Biopharmaceutical Section

American Statistical Association

Biopharmaceutical Report

Volume 18, No. 2 Fall 2011

Chair: ***Steve Wilson*** **Editors:** *Jose Alvir, Deborah Panebianco, Amit Bhattacharyya*

Note from the Editors

In this issue we feature two articles. The first, by Danel J. Zaccaro and Leela M. Aertker (Rho Inc) discusses the use of zero-inflated mixture models in vaccine testing to better characterize subjects whose antibody levels fall below the lower limit of quantitation. They apply this methodology to data from an H1N1 vaccine trial.

John J. Peterson (GlaxoSmithKline Pharmaceuticals) and Ron S. Kenett (The KPA Group) argue that process improvement and optimization in pharmaceutical development and manufacturing can be enhanced through greater use of stochastic process modeling. They identify and provide a thorough review of two such modeling opportunies, multivariate predictive regression and Bayesian Networks.

We also welcome Yongming Qu (Eli Lilly & Co.) as new Associate Editor. Yongming will be replacing Amit Bhattacharyya, who is stepping down from his editorial position in 2012 to assume his new role as section Chair-Elect. ■

Contents

Letter from the Chair…

The Steering Committee for this year's ***ASA/Biopharm*** *FDA/Industry Meeting* at the Marriott Wardman Park Hotel in Washington, DC, had a unique challenge in meeting the demands of an overwhelmingly full program (37 sessions!) – how to "kick-off" the meeting when they were starting with 5 parallel sessions? In responding to this challenge, the decision was made to have Committee members in each of the 5 rooms simultaneously "open" the meeting with a common slide deck and some of their own words.

In thinking about what to say, Brenda Crowe from Eli Lilly, the Industry Co-Chair for the meeting sent me an e-mail that said,

> *"I am reviewing what I want to say in my introductory comments. I plan to remind people to join the Biopharm Section and thought it might be more powerful if I told them of some of the benefits. Do you have thoughts on this?"*

After 9 months of serving as your Chair, working with this Section's tremendous elected/appointed officers and volunteers, putting together the BIOP ENAR and JSM Executive Committee Meetings, thinking about Section business, answering many, many e-mails, learning about all of the Section's activities and history…the thoughts and ideas tumbled out of my head (my apologies to Brenda who was just looking for a few simple bullets).

With this year's experience I have come to the realization that my advice is a "no-brainer" – if it is important to a statistician to be a fully involved professional (in all aspects), dedicated to the development of new medical (animal and human) products for the Public's good, then it is important that he/she be a supporting member of the *ASA Biopharm Section*.

However, realizing that Brenda just wanted a few bullets for a 30-second timeslot, I boiled it down to the following:

The Biopharm Section Works for all of us…developing numerous professional and leadership opportunities:

Professional Visibility, Networking, Continuous Education, Volunteer Opportunities and Career Support

- ASA/Biopharm FDA/Industry Workshop
- JSM
  - —Invited Sessions
  - —Contributed Paper Competition
  - —Tutorials
  - —Student Paper
  - —Business Meeting/Mixer
  - —Webinars
  - —Fellows Committee

Communication and Collaboration

- ASA publications
- Collaborative Web Community of colleagues
- Biopharm Newsletter

(Please feel free to use these bullets the next time a colleague asks you "Why should I pay an extra eight bucks per year to be a member of the Biopharm Section?")

According to everybody I have talked to, this year's *ASA/Biopharm FDA/Industry Meeting* was a great success!

- The biggest ever, with 820 participants;
- A terrific venue – "on the Red Line," new meeting rooms, internet access, good food, etc.; and
- So many well-designed/useful sessions to choose from – something for everyone.

All of this made possible by the many Section volunteers and the dedicated work of the ASA meeting support staff. A quick count of persons involved in the meeting (serving on the organizing committee, managing the roundtables, organizing sessions, chairing sessions, teaching, speaking, discussing, etc.) reveals that nearly 300 professionals, served as volunteers for this important ASA Biopharm activity—we did it, we put this meeting together for ourselves. The Workshop (the most democratic and collaborative of all of our meetings) is what the Section is all about – professional leadership, volunteerism, collaboration, career support and continuous education. Many kudos and thanks to the meeting co-chairs, Brenda from Eli Lilly and Joan Buenconsejo from FDA/CDER, and to all of the dedicated professionals who made this meeting happen.

So what do I do with 3 whole months remaining as your Chair? We are good, but (me being me) I feel that we can get better. How do we strategically plan for our future?

One of the few "powers" a BIOP Chair enjoys is the ability to create ad hoc committees to serve the needs of the Section, so I have assigned myself the task of creating and nurturing three new ones:

Leadership Committee – dedicated to examining how the Section can contribute to the recognition and the growing number of efforts in industry, academia and government to develop statisticians as leaders;

Continuing Education Committee – formed with a mandate to figure out how we can best describe, structure, grow and plan all of our various (current and future) CE and collaboration efforts; and

Specialty Committee – to examine and assure ourselves that we have the right policies and activities in place to meet the "specialized needs" of all of our members. Our *Charter* tells us that "The special interest of the Biopharmaceutical Section is the application of statistics to the development and use of therapeutic drugs, biologics, and devices in humans and animals." Are we doing all that we could/should? Can we do it better? Do all of our members feel equitably well-served by the Section's efforts and activities?

I have already recruited a number of you to help with these efforts and will more fully describe these committees at the upcoming *Transition Meeting* at the end of October. If you are interested in joining any of these groups—to think about our future—please get in touch with me (*stephen.wilson@fda.hhs.gov*). Hopefully, this will give me plenty to do during 2012 as your "Past Chair."

I am starting to realize that this has been a very short year—serving as your Chair…how could my term be nearly over when I am still learning how to do the job? The story of many of our lives: Too much to do, too many ideas and too little time.

Cheers.
Steve Wilson
2011 Chair, ASA BIOP Section

Summary of the Minutes of the ASA Biopharmaceutical Executive Committee Meeting Held August 1, 2011 in Miami, FL

Submitted by Rick Caplan

- Matilda Sanchez, Treasurer, reported that the Section has $359,653.43 as of July 25th. The proposed 2012 budget and the budget for the Biopharmaceutical FDA/Industry Workshop were reviewed. The Workshop is sponsored by the Biopharmaceutical Section.
- Alex Dmitrienko, David Breiter and Stephine Keeton gave the Council of Sections report. There will be an update to Sections' charters. A BIOP Ad Hoc Committee will handle this. The COS is proposing an earlier deadline, February 19, for decisions about student paper awards. There was a proposal for a new ASA Section on Imaging. ASA is organizing a Conference on Statistical Practice in February in Orlando.
- Devan Mehrotra, Publications Officer, reported that all items for the *Amstat News* are on track.
- *Biopharm Report* editors, Jose Alvir, Deborah Panebianco and Amit Bhattacharyya, reported that the Spring issue was published, the Summer issue will soon be published, and there are plans for a Winter edition.
- Fellows Committee Chair, Neal Thomas, reported that there were 5 successful nominations of Biopharm Section members this year.
- JSM Program Chair and co-chair, Jeff Maca and Carmen Mak, reported that the Biopharm Section has more sessions than any other Section. It was a very successful meeting.
- Mani Lakshminarayanan and Venkat Sethuraman reported on web-based training. There have been 8 to 10 webinars per year for the past 2 years. Recently, webinars have been organized in teams of 2; and that's worked well. The webinar program is sponsored by the Biopharmaceutical Section.
- Steve Snapinn, Editor, gave information about the ASA-sponsored journal, Statistics in Biopharmaceutical Research. It is an electronic-only journal, though a special hardcopy issue was recently commissioned by UNC to honor Gary Koch. There will be future special issues on biomarkers, non-clinical statistics, and an issue honoring Bob O'Neill's tenure at the FDA.
- Yongming Qu, Jingli Song, Jerry Wang, Poster Competition Committee, reported the winners of this year's contest. The following first, second and third place lead author winners are:
    1. Martin O. Carlsson: A Comparison of Methods for Adjusting for the Baseline Measure.
    2. Kelly H. Zou: Cross-Sectional and Longitudinal Joint Modeling of Repeated Measures of Quasi-Continuous Patient-Reported Outcome and Binary Response Data.
    3. Yufan Zhao: Reinforcement Learning Strategies for Lung Cancer Clinical Trials.
- Joan Buenconsejo and Brenda Crowe reported that planning for the 2011 FDA/Industry Biopharmaceutical Workshop is going well. It will be September 19-21 at the Marriott Wardman Park, Washington, DC.

- John Johnson, Qi Jiang, Veronica Taylor, Contributed Paper Award Committee, reported last year's winners, which were announced in the Spring 2011 edition of the *Biopharm Report*.
- Christie Clark announced the winner of the Student Paper Competition. The winner is David Vock: Mixed Model Analysis of Censored Longitudinal Data with Flexible Random Effects Density.

Biopharmaceutical Section Poster Awards at 2012 Joint Statistical Meeting

If you plan to attend the 2012 JSM and plan to present a poster, you may consider participating in the Poster Competition sponsored by the ASA Biopharmaceutical Section. All authors who present posters sponsored by the Biopharmaceutical Section are qualified to compete for this award. The entry criteria for the Poster Awards are:

- Topics in statistics which are applicable to biopharmaceutical research. Suitable topics include but are not limited to methodological issues in preclinical or clinical trials, epidemiology studies of drug safety (device or biological), genetic studies predicting drug (or biological) response, laboratory and toxicological data analyses, methods for high-dimensional data from high-throughput screening, and non-linear pharmacokinetic modeling.
- All JSM attendees (not restricted to members of the Biopharmaceutical Section) can participate in the competition.
- Posters will be evaluated based on the following criteria:
  - —Innovation
  - —General applicability in pharmaceutical research
  - —Appropriate example(s)
  - —Effectiveness of presentation (well written, well organized, etc)
- Authors who compete for the Poster Awards cannot also compete for the Students Paper Awards.

Three awards with cash prizes of $1000, $600 and $400 will be given for 1st, 2nd and 3rd place, respectively.

The process is as follows:

1. Submit an abstract through the Biopharmaceutical Section by the JSM abstract submission deadline.
2. Submit your poster to Jerry Wang, Chair for the Poster Awards through email (*junyuan.wang@bms.com*) by May 1, 2012.
3. Each poster will be reviewed by two reviewers and an average score will be assigned.
4. Posters with the highest scores will be the winners.
5. Ribbons will be put on the corner of the posters to indicate the winners during the poster presentation at JSM.
6. Winners will be announced with certificates at the Biopharmaceutical Section Mixer at the 2012 JSM.

