to ensure that experimental results do not occur through sheer happenstance, one *replicates* or repeats the experiment several times. But as stated earlier, it was because of the variable response to Vioxx within the population that the risks of the drug were not fully appreciated. In order to maintain the homogeneous conditions necessitated by an RCT, the drug manufacturers left out data pertaining to those who would have been at greatest risk — an older demographic with previous history of heart disease. Given FDA's enshrinement of RCTs as the gold standard for substantial evidence supporting claims of drug safety, it is not difficult to see why false negatives — as in the Vioxx case — were inevitable.

Shortly after Vioxx was taken off the market, the National Academy of Science's Institute of Medicine (IOM) issued a report clearly stating what others had been saying for some time: that FDA's practices were unlikely to detect rare but serious drug risks [18]. Before drug approval by the FDA, RCTs simply do not have the statistical power to generate the information needed to assess risks that arise when the general population is exposed to a drug. After drug approval by the FDA, FDA did not possess the statutory authorities needed to implement a nation-wide system to continue gathering this information. The IOM report advocated assessing safety over a drug's life-cycle, in which data were to be continuously gathered from multiple sources for ongoing analyses.

In 2007, Congress passed the Food and Drug Administration Act, which corrected the FDA's over-reliance on RCTs and statistical p-value tests to evaluate drug safety. First, Congress directed FDA to establish a network of data systems to integrate any and all information that can be used to evaluate drug risks. Second, it provided FDA with new, extensive authorities to require continuous submission of risk information from drug companies.

Given the newness of the FDA Act of 2007, as well as the unprecedented use of genomic information to inform regulatory decisions, it remains to be seen how courts will rule when these decisions are challenged based on a lack of "rational basis."

6   Conclusion

For the first time, technology exists to monitor the biological state of an organism at multiple levels. It is now possible to detect which genes are activated or deactivated when exposed to a chemical compound; to measure how these changes in gene expression cause the concentrations of cell metabolites to increase or decrease; to record whether these changes influence the over-all health of the organism. By integrating all this information, it may be possible not only to explain how a person's genetic make-up might enhance her susceptibility to disease, but also to anticipate how drug therapy might affect that individual in a particularized manner.

But two related uncertainties obscure the path forward in using these advances to make regulatory decisions. These uncertainties relate to the unsettled notion of the term "evidence" — both from a scientific and legal perspective. From a scientific perspective, as models based on genomic information are developed using multiple datasets and multiple studies, the weight of scientific evidence will need to be established not only on long established protocols involving p-values, but will increasingly depend on still evolving Bayesian measures of evidentiary value. From a legal perspective, new legislation for the Food and Drug Administration has only recently made it possible to consider information beyond randomized, clinical trials when evaluating drug safety. More generally, regulatory agencies are mandated to issue laws based on a "rational basis," which courts have construed to mean that a rule must be based, at least partially, on the scientific evidence. It is far from certain how judges will evaluate the use of genomic information if and when these rules are challenged in court.

References

1. Baragatti, M.: Bayesian Variable Selection for Probit Mixed Models Applied to Gene Selection. Bayesian Analysis 6(2) 209–229 (2011)
2. Basu, N.: Proposal for EPA Grant. On file with author (2011)
3. Butler, D.: Science after the sequence. Nature 465 1000–1001 -(2010)
4. Cohen, J.: The Human Genome, a Decade Later. Technology Review, January/February (2011)
5. Ferrell Jr, J.E.: Q&A: Systems Biology. Journal of Biology 8(2) Article 2 (2009)
6. Grosser, T., Yu, Y., et al: Emotion Recollected in Tranquility: Lessons Learned from the COX-2 saga. Annual Review of Medicine 61 17–33 (2010)
7. Hayden, E.C.: Human Genome at Ten: Life is Complicated. Nature 464 664–667 (2010)
8. Imai, K., van Dyk, D.A.: A Bayesian Analysis of the Multinomial Probit Model using Marginal Data Augmentation. Journal of Econometrics 124(2) 311–334 (2005)

