Approaches for evaluating veterinary epidemiologic models: verification, validation, and their limitations

Aaron Reeves*, Mo D. Salman, and Ashley E. Hill

Animal Population Health Institute, Department of Clinical Sciences, College of Veterinary Medicine and Biomedical Sciences, Colorado State University, Fort Collins, CO 80523, USA

* Corresponding author. Tel: +1-970-297-5140. Fax: +1-970-297-5228. E-mail address: Aaron.Reeves@colostate.edu

Submitted for publication in *Rev Sci Tech OIE*

Summary

The evaluation of models the spread and control of animal diseases is a crucial undertaking if such models are to be used to inform decisions regarding the control or management of such diseases. Two key steps in the evaluation of epidemiologic models are model verification and model validation. Verification is the demonstration that a computer-driven model is operating correctly, and conforms to its intended design. Validation refers to the process of determining how well a model corresponds to the system that it intended to represent. For a veterinary epidemiologic model, validation would address issues such as how well the model represents the dynamics of the disease in question in a population to which the model is applied, and how well the model represents the application of different measures for disease control.

Just as the development of epidemiologic models is a subjective, ongoing process subject to change and refinement, so too is the evaluation of models:. The purpose of model evaluation is not to demonstrate that a model is a “true” or an “accurate” representation of a system, but to subject a model to sufficient – and continuing – scrutiny so that it may be used with an appropriate degree of confidence as an aid to the decision-making process.

Among the steps that can be taken by epidemiologic modelers to facilitate the processes of model verification are to clearly state the purpose, assumptions, and limitations of a model; to provide a detailed description of the conceptual model for use by everyone who might be tasked with evaluation of a model; document steps already taken to test the model; and thoroughly describe the data sources and the process used to produce model input parameters from data.

Keywords

Model verification – model validation – model credibility – evaluation of models

1. Introduction

Computer-driven epidemiologic modeling is an increasingly common technique for the assessment of the potential for spread and for the potential consequences of animal diseases. Modeling of animal diseases has been used to estimate the possible magnitude of an outbreak and the resources needed for response, and to inform policy decisions regarding measures for disease control (4, 6, 13, 14, 17, 28, 29, 46, 57). Epidemiologic models may take any of several forms. Some are based on analytical formulas that describe the system of interest in a rigorously mathematical way (13, 14, 28, 29, 63). Other models employ computer-driven simulation in order to mimic the actual mechanistic processes at work within a system (5, 15, 22).

Regardless of their form, all models – especially models which are intended for use by response planners and policy makers – require careful evaluation. For models to be effectively used in these instances, a sufficiently high level of credibility of the model and its results must be achieved such that decision makers and other stakeholders can have a justifiable degree of confidence in their application. By the same token, the careful evaluation of models can elucidate their limitations and weaknesses, can temper tendencies toward overreliance on apparently "objective" model-produced outcomes, and can minimize the misapplication of models.

Methods for model evaluation are quite diverse; as several authors have noted, there is no single standard or approach that can be applied to all models (32, 41). At a very basic level, as the mathematical or computational complexity of epidemiologic

models increases, it is essential to demonstrate that the mathematical framework or software used for a model is free from major errors which would threaten the accuracy of the calculations that the model produces. Some approaches for the evaluation of models are by necessity qualitative: any assessment of the conceptual quality of a model, for example, is fundamentally qualitative in nature. In some instances, it may be possible to use quantitative or statistical approaches to demonstrate correspondence between a model and a natural system, although the use of such quantitative methodologies does not necessarily ensure that a model is conceptually sound.

The aim of this paper is to describe approaches for the evaluation of epidemiologic models intended to inform management or policy decisions regarding diseases of animals, with an emphasis on two approaches that have been called verification and validation. Our specific objectives are as follows:

- To briefly define and describe the processes of model verification and validation;
- To discuss several approaches used to address the challenging issue of validation of epidemiologic models intended to inform emergency response plans;
- To illustrate practical approaches to model verification and validation based on our experiences as members of the research team behind the *North American Animal Disease Spread Model* (*NAADSM*: 22);
- And finally, to present a set of suggestions for steps that could be taken to improve the credibility and acceptance of epidemiologic models for the management of animal diseases.

2. Model context, development, and evaluation

Figure 1 illustrates a conceptual series of steps in the process of model development and application. Several of these steps deal explicitly with the evaluation of models, but almost every stage in the figure implies some form of appraisal of the model under development. Decisions made at the outset of model development regarding the specific purpose of a model and the questions for which it is being designed to answer will affect the ways in which the model's utility and credibility are assessed.

First and foremost, models must be evaluated in the context of the problems that they are intended to answer (35, 43, 52): the criteria used to judge a model intended to inform broad questions in a qualitative way will be quite different from those used to evaluate a model that claims to offer specific predictive capabilities.

Second, for results of modeling investigations to be credible, the models must be built upon reliable data (50, 62). Models based on incomplete or theoretical input data may yield useful hypotheses for further research and evaluation, but the limitations of such models should be clearly and expressly stated. The more complete the input data for a model is, the more likely it is that its output will be credible.

Third, just as the conceptual development of models is, in many respects, a subjective undertaking, so too is the evaluation of models. Individual modelers must weigh the relative importance of different aspects of epidemiologic systems, and may come to different conclusions about how to represent different processes in their models, or even which processes to represent. Any assessment of the credibility of a model must consider these subjective design decisions.

Fourth, Figure 1 makes the distinction between a conceptual model or model framework, and a specific model that applies a particular conceptual framework together with a particular dataset or set of parameter values to represent a specific situation. *NAADSM*, for example, is a framework for the development of epidemiologic simulation models, which has been used to build specific models of a variety of diseases in different settings and populations, such as foot-and-mouth disease (FMD) (45, 67), pseudorabies (46), and highly pathogenic avian influenza (HPAI) (19, 44), among others. Both the conceptual framework and the particular instances in which the framework are used need to be evaluated. The utility of the former does not necessarily rely upon the latter, but the quality of specific models is highly dependent on both the conceptual framework as well as on the data used for construction of specific models.

Finally, Figure 1 illustrates that the process of model development and evaluation is cyclical and iterative. Evaluation is not a single, discrete step, and "is not something to be attempted after the simulation model has already been developed, and only if there is time and money remaining" (35). Model evaluation should instead be considered ongoing: model assumptions should be reassessed continually as new sources of information become available.