Modeling Opportunities for Statisticians Supporting Quality by Design Efforts for Pharmaceutical Development and Manufacturing

John J. Peterson (GlaxoSmithKline Pharmaceuticals) and Ron S. Kenett (The KPA Group)

Introduction

Variability and randomness in product quality and process performance present the pharmaceutical industry with day to day challenges. A key component to the design and quality improvement of products and processes is understanding both deterministic and stochastic (random) variation (Kenett and Kenett, 2008). This broad perspective holds for the pharmaceutical industry as well as any other industries and services such as automotive, electronics, healthcare, banking, etc. that have made substantial improvements by exploiting knowledge of variation. Historical examples involve applications of classical methodologies such as design of experiments (DoE) and statistical process control (SPC). DoE is used to separate (deterministic) signal from noise (uncontrolled variability) by using models to analyze data collected in well laid out experimental runs (Kenett and Zacks, 1998). Purposeful deterministic variation introduced in experimental arrays has been used to find optimal mean responses and to determine how different factors act together to affect quality attributes. SPC is used to quantify stochastic variation of a process over time, to detect process drift, and "special" causes of variation, both deterministic and stochastic. However, the clear success of these tools for process improvement generates new questions such as "Can more be done with exploiting variation understanding for process improvement in pharmaceutical development and manufacturing?". We believe the answer is "Yes", particularly if we broaden our viewpoint to include stochastic modeling and an improved understanding of the interplay of stochastic distributions that propagate through a process.

In this review paper, we present two opportunities for stochastic process modeling that can be used to improve process understanding and optimization in the context of Quality by Design initiatives, and beyond. One opportunity we discuss involves multivariate predictive regression modeling of laboratory and manufacturing processes (Peterson, 2007). Specific applications may include those to multiple-response-surface optimization, product risk assessment, ICH Q8 design space development, and assessment of assay ruggedness & system suitability. The other opportunity involves Bayesian Network modeling of cause and effect relationships between process and product variables (Ben Gal, 2007). Here, we give a specific example involving bioreactor optimization. These tools exploit concepts of stochastic distributions and the laws of probability to help us make better decisions about the complex systems we encounter in modern pharmaceutical development and manufacturing.

Predictive Distribution Regression Modeling

Most processes, including those in pharmaceutical development and manufacturing, are inherently stochastic processes (see for example Biwer et al., 2005). This is basically the case whether or not they are batch or continuous processes. By this we mean that even if measurement error could be completely removed, these processes would still exhibit some sort of common cause variation (e.g. from batch to batch or as the process unfolds as in dissolution or mixing). As such, quantitative descriptions of such processes cannot be completely described by mean profiles or other deterministic functions. The best way to describe such processes is by a probability distribution that changes according to various process control conditions and perhaps over time.

The noted statistician and quality guru, W. Edwards Deming, states that quality improvement involves reduction of variation about a target value (Wayne, *http://www.q-skills.com/Deming6sigma.htm*). For processes with multiple responses we can generalize this to a moving and shrinking of a probability distribution of quality responses towards a multivariate target. This will increase the likelihood that the multiple quality specifications will be jointly met.

One way to quantify the reliability or level of assurance of meeting multiple quality specifications is to construct a predictive distribution for the process at hand. Such a predictive distribution based on a model over a region of process control conditions, and possibly over time, can be used to quantify the probability of meeting quality specifications (Peterson, 2004).

A key concept in pharmaceutical manufacturing quality assessment is the concept "design space" as defined in the ICH Q8 regulatory guidance (ICH, 2009). ICH Q8 defines "design space" as "The multidimensional combination and interaction of input variables (e.g.,material attributes) and process parameters that have been demonstrated to provide assurance of quality." Translating the ICH Q8 design space definition into a quantitative region with associated "assurance of quality" is straightforward if one uses predictive distributions for a process. Peterson (2008) and Peterson et al. (2009) propose a Bayesian probabilistic definition as follows. A quantitative definition for design space is the set of process controllable factors, DS, such that

$$\mathrm{DS} = x \in E : \Pr Y \in S \,|\, x, \mathrm{data} \geq R \,, \quad (1)$$

where x is a vector of process control factors, E is the experimental region, Y is a vector of quality responses, and S is a specification region corresponding to the quality responses in Y. Here, Pr is a probability measure based upon a posterior predictive distribution and R is a prespecified reliability level for the design space to provide "assurance of quality", based upon the predictive model embedded in the distribution of Y as a function of x defined on E (conditional on the experimental data used to build the model). As an example, Figure 1 is a contour plot of the Design Space for an early phase synthetic chemistry example presented in Stockdale and Cheng (2009), The design space approach in (1) above was used in the biopharmaceutical industry's A-Mab case study (available at: *www.ispe.org/PQLI_A_Mab_Case_Study_Version_2_1.pdf*). Preliminary attempts to quantify ICH Q8 design space involved overlapping mean response surfaces, which provide poor assurance for quality (Peterson and Lief, 2010).

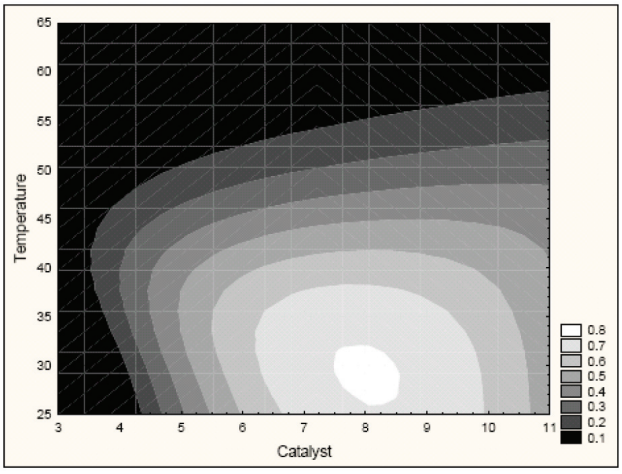


Figure 1. Contour plot of the Design Space for an early phase synthetic chemistry example presented in Stockdale and Cheng (2009) (Used with permission from the Association for Quantitative Management)

The probability of conformance can also be used as an optimization strategy for multiple response processes (Peterson, 2004, 2007, Peterson et al., 2009) and as a method for processes ruggedness and system suitability assessment (Peterson and Yahyah, 2009). This approach has been applied to several GlaxoSmithKline processes involving the development of important company assets (e.g. Castagnoli et al. 2010). In addition, this probability of conformance approach easily adapts to the presence of noise variables, thereby allowing for robust parameter design optimization. See for example Miró-Quesada et al. (2004) and del Castillo (2007, pp342-346).

Additional opportunities for process optimization and design space development involve use of information about raw material properties. Use of information about raw material properties can be used to develop a "dynamic" design space that depends upon the predictive information in the raw materials. Different raw material properties then give rise to slightly different design space regions for the process controllable factors. MacGregor and Bruwer (2008) and Polizzi and García-Muñoz (2011) present a strategy for the utilization of raw material properties and their relation to design space. MacGregor and Bruwer (2008) in particular recognize the important issue of process capability (to provide assurance for the design space) but do not provide a clear, quantitative formulation that one can use to develop a dynamic design space involving multiple quality responses. However, using a Bayesian predictive distribution approach would help with regard to assessing what configuration of control factor levels and raw material properties would be likely to meet quality specifications with a satisfactory degree of assurance. The quantitative key here is to develop good predictive models that incorporate both controllable process factors as well as raw material properties and possibly other ambient influences (such as process noise variables). Such predictive models will likely involve latent variables because (raw material) principal property scores will often be high dimensional. Here, we begin to get to the cutting edge of Bayesian modeling knowledge, but some recent work has been done. See for example Chen et al. (2009).

A practical point worth mentioning is that a Bayesian predictive distribution, obtained from a stochastic model of a process, can be used to estimate how much additional data may be needed in order to reduce uncertainty about unknown model parameters, and thereby obtain more accurate predictions. Furthermore, one may be able to estimate how much process variation may need to be reduced in order to adequately increase the probability that future process quality responses will meet specification. See Peterson (2004) for a discussion within the context of a simple multivariate linear model. This issue may become important in that, as statisticians, we are often asked to assess how much data is needed to properly quantify the risks associated with a pharmaceutical process. The Bayesian approach, using a weakly informative prior, can start with a modest amount of data and then determine how much more is needed. If much more data is needed, the predictive distribution will be rather spread out, reflecting the need for more information on the model parameters and/or possibly the need to reduce the process variability. Bayesian design of experiments can also be a helpful methodology (Lunney et al. 2008).

Pharmaceutical scientists and chemical engineers are now starting to see the utility of using a Bayesian predictive approach to quality improvement and design space development. DynoChem's Design Space and QbD blog have posted on the utility of Bayesian methods for design space (see for example: *http://designspace-qbd.blogspot.com/2009_01_01_archive.html*). See Castagnoli et al. (2010) for a Bayesian application to process robustness. Stamatis (2011), on behalf of the National Institute for Pharmaceutical Technology and Education (NIPTE), has applied Bayesian predictive distributions to mechanistic shelf-life models to assess the risks associated with various manufacturing conditions. In addition, the upcoming "Comprehensive Quality by Design in Pharmaceutical Development and Manufacture" sessions at the American institute of Chemical Engineers 2011 annual conference this October will have at least three talks on Bayesian applications to pharmaceutical process optimization. Mockus et al. (2011) provide a Bayesian predictive analysis of lyophilization cycle parameters. See Blau et al. (2008) and Hsu et al. (2009) for general overviews of Bayesian analysis for mechanistic models in chemical engineering. As such, nonclinical statisticians who support primary and secondary pharmaceutical product development will probably need to acquire the tools and skills necessary for Bayesian modeling. Or, at the very least, be able to assess process risk using statistical procedures and predictive distributions that take into

account the uncertainty of unknown model parameters (Bayesian procedures can do this). In some cases, a simple parametric bootstrap approach may be helpful as a start (Peterson, 2009, available at: *http://www.pharmaqbd.com/qbd_classics_what_your_design_space_needs*)

As statisticians, it is important to help our clients better understand the importance of modeling sources of variation for their processes. This is particularly important for pharmaceutical process optimization and design space development where batch-to-batch variation and other sources of common cause variability can be considerable. (See for example: *http://www.pharmamanufacturing.com/articles/2011/074.html*). Bayesian methods can also be very useful for statistical inference about variance components (Wolfinger and Kass, 2000).

