Comment

Bayesian Network Integrated Testing Strategy and beyond

Summary

In a recent series of papers written by Jaworska with different coauthors, compelling reasons for adopting a probabilistic approach to Integrated Testing Strategies were detailed. In a case study on skin sensitization, a Bayesian Network proved to be effective in adapting testing strategies to the available evidence. There is no doubt that probabilistic Integrated Testing Strategies are one way to pursue the goals of 3Rs effectively; nevertheless, some issues deserve further comment to pinpoint statistical criticalities and to widen the methodological perspective towards Bayesian graphical models.

1 Introduction[1]

In their seminal book, published in 1959, Russell and Burch lamented the delay in the use of some statistical methods (tests) that "... have probably not been exploited to the full, even in research immediately after their provision." While explaining the importance of the design of experiments, they emphasized that "Every time any particle of statistical method is properly used, fewer animals are employed than would otherwise have been necessary." Equipped with their words, we make statistical remarks on three recent papers dealing with the integration of testing strategies.

In a recent paper, Jaworska and Hoffmann (2010b) stated that Integrated Testing Strategies (ITS) may be considered combinations of test batteries covering relevant mechanistic steps, organized in a logical and hypothesis driven decision scheme, with the aim of providing a comprehensive information basis for making decisions on chemical hazard and risk management. In the same paper, the authors reviewed conceptual requirements for ITS and defined properties that ITS should have to meet the identified requirements.

Among the issues addressed by the authors, the need for context and interpretation to transform data into information, and therefore knowledge, seems to us preeminent for several reasons. The systematic analysis and use of the existing wealth of multifaceted biological data is a requisite for advances in the understanding of life processes and risks, and it appears to be impossible without setting proper contexts. The context determines, to a large extent, which statistical techniques are suited to extract information from data, but such information becomes knowledge only after updating the personal system of scientific beliefs, a step in which interpretation is essential. This process is inherently affected by uncertainty, called epistemic uncertainty, which has to be considered together with aleatory uncertainty due to randomness, and with uncertainty induced by the measurement process before distilling knowledge from data.

From this standpoint, Bayesian statistics has much to offer to ITS because it is the methodological field devoted to the study of uncertainty (Lindley, 2000, 2006, for an informal introduction), as suggested in Jaworska et al. (2010a, p. 160) both for data analysis and reasoning with evidence. This is a need recognized by Jaworska and Hoffmann (2010b) who stated: "... probabilistic methods provide a formal approach for quantifying uncertainty from heterogeneous input sources, relationships between them, and overall target uncertainty", and also "... an operational framework (for ITS) that needs to be probabilistic, even better Bayesian and adaptive." Major pitfalls in statistical modeling, such as incomplete data and conflicting evidence, may be properly faced in the Bayesian paradigm, with the guarantee of full agreement with the principles of logic and rationality (Lindley, 2000). An explicit mention is made by the authors about the possibility of discovering weak signals of high importance, like those arising as a consequence of complex feedback mechanisms in biological signaling, using "... prior knowledge about the target of interest" (Jaworska and Hoffmann, 2010b): this attitude towards the elicitation of the degree of belief in

[1] **Abbreviations**
BN, Bayesian Network; BNITS, Bayesian Network Integrated Testing Strategy; CI, Conditional Independence; DAG, Directed Acyclic Graph; FEPEs, fast and exact propagation of evidences; ID, Influence Diagram; ITS, Integrated Testing Strategies; MCMC, Markov Chain Monte Carlo; SCM, Structural Causal Model

a quantitative way is at the core of Bayesian model building (Garthwaite et al., 2005).

The probabilistic approach to ITS seems first to be discussed in Jaworska et al. (2010a), where ITS was criticized for the lack of a principled information processing framework able to incorporate all relevant information while updating uncertainty in a coherent way. The authors built an information-theoretic approach strongly rooted in probabilistic modeling, called the "ITS inference framework," where Bayesian networks were proposed as the software tool to make the ITS inference framework operational. The framework proposed in (Jaworska et al., 2010a) complies with the OECD (2008) recommendations that ITS development should be structured, consistent, transparent, and hypothesis-driven. Many of the above desiderata, if not all, are achieved by using probabilistic analysis and reasoning techniques that find their methodological foundations in the Bayesian paradigm. From the standpoint of applications, besides pursuing hypothesis-driven inferences, the approach supports the assessment of the value of collected information, with the possibility of calculating the expected value of information gain provided by alternative testing procedures.