The assessment of the computational correctness of a model has been called "verification". Verification deals with questions such as "Does the computer program perform all calculations correctly?", and "Does the program match exactly what the designers intended?" The assessment of how well a model conforms to or exemplifies the system that it is intended to represent is sometimes referred to as "validation" (32, 52, 56). Validation efforts are intended to address the question "Is a model an adequate

representation of the real system?" [For the remainder of this paper, we will follow these definitions for "verification" and "validation", but note that these definitions are not universally applied: for example, Oreskes *et al.* (42) use the terms "verification" and validation" to denote somewhat different concepts.] Together, verification and validation efforts can help investigators to ascertain the overall quality and credibility of a model.

3. Model verification

Model verification refers to the process of determining whether the model, as implemented in software, conforms to the desired conceptual model (52). In other words, verification provides an assessment of whether the software implementation of the model is working correctly. Among the criteria by which a model's verification status might be assessed include its correctness (the "extent to which a model meets its specifications") and its reliability (the "extent to which a model can be expected to perform its intended function with required precision") (40, 58). Any model used for scientific research or for decision support should be expected to meet a high standard for such characteristics.

Model verification, although straight-forward in concept, can be time-consuming, particularly as models become more complex. Sargent (52) and Scheller *et al.* (55) presented useful discussions of some of the software engineering practices that can facilitate the construction of verified models, particularly for larger projects, and several authors have provided detailed descriptions of approaches to verification (33, 70). Here, we focus on two central aspects of model verification that directly impact the credibility of epidemiologic models regardless of their form, size, or scope: the production of documentation that describes the conceptual model in detail, and thorough testing to ensure that the model is performing as intended.

3.1. Describing the conceptual model

As shown in Figure 1, design of the conceptual model is an early stage in model development. There is a great deal of value in explicitly documenting this conceptual model: such documentation can be used in an assessment of the conceptual validity of the model (see below), but at a more basic level, it can provide a standard by which the correctness of a model can be judged (33, 55). The purpose of a written model specification is to describe, in clear, accessible language, the purpose, requirements, and conceptual details of a model. The intended audience of such a document includes the modelers themselves as well as any technical personnel who will be involved in implementing the model, among others (see Section 4.2.1 below). The model specification can also provide a basis for model testing (24, 55).

In the case of *NAADSM*, the model specification document (24) describes every component of the modeling framework in detail: it is the authoritative source that describes how the conceptual model should operate, and is the standard by which the software implementation of the conceptual model is judged. Although the specification may be updated as needed to correct ambiguities or to incorporate new features, the complete history of the specification is tracked, and every version is available for reference and evaluation by independent researchers (23, 24).

3.2. Model testing

Fairley (11) and Whitner and Balci (70) distinguish between two forms of model testing, which they refer to as “static” and “dynamic”. For simple models, static testing may be sufficient: such an approach involves a structured examination of the formulas, algorithms, and code used to implement a model, preferably by several reviewers who

were not directly involved in writing the implementation themselves. Garner and Beckett (15) described the use of this approach in the development of *AusSpread*, a simulation platform designed initially to model the spread and mitigation of FMD.

For more complex models, dynamic testing is often useful: during dynamic testing, a computer program is run repeatedly under different conditions to ensure that the output it produces is correct according to the conceptual model and consistent with expectations. Often, such tests are established to be run repeatedly and automatically, to ensure that any changes to the software implementation did not inadvertently introduce errors; this process is referred to as regression testing. Scheller *et al.* (55) describe several levels of testing, from simple unit tests that evaluate specific, individual functions; to system testing that assess the interaction of all of the components of a model. We will illustrate these approaches in the following sections with examples from the development of *NAADSM*.

3.2.1. Automated software testing of the NAADSM framework

In order to ensure that the *NAADSM* application correctly implements the conceptual model specification, *NAADSM* relies upon an automated regression testing approach. Simple models have been constructed to test every aspect of the *NAADSM* application. There are currently well over 1000 individual models in this suite of tests, and new tests are continually being developed. When the *NAADSM* application is compiled from program source code, every test is automatically run and results are tracked using a freely available framework for software testing (54). Prior to the public release of any new version of *NAADSM*, every test in the suite must be passed. Every

simple model developed for testing is published along with the complete source code for the *NAADSM* application.

3.2.2. Manual testing of NAADSM

In addition to automated use of simple tests, manual testing using more complex situations has been carried out for the *NAADSM* framework. Every aspect of the model framework is examined by analysts working independently of the programmers themselves to confirm that the model conforms to the published specification. Any errors identified during manual testing are noted and must be corrected prior to public release.

3.3. The limitations of model verification

Model verification procedures can be quite objective and thorough. Many techniques developed in the field of software engineering can be rigorously applied to the programming of models (7, 55). Model verification offers no answer, however, to the crucial questions "Is the model useful?" and "Is the model adequate for the purposes for which it was designed?" Questions like these can be addressed by a variety of approaches that fall under the general heading of model validation.

4. Model validation

Validation refers to the process of determining whether a model is an acceptable representation of the system that it is intended to represent, given the purpose of the model or study (35, 52). A more elaborate definition is provided by Schlesinger *et al.* (56): model validation is the "substantiation that a... model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model". It is important to note that "acceptable representation" in the

definition above does not connote an "accurate" or a "true" representation: Oreskes *et al.* (42) convincingly argued that it is impossible to establish that any particular model is an accurate representation of a natural system, and that the use of the term "validation" in this sense is highly misleading.

4.1. The problem of model validation

In contrast to the process of model verification, establishing the validity of models is not clear-cut, and can be quite problematic. As McCarl (41) observed, "there is not, and never will be, a totally objective and accepted approach to model validation." The standards by which a model's validation is judged are partly dependent upon the purpose of the model. The validation of models designed strictly to address research questions (for example to generate and test hypotheses concerning population or disease dynamics or to identify new areas of research) does not have to be as stringent as the evaluation of models that will be used to inform operational management decisions. When such decisions will be made on the basis of the results of modeling studies, we would wish to know that these studies are appropriate, accurate, and correct. Given the difficulties associated with the study of very complex multifactorial problems, the subjective elements of modeling itself, and philosophical issues like those presented by Oreskes *et al.* (42), the threshold for the acceptance of models cannot be that of "proof" of their accuracy or validity. Rather, this threshold should be that of reasonable confidence in the results produced by models. As Holling (25) stated, "provisional acceptance of any model implies not certainty, but rather a sufficient degree of belief to justify further action." The task of model validation, as described here, is that of evaluating models in

order to have a justifiable level of confidence in their results before they influence policy or management decisions.