There are several process modeling opportunities for pharmaceutical product development where statisticians can help in the development of predictive distributions for process optimization, risk assessment, and in some cases ICH Q8 design space formulation. A good place to begin familiarizing oneself with Bayesian tools is with univariate linear models associated with experimental designs for process factor screening (Allen and Rajagopalan, 2011) and optimization (del Castillo, 2007). Some good basic books on Bayesian statistics are: Congdon (2006), Ntzoufras (2009), Krusche (2010), and Christensen et al. (2011). See also the recent Wiley Encyclopedia of Statistics in Quality and Reliability (Ruggeri et al., 2007, 2008). A possible next step is to move to multivariate linear models which have very nice applications to multiple response process optimization (Peterson, 2004, Peterson et al. , 2009) and assay ruggedness and system suitability assessment (Peterson and Yahyah, 2009). WinBUGS software can be helpful for producing predictive distributions for nonlinear regression (mechanistic) models that occur in active pharmaceutical ingredient (API) modeling and sometimes in stability or dissolution modeling (LeBlond et al. 2011). See Appendix A for some Bayesian programs for building for predictive regression models.

Two very promising areas that require much attention by statisticians and engineers involve biopharmaceutical growth curve modeling and predictive models for API synthesis that involve raw material properties. Biopharmaceutical growth curve predictions may benefit greatly from Bayesian dynamic models (Gamerman and Lopes, 2006, pp63-68, 172-176), while as stated above, latent variable modeling will be needed for modeling the effects of raw material properties on API quality responses. More complex systems, involving multiple interconnected stages, may be able to benefit from Bayesian Network methodology. This is presented next.

Bayesian Network Modeling

Bayesian Networks (BN) implement a graphical model structure known as a *directed acyclic graph* (DAG) that is popular in Statistics, Machine Learning and Artificial Intelligence. BN are both mathematically rigorous and intuitively understandable. They enable an effective representation and computation of the joint probability distribution over a set of random variables (Pearl, 2000). The structure of a DAG is defined by two sets: the set of nodes and the set of directed edges. The nodes represent random variables and are drawn as circles labeled by the variables names. The edges represent direct dependencies among the variables and are represented by arrows between nodes. In particular, an edge from node X_i to node X_j represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i. Node X_i is then referred to as a 'parent' of X_j and, similarly, X_j is referred to as the 'child' of X_i. An extension of these genealogical terms is often used to define the sets of 'descendants', the set of nodes from which the node can be reached on a direct path. The structure of the acyclic graph guarantees that there is no node that can be its own ancestor or its own descendent. Such a condition is of vital importance to the factorization of the joint probability of a collection of nodes. Although the arrows represent direct causal connection between the variables, the *reasoning process* can operate on a BN by propagating information in any direction. A BN reflects a simple conditional independence statement, namely that each variable is independent of its non-descendants in the graph given the state of its parents. This property is used to reduce, sometimes significantly, the number of parameters that are required to character-

ize the joint probability distribution (JPD) of the variables. This reduction provides an efficient way to compute the posterior probabilities given the evidence present in the data (Lauritzen et al, 1988, Pearl, 2000, Jensen, 2001). In addition to the DAG structure, which is often considered as the "qualitative" part of the model, one needs to specify the "quantitative" parameters of the model. These parameters are described by applying the Markov property, where the conditional probability distribution (CPD) at each node depends only on its parents. For discrete random variables, this conditional probability is often represented by a table, listing the local probability that a child node takes on each of the feasible values – for each combination of values of its parents. The joint distribution of a collection of variables can be determined uniquely by these local conditional probability tables (CPT).

As an example consider 4 bioreactors operating in parallel, over up to 21 days. Several amino acids in the medium composition are tracked periodically. These include: Taurine, Aspartic acid, Hydroxyproline, Threonine, Serine, Asparagine, Glutamic acid, Glutamine Proline, Glycine, Alanine, Valine, Cystine, Methionine, Isoleucine, Leucine, Tyrosine, Phenylalanine, Ornithine, Lysine, Histidine and Arginine. The control parameters include: IGF and levels of two control factors, A and B. The target variables consist of: Volumetric productivity, Ps, Titer, Max Cell and Diamid%.

A bioreactor monitored by n variables produces responses that can be considered random variables, $X_1, \ldots, X_n$. Some of these variables, say q of them, are considered target variables. As mentioned, examples of target variables include: Volumetric productivity, Ps, Titer, Max Cell and Diamid%. Variables, such as the amino acid composition, $X_1, \ldots, X_k$, k = n-q, can be analyzed under the hypotheses that they are positively dependent with target variables. The combinations (X_i, X_j), $X_i \in X_1, \ldots, X_{n-q}$, $X_j \in X_{n-q}+1, \ldots, X_n$ are either positive dependent or independent, for each pair of variable (X_i, X_j), $i \leq n\text{-}q$, $n\text{-}q < j \leq n$. In general, dependency patterns can be extracted from data by using statistical models and data mining techniques (Hand et al. 2001). This section describes how to use Bayesian Networks for mapping such patterns. In constructing a Bayesian Network, several learning algorithms can be implemented to set up the structure of the DAG. Again, one main advantage of the BN is that it allows the combination of structural components derived from expert opinion with components learned from the data. Several software programs that implement algorithms and models for constructing BN are listed in Appendix B.

We return to the bioreactor example to demonstrate the potential in using Bayesian Networks to analyse such data. Our objective is to generate insights on the behavior of the bioreactors for improved operation and monitoring. If we better understand how the control factors affect the target response variables we will know how to optimize the process and generate early warning signals during production for mid-course corrections.

Figure 2 presents a Bayesian Network of the bioreactor data produced with the GeNie software. Each node represents a discretized variable. Some are naturally discrete such as the bioreactor number or the day of operation. The variable "Days" has further aggregated the day of operation in 4 stages. Stage I consists of the first 5 days, Stage IV the final 5 days of operation.

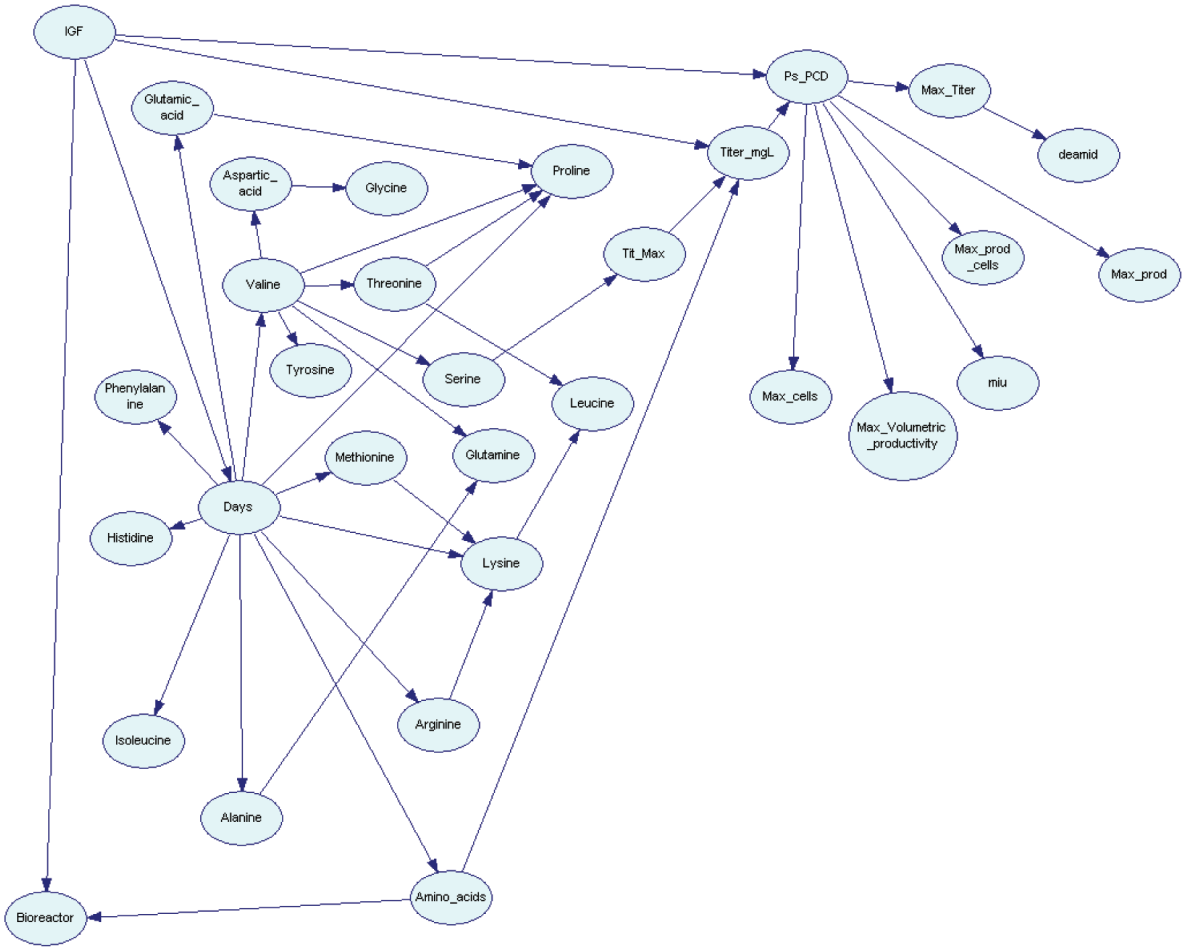


Figure 2. Bayesian Network of bioreactor variables

The network has been automatically learned from the data, without any outside intervention. In learning a Bayesian Network several algorithms can be implemented. The *bnlearn* R package implements five constraint based learning algorithms (see Appendix B). The GeNie package generated Figure 3 using the Greedy Thick Thinning algorithm. In learning the network one can include white lists of forced causality links imposed by expert opinion and black lists of links that are not to be included in the network, again using inputs from content experts. One can see that the variable "Days" is affecting the composition of many of the amino acids and that IGF and the amino acid used as control parameter characterize the bioreactor number. In this case a full factorial experiment was conducted with different combinations of IGF levels and control amino acid as factors, so that the links from IGF and amino acid to bioreactor number reflect the experimental design set up. In Figure 3 we condition the network on the first and last stage of operation, and show the distribution of the composition of the various amino acids. The discretized values are presented as ordinal categories with blue and purple standing, respectively, for the lowest and highest categories.