Jaworska et al. (2011) put the above concepts into practice by developing the "Bayesian Network Integrated Testing Strategy" (BNITS) to estimate skin sensitization hazard. The proof of concept case proved BNITS to be effective in adapting testing strategies to available evidence while combining *in silico*, *in chemico*, and *in vitro* data related to skin penetration, peptide reactivity, and dendritic cell activation. A key issue highlighted by the authors is that the search for an unlikely gold-substituting *in vitro* test or best testing strategy should be substituted by the optimal decision in face of the available experimental evidence: this approach was even successful in a case study where missing values (data gaps) amounted to 50% of database records.

The above mentioned papers (Jaworska et al., 2010a, 2011; Jaworska and Hoffmann, 2010b) provided a wide-ranging account of the role played by probabilistic inference in ITS, masterly framed within the perspective of validation strategies. Overall, we endorse the grand vision depicted as "ITS inference framework," because several of its nice features follow from methodological results proper of the Bayesian field (Robert, 1994; O'Hagan, 1994; Bernardo and Smith, 1994, for comprehensive accounts). Nevertheless, some issues deserve further refinement to unleash the full power of the Bayesian paradigm in ITS and to put the "Bayesian Network Integrated Testing Strategy" in perspective.

2 Bayesian Network Integrated Testing Strategy

The term "Bayesian Networks" (BNs) typically is used to indicate a class of statistical models in which the joint probability distribution of a vector made by discrete random variables is represented as a product of conditional distributions like $p(x_v|x_{pa(v)})$, where v is a node in the Directed Acyclic Graph (DAG) G defining the network of random variables and $pa(v)$ is the collection of its parent nodes. Figure 2 in Jaworska et al. (2010a) shows a DAG of three nodes, *Carcinogenic*, *T1Ames* and *T2MLA*, where directed edges are: *Carcinogenic* → *T1Ames*, *Carcinogenic* → *T2MLA* and *T1Ames* → *T2MLA*. It follows that the joint distribution of those three variables is decomposed into the product:

$$p(Carcinogenic) \bullet p(T1Ames\,|\,Carcinogenic) \bullet \\ p(T2MLA\,|\,Carcinogenic, T1Ames)$$

because the required conditional distributions are straightforwardly read from the DAG (Cowell et al., 1999).

A secondary meaning of BN refers to the software implementation of one model (a specific instance in the class of BNs). Commercial, free, and open source software exist to support the creation and use of BNs through a graphical user interface. Calculations like conditioning and marginalization are exact (without approximations besides those due to floating point computations) and fast (performed by highly optimized algorithms), the so called fast and exact propagation of evidences (FEPEs). Current software programs to implement BNs, besides FEPEs, also offer tools to infer the DAG structure and to estimate parameters of conditional distributions using actual observations, two tasks respectively called structural learning and parameter learning.

Probabilistic reasoning with BNs and BN learning are distinct tasks, and they require a quite different degree of statistical expertise because the first one is attainable after limited training without leaving the toxicological context, while BN learning involves far more statistical skills. Latent variables, the imputation of missing values, and model averaging over several DAG structures (structural uncertainty) are issues often present in actual applications, as they are in Jaworska et al. (2011). Last but not least, records of a database must be exchangeable for being properly processed by common software learning procedures.

Many of the above critical issues are present in (Jaworska et al., 2011), so they are all but immaterial, but the many ways BN learning can fail deserve proper consideration. For the sake of brevity, just a few critical issues are highlighted here. A model with latent variables may suffer from partial identification, thus the elicitation of prior information is highly recommended in such cases. Nevertheless, BN software often allows the specification of limited or no dependence among model parameters during the elicitation of the prior distribution and learning. The numerical optimization of the likelihood function, or other likelihood-related scores, may end to suboptimal points, especially in the huge search spaces of DAG structures. For example, six variables define a set of 3.781.503 different DAGs. From this standpoint the statement "... the structure of the BN and the probabilistic relationships between variables were extracted directly from the data" (Jaworska et al., 2011, p. 213, right column) may be perceived as the indication of an automatic and blind learning procedure, a dangerous attitude if wrongly pursued. Missing values and latent variables typically increase the uncertainty about an estimated BN, and the analysis of a dataset after single imputation (called data gaps filling in Jaworska et al. 2011) with estimated values may lead to overstated conclusions due to the single imputation. The uncertainty about structure and parameter values of a BN plays the same role as the uncertainty present in the toxicological problem domain. Full probabilistic coherence is achieved only if all relevant sources of uncertainty are properly taken into account, for example, by avoiding the substitution of unknown quantities through the plug-in of their point estimates, an issue apparently neglected in Jaworska et al. (2011). Sensitivity analysis, as performed in Jaworska et al. (2011, p. 223, left column), is useful to evaluate performances of variants of the DAG (different networks), but in general it does not substitute model averaging over DAGs while performing probabilistic predictions.