It is often constructive to think of a model in a way similar to a scientific hypothesis. An epidemiologic model, for example, represents the modelers' hypotheses regarding the interactions among members of a population, the dynamics of disease in that population, mechanisms of disease spread, and the efficacy of different disease control measures. As with any hypothesis, models should be tested and challenged. As models are subjected to and withstand increasing levels of scrutiny in diverse situations, their credibility is increased. Such models can then be applied to problems of management and policy with greater confidence, provided that it is always clearly understood that no model truly represents physical reality, and that the acceptance of any model must be subject to ongoing evaluation.

What follows is not a set of methods that will prove that a model represents a real system, but rather a set of activities that might be undertaken to provide evidence which may either support or refute the hypothesis presented by a model. Several authors present descriptions and detailed taxonomies of the methods used to assess model validity (33, 34, 50, 52). Our intention in the following sections is to present and discuss the utility of some of these methods, together with examples of their application both from our own experience and from other published reports of animal disease modeling. We would also refer readers to several excellent discussions of model validation, including those presented by Oreskes *et al.* (42), Rykiel (50), and Taylor (62).

4.2. Conceptual validity

A particularly useful – and a foundational – criterion for the validation of an epidemiologic model is the answer to the question "Does the structure of a model make logical and biological sense?" This has been referred to as "conceptual validity" (50, 52). For a model to have conceptually validity, its theoretical underpinnings should be shown to be based on known and scientifically accepted properties of the system of interest, or at least on reasonable and justifiable assumptions about such properties. Among some of the questions that might be addressed in assessing the conceptual validity of a model are the following:

- Does the model fit the purpose or purposes for which it was designed?
- Does the structure of the model sufficiently capture the relationships and interactions among components of the system being modeled?
- Given the purpose of the model, are key components of the system absent from the model, or oversimplified? Is additional detail necessary for any component?
- Based on existing knowledge and experience, are the outcomes produced by a model reasonable?

Review by independent subject matter experts (sometimes referred to as establishing "face validity": 50) can be used as a means of assessment. In order to facilitate such review, it is quite helpful to have a detailed document that describes the conceptual model, as described in Section 3.1. Such a document can provide a basis for discussion and evaluation of the details of model operation. The publication of model descriptions (5, 22, 26, 60) greatly facilitates the assessment of the conceptual validity of models.

Reliance on the peer-reviewed literature provides one avenue for the conceptual assessment of epidemiologic models. The *NAADSM* development team has also taken a more direct approach, and has sponsored a series of meetings of subject matter experts, including epidemiologists, virologists, economists, policy makers, and other modelers, for the purpose of review of the *NAADSM* modeling framework (10, 64, 65). The structure and assumptions of the modeling platform have been described in detail during these workshops, and discussion, suggestions, and advice are solicited from all participants. The results of these expert panel evaluations are then used to guide future research and development.

4.3. The utility of data in model validation

As noted in Section 2, it is possible to assess the conceptual framework apart from the data used to inform a model. Empirical data are generally used in two ways during modeling: 1) input data are used to develop parameters that will influence model outcomes, and 2) data that represent the outcomes or results of a system (output data) are used to provide a basis for comparison with model-produced outcomes. In a few cases, particularly for endemic disease situations, large amounts of both types of data may be available for models of disease spread in populations. In many instances, however, we have access to information pertaining to only a single outbreak of disease in a particular set of circumstances. Information collected during the 2001 outbreak of FMD in the United Kingdom, which has been widely used for modeling studies (13, 14, 28, 29, 53), represents one such dataset. In still other cases, models are developed to explore hypothetical situations (5, 6, 15, 19, 44). In these cases, some information is generally

established to inform model inputs, but there can be no data regarding the (nonexistent) system outcomes.

Whatever the form or source of data used to inform models, their correctness and validity also should be considered. As Rykiel (50) pointed out, there is no guarantee that available data necessarily provide a better or more accurate depiction of a real system than a conceptual model. The process of ensuring so-called data validity (50, 52) can by itself be complex.

Several authors have emphasized the notion that, in order to demonstrate validity, models should be tested against data not used during their construction (31, 59). Green and Medley (20) indicated that such a step should be a requirement before a model is used to inform policy decisions. This is one of several possible approaches that falls into the general category of "operational validation" (52).

Although this suggestion seems straight-forward, its implementation for incompletely understood biological and epidemiologic systems is problematic. First, it implies that reliable, valid data exist for at least two situations, for both the development of parameters and for comparison to actual system outcomes. Second, this approach would require the existence of a suitable means of evaluation by which the similarity of model-produced outcomes to system outputs can be assessed. Third, it implies that these situations are sufficiently dissimilar from one another that they represent unique tests of a model, but are still similar enough that exactly the same approach to modeling developed for one situation can be legitimately applied to the others. We have already mentioned the first difficulty. The remaining two problems are discussed below.

A variety of quantitative, statistical approaches to show the correspondence between model-produced outputs and outcomes generated by biological systems have been devised and applied in a few situations (12, 32, 36, 39, 47, 48, 49, 69). Most of these approaches to what has been called statistical validation rely upon the existence of a large amount of data (*i.e.*, many observations) pertaining to the outcome of the natural system, which limits their applicability to most situations of interest to animal disease modelers.

Waller *et al.* (69) proposed the use of Monte Carlo hypothesis tests, which in essence compare a single set of outcome data from a real system to multiple model-generated outcome datasets, and seek to answer the question "Do the observed data appear consistent with the model?" rather than the more typical question "Does the model appear consistent with the observed data?" Although this approach is not without utility, it raises an additional question: how representative is any single outcome? Among recent outbreaks of FMD in the United Kingdom, for example, is the 2001 outbreak, which resulted in the infection of over 2000 herds (1), more or less representative than the 2007 outbreak, which produced only eight infected herds (2)? How "consistent" would each of these two outcomes have to be with model-produced data in order to conclude affirmatively that the data are consistent with the model? Efforts made to compare outcomes from epidemiologic models to data generated by individual outbreaks should be undertaken with care: such comparisons are potentially informative, but an overreliance on quantitative approaches for the evaluation of models may well be misleading.