As an example, on the left panel of Figure 3 we can see that, according to the BN model, at Stage I the highest compositions of Isoleucine, Alanine and Arginine correspond to 63%, 12% and 62% respectively. As we move to Stage IV (right panel), these numbers become, respectively, 13%, 25%, 12% with a dramatic drop in high values of Isoleucine and Arginine and an increase of 100% in the high values of Alanine. This demonstrates how conditioning the network on the bioreactor stage is demonstrating the strength of the model as a predictive tool. Conversely, by conditioning the network on

target variables or end results such as the highest maxim volumetric productivity or cell production can help diagnose the conditions which produced these record numbers. In other words, Bayesian Networks can be used to predict an outcome when a process is set at a certain set of parameters or diagnose what can cause a result we want to prevent or duplicate.

The use of Bayesian and multivariate methods, in the context of Quality by Design biopharmaceutical initiatives, is described in Kenett and Kenett (2008). Sensitivity analysis for determining robustness of the network structure have been proposed in Cornalba et al (2007). The main disadvantages of GeNie and *bnlearn* are that they do not allow the mixing of continuous and categorical variables. Some existing libraries handle networks with mixed variables; however their learning procedure is still experimental and hardly applicable to complex models and large datasets (Bottcher and Dethlefsen, 2003). For more on BN in general, and in the context of operational risks and health care, see Ben Gal (2007), Kenett (2007), Kenett and Raanan (2010) and Kenett (2012). An introductory book on Bayesian Networks is by Koski and Noble (2009).

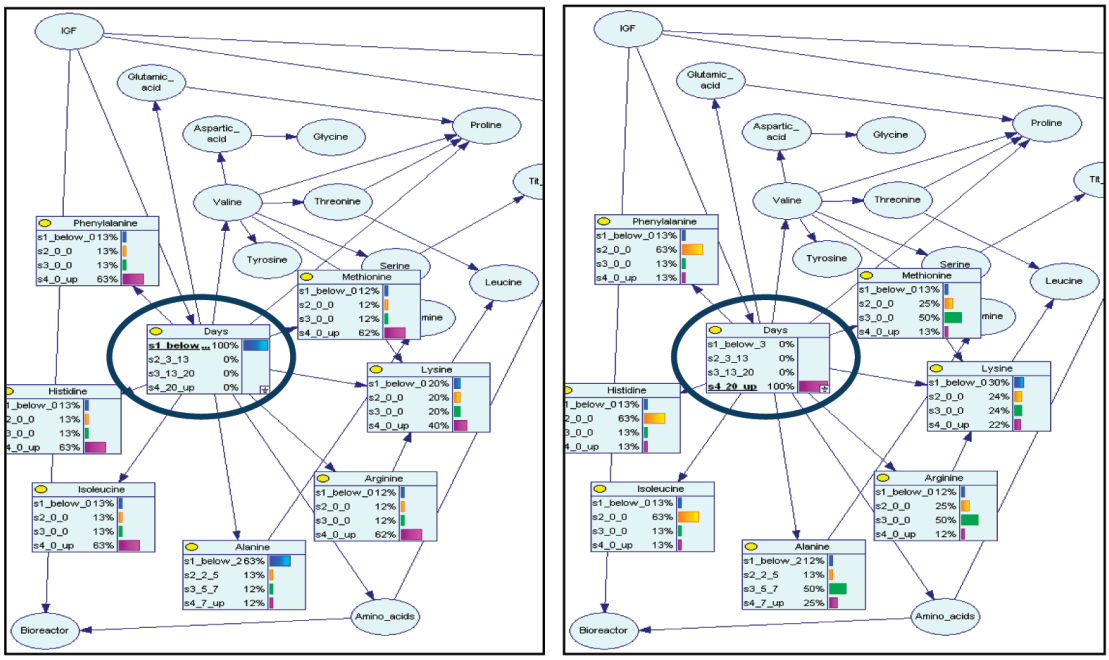


Figure 3. Bayesian Network conditioned on operation day (left: Stage I, right: Stage IV)

Discussion

Using the laws of probability and Monte Carlo simulations, statisticians can help their scientific colleagues to better understand the uncertainties and risks involved in making decisions about increasingly complex pharmaceutical processes. A successful stochastic predictive model provides evidence that the drug sponsor understands the process with regard to how process risk will vary under different manufacturing conditions. A Monte Carlo predictive distribution approach appears to have a promising future with regard to quantitative decision making. See for example the book by Savage (2010). Of course, not all risks can be quantitatively modeled (see for example Taleb, 2010, Kenett and Tapiero, 2009, and Kenett and Raanan, 2010), but for many processes deeper insight can be obtained with careful stochastic modeling. This review was designed to provide an introduction and a perspective of how such modeling is gradually impacting modern pharmaceutical research and development.

References

Allen, T.T. and Rajagopalan, R. (2011). A Bayesian Plotting Method for Fractional Factorial Data Analysis, *Journal of Quality Technology*, **43**, 224-236.

Ben Gal, I. (2007). Bayesian Networks, in *Encyclopaedia of Statistics in Quality and Reliability*, Ruggeri, F., Kenett, R. S. and Faltin, F. (editors in chief), Wiley, UK.

Biwer, A., Griffith, S., and Cooney, C. (2005). Uncertainty Analysis of Penicillin V Production Using Monte Carlo Simulation, *Biotechnology and Bioengineering*, 90, 167-179.

Blau, G., Lasinski, M., Orcun, S., Hsu, H-S, Caruthers, J., Delgass, N., Venkatasubramanian, V. (2008). High fidelity mathematical model building with experimental data: A Bayesian approach, *Computers and Chemical Engineering*, **32**, 971-989.

Boettcher SG, Dethlefsen C (2003). *deal:* A Package for Learning Bayesian Networks, *Journal of Statistical Software*, 8(20), 1-40. *http://www.jstatsoft.org/v08/i20*.

Castagnoli, C., Yahyah, M., Cimarosti, Z., and Peterson, J. J. (2010). Application of Quality by Design Principles for the Definition of a Robust Crystallization Process for Casopitant Mesylate, *Organic Process Research and Development*, **14**(6), 1407–1419.

Chen, H, Bakshi, B.R., and Goel, P.K. (2009). Bayesian latent variable regression via Gibbs Sampling: methodology and practical aspects, *Journal of Chemometrics*, 21, 578–591.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, Chapman & Hall, CRC, Boca Raton, FL.

Congdon, P. (2006). *Bayesian Statistical Modeling* (2nd ed), John Wiley & Sons, Inc., Hoboken, NJ.

Cornalba, C., Kenett, R.S. and Giudici, P. (2007) Sensitivity Analysis of Bayesian Networks with Stochastic Emulators, ENBIS-DEINDE proceedings, University of Torino, Turin, Italy, *http://web.econ.unito.it/deinde07*.

del Castillo, E. (2007), *Process Optimization: A Statistical Approach*, Springer-Verlag, New York, NY.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, (2nd ed), Chapman & Hall/CRC, Baca Raton, FL.

Hand, D., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining (Adaptive Computation and Machine Learning)*, The MIT Press.

Hsu, S-H, Stamatis, S.D., Caruthers, J. M., Delgass, W.N., Venkatasubramanian, V., Blau, G, and Lasinski, M. and Orcun, S. (2009). Bayesian Framework for Building Kinetic Models of Catalytic Systems, *Industrial Engineering and Chemistry Research*, **48**, 4768–4790

ICH (2009). ICH Harmonized Tripartite Guideline: Pharmaceutical Development, Q8 (R2).

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*, Springer.

Kenett, R.S. and Zacks, S. (1998). Modern Industrial Statistics : *Design and Control of Quality and Reliability* Duxbury press, 1998, Spanish edition, 2002, 2nd paperback edition, 2002, Chinese edition, 2004.

Kenett, R.S. (2007). Cause and Effect Diagrams, in *Encyclopedia of Statistics in Quality and Reliability*, Ruggeri, F., Kenett, R.S. and Faltin, F. (editors in chief), Wiley, Online version, Wiley InterScience, UK.

Kenett, R.S. and Kenett, D.A. (2008). Quality by Design Applications in Biosimilar Technological Products, ACQUAL, *Accreditation and Quality Assurance*, Springer, Vol. 13, No 12, pp. 681-690.

Kenett, R.S. and Tapiero, C.S. (2009). Quality, Risk and the Taleb Quadrants, *Risk and Decision Analysis*, 1(4), 231-246.

Kenett, R.S. and Raanan, Y. (2010). *Operational Risk Management: a practical approach to intelligent data analysis*, John Wiley and Sons.

Kenett, R.S. (2012). Risk Analysis in Drug Manufacturing and Healthcare, in *Statistical Methods in Healthcare*, Faltin, F., Kenett, R.S. and Ruggeri, F. (editors in chief), John Wiley and Sons, 2012.

Lauritzen, S.L., Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, Series B (Methodological), 50(2):157-224.

Koski, T. and Noble, J. (2009) *Bayesian Networks – An Introduction*, John Wiley & Sons, Ltd, West Sussex, UK.

Krusche, J. K. (2010), *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*, Academic Press, Oxford, UK.

LeBlond, D., Peterson, J.J., Altan, S. (2011). Posterior Probability of Dissolution Equivalence, presented at the *Midwest Biopharmaceutical Statistics Workshop*, May 25th, Muncie, IN.

Lunney, P.D., Cogdill, R.P. and Drennen, J.K. (2008). Innovation in Pharmaceutical Experimentation Part I: Review of Experimental Designs Used in Industrial Pharmaceutics Research and Introduction to Bayesian D-Optimal Experimental Design, *Journal of Pharmaceutical Innovation*, **3**, 188-203.

MacGregor, J.F. and Bruwer, M-J. (2008). A Framework for the Development of Design and Control Spaces, *Journal of Pharmaceutical Innovation*, **3**, 15-22.

Miro-Quesada, G., del Castillo, E., and Peterson, J.J. (2004). A Bayesian Approach for Multiple Response Surface Optimization in the Presence of Noise Variables, *Journal of Applied Statistics*, **31**, 251-270

Mockus, L., LeBlond, D. Basu, P. K., Shah, R. B., and Khan, M. A. (2011). A QbD Case Study: Bayesian Prediction of Lyophilization Cycle Parameters, *AAPS PharmSciTech*, 12(1), 442-448.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*, John Wiley and Sons, Inc. Hoboken, NJ.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, UK.

Peterson, J.J. (2004). A Posterior Predictive Approach to Multiple Response Surface Optimization, *Journal of Quality Technology*, **36**, 139-153.