The above discussion points towards the conclusion that Bayesian networks, as a probabilistic framework characterized by exact computation with discrete variables, is unnecessarily restrictive and of limited learning abilities, at least in most of the software currently available. Are online fast calculations really needed, as it happens in emergency departments? Is the discretization of continuous variables causing minor loss of information? Are all relevant variables natively discrete? Is the uncertainty about model parameters negligible? Is the net structure (DAG) based on strong prior information or estimated using very large databases? If all the answers to the above questions are "yes," then BNs are likely to be the right tool; otherwise, the adoption of a wider framework, often indicated as Bayesian graphical models (Buntine, 1994), is recommended. This is not a substitution but an extension of the probabilistic approach provided by BNs, which keeps all the key features highlighted in Jaworska et al. (2011, p. 222, left column), such as the ability to deal with uncertainty in biological knowledge, to combine heterogeneous pieces of evidence, and to quantify uncertainty about target and relationships.

Bayesian graphical models include Bayesian networks as specialized instances, those in which the factorization of the joint distribution is determined by a DAG and where variables are discrete. Parameters may be included within the DAG, if they are affected by uncertainty and missing values are considered at the same level as model parameters. There is virtually no limitation on the kind of variables the scientist may use to properly represent his/her beliefs, collected observations, and specific features of a problem domain. FEPEs often is no longer possible in general, but conditioning and marginalization may still be performed through approximated computations. In Jaworska et al. (2010a, p. 164, left column), the authors made an ambiguous statement because Gaussian Bayesian networks in which all variables are normal admit FEPEs. Similarly, mixed Bayesian networks made by discrete and Gaussian variables admit FEPEs if Gaussian variables are never parents of discrete variables (Cowell et al., 1999).

The generality of Bayesian graphical models comes at the price of a higher computational cost. The royal road to Bayesian computation is Monte Carlo simulation, possibly Markov Chain Monte Carlo (MCMC) (Brooks, 1998). The typically difficult integrals required by the application of the Bayes rule are avoided by sampling from the (unnormalized) posterior distribution of all the unknown quantities (parameters, missing values, and latent variables). The degree of approximation depends mainly on the quality of the sampler and on the sample size; thus, it is largely under the control of the scientist and it may be increased as needed. The availability of open source software to perform MCMC (Spiegelhalter et al., 2003; Plummer, 2003; Stan Development Team, 2013) makes model specification very fast, with templates already available for a large class of standard statistical models. In particular Winbugs (Spiegelhalter et al., 2003) may take a DAG as starting input during model specification. In this framework, FEPEs is substituted by calculations performed on the predictive distribution of the next (multivariate) observation, given cases already considered during learning.

At the very end, the nice properties of probabilistic reasoning with Bayesian networks neither depend on the discrete nature of variables nor on the existence of FEPEs. The potential confusion of model properties and features of the available software should be avoided.

3 Graphs for probabilistic and causal reasoning

Probabilistic and causal models can be represented by graphical models. This point is acknowledged in Jaworska et al. (2010a, p. 161, left column), where the authors stated that "BNs ... are defined as graphical models of probabilistic relationships between variables of interest ..." and a few lines after "BNs can be regarded as decision-support frameworks because of their ability to explain causal relationships and to serve as prediction models." A deeper appreciation of the above two aspects is possible by establishing an explicit connection between them.

Bayesian inference defines how the expert should rationally change his/her beliefs in face of new evidence, whether randomized controlled experiments or observational studies are performed, with the extreme circumstance represented by the design of an experiment where only prior information is exploited. Conditional Independence (CI) (Dawid, 1979) and the Bayes rule are pillars of the probability calculus by which inferential answers are produced. Qualitative reasoning about CI relationships may be performed using DAGs – that is without dwelling on algebraic manipulations of probability distributions but exploiting graph separation theorems (Cowell et al., 1999). Here two remarks are mandatory, the first to emphasize that not all the CI relationships in a distribution can always be represented by a DAG, therefore the need for more general graphical representations follows. The second remark is to make precise that in a DAG of a Bayesian network, two variables X_a and X_b are represented as conditionally independent given X_c if such relationship holds for all possible values $c_1,c_2,\ldots,c_k$ of the conditioning variable X_c. Thus only strong CI relationships are explicitly represented by a DAG.