9. Ioannidis, J.P.A., Panagiotou, O.A.: Comparison of Effect Sizes Associated With Biomarkers Reported in Highly Cited Individual Articles and in Subsequent Meta-analyses. Jama-Journal of the American Medical Association 305(21) 2200–2210 (2011)

10. Kwon, D., Landi, M.T., et al.: An Efficient Stochastic Search for Bayesian Variable Selection with High-dimensional Correlated Predictors. Computational Statistics & Data Analysis 55(10) 2807–2818 (2011)

11. Lander, A.: The Edges of Understanding. BMC Biology 8(1) 40 (2010)

12. Lenormand, T., Roze, D., et al.: Stochasticity in Evolution. Trends in Ecology & Evolution 24(3) 157–165 (2009)

13. Lewin, A. and Richardson, S.: Bayesian Methods for Microarray Data. In: Balding, G.J., Bishop, M., Cannings, C. (eds.) Handbook of Statistical Genetics, 3rd Edition, Wiley, pp. 267–295 (2007)

14. Matthews, J.N.A.: Moore Looks Beyond the Law. Physics Today 61 20 (2008)

15. National Human Genome Research Institute (NHGRI): DNA Sequencing Costs. Accessed Nov. 11, 2011 at http://www.genome.gov/sequencingcosts/

16. National Research Council (NRC): Toxicity Testing in the 21st Century: A Vision and a Strategy. National Academies Press, Washington, DC (2007)

17. Pascual, P.: Evidence-based Decisions for the Wiki World. International Journal of Metadata, Semantics and Ontologies 4(4) 287–294 (2009)

18. Pray, L.A., Robinson, S., et al.: Challenges for the FDA: the Future of Drug Safety: Workshop Summary. National Academies Press, Washington, DC (2007)

19. Ritter, J.M., Harding, I., et al.: Precaution, Cyclooxygenase Inhibition, and Cardiovascular Risk. Trends in Pharmacological Sciences 30(10) 503–508 (2009)

20. Rusyn, I., Daston, G.P.: Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century. Environmental Health Perspectives 118(8) 1047 (2010)

21. Sha, N.J., Vannucci, M., et al.: Bayesian Variable Selection in Multinomial Probit Models to Identify Molecular Signatures of Disease Stage. Biometrics 60(3) 812–819 (2004)

22. St Germaine, C.G., Bogaty, P., et al.: Genetic Polymorphisms and the Cardiovascular Risk of Non-Steroidal Anti-Inflammatory Drugs. American Journal of Cardiology 105(12) 1740–1745 (2010)

23. Stingo, F.C., Chen, Y.A., et al.: Incorporating Biological Information into Linear Models: a Bayesian Approach to the Selection of Pathways and Genes. The Annals of Applied Statistics 5(3) 1978–2002 (2011)

24. Valentinuzzi, M.E.: Neuroendocrinology and its Quantitative Development: A Bioengineering View. Biomedical Engineering Online 9 (2010)

25. West, M.: Bayesian Factor Regression Models in the "Large p, small n" Paradigm. Bayesian Statistics 7 723–732 (2003)

26. Zeggini, E., Ioannidis, J.P.A.: Meta-analysis in Genome-wide Association Studies. Pharmacogenomics 10(2) 191–201 (2009)

27. Zhang, X., Boscardin, W.J., et al.: Bayesian Analysis of Multivariate Nominal Measures using Multivariate Multinomial Probit Models. Computational Statistics & Data Analysis 52(7) 3697–3708 (2008)

DISCUSSION

Speaker: Pasky Pascual

Brian Smith : You expressed the issue of injury-in-fact versus probability of the event as a concern with legal issues. Why is cost not a part of the issue, or why is it not discussed?

Pasky Pascual : The only reason why I did not specifically discuss the issue of cost was because of time constraints. As a matter of law, each major regulation that is issued must first undergo a cost-benefit analysis which is then submitted to the White House's Office of Management and Budget. So, when issuing a regulatory decision that is based on genomic information, an agency must also be able to estimate the monetary value of the costs and the benefits associated with a particular public health or environmental law.

Maurice Cox : My understanding of your main thesis is as follows. Scientific evidence is out there, usually in the form of data. You explain your assumptions and the statistical or computational model you are using. Then, if you have done your job properly, you should be able to convince the court. But, the court might question the validity of that data in terms of its reliability and consistency. I would welcome your comments.