Peterson, J.J. (2007). A Review of Bayesian Reliability Approaches to Multiple Response Surface Optimization, in *Bayesian Process Monitoring, Control, and Optimization*, (Chapter 10), (Colosimo, B. M. and del Castillo, E., Eds), Chapman & Hall/CRC, Boca Raton, FL.

Peterson, J.J. (2008). A Bayesian Approach to the ICH Q8 Definition of Design Space, *Journal of Biopharmaceutical Statistics*, **18**, 959-975.

Peterson, J.J., Miro-Quesada, G., del Castillo, E. (2009). A Bayesian Reliability Approach to Multiple Response Optimization with Seemingly Unrelated Regression Models, *Journal of Quality Technology and Quantitative Management* **6**(4), 353-369.

Peterson, J.J. and Yahyah, M. (2009). A Bayesian Design Space Approach to Robustness and System Suitability for Pharmaceutical Assays and Other Processes, *Statistics in Biopharmaceutical Research* **1**(4), 441-449.

Peterson, J.J. (2009). What Your ICH Q8 Design Space Needs: A Multivariate Predictive Distribution, *Pharmaceutical Manufacturing*, **8**(10) 23-28.

Peterson, J.J. Snee, R.D., McAllister, P.R., Schofield, T.L., and Carella, A.J. (2009). Statistics in the Pharmaceutical Development and Manufacturing (with discussion), *Journal of Quality Technology*, **41**, 111-147.

Peterson, J.J. and Lief, K. (2010). The ICH Q8 Definition of Design Space: A Comparison of the Overlapping Means and the Bayesian Predictive Approaches, *Statistics in Biopharmaceutical Research*, **2**, 249-259.

Polizzi, M.A. and García-Muñoz, S. (2011). A framework for in-silico formulation design using multivariate latent variable regression methods, *International Journal of Pharmaceutics*, doi:10.1016/j.ijpharm.2011.04.064

Ruggeri, F., Kenett, R.S. and Faltin, F. (2007), *Encyclopedia of Statistics in Quality and Reliability*, John Wiley and Sons, Chichester, UK.,

Ruggeri, F., Kenett, R.S. and Faltin, F. (2008), *Encyclopedia of Statistics in Quality and Reliability* Online version, *http://mrw.interscience.wiley.com/emrw/9780470061572/home*, Wiley InterScience, UK.

Stamatis, S.D. (2011). NIPTE-FDA Collaborative Case Study On Model-based Design Space Development Across Scales & with Stability Considerations, *NIPTE Research Conference: Critical Path Research for Pharmaceutical Process Scale-Up and Stability*, June 15-16, 2011, Rockville, MD

Savage, S. (2010). *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*, John Wiley & Sons, Inc., Hoboken, NJ.

Stockdale, G. and Cheng, A. (2009). Finding Design Space and a Reliable Operating Region using a Multivariate Bayesian Approach with Experimental Design, *Quality Technology and Quantitative Management* **6**(4), 391-408.

Taleb, N.N. (2010). *The Black Swan-The Impact of the Highly Improbable*, (2nd ed) Random House, N.Y., N.Y.

Wolfinger, R.D. and Kass, R.E. (2000). Noncongugate Bayesian Analysis of Variance Component Models, *Biometrics*, **56**, pp786-774.

Appendix A: Bayesian Programs for Building for Predictive Models

WinBUGS: A very flexible Bayesian software package that can be used to model linear or nonlinear, univariate or multivariate, regression models, with mixed effects if desired. WinBUGS can be called from R via the R2WinBUGS package. A nice book on this software tool is by Ntzoufras (2009). The WinBUGS web site is *http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml*.

bayesm: An easy-to-use R package that contains a variety of Bayesian analysis functions. In particular, the rsurGibbs function computes samples from the posterior distribution of multivariate linear models with possibly different functional forms for each response-type, i.e. the seemingly unrelated regressions (SUR) model. The bayesm package is available at the CRAN R web site: *http://cran.r-project.org/web/packages/#available-packages-M*.

MCMCglmm: an R package for Bayesian univariate or multivariate response generalized linear mixed models. Mixed models are useful for modeling batch effects or split-plot designs. The MCMCglmm package is also available at the R CRAN web site.

Appendix B: Bayesian Network Programs

GeNIe (Graphical Network Interface) is the graphical interface to SMILE (Structural Modelling, Inference, and Learning Engine), a fully portable Bayesian inference engine developed by the Decision Systems Laboratory and thoroughly field tested since 1998. GeNIe can be freely downloaded from *http://genie.sis.pitt.edu* with a user guide and related documentation.

Hugin (*http://www.hugin.com/index.php*) is a commercial software which provides a variety of products for both research and non-academic use. The Hugin GUI (Graphical User Interface) allows building BN, learning diagrams, etc.

IBM SPSS Modeller (*http://www.spss.com*) includes several tools which enable the user to deal with a list of features and statistical methods such as BN. IBM SPSS is not free software.

The R *bnlearn* package is powerful and free. Compared with other available BN software programs, it is able to perform both constrained-based and score-based methods. It implements five constraint based learning algorithms (Grow-Shrink, Incremental Association, Fast Incremental Association, Interleaved Incremental association, Max-min Parents and Children), two scored based learning algorithms (Hill-Climbing, TABU) and two hybrid algorithms (MMHC, Phase Restricted Maximization).

The main disadvantage in most available BN programs is that they do not allow the mixing of continuous and categorical variables. Some experimental libraries handle networks with mixed variables; however their learning procedures are not yet applicable to complex models and large datasets.

Use Of Zero-Inflated Mixture Models to Compare Antibody Titers in Response to H1N1 Vaccination

Daniel J. Zaccaro[1], M.S., Leela M. Aertker[1], M.P.H.
[1]Rho Inc, Chapel Hill, NC

Abstract

Pandemic H1N1 vaccine was administered to participants with mild/moderate and severe asthma to investigate quantitative and qualitative differences in immunogenicity among subgroups. H1N1 antibody titers were measured pre-vaccination (Day 1) and post-vaccination (Days 8, 21, 28 and 41). A second vaccination of the same dose (15 or 30 mcg) was administered after blood samples were taken on Day 21. H1N1 antibodies at Day 1 and three weeks post-vaccination (Day 21) were of primary interest for the current article. A preponderance of titers below the lower detection limit was observed (36% - 75% of observations at Day 1, depending on subgroup). Titers above the upper detection limit (6 - 52% at Day 21) were also observed. Because of this preponderance of censored values, assumptions of Gaussian data are not appropriate, and traditional modeling approaches could produce biased estimates of differences in immunogenicity. Zero-inflated log-normal models that accounted for left- and right-censoring and a "point mass" below the lower limit of detection were utilized to compare subgroups with respect to antibody titers. Results derived from traditional analytical methods such as imputation of censored values were compared to results from zero-inflated methods. By formal criteria, zero-inflated models provided a better fit to data and yielded results that were qualitatively and quantitatively different from traditional models. Zero-inflated models yielded geometric mean titers that were 2- to 3-fold different from traditional models and elucidated differences among subgroups that were obscured by traditional methods.

Introduction

Investigations of immunological processes often produce a preponderance of measurements which are below a lower limit of detection (LLOD). Observations below the LLOD may consist of values which are truly zero, values which are non-zero but still below the LLOD (left-censored), or a mixture of both. Examples include antibody levels in response to vaccine,[1] levels of HIV mRNA,[2] and serum-specific IgE antibodies.[3] Given that distributions with a preponderance of censored values are typically not Gaussian, traditional solutions such as imputation of censored values (e.g., LLOD/2), may produce biased estimates of the parameters of interest.[1] Moreover, the presence of a "point mass" of individuals with undetectable measurements, suggesting more observations below the LLOD than accounted for by left-censoring alone, may suggest an important biological phenomenon that would be overlooked with traditional imputation methods. For example, a point mass of individuals with undetectable antibody levels may suggest a subgroup with no prior exposure to relevant antigens or an inability to produce antibodies due to presence of disease which impacts vaccine immunogenicity. To address this, we illustrate the zero-inflated modeling approach described by Moulton and Halsey[1] and others.[4,5,6] The zero-inflated model consists of two components: 1) a binary (Bernoulli) component which assigns each individual to either a point mass or a log-normal distribution with probability p, depending on the individual's covariates; and 2) conditional on an individual's assignment to the log-normal distribution, the model consists of a traditional linear component with normal errors, mean (μ), and variance (σ^2). We consider the following approaches: 1) the naïve model with imputation of left-censored antibody titers, 2) the left-censoring model with no point mass, and 3) the left-censoring model including a point mass of individuals who have antibody titers (potentially) equal to zero. The left-censoring, zero-inflated

modeling approach could be extended to account for right-censored observations, which are above an upper limit of detection (ULOD), but we omit this modeling strategy in the present article to narrow our focus on the utility of left-censoring and inclusion of a point mass. We observed left- and right-censoring and an excess zero phenomenon in the recently completed NIAID/NHLBI-sponsored clinical trial of safety and immunogenicity of the H1N1 vaccine in individuals with asthma, part of the Severe Asthma Research Program (SARP).[7] These phenomena are demonstrated in the following figures. Figure 1 illustrates the distributions of baseline geometric mean titers (GMTs) from the H1N1/SARP trial for participants who indicated no receipt of 2009 seasonal influenza vaccine (A), as well as those who did indicate receipt of the 2009 seasonal influenza vaccine (B). Note the preponderance of values below the LLOD (GMT values <10). Figure 2 illustrates the distinction between left-censoring and the point mass phenomenon with a hypothetical distribution of titers. An excess zero phenomenon is indicated when more observations are left-censored than predicted from a log-normal distribution.