348 The disparity between these two recent FMD outbreaks in the UK also illustrates
349 the third potential problem raised above: the dissimilarities among outbreaks of even the
350 same diseases in generally the same types of populations make it difficult to test a model
351 against data not used during its construction. As described in Section 2, the use of data is
352 integral to model construction. Although a conceptual framework of a model and the
353 data used to inform a model are distinct and can (and should) be evaluated individually,
354 output generated by a model is inseparable from the combination of these two elements.
355 The correspondence of model output to a natural system cannot be evaluated without
356 considering the conceptual model and the source data simultaneously.
357 Although it is generally helpful to have detailed data, it may not be strictly
358 necessary for a model to be built upon detailed empirical information for it to be useful as
359 a source of information for policy development. For example, Green *et al.* (19)
360 developed a model of HPAI in domesticated poultry in Manitoba, Canada. They
361 compared several disease control strategies, one of which was the use of concentric zones
362 with radii of several kilometers around infected premises, for disease surveillance,
363 restriction of movement, and depopulation. Although most of the parameters used in the
364 models for this study were derived from expert opinion rather than empirical data, model
365 results strongly suggested that the utility of such relatively small zones would be
366 insufficient in the population and setting for which the study was conducted. Results of
367 this study are not definitive and do not prescribe a specific policy formulation, but they
368 provided sufficient justification for, and have already prompted, a re-evaluation of
369 existing approaches to surveillance (C. Green, personal communication).

Comment: The manuscript by Green *et al.* is in the final stages of preparation, and we fully expect to have it submitted for publication before this article appears in print.

4.4. Validation of model components

Although it is difficult to demonstrate the validity of an entire model by the means described above, especially in the absence of relevant data, it may be possible to assess the validity of some individual components of a more complex model. This component-based approach to validation is sometimes recommended (38). An example is a recently completed validation of the process used in *NAADSM* to simulate animal movements and contacts among farm premises (C. Dubé, personal communication).

Briefly, the objective of this study was to validate the contact component used in *NAADSM* by comparing simulated movements to real-world, farm-to-farm movements that had been recorded for adult milking cows in Ontario, Canada. The study concluded that the approach used in *NAADSM* performed reasonably well in simulating average network characteristics observed in real-world movement data, but did not perform as well in simulating extreme upper percentiles of movement network components, involving rare but observed farms with excessively high shipment frequencies. The results of this study will be used to inform future development, with the objective of providing better representations of actual events and greater confidence in the results of modeling studies.

4.5. Comparison of models

Comparison of several independently developed models may be used to improve the level of confidence in the models tested. This process of comparing the results of several models has been called “relative validation” (9).

Dubé *et al*. (9) conducted a comparison of three simulation models using several relatively simple disease scenarios. Among the findings of this comparison was that,

although statistically significant differences were observed among model outputs, results from all three models supported the same or very similar conclusions regarding approaches for disease control. This finding could be used to increase the confidence of end users and decision makers in modeling results (9). The results of a follow-up investigation that considered more complex scenarios are reported elsewhere in this issue (51).

Comment: My understanding is that Robert Sanson is authoring a paper for the same issue of Rev Sci Tech OIE. If this is incorrect, this sentence should be removed.

Several similar comparisons of models of the spread and control of animal disease have also been undertaken. Vigre (68) reported on a comparison of mathematical and simulation-based models. The differences identified were more substantial than those reported by Dubé *et al*. (9), and may reflect the broader distinctions between the fundamental assumptions made by the individual models. Continued investigation in this vein would be quite helpful. Gloster *et al*. (18) also recently reported on the comparison of several models of airborne dispersion of FMD virus. Like Dubé *et al*. (9), they reported that the results of the models evaluated were broadly similar, but of course, highly dependent on the assumptions made and the data used by different groups of modelers.

Loehle (36) identified the comparison of models as a component of the larger process of what he called structural analysis, or the evaluation of the inherent assumptions and the identification of deficiencies in different models. Loehle argued that, because of the existence of such structural differences among models, and because the comparison of multiple models is the most effective way to identify and determine the effects of such differences, it is essential to have multiple modeling efforts directed toward addressing any important policy or management problem.

4.6. Sensitivity analysis as a form of validation

When data from real systems are limited, sensitivity analysis is sometimes suggested and used to inform model validation efforts (6, 27, 32). Sensitivity analysis is used to determine the amount of influence that particular parameters have on model-produced outcomes. Sensitivity analysis can also be used to assess the conceptual validity of a model: if certain parameters are expected to be important in a system based on prior knowledge of that system, then sensitivity analysis should bear out these expectations (32).

Of greater utility is the use of sensitivity analysis to determine which parameters in a model are important. If a model includes parameters about which there is a high degree of uncertainty but which are shown by sensitivity analysis to have a substantial impact on model results, such parameters are good targets for additional research. A broader discussion of the application of sensitivity analysis for animal disease modeling can be found elsewhere in this issue (61).

Comment: My recollection is that Mark Stevenson is writing a paper that discusses sensitivity analysis for the same issue of OIE Sci Tech Rev. If this statement is incorrect, it should be removed.

5. Suggestions for the construction of useful, credible models of animal disease

As we have discussed in the preceding sections, the primary objective of model verification and validation efforts is not to demonstrate that a model is a true or even a highly accurate representation of a real system, but rather to provide a set of approaches and criteria by which a model can be evaluated. For models that might be used as a partial basis for policy or management decisions, it is essential that such an evaluation establishes a foundation of support and credibility. To that end, we suggest the following practical steps that members of the veterinary epidemiologic community can take to

produce credible, useful models of the spread and control of disease in animal populations. These suggestions are drawn from our own experience, as well as from many of the other valuable sources cited throughout this article, in particular those written by Bart (3), Rykiel (50), Law and McComas (35), and Sargent (52).

Clearly and precisely state the purpose for which a model was designed

The importance of the first step illustrated in Figure 1, that of determining and then clearly and precisely stating the questions to be asked of a model, may seem self-evident, but this step is often overlooked (3). Overton (43) remarked that "the great majority of criticisms of models relate to a capacity for which the model was not designed in the first place." A clear understanding of the purpose of a model is a prerequisite for any further evaluation.