A DAG has to be specified well before the numerical details pertaining to conditional distributions of a BN. Nevertheless, the representation of CI relationships do not cover all the needs in ITS, as clearly stated in Jaworska et al. (2011, p. 222, left column), where the authors expounded that "Determining the causal structure is a key for mechanistic interpretation capability of ITS"; in Jaworska et al. (2011, p. 214, top left column), the authors stated that "... the value of using the network is far more than a prediction framework. The network represents key steps of the skin sensitization process ..." The general theory of causation based on the Structural Causal Model (SCM) is due to Pearl (Pearl, 2009, for an introduction), who developed a mathematical foundation in the analysis of causes and counterfactuals. Narrative summaries of causal knowledge are substituted by DAGs and other diagrams that are useful to communicate causal assumptions, to decide if they are sufficient for obtaining estimates of the desired target quantities, to derive closed-form expression of such quantities and to suggest the observations that, if collected, would make target quantities estimable (Pearl, 2000, 2009).

Over and above processing uncertainty/information in a coherent way, causal relationships are modeled to predict a system under external intervention that is used to calculate the probability distribution of some random variables that would result if some other variables were forced to take certain values. This kind of information typically is obtained by randomized controlled experiments, when manipulation of a system or process is feasible in controlled conditions. Nevertheless, using SCMs it is, in principle, possible to evaluate the effect of interventions on systems or processes that were passively observed, without manipulation. Pearl's SCM does not put constraints on the nature of random variables, and it does not force the scientist to think in terms of parameterized distributions, either on discrete variables or not. SCM embeds deterministic relationships, which is an advantage if causal relationships among variables are natively characterized in this way.

Other approaches to causal modeling under active development and use include Rubin's Potential Outcomes (PO), which extends the framework of randomized experiments proposed by Fisher and Neyman (Mealli et al., 2011, for an introduction). A toxicologist might prefer SCM because the PO framework does not natively exploit graphs to represent assumptions and because it forces the restructuring of a causal inference problem as a problem of missing data. Pearl stated that SCM is a general theory that has PO as a specialized instance and that the two approaches lead to the same calculations. It is worthy of special attention that the framework proposed by Dawid (2002), where DAGs are augmented by other types of nodes to represent parameters, decision strategies, and utilities, without introducing concepts outside those already standard in the probabilistic framework. In particular, Influence Diagrams (ID) are proposed as graphical models with "... just the right degree of expressive power" (Dawid, 2002) to handle the estimation of effects due to the considered causes.

Causal relationships are top quality information and are absolutely relevant for a toxicologist because the main interest focuses on what happens to a system (a cell, tissue, or organism) subject to intervention (manipulation-perturbation), for example after applying a given cosmetic to the skin. The point of contact between pure probabilistic and causal modeling is DAG modularity-stability, that is, the property that intervention produces local changes in manipulated variables, thus leaving all other variables and relationships unchanged. Using the words of Pearl (Pearl, 2009, p. 118) "The new ingredient that causal analysis brings to this tradition is the necessity of obtaining explicit judgments, not about properties of the distributions but about the invariants of a distribution..."

The causal interpretation of a DAG or ID has to be justified by substantive reasons, especially if actual intervention studies are not feasible and the graph structure is inferred using observational data. Here the context plays a major role because a causal model is meaningless without a proper defining context. The context includes things like the specification of the protocol-equipment involved in the intervention and the collection of considered variables, those appearing as nodes of a DAG. This is not a trivial choice, because, for example, the collection of considered variables determines the model granularity, that is the level of detail under consideration, and thus two DAGs at different levels of model granularity may show two different sets of direct causes for the same variable. A detailed discussion including covariates in observational studies is provided by Pearl (2000).

4 Conclusions

For reasons considered in the above sections, the operational framework indicated as Bayesian Network Integrated Testing Strategy can and should be broadened to include more general Bayesian graphical models. Most important, DAGs, IDs, and other graphical representations enable toxicologists to reason on important causal and probabilistic model features without resorting to specific model parameterizations or numerical details that typically require extensive statistical training. The intent was not to suggest that there is something wrong with BNs, as we applied BNs in fields as diverse as forensic science (Corradi et al., 2003) and breast cancer biomarkers (Stefanini et al., 2009). By recognizing DAGs and BNs as distinct tools, it becomes natural to consider other useful graphical representations and to emphasize that DAGs are important tools in and of themselves (Luciani and Stefanini, 2012, for an example in medical knowledge engineering). Here, it was not possible to cover Bayesian graphical models in full depth, and Bayesian structural learning under sparse prior information on structure seems one among the most important exclusions (Stefanini, 2012).