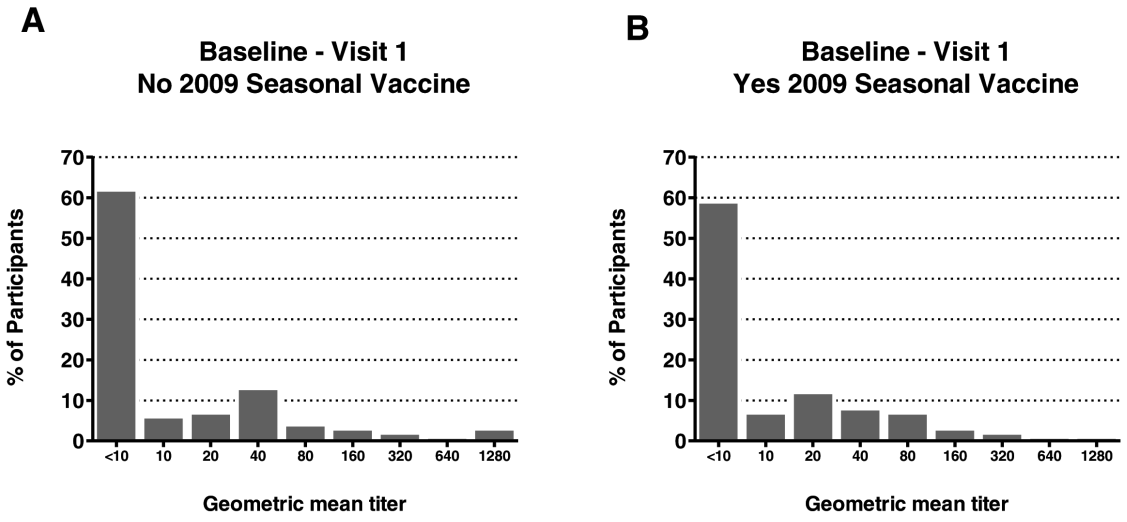


Figure 1. H1N1 antibody geometric mean titers at baseline (pre-vaccination) for participants who indicated no prior receipt of the 2009 seasonal influenza vaccine (A) vs. those who indicated prior receipt of the 2009 seasonal influenza vaccine (B).

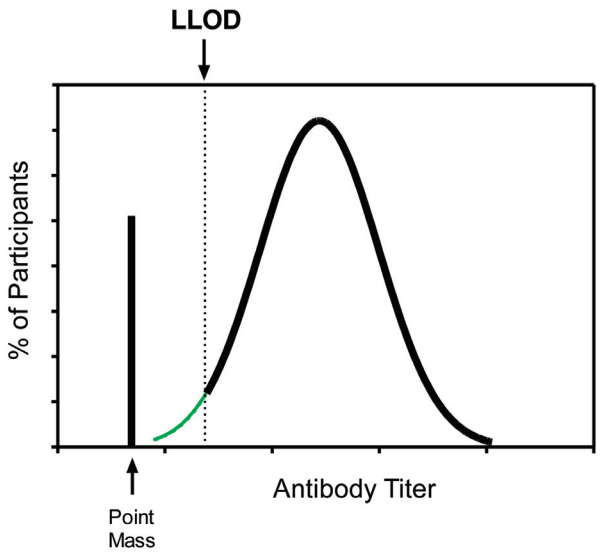


Figure 2. Hypothetical distribution of antibody titers to illustrate a theoretical log-normal distribution of titers with left-censored values (green area of curve), shown here to be a distinct subset from the point mass.

H1N1/SARP Trial

The primary analysis of the H1N1/SARP trial indicated that overall seroprotection levels (defined as a post-vaccination titer ≥1:40) were adequate, but that participants with severe asthma required the higher dose of vaccine (30mcg) as they did not respond adequately to the lower dose level. Exploratory analyses also indicated that recipients of the 2009 seasonal influenza vaccine had lower seroprotection levels than those who did not receive the 2009 seasonal influenza vaccine. In this article, we examine potential differences in H1N1 antibody titers among subgroups of participants while illustrating the utility of the zero-inflated modeling approach. We utilized zero-inflated methods to address two questions: 1) do differences exist among subgroups with respect to the proportion of participants who are in a point mass of observations versus the log-normal distribution (e.g., are subjects ≥60 years of age more likely to have titers in the point mass when compared to subjects ages 12 - 17?); and 2) given the subset of participants with values in the log-normal distribution (and not in the point mass), are there quantitative differences among subgroups (e.g., age groups, receipt of 2009 seasonal influenza vaccine) with respect to expected H1N1 antibody titers?

This article describes the results of an exploratory analysis of the 385 participants vaccinated at least once and who provided blood for at least one post-vaccination titer measure. Participants were divided into two asthma severity groups - mild/moderate and severe. Vaccine dose (15 versus 30 mcg) was randomly assigned to participants stratified by severity group, clinical site, and age group (i.e., 12-17, 18-60, and > 60).[7] Baseline (Day 1) and post-vaccination (Day 21) H1N1 antibody titers were the focus of the current analysis.

Methods

Ordinary least squares (OLS) regression (with imputation of unobserved, left-censored titers) and zero-inflated models were used to describe associations between H1N1 antibody titers and participant characteristics, including vaccine dose, asthma severity group, age, gender, 2009 seasonal influenza vaccination, and the fluticasone-equivalent dose of inhaled corticosteroid (ICS) use. The zero-inflated

model does not rely on imputation of left-censored titer values, but rather includes both a binary component to the likelihood (i.e., is the participant in the point mass or is the participant in the log-normal distribution?), as well as a log-normal component to the likelihood (i.e., given that the participant is in log-normal distribution, what is the expected titer value?). The zero-inflated model fits both of these components simultaneously using all available data. In this article, we illustrate that zero-inflated models produce different estimates of expected antibody titers at baseline and post-vaccination when compared to OLS methods using imputation (i.e., LLOD/2 in place of left-censored titers).

Relative goodness of model fits were evaluated using Akaike information criteria (AIC). SAS PROC NLMIXED was utilized for all computations of model estimates. Geometric mean titer (GMT) values were $\log_{10}$ transformed for all modeling computations. The LLOD (y_L) was therefore $\log_{10}(10) = 1.0$ and the ULOD (y_U) was log10 (1280) = 3.10721. Summary frequencies of the GMT values by subgroup for both observed (i.e., not censored) and unobserved (i.e., left- or right-censored) values are provided in Table 1 for Day 1 and Day 21.

Table 1—Summary of H1N1 antibody geometric mean titers by subgroups at baseline (Day 1) and post-vaccination (Day 21)

	Baseline, Day 1				Post-vaccination, Day 21			
		GMT[1] ≥ 10, < 1280				GMT[1] ≥10, < 1280		
Subgroup (n)	GMT[1] < 10 n (%)	n (%)	GMT[2]	GMT[1] ≥ 1280 n (%)	GMT[1] < 10 n (%)	n (%)	GMT[2]	GMT[1] ≥ 1280 n (%)
Seasonal Vaccine 2009								
Yes (204)	121 (59)	82 (40)	53	1 (0)	11 (5)	133 (65)	193	60 (29)
No (181)	112 (62)	63 (35)	61	6 (3)	5 (3)	107 (59)	242	69 (38)
Asthma Severity, Vaccine Dose								
Mild/Moderate, 15 mcg (107)	70 (65)	35 (33)	63	2 (2)	5 (5)	71 (66)	203	31 (29)
Mild/Moderate, 30 mcg (107)	62 (58)	44 (41)	50	1 (1)	3 (3)	56 (52)	267	48 (45)
Severe, 15 mcg (86)	50 (58)	33 (38)	56	3 (3)	6 (7)	60 (70)	160	20 (23)
Severe, 30 mcg (85)	51 (60)	33 (39)	60	1 (1)	2 (2)	53 (62)	248	30 (35)
Age								
12-20 (75)	27 (36)	42 (56)	98	6 (8)	0 (0)	36 (48)	355	39 (52)
21-39 (95)	61 (64)	34 (36)	52	0 (0)	2 (2)	53 (56)	255	40 (42)
40-59 (152)	98 (64)	53 (35)	49	1 (1)	5 (3)	101 (66)	204	46 (30)
≥60 (63)	47 (75)	16 (25)	27	0 (0)	9 (14)	50 (79)	134	4 (6)
Gender								
Male (156)	84 (54)	68 (44)	61	4 (3)	4 (3)	93 (60)	247	59 (38)
Female (229)	149 (65)	77 (34)	53	3 (1)	12 (5)	147 (64)	194	70 (31)
Fluticasone-Equivalent Doses								
0 (53)	33 (62)	19 (36)	53	1 (2)	0 (0)	31 (58)	239	22 (42)
≤250 (79)	49 (62)	30 (38)	53	0 (0)	5 (6)	51 (65)	215	23 (29)
251-499 (75)	44 (59)	29 (39)	52	2 (3)	4 (5)	49 (65)	238	22 (29)
500-999 (72)	44 (61)	26 (36)	62	2 (3)	2 (3)	37 (51)	200	33 (46)
≥1000 (106)	63 (59)	41 (39)	62	2 (2)	5 (5)	72 (68)	193	29 (27)

[1]GMT = Geometric mean titers are back-transformed from the mean of $\log_{10}$ transformed H1N1 antibody titers, within a participant and visit.

[2]GMT = Summary geometric mean titers are back-transformed from the average of the mean of $\log_{10}$ transformed H1N1 antibody titers, across participants for the indicated visit.

Similar to Moulton and Halsey[1], we consider a Bernoulli (binary outcome) component in the likelihood; i.e., covariates predict whether an observation is in the log-normal component or in the point mass. Once an observation is predicted to reside in the log-normal component, then we consider the potential for left-censored values which are theoretically in the log-normal component, but are not directly observed. For simplicity in our discussion, we provide a general prediction equation for the expected values for both Bernoulli and log-normal components. The Bernoulli component is given by:

$$(1) \qquad Ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i \chi_i$$

where:

p = conditional probability of the observation falling into the log-normal distribution (and not in the point mass) given the independent variables

$x_1, x_2, \ldots, x_k$ = independent variables used to predict whether an observation is in the log-normal distribution and not in the point mass (e.g., age group indicator variables)

$\beta_0, \beta_1, \ldots, \beta_k$ = parameters to be estimated in the Bernoulli component, which are used to predict whether an observation is in the log-normal distribution and not in the point mass

Consequently, $(1 - p)$ denotes the probability of an observation residing in the point mass. Note that the model above could have easily been constructed so that covariates were used to predict the probability of an individual residing in the point mass and not in the log-normal distribution.

Conditional on an observation residing in the log-normal distribution, the expected value (mean) is given by:

$$(2) \qquad \mu = \gamma_0 + \sum_{j=1}^{m} \gamma_j z_j$$

where:

μ = conditional expected value (mean) of observations in the log-normal distribution, given the independent variables

$z_1, z_2, \ldots, z_m$ = independent variables used to predict the expected value (mean) of observations in the log-normal distribution

$\gamma_0, \gamma_1, \ldots, \gamma_m$ = parameters to be estimated, used to predict the expected value (mean) of observations in the log-normal distribution

σ^2 = conditional variance of the log-normal distribution, assumed constant across all levels of independent variables

We considered several models to describe associations between titers and baseline characteristics, but only present three models in this article. Other models assume a combination of left-censoring, right-censoring, and point mass, or all three (see Appendix I). Models considered for this article are as follows:

Traditional Model: Log-normal distribution with imputation (values < LLOD set to LLOD/2) using OLS regression

Left-Censoring Model: Log-normal distribution with left-censoring, but no point mass

Left-Censoring + Point Mass Model: Log-normal distribution with left-censoring and point mass

Assuming equations (1) and (2) above for expected values, likelihood expressions are described below for each model. Expressions are provided for contributions from a single observation to each likelihood. Appendix I contains a complete list of Ln likelihoods for all models, but only results from the above three models are illustrated in this article.