Provide a detailed description of the conceptual model, and documentation concerning the assumptions and limitations of the model

Virtually every paper on techniques for the verification and validation of models stresses the importance of documentation for the conceptual model (3, 33, 35, 52, 55). A model description should not be produced solely, or even primarily, for the developers of an individual model. Those who will derive the most benefit from the existence of such documents will be other model users, in the broadest sense of the term: other researchers, analysts, and decision makers who will be tasked with the application or evaluation of the model and its results. It is particularly useful when model documentation includes discussions of assumptions and limitations, presented in ways that are clear and biologically relevant (21).

Provide details of steps taken for model verification

At its most basic level, the credibility of a model relies upon the demonstration that the model, as implemented in software, does what it is supposed to do. Anyone tasked with the evaluation of a model, particularly if it will be used to influence policy, should have sufficient access to a computational implementation of a model, details of the verification procedure employed, and to any tests used for verification purposes such that he or she can reproduce and evaluate the computational correctness of the model.

Describe the data used to develop model parameters, and provide documentation for the approaches and assumptions used to produce model parameters from data

The process of translating raw data into parameters suitable for use in models is seldom straight-forward. An understanding of this process, however, is essential for reviewers to have an adequate basis by which to judge the results of a model. Two recent reports illustrate this suggestion quite nicely: Mardones *et al.* (37) conducted a meta-analysis based on 21 research papers and documented in detail the procedures that they used to estimate the durations of different disease states for FMD for the purpose of disease modeling. In a different study, Patyk *et al.* (44) produced a model of the spread and control of HPAI in the US state of South Carolina. This study included an online supplement that described in detail all of the sources of information used for the study, as well as the several computational tools that they developed and used for parameter development.

In order to further support the objective of transparently presenting the process of parameter development, the research team behind *NAADSM* recently implemented an online, collaborative resource for the storage and organization of information used to develop simulation models. In particular, this resource is designed to make it possible for

researchers to document and share the approaches, tools, and assumptions that they use to produce the input values that they use. This parameter library, which is freely and publicly available via the Internet at *http://www.naadsm.org/parameters*, is currently optimized for users of the *NAADSM* modeling framework, but we hope that, as additional information becomes available for other models, we will be able to incorporate other approaches to modeling into the parameter library.

Comment: This site is currently in the final stages of testing, and will be publicly available by press time.

Involve independent experts in the evaluation of models and their outcomes

Veterinary epidemiologic modeling is an interdisciplinary undertaking. Modelers can take advantage of a great deal of expertise in different fields by involving experts from these fields. For models to be used for decision making, it is also essential to involve other stakeholders, for example, those who will be responsible for decision making or for implementing policies in the field, in this process. In our own experience with *NAADSM*, we have found that, through its widespread application, we have benefited substantially from the efforts of others to use and evaluate it.

A variety of forums for the sharing and discussion of veterinary epidemiologic modeling work have been available over the last several years (10, 64, 65, 66). We encourage anyone involved with the construction, use, or evaluation of models to seek out and take advantage of such opportunities when they occur.

When possible, use existing information for data-driven validation of models or their components

We have discussed the limitations and advantages of this approach in Sections 4.3 and 4.4 above. Such approaches should be undertaken with care, and with the recognition that the results will not be definitive: a poor conceptual model may still

produce a good fit to observed data, and *vice versa*. In situations where appropriate information is available, however, the comparison of model-produced outcomes to real data can still be enlightening. Retrospective analysis of past outbreaks is critical to understanding them, and modeling can be a very useful tool in this pursuit (16, 30).

Present a range of possible outcomes, including "best case" and "worst case" scenarios

As we have discussed, models are not definitive representations of reality. We are often uncertain about the ways in which at least some components of our systems operate, and also about specific parameter values. Presenting a range of results is one way to capture some of this uncertainty.

Use sensitivity analysis to determine the importance of parameters used in a model

In addition to the benefits discussed in Section 4.6, the evaluation of the importance of model parameters – especially those for which data are limited – can be used to evaluate the potential effects of parameters about which we are uncertain.

Compare the purposes, conceptual bases, and outcomes of different models

During the modeling process, different modelers make different subjective decisions and assumptions. Qualitative agreement among several models may lend credibility to the conclusions drawn from model-based studies. Areas of disagreement among models should prompt additional research and investigation to improve our level of understanding of the system components in question.

Finally, treat model evaluation as an ongoing process, not as settled fact

Every epidemiologic model is a work in progress, informed and updated by existing and new knowledge about the dynamics of disease; changes in agricultural and social practices; and changes in the forms, sources, and quality of available data. The

validity of any epidemiologic model should be continually reassessed under new conditions or as the state of our knowledge is improved.

6. Conclusions

The careful evaluation of any model intended to inform management or policy decisions is a critical activity. Two key steps in the assessment of the quality and utility of epidemiologic models are verification and validation. Unfortunately, there are no purely quantitative, strictly objective means by which to evaluate models. Each model, and each situation to which modeling will be applied, is at least somewhat unique, and unique means may be necessary to evaluate a model and its particular applications.

Holling (25) pointed out that "provisional acceptance of any model implies not certainty, but rather a sufficient degree of belief to justify further action." We have outlined a set of recommendations that can be used by epidemiologic modelers to continue to cultivate a level of confidence in the application of the technique to important problems in animal population health. Individual models will continue to be developed and compared, and will evolve as they are scrutinized. Through these exercises, our collective objective of providing useful tools to assist in decision-making processes can be met.

In order to achieve a sufficient level of credibility in model outcomes, it is essential to involve not just modelers in their evaluation. As Rykiel (50) observed, "to the extent that a model is a scientific experiment and theoretical development, its testing and validation are within the purview of the scientific community." We agree, and would add that, in the case of models for animal diseases, the evaluation of models is also within

the purview of field epidemiologists and veterinary practitioners, policy planners and decision makers, and animal industry representatives.

Acknowledgements

We thank Ric Hupalo and Marian Talbert for their contributions to the literature search that accompanied the writing of this paper. We would also like to thank the many participants in our modeling courses and workshops who made valuable contributions to discussions of model evaluation and validation. Partial financial support for this work was provided by the Colorado State University – Program of Economically Important Infectious Animal Diseases, which is funded by a special grant from the US Department of Agriculture – National Institute of Food and Agriculture. This work was conducted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Clinical Sciences from Colorado State University (AR).