Pasky Pascual : That's quite right. But part of the problem lies in the fact that the courts may not have the appropriate scientific training to evaluate the scientific evidence with which it is presented. As a matter of law, the courts will be deferential to agencies, particularly in areas that fall within agency's technical expertise and competence. But when a decision is challenged, the courts will subject the "rational basis," which includes the scientific basis, of an agency to a critical review. These reviews are necessarily *ad hoc* and depend on the particularities of the case. But few guidelines, if any, exist to assist the court in conducting this review.

Jeffrey Fong : Does the EPA have a policy statement on the minimum reliability of informatics data that is acceptable? If not, does the speaker have a personal opinion on this question?

Pasky Pascual : My personal opinion is that rather than have a standard score of reliability that then determines acceptability, I would find transparency of the informatics data and the analysis through which the data are used to derive inferences to be more useful. If I were to tell you, for example, that a particular dataset is 99% reliable, what would that mean? Perhaps it is unavoidable that people will demand some kind of seal of approval for a dataset or model that is used to make a decision, but I would want to make sure that the process through which this evaluation occurs is also communicated.

Tony O'Hagan : This is a conference on uncertainty quantification. You've talked a lot about the law. My understanding is that these two things don't go

together. Lawyers hate uncertainty, unless they can use it as a weapon against someone foolish enough to admit uncertainty. They hate uncertainty quantification even more. For instance, you pointed out that the law would much prefer anecdotal evidence of actual harm to scientific reasoning of probabilistic harm. What do you feel can be done about this?

Pasky Pascual : I agree that the law tends to operate on binary terms — you comply with a rule or you don't; a drug is safe to market or it is not. So the legal decisions that ultimately get made based on scientific evidence do tend to eschew uncertainty. But at least within a regulatory context, these decisions do consider the uncertainties of science. For example, when EPA issues a regulation these days, it will generally conduct formal uncertainty analyses in order to better understand sources of uncertainty in the science. It may be something as simple as conducting Monte Carlo draws in order to derive a distribution of outputs from a model, rather than a single estimate. So quantification of uncertainty occurs at that phase of rule development. The decision itself may be binary — the Agency regulates or does not regulate a compound — but the analysis that enters into the decision is not. Moreover, when EPA does conduct formal uncertainty analyses when it proposes a rule, these analyses are generally discussed in the documents that accompany the issuance of a rule.

William Oberkampf : Given the strong aversion to uncertainty in the legal and judicial system, how will the EPA deal with more sophisticated uncertainty quantification methods in the future?

Pasky Pascual : My personal opinion is that uncertainty quantification is not going to go away. We will see more, rather than less, of it. It is in the best interest of regulatory agencies to be transparent in their analyses. Transparency is what leads to more defensible decisions. And part of analytical transparency is transparency about sources of uncertainties — both epistemic and alleatory.

Antonio Possolo : In relation with your stated goal of replacing in vivo animal experimentation with studies of differential gene expression: in 2005, colleagues and I published an article in Toxicological Sciences suggesting that studies of differential gene expression in vitro, using live rat and human liver cells, was an effective proxy for studies involving live animals, and also much more expeditious (days, including microarray processing and data analysis, versus the years that it takes for malignancy indications to express themselves), induced by PCBs. Why is it taking the EPA so long to put these and similar scientific, peer-reviewed results to widespread use?

Pasky Pascual : Part of it lies in the complexity of the organism. As we are realizing more and more, metabolic pathways rarely proceed in a linear fashion. For example, if we know that a gene that triggers a series of reactions is deactivated, there still may be other genes to ensure that the reactions will occur. So, the ways that a gene may relate to the manifestation of an observed harm is organized through networks of information rather than simple discrete pathways. And figuring what those networks are and how they operate is extremely hard,

I think. Also, it's still not clear, to me anyway, what the evidentiary threshold has to be, before we can say — in a way that is legally defensible — that the behavior of this particular set of biomarkers are a reliable indicator that the likelihood of harm is increased to a level that warrants regulatory action.