The hope is to have provided an expanded perspective on BNITS to motivate many toxicologists to seriously consider Bayesian graphical models as a major methodological opportunity to strengthen ITS even further towards the fulfillment of Russell and Burch's 3Rs: Replace, Reduce, and Refine.

Acknowledgements

The author thanks Chiara Urani for fruitful discussions on toxicology *in vitro* and for comments that improved the clarity of the manuscript.

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York, USA: John Wiley & Sons.

Brooks, S. P. (1998). Markov Chain Monte Carlo method and its application. *J Roy Stat Soc D Sta 47*, 69-100.

Buntine, W. L. (1994). Operations for learning with graphical models. *J Artif Intell Res 2*, 159-225.

Corradi, F., Lago, G., and Stefanini, F. M. (2003). The evaluation of DNA evidences in pedigrees requiring population inference. *J Roy Stat Soc A Sta 166*, 425-440.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., et al. (1999). *Probabilistic Networks and Expert Systems*. Heidelberg, Germany: Springer Verlag.

Dawid, A. P. (1979). Conditional independence in statistical theory. *J Roy Stat Soc B Sta 41*, 1-31.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *Int Stat Rev 70*, 161-189.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *J Am Stat Assoc 100*, 680-701.

Jaworska, J., Gabbert, S., and Aldenberg, T. (2010a). Towards optimization of chemical testing under REACH: A Bayesian network approach to Integrated Testing Strategies. *Regul Toxicol Pharm 57*, 157-167.

Jaworska, J. and Hoffmann, S. (2010b). Integrated Testing Strategy (ITS) – Opportunities to better use existing data and guide future testing in toxicology. *ALTEX 27*, 231-242.

Jaworska, J., Harol, A., Kern, P. S., et al. (2011). Integrating non-animal test information into an adaptive testing strategy – Skin sensitization proof of concept case. *ALTEX 28*, 211-225.

Lindley, D. V. (2000). The philosophy of statistics. *J Roy Stat Soc D Sta 49*, 293-337.

Lindley, D. V. (2006). *Understanding Uncertainty*. New York, USA: John Wiley & Sons.

Luciani, D. and Stefanini, F. M. (2012). Automated interviews on clinical case reports to elicit directed acyclic graphs. *Artif Intell Med 55*, 1-11.

Mealli, F., Pacini, B., and Rubin, D. B. (2011). Statistical inference for causal effects. In R. S. Kenett and S. Salini (eds.): *Modern Analysis of Customer Surveys: With Applications Using R* (173-192). Chichester, UK: John Wiley & Sons.

O'Hagan, A. (1994). *Bayesian Inference. Kendall's Advanced Theory of Statistics*. London, UK: Edward Arnold.

OECD (2008). Workshop on integrated approaches to testing and assessment. OECD Environment Health and Safety Publications. *Series On Testing And Assessment No. 88*. Paris, France: OECD.

Pearl, J. (2000). *Causality. Cambridge*, UK: Cambridge University Press.

Pearl, J. (2009). Causal inference in statistics: an overview. *Stat Surv 3*, 96-146.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis (eds.): *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (1-10). Vienna, Austria: Technische Universität Wien. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/

Robert, C. P. (1994). *The Bayesian Choice: a Decision-Theoretic Motivation*. New York, USA: Springer-Verlag.

Russell, W. M. S. and Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. http://altweb.jhsph.edu/pubs/books/humane_exp/het-toc

Spiegelhalter, D., Thomas, A., Best, N., et al. (2003). *WinBUGS User Manual, Version1.4*. http://www.mrc-bsu.cam.ac.uk/bugs

Stan Development Team (2013). *Stan: A C^{++} Library for Probability and Sampling, Version 1.1*. http://mc-stan.org/

Stefanini, F. M., Coradini, D., and Biganzoli, E. (2009). Conditional independence relations among biological markers may improve clinical decision as in the case of triple negative breast cancers. *BMC Bioinformatics 10*, S13.

Stefanini, F. M. (2012). Graphical models for eliciting structural information. In G. Antonio, G. Ritter, and M. Vichi (eds.), *Classification and Data Mining* (139-146). Heidelberg, Germany: Springer Verlag.

Federico M. Stefanini
Dipartimento di Statistica, Informatica,
Applicazioni 'G. Parenti' – DiSIA
Università degli Studi di Firenze
Viale Morgagni 59, I-50134, Firenze, Italy
Fax: +39 055 4223560
Phone: +39 055 4237266
e-mail: stefanini@disia.unifi.it