References

1. Anderson, I. (2002). – Foot and mouth disease 2001: lessons to be learned inquiry report. Available online at *http://archive.cabinetoffice.gov.uk/fmd/fmd_report/documents/index.htm*. Last accessed September 10, 2010.
2. Anderson, I. (2008). – Foot and mouth disease 2007: a review and lessons learned. Available online at *http://archive.cabinetoffice.gov.uk/fmdreview/*. Last accessed September 10, 2010.
3. Bart, J. (1995). – Acceptance criteria for using individual-based models to make management decisions. *Ecol. Appl.* **5** (2), 411-420.

4. Bates, T.W., Carpenter, T.E., & Thurmond, M.C. (2003). – Benefit-cost analysis of vaccination and preemptive slaughter as a means of eradicating foot-and-mouth disease. *Am. J. Vet. Res.* **64** (7), 805-812.

5. Bates, T.W., Thurmond, M.C., & Carpenter, T.E. (2003). – Description of an epidemic simulation model for use in evaluating strategies to control an outbreak of foot-and-mouth disease. *Am. J. Vet. Res.* **64** (2), 195-204.

6. Bates, T.W., Thurmond, M.C., & Carpenter, T.E. (2003). – Results of epidemic simulation modeling to evaluate strategies to control an outbreak of foot-and-mouth disease. *Am. J. Vet. Res* **64** (2), 205-210.

7. Baxter, S.M., Day, S.W., Fetrow, J.S., & Reisinger, S.J. (2006). – Scientific software development is not an oxymoron. *PLoS Comput. Biol.* **2** (9), e87.

8. Dent, J.B. & Blackie, M.J. (1979). – Systems simulation in agriculture. Applied Science Publishers, London, England.

9. Dubé, C., Stevenson, M.A., Garner, M.G., Sanson, R.L., Corso, B.A., Harvey, N., Griffin, J., Wilesmith, J.W., & Estrada, C. (2007). – A comparison of predictions made by three simulation models of foot-and-mouth disease. *New Zeal. Vet. J.* **55** (6), 280-288.

10. Dubé, C., Geale, D., & Sanchez, J. (2008). – *NAADSM* orientation workshop and project plan for pilot studies, software development, and oversight of *NAADSM* application in South America. Canadian Food Inspection Agency, Ottawa, Ontario, Canada.

11. Fairley, R.E. (1976). – Dynamic testing of simulation software. *In* Proceedings of the 1976 summer computer simulation conference. Society for Modeling and Simulation International, San Diego, California, 40-46.

12. Fay, T.H., Greeff, J.C., Eisenberg, B.E., & Groeneveld, H.T. (2006). – Testing the model for one predator and two mutualistic prey species. *Ecol. Model.* **196**, 245-255.

13. Ferguson, N.M., Donnelly, C.A., & Anderson, R.M. (2001). – The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science* **292**, 1155-1160.

14. Ferguson, N.M., Donnelly, C.A., & Anderson, R.M. (2001). – Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* **413**, 542-548.

15. Garner, M.G. & Beckett, S.D. (2005). – Modelling the spread of foot-and-mouth disease in Australia. *Aus. Vet. J.* **83**, 758-766.

16. Garner, M.G., Dubé, C., Stevenson, M.A., Sanson, R.L., Estrada, C., & Griffin, J. (2007). – Evaluating alternative approaches to managing animal disease outbreaks - the role of modelling in policy formulation. *Veter. Ital.* **43** (2), 285-298.

17. Garner, M.G., Cowled, B., East, I.J., Moloney, B.J., & Kung, N.Y. (2010). – Evaluating the effectiveness of early vaccination in the control and eradication of equine influenza - A modelling approach. *Prev. Vet. Med.* In press.

18. Gloster, J., Jones, A., Redington, A., Burgin, L., Sørensen, J.H., Turner, R., Dillon, M., Hullinger, P., Simpson, M., Astrup, P., Garner, G., Stewart, P., D'Amours, R.,

Sellers, R., & Paton, D. (2010). – Airborne spread of foot-and-mouth disease - model intercomparison. *Vet. J.* **183**, 278-386.

19. Green, C., Whiting, T., Duizer, G., Douma, D., Kloeze, H., Lees, W., & Reeves, A. (Submitted for publication). – Modeling the potential impact and the efficacy of control strategies for a highly pathogenic avian influenza outbreak in Manitoba, Canada.

20. Green, L.E. & Medley, G.F. (2002). – Mathematical modelling of the foot and mouth disease epidemic of 2001: strengths and weaknesses. *Res. Vet. Sci*. **73**, 201-205.

21. Guitian, J. & Pfeiffer, D. (2006). – Should we use models to inform policy development? *Vet. J.* **172**, 393-395.

22. Harvey, N., Reeves, A., Schoenbaum, M.A., Zagmutt-Vergara, F.J., Dubé, C., Hill, A.E., Corso, B.A., McNab, W.B., Cartwright, C.I., & Salman, M.D. (2007). – The *North American Animal Disease Spread Model*: A simulation model to assist decision making in evaluating animal disease incursions. *Prev. Vet. Med*. **82**, 176-197.

23. Harvey, N. & Reeves, A. (Eds.) (2008). – Model description: *North American Animal Disease Spread Model* 3.1, version 1.1.1. Available online at *http://www.naadsm.org/documentation/specification*. Last accessed September 10, 2010.

24. Harvey, N. & Reeves, A. (Eds.). (2010). – Model description: *North American Animal Disease Spread Model* 3.1, version 1.1.2. Available online at *http://www.naadsm.org/documentation/specification*. Last accessed September 10, 2010.

Comment: The manuscript by Green *et al.* is in the final stages of preparation, and we fully expect to have it submitted for publication before this article appears in print.

638 25. Holling, C.S. (1978). – Adaptive environmental assessment and management. John
639 Wiley and Sons, New York, New York, USA.

640 26. Jalvingh, A.W., Nielen, M., Maurice, H., Stegeman, A.J., Elbers, A.R.W., &
641 Dijkhuizen, A.A. (1999). – Spatial and stochastic simulation to evaluate the impact
642 of events and control measures on the 1997-1998 classical swine fever epidemic in
643 the Netherlands. I. Description of simulation model. *Prev. Vet. Med.* **42**, 271-295.

644 27. Karsten, S., Rave, G., & Krieter, J. (2005). – Monte Carlo simulation of classical
645 swine fever epidemics and control. II. Validation of the model. *Vet. Microbiol.* **108**,
646 199-205.