Traditional (OLS) Model with Imputation

The likelihood for an OLS linear regression model for which left-censored values have been assigned values of LLOD/2 is given by:

$$L\left(\gamma_0,\gamma_1,\ldots\gamma_m,\sigma\middle|Z_1,\ldots Z_m\right)=\phi\left(y\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$$

where:

y = $\log_{10}$ (*GMT*)

μ, σ = parameters of a normal distribution (i.e., mean and standard deviation)

γ_0, γ_1, …, γ_m = parameters to be estimated that are used to predict the expected value (mean) of observations in the log-normal distribution, and

$$\phi\left(y\right)=prob\left[GMT=y\middle|\mu,\sigma\right]$$

Left-Censoring Model

The likelihood for a single observation for the model with left-censoring and no point mass is given by:

$$L\left(\gamma_0,\gamma_1,\ldots\gamma_m,\sigma\middle|Z_1,\ldots Z_m,\delta\right)=\left\{\Phi\left[y_L\right]\right\}^{\delta}\left\{\phi\left(y\right)\right\}^{1-\delta}$$

where:

y_L= $\log_{10}$ (LLOD) = $\log_{10}(10)$ = 1

δ= 1 if $y < y_L$; δ= 0 if $y \geq y_L$

$$\Phi\left[y_L\right]=prob\left[GMT<y_L\right]$$

The component of the likelihood expression under the δ indicator allows for unobserved values to be left-censored while still originating from the log-normal distribution. Observed values from the log-normal distribution are included under the 1- δ indicator.

Left-Censoring + Point Mass Model

Finally, the likelihood using a mixture model for a single observation, which includes components for left-censoring as well as a point mass is given by:

$$L\left(\beta_0,\beta_1,\ldots,\beta_k,\gamma_0,\gamma_1,\ldots,\gamma_m,\sigma\middle|X_1,\ldots,X_k,Z_1,\ldots,Z_m,\delta\right)=\left\{\left(1-p\right)+p\Phi\left[y_L\right]\right\}^{\delta}\left\{p\phi\left(y\right)\right\}^{1-\delta}$$

Recall that p is the conditional probability of an observation falling into the log-normal distribution (and not in the point mass) given the independent variables.

This model specification assumes that a subset of individuals with unobserved values are either in a sub-population with antibody titers (possibly) equal to 0 (with probability $1-p$) while others presumably would have been observed if the measurement technique was more precise (with probability $p\Phi[y_L]$).

Results

Inclusion of left censoring and a point mass provided a better fit to the data than the traditional method of imputation using OLS or left-censoring alone (see Table 2). Therefore, we narrow our focus on the comparison between the model with the best fit, the left-censoring model with point mass, and the naive model with imputation.

Table 2—Summary of model-fitting criteria using naïve (imputation of values < LLOD) as well as zero-inflated models at baseline and post-vaccination

	Akaike Information Criteria (AIC) by Model[2]		
Subgroups[1]	Traditional Model with Imputation	Left-Censoring Model	Left-Censoring + Point Mass Model
Baseline, Day 1			
Seasonal Vaccine 2009	758	742	737
Severity/Dose Groups	762	745	745
Age	705	700	695
Gender	753	736	735
Fluticasone-Equivalent Doses	764	747	750
Post-vaccination, Day 21			
Seasonal Vaccine 2009	747	755	709
Severity/Dose Groups	739	746	703
Age	683	691	650
Gender	748	756	710
Fluticasone-Equivalent Doses	752	760	717

[1]Subgroups are included in model as covariates for both Bernoulli and log-normal components.

[2]Lower AIC value indicates a better fit.

Of primary (biological) interest to this article were the differences in H1N1 antibody titers between participants who did/did not receive the 2009 seasonal influenza vaccine. Participants who received the 2009 seasonal influenza vaccine had a baseline GMT half the magnitude of those who did not receive the 2009 seasonal influenza vaccine (21versus 42, respectively; see Table 3), and also exhibited fewer estimated participants in the point mass (38% versus 52%, respectively). These differences were obscured in the traditional model, which indicated virtually identical GMT values at baseline between participants who did/did not receive the 2009 seasonal influenza vaccine (14 versus 15, respectively; see Table 3). We examined Day 21 GMT differences between 2009 seasonal influenza vaccine recipients/non-recipients by adjusting for H1N1 baseline titers using a zero-inflated model. Participants who received the 2009 seasonal influenza vaccine exhibited lower GMT values than those who did not receive the 2009 seasonal influenza vaccine in the log-normal component before ($p=0.0202$) and after ($p=0.0195$) adjustment for baseline titers.

Using both a traditional model with imputation and a zero-inflated model with left-censoring and a point mass, we explored other potential effects of participant subgroups on baseline and Day 21 antibody titers (see Table 3). When compared to other severity/dose groups, the severe asthma 15 mcg group exhibited similar levels of antibody response and percentage of participants in the point mass at baseline, yet exhibited the lowest antibody titers and the highest estimated percentage of participants in the point mass at Day 21 compared to the other asthma severity/dose groups. Participants with severe asthma receiving 15 mcg could have more "true zeros", which may be partly due to an effect of age (i.e., severe asthma participants were older, on average, than mild/moderate participants; $p<0.001$). Using the left-censoring and point mass model, we found that at Day 21, 14% of older participants

(≥ 60yrs) were in the point mass compared to 0 - 3% of younger participants (< 60yrs). Males tended to exhibit higher antibody titers and lower percentages in the point mass compared to females, but differences in GMTs were not affected by the modeling approach. Additionally, there were no differences among fluticasone-equivalent ICS dose groups, either in the point mass estimates or in the GMT estimates.

Table 3—Modeling Results to Describe H1N1 Antibody Titers By Subgroups of Interest at Baseline and Post-Vaccination Using Zero-Inflated Mixture Model With Point Mass and Left-Censoring

Subgroup (n)	Baseline, Day 1			Post-vaccination, Day 21		
	Point Mass (Estimated %)	GMT[1] for log-normal	GMT[1] by naive approach	Point Mass (Estimated %)	GMT[1] for log-normal	GMT[1] by naive approach
Seasonal Vaccine 2009						
Yes (204)	38	21	14	5	345	276
No (181)	52	42	15	3	463	410
Asthma Severity, Vaccine Dose						
Mild/Moderate, 15 mcg (107)	55	36	13	5	353	291
Mild/Moderate, 30 mcg (107)	35	19	14	3	550	483
Severe, 15 mcg (86)	45	35	16	7	266	204
Severe, 30 mcg (85)	45	29	14	2	447	405
Age						
12-20 (75)	32	112	42	0	691	691
21-39 (95)	52	27	12	2	509	463
40-59 (152)	52	26	12	3	362	316
≥60 (63)	21	5	8	14	154	96
Gender						
Male (156)	39	34	17	2	466	417
Female (229)	49	24	12	5	354	286
Fluticasone-Equivalent Doses						
0 (53)	46	25	13	0	480	480
≤250 (79)	40	18	13	6	373	285
251-499 (75)	42	27	14	5	399	318
500-999 (72)	50	39	15	3	479	423
≥1000 (106)	46	33	15	5	330	273

[1]GMT = Geometric mean titers are back-transformed from model estimates; all model estimates are derived from $\log_{10}$ transformed H1N1 antibody titers.

Discussion

One of the unexpected results from the original analysis[7] was that participants who received the 2009 seasonal influenza vaccine had significantly lower seroprotection rates compared to participants who did not receive the 2009 seasonal influenza vaccine. Earlier studies have also shown that recipients of previous influenza vaccines have lower GMTs compared to those who did not receive previous influenza vaccines.[7,8] The use of a zero-inflated model accounting for a point mass and left-censoring resulted in a slight attenuation of fold differences at Day 21 between GMTs of those who received the 2009 seasonal influenza vaccine compared to those who did not receive the 2009 seasonal influenza vaccine. Most

interestingly, the zero-inflated model revealed a two-fold difference between baseline GMTs of these two groups. Referencing baseline distributions (Figure 1) as well as model estimates (Table 3), it appears that recipients of the 2009 seasonal influenza vaccine had a distribution of baseline titers shifted to the left (lower) and hence yielded more observations that were left-censored and fewer in the point mass. Conversely, participants who did not receive the 2009 seasonal influenza vaccine had higher titers at baseline, fewer left-censored observations, and hence more observations estimated to reside in the point mass (i.e., "true zeros"). These results potentially indicate two different phenomena: 1) those who received the 2009 seasonal influenza vaccine represent a distinct subpopulation, and 2) receipt of the 2009 seasonal influenza vaccine may potentially affect immunogenicity of the H1N1 vaccine. Although, the zero-inflated model does not fully explain why recipients of the 2009 seasonal influenza vaccine were less responsive to the H1N1 vaccine, it is apparent from zero-inflated modeling results that the two subgroups began the trial with different baseline titer distributions.

The previous paper also revealed that the severe asthma 15 mcg group had the lowest GMTs post-vaccination compared to the mild/moderate 15 and 30 mcg groups and the severe asthma 30 mcg group.[7] We confirm this trend with zero-inflated models, as the GMT values were generally higher for higher doses, and percentages in the point mass were higher for lower doses at Day 21 (Table 3).

These results demonstrate the utility of zero-inflated models when there is a preponderance of observations below the LLOD, which could be truly zero and/or non-zero, but left-censored. By accounting for left-censored observations and a point mass, the zero-inflated model revealed a difference in baseline titer distributions between 2009 seasonal influenza vaccine recipients and non-recipients. The zero-inflated model also pointed to a sizeable proportion of participants ≥ 60yrs residing in the point mass who had no response (i.e., potentially zero antibody titers) to the first vaccine administration. These two observations were obscured by traditional models and suggest the utility of zero-inflated modeling to discern differences in vaccine immunogenicity that are not immediately apparent from traditional approaches.

References

1. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995; 51: 1570-1578.