647 28. Keeling, M.J., Woolhouse, M.E.J., Shaw, D.J., Matthews, L., Chase-Topping, M.,
648 Haydon, D.T., Cornell, S.J., Kappey, J., Wilesmith, J., & Grenfell, B.T. (2001). –
649 Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a
650 heterogeneous landscape. *Science* **294**, 813-817.

651 29. Keeling, M.J., Woolhouse, M.E.J., May, R.M., Davies, G., & Grenfell, B.T. (2003).
652 – Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136-
653 142.

654 30. Kitching, R.P., Thrusfield, M.V., & Taylor, N.M. (2006). – Use and abuse of
655 mathematical models: an illustration from the 2001 foot and mouth disease epidemic
656 in the United Kingdom. *Rev. Sci. Tech. Off. Int. Epiz.* **25**, 293-311.

657 31. Kitching, R.P., Taylor, N.M., & Thrusfield, M.V. (2007). – Vaccination strategies for
658 foot and mouth disease. *Nature* **445**, E12-E13.

659 32. Kleijnen, J.P.C. (1999). – Validation of models: statistical techniques and data
660 availability. *In* Farrington, P.A., Nembhard, H.B., Sturrock, D.T., & Evans, G.W.
661 (Eds.), Proceedings of the 1999 Winter Simulation Conference, 647-654.

662 33. Knepell, P.L. & Arangno, D.C. (1993). – Simulation validation: a confidence
663 assessment methodology. IEEE Computer Society Press, Los Alamitos, California,
664 USA.

665 34. Law, A.M. & Kelton, W.D. (2000). – Simulation modeling and analysis. McGraw-
666 Hill, Boston, Massachusetts, USA

667 35. Law, A.M. & McComas, M.G. (2001). – How to build valid and credible simulation
668 models. *In* Peters, B.A., Smith, J.S., Medeiros, D.J., & Rohrer, M.W. (Eds.),
669 Proceedings of the 2001 Winter Simulation Conference, 22-29.

670 36. Loehle, C. (1997). – A hypothesis testing framework for evaluating ecosystem model
671 performance. *Ecol. Model*. **97**, 153-165.

672 37. Mardones, F.O., Perez, A.M., Sanchez, J., AlKhamis, M.A., & Carpenter, T.E.
673 (2010). – Parameterization of the duration of infection stages of serotype O foot-and-
674 mouth disease virus: an analytical review and meta-analysis with application to
675 simulation models. *Vet. Res.* **41**, 45.

676 38. Martin, S.W., Meek, A.H., & Willeberg, P. (1987). – Veterinary epidemiology:
677 Principles and methods. Iowa State University, Ames, Iowa, USA.

678 39. Mayer, D.G. & Butler, D.G. (1993). – Statistical validation. *Ecol. Model.* **68**, 21-32.

679 40. McCall, J.A., Richards, P.K., & Walters, G.F. (1977). – Factors in software quality.
680 Air Development Center, Rome, Italy.

41. McCarl, B.A. (1984). – Model validation: an overview with some emphasis on risk models. *Rev. Market. Agric. Econ.* **52**, 153-173

42. Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). – Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **263**, 641-646.

43. Overton, W.S. (1977). – A strategy of model construction. *In* Hall, C.A.S. & Day, J.W. Jr. (Eds.), Ecosystem modeling in theory and practice. John Wiley and Sons, New York, New York, USA, 49-74.

44. Patyk, K.A., Helm, J., Martin, M.K., Forde-Folle, K.N., Olea-Popelka, F.J., Hokanson, J., Fingerlin, T., & Reeves, A. (Submitted for publication). – Establishing input parameters for an epidemiologic simulation model scenario of the spread and control of highly pathogenic avian influenza (H5N1) among commercial and backyard poultry flocks in South Carolina.

45. Pendell, D.L., Leatherman, J., Schroeder, T.C., & Alward, G.S. (2007). – The economic impacts of a foot-and-mouth disease outbreak: a regional analysis. *J. Agric. Resour. Econ.* **39**, 19-33.

46. Portacci, K., Reeves, A., Corso, B., & Salman, M. (2009). – Evaluation of vaccination strategies for an outbreak of pseudorabies virus in US commercial swine using the *NAADSM*. *In* Proceedings of the 12th International Symposium on Veterinary Epidemiology and Economics, Durban, South Africa, 78.

47. Power, M. (1993). – The predictive validation of ecological and environmental models. *Ecol. Model.* **68**, 33-50.

48. Reynolds Jr., M.R., Burkhart, H.E., & Daniels, R.F. (1981). – Procedures for statistical validation of stochastic simulation models. *Forest Sci.* **27** (2), 349-364.

49. Robinson, A.P. & Froese, R.E. (2004). – Model validation using equivalence tests. *Ecol. Model.* **176**, 349-358.

50. Rykiel, E.J. (1996). – Testing ecological models: the meaning of validation. *Ecol. Model.* **90**, 229-244.

51. Sanson, R.L., Harvey, N., Garner, M.G., Stevenson, M.A., Davies, T.M., Hazelton, M.L., O'Connor, J., Dubé, C., Forde-Folle, K.N., & Owen, K. (Submitted for publication). – Foot-and-mouth disease model verification and 'relative validation' through a formal model comparison.

52. Sargent, R.G. (2009). – Verification and validation of simulation models. *In* Rossetti, M.D., Hill, R.R., Johansson, B., Dunkin, A., & Ingalls, R.G. (Eds), Proceedings of the 2009 winter simulation conference, 162-176.

53. Savill, N.J., Shaw, D.J., Deardon, R., Tildesley, M.J., Keeling, M.J., Woolhouse, M.E.J., Brooks, S.P., & Grenfell, B.T. (2007). – Effect of data quality on estimates of farm infectiousness trends in the UK 2001 foot-and-mouth disease epidemic. *J. R. Soc. Interface* **4**, 235-241.

54. Savoye, R. (2004). – DejaGnu: the GNU testing framework. Free Software Foundation, Boston, Massachusetts, USA. Available online at *http://www.gnu.org/software/dejagnu/manual/*. Last accessed September 10, 2010.

722 55. Scheller, R.M., Sturtevant, B.R., Gustafson, E.J., Ward, B.C., & Mladenoff, D.J.
723 (2010). – Increasing the reliability of ecological models using modern software
724 engineering techniques. *Front. Ecol. Environ.* **8** (5), 253-260.

725 56. Schlesinger, S. (1979). – Terminology for model credibility. *Simulation* **32**, 103-
726 104.

727 57. Schoenbaum, M.A. & Disney, W.T. (2003). – Modeling alternative mitigation
728 strategies for a hypothetical outbreak of foot-and-mouth disease in the United States.
729 *Prev. Vet. Med*. **58**, 25-52.

730 58. Scholten, H. & Udink ten Cate, A.J. (1999). – Quality assessment of the simulation
731 modeling process. *Comput. Electron. Agr*. **22**, 199-208.

732 59. Spedding, C.R.W. (1988). – An introduction to agricultural systems. Elsevier
733 Applied Science, London, England.

734 60. Stärk, K.D.C., Pfeiffer, D.U., & Morris, R.S. (2000). – Within-farm spread of
735 classical swine fever virus - a blueprint for a stochastic simulation model. *Vet. Quart.*
736 **22**, 36-43.

737 61. Stevenson, M.A. (Submitted for publication). – Sensitivity analysis.

738 62. Taylor, N. (2003). – Review of the use of models in informing disease control policy
739 development and adjustment. UK Department of Environment, Food, and Rural
740 Affairs. Available online at *http://epicentre.massey.ac.nz/resources/acvsc_grp/docs/*
741 *Taylor_2003.pdf*. Last accessed September 10, 2010.

742 63. Thornley, J.H.M. & France, J. (2009). – Modelling foot and mouth disease. *Prev.*
743 *Vet. Med*. **89**, 139-154.

Comment: This reference needs to be updated.

64. United States Department of Agriculture – Animal and Plant Health Inspection Service – Veterinary Services – Centers for Epidemiology and Animal Health. (2002). North American Animal Health Committee Conference. USDA-APHIS-VS-CEAH, Fort Collins, Colorado, USA.

65. United States Department of Agriculture – Animal and Plant Health Inspection Service – Veterinary Services – Centers for Epidemiology and Animal Health. (2004). Validation of *SpreadModel*: minutes of the subject matter expert team meeting. USDA-APHIS-VS-CEAH, Fort Collins, Colorado, USA.

66. United States Department of Agriculture – Animal and Plant Health Inspection Service – Veterinary Services – Centers for Epidemiology and Animal Health & Office International des Épizooties. (2008). – CEAH-OIE Epidemiological Modeling Workshop, August 11 – 13, 2008, Fort Collins, Colorado, USA.

67. United States Department of Agriculture – Animal and Plant Health Inspection Service – Veterinary Services – Centers for Epidemiology and Animal Health. (2009). – Vaccine bank requirements for foot-and-mouth disease in southwest Kansas: a preliminary report. USDA-APHIS-VS-CEAH, Fort Collins, Colorado, USA.

68. Vigre, H. (2008). – A comparison of three simulation models - the EpiLab project. *In* United States Department of Agriculture – Animal and Plant Health Inspection Service – Veterinary Services – Centers for Epidemiology and Animal Health & Office International des Épizooties. CEAH-OIE Epidemiological Modeling Workshop, August 11 – 13, 2008, Fort Collins, Colorado, USA.

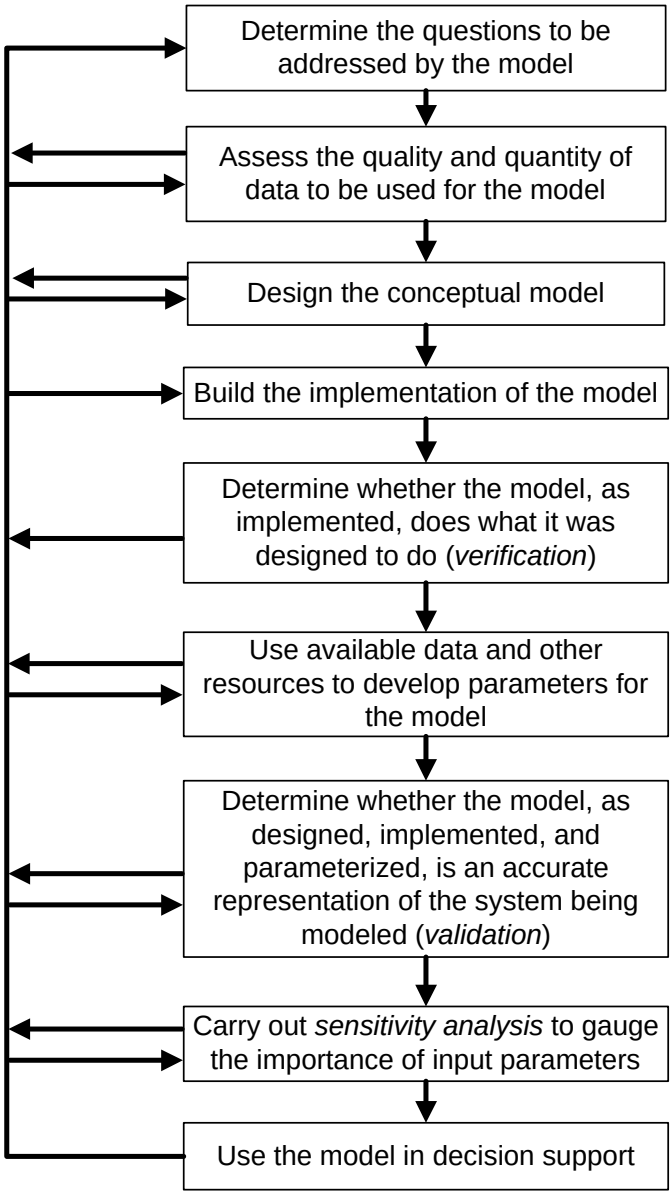
766 69. Waller, L.A., Smith, D., Childs, J.E., & Real, L.A. (2003). – Monte Carlo
767 assessments of goodness-of-fit for ecological simulation models. *Ecol. Model.* **164**,
768 49-63.

769 70. Whitner, R.B. & Balci, O. (1989). – Guidelines for selecting and using simulation
770 model verification techniques. *In* MacNair, E.A., Musselman, K.J., & Heidelberger,
771 P. (Eds), Proceedings of the 1989 winter simulation conference, 559-568.

772

773 **Figure 1.** Schematic diagram of the stages of model development, evaluation, and

774 application. Adapted from Dent and Blackie (8), Martin *et al.* (38), and Taylor (62).



775