2. Moulton LH, Curriero FC, Barroso PF. Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research* 2002; 11: 317-325.

3. Maloney JM, Rudengren M, Ahlstedt S, Bock SA, Sampson HA. The use of serum-specific IgE measurements for the diagnosis of peanut, tree nut and seed allergy. *Journal of Allergy and Clinical Immunology* 2008; 122: 145-151.

4. Lachenbruch PA. Analysis of data with excess zeros. *Statistical Methods in Medical Research* 2002; 11: 297-302.

5. Maller RA, Zhou S. Testing for the presence of immune or cured individuals in censored survival data. *Biometrics* 1995; 51: 1197-1205.

6. Taylor DJ, Kupper LL, Rappaport SM, Lyles RH. A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics* 2001; 57: 681-688.

7. Busse WW, Peters SP, Fenton MJ, Mitchell H, Bleecker ER, Castro M, Wenzel S, Erzurum SC, Fitzpatrick AM, Teague WG, Jarjour N, Moore WC, Sumino K, Simeone S, Ratanamaeechat S, Penugonda M, Gaston B, Ross TM, Sigelman S, Schiepan JR, Zaccaro DJ, Crevar CJ, Carter DM, Togias A. Vaccination of patients with mild and severe asthma with a 2009 pandemic H1N1 influenza virus vaccine. *Journal of Allergy and Clinical Immunology* 2011; 127(1): 130-137.

8. Keitel WA, Cate TR, Couch RB. Efficacy of sequential annual vaccination with inactivated influenza virus vaccine. *American Journal of Epidemiology* 1988; 127: 353-64.

9. Treanor J, Keitel W, Belshe R, Campbell J, Schiff G, Zangwill K et al. Evaluation of a single dose of half strength inactivated influenza vaccine in healthy adults. *Vaccine* 2002; 20: 1099-105.

Appendix I. Ln likelihood expressions used for SAS PROC NLMIXED

Model 1: Ordinary least squares models- imputation of censored values

The Ln likelihood for an observation from a log-normal density (with no point mass or censoring, and censored values are imputed) is given by:

$$\ln L\left(\gamma_0,\gamma_1,\ldots\gamma_m,\sigma\middle|Z_1,\ldots Z_m\right)=-\ln\left[\sigma\sqrt{2\pi}\right]-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$$

Model 2. Left-censoring only, no point mass

The Ln likelihood for an observation from a log-normal density with left-censoring (but no point mass) is given by:

$$\ln L\left(\gamma_0,\gamma_1,\ldots\gamma_m,\sigma\middle|Z_1,\ldots Z_m,\delta\right)=\delta\ln\left\{\Phi\left[y_L\right]\right\}+\left(1-\delta\right)\ln\left\{\phi\left(y\right)\right\}$$

Model 3. Left- and right-censoring, no point mass

The Ln likelihood for log-normal data with left- and right-censoring (but no point mass) is given by:

$$\ln L\left(\gamma_0,\gamma_1,\ldots\gamma_m,\sigma\middle|Z_1,\ldots Z_m,\delta\right)=\left(\delta_1\right)\ln\left\{\Phi\left[y_L\right]\right\}+\left\{1-\delta_1\right\}\left\{1-\delta_2\right\}\ln\left\{\phi\left(y\right)\right\}+\left(\delta_2\right)\ln\left\{\left[1-\Phi\left(y_U\right)\right]\right\}$$

Model 4. Point mass only

The Ln likelihood of the log-normal model, which considers a point mass (but no censoring) in addition to a separate log-normal component is given by:

$$\ln L\left(\beta_0,\beta_1,\ldots\beta_k,\gamma_0,\gamma_1,\ldots\gamma_m,\sigma\middle|X_1,\ldots X_k,Z_1\ldots Z_m,\delta\right)=\delta\ln\left\{1-p\right\}+\left(1-\delta\right)\ln\left\{p\phi\left(y\right)\right\}$$

Model 5. Left-censoring, point mass

Combining the features of Models 2 and 4 yields the Ln likelihood for the model with a point mass and left-censoring:

$$\ln L\left(\beta_0,\beta_1,\ldots,\beta_k,\gamma_0,\gamma_1,\ldots,\gamma_m,\sigma\middle|X_1,\ldots,X_k,Z_1\ldots,Z_m,\delta\right)=\delta\ln\left\{\left(1-p\right)+p\Phi\left[y_L\right]\right\}+\left(1-\delta\right)\ln\left\{p\phi\left(y\right)\right\}$$

Model 6. Left- and Right-censored, point mass

The Ln likelihood using a mixture model for a single observation which includes components for left- and right-censoring as well as a point mass is given by:

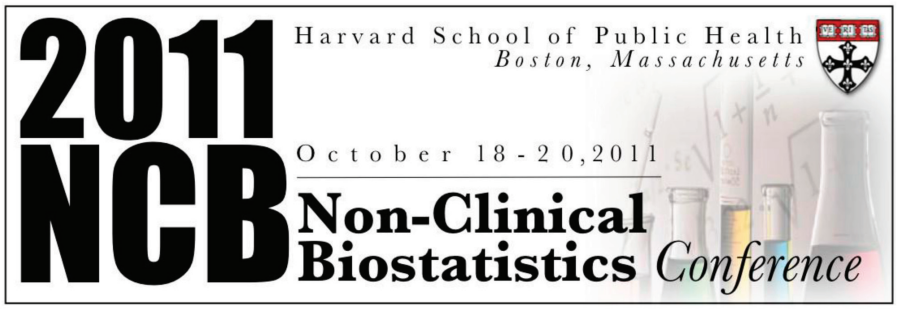
$$\begin{aligned}&\ln L\left(\beta_0,\beta_1,\ldots,\beta_k,\gamma_0,\gamma_1,\ldots,\gamma_m,\sigma\middle|X_1,\ldots,X_k,Z_1,\ldots,Z_m,\delta\right)\\&=\delta_1\ln\left\{\left(1-p\right)+p\Phi\left[y_L\right]\right\}+\left(1-\delta_1\right)\left(1-\delta_2\right)\ln\left\{p\phi\left(y\right)\right\}+\delta_2\ln\left\{p\left[1-\Phi\left(y_U\right)\right]\right\}\end{aligned}$$

Biopharmaceutical Section
ASA Fellow Nominations Requested

Do you know someone in the Biopharmaceutical Section who you think deserves to become an ASA Fellow? The ASA Fellows committee recognizes there are numerous ways to make important contributions to the profession and does not restrict itself to review of research publications. Successful candidates have distinguished themselves by organizing educational activities at the secondary and collegiate level, organizing professional meetings and sections within meetings, effectively managing large groups of statisticians and supporting their external involvement in the profession, fundraising to support a statistical organization, strengthening local ASA chapters through innovative leadership and participation, influencing regulatory or corporate statistical policy through research and communication. A list of criteria for rating nominees is available at *http://www.amstat.org/careers/fellows.cfm*.

If you know an appropriate candidate, please go to the ASA web page on ASA Fellows (*http://www.amstat.org/careers/fellowslist.cfm*) to determine whether the person you have in mind is already an ASA Fellow. If not, please send the name(s) of the person(s) to the Biopharmaceutical Section Fellows Committee through Neal Thomas (*snthomas99@yahoo.com*). The committee will evaluate each recommendation for this prestigious award. The committee does not typically sponsor candidates, but we can help them identify a sponsor and we can supply a letter of support from the section.

Meeting Announcements



2011 Non-Clinical Biostatistics Conference
October 18 – 20, 2011, Boston

Advancing Discovery, Preclinical and CM&C Drug Development through Statistical Science

We are pleased to announce the second U.S. conference dedicated entirely to Non-Clinical Biostatistics. It is organized jointly by regulatory and pharmaceutical/biotech statisticians in collaboration with the *Department of Biostatistics at the Harvard School of Public Health*. The conference will take place October 18 - 20, 2011, at the Harvard Medical School's Joseph B. Martin Conference Center in Boston.

Members of the non-clinical/pre-clinical statistics community are invited to submit proposals for presentations and posters discussing significant scientific and regulatory issues. Attendees will have ample opportunity to network, share experiences and discuss current scientific issues with leaders in the field. Submissions will be accepted up to June 1, 2011.

Registration and a call for abstracts are open on the conference website: *www.ncb2011.org*.

PRELIMINARY PROGRAM

Featured Speakers: ***Bob O'Neill (FDA) & Robert Rodriquez (ASA President Elect)***

- Half-day short course on linear and nonlinear models presented by Andrew Gelman (Columbia University)
- Invited Speakers:

  —Discovery/Early Development/-omics
  Richard Bourgon (Genentech), Anne Carpenter (Broad Institute), Richard Simon (NCI)

  —Pharmacology/Safety/Toxicology/pK
  Mohammad Atiar Rahman (FDA), Vikram Sinha (Eli Lilly), John Szumiloski (Merck)

  —CM&C/Manufacturing
  Rick Burdick (Amgen), Walter Hauck (USP), Meiyu Shen (FDA)
- Poster Session & Reception, Roundtable discussions, Vendor presentations and courses
- ASA Presidential Address Reception

The 67th Deming Conference on Applied Statistics December 5 – 9, 2011 at Atlantic City, New Jersey

Sponsored by Metropolitan Section, ASQ and Biopharmaceutical Section, ASA

The conference's purpose is to provide a learning experience on recent developments in statistical methodologies. The three-day conference is followed by two parallel two-day short courses. The conference is composed of twelve three-hour tutorials on current applied statistical topics of interest. Recognized experts in the field of applied statistics are invited to give the lectures and short courses based on their recently published books. The conference makes these books available for sale at an approximately 40% discount. Attendees will receive bound proceedings of the presentations. The full program will be available on www.demingconference.com before June and all other information on the website, including fees, remains valid. The conference will be held in the state-of-the-art Havana Tower of the Tropicana Casino Resort. Walter Young has chaired this conference for 42 consecutive years.

For more info click here: *www.demingconference.com*.

Let's Hear from You!

If you have any comments or contributions, please contact the Editors: Jose Alvir (*Jose.Alvir@pfizer.com*), Deborah Panebianco (*deborah_panebianco@merck.com*), or Yongming Qu (*QU_YONGMING@LILLY.COM*).

As we have stated in previous issues, the *Report* is a joint effort of the editors and the members of the Biopharmaceutical Section. Volunteers are welcome to write articles of interest to our members. This is an excellent opportunity to "publish and flourish" by sharing your expertise with Section members and a larger audience. Do not hesitate to get in touch with us should you consider contributing an article or know of someone who would like to do so.

Letters to the Editors are also welcome.

We look forward to hearing from you.