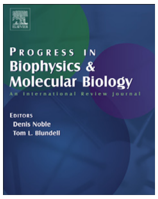
Progress in Biophysics and Molecular Biology xxx (2016) 1–7

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio



Progress in Biophysics & Molecular Biology

An International Review Journal

Editors

Denis Noble

Tom L. Blundell

A framework to establish credibility of computational models in biology

Eann A. Patterson [a,*], Maurice P. Whelan [b]

[a] School of Engineering, University of Liverpool, Liverpool, UK
[b] European Commission Joint Research Centre, Italy

ARTICLE INFO

Article history:
Received 16 April 2016
Received in revised form
18 July 2016
Accepted 1 August 2016
Available online xxx

Keywords:
Computational biology
Computational models
Credibility
Validation
In silico medicine
Biomedical modelling

ABSTRACT

Computational models in biology and biomedical science are often constructed to aid people's understanding of phenomena or to inform decisions with socioeconomic consequences. Model credibility is the willingness of people to trust a model's predictions and is often difficult to establish for computational biology models. A 3 × 3 matrix has been proposed to allow such models to be categorised with respect to their testability and epistemic foundation in order to guide the selection of an appropriate process of validation to supply evidence to establish credibility. Three approaches to validation are identified that can be deployed depending on whether a model is deemed untestable, testable or lies somewhere in between. In the latter two cases, the validation process involves the quantification of uncertainty which is a key output. The issues arising due to the complexity and inherent variability of biological systems are discussed and the creation of 'digital twins' proposed as a means to alleviate the issues and provide a more robust, transparent and traceable route to model credibility and acceptance.

Contents

1. Introduction

Whenever a model is developed, a primary concern of the modeller is the credibility of their model. Credibility has been described by Schruben (Schruben, 1980) as reflecting 'the willingness of persons to base decisions on information obtained from the model'. So, the issue becomes a matter of providing sufficient evidence of the model's fitness for purpose to induce this willingness. Rudner (1953) postulated that our judgement on the strength of the evidence depends on the importance or consequences of making a mistake, which implies that modellers need to consider the intended uses of their model when identifying the evidence required to underpin credibility.

Often in biology, as in other areas of pure science, the primary value of computational models is heuristic (Oreskes et al., 1994). They are representations of reality that are valuable for understanding and guiding further research or study. In these circumstances, when the role of the model is not associated with decision-making, its absolute accuracy is not the essential issue. Rather, it is more appropriate to consider computational models as the

[*] Corresponding author.
E-mail address: eann.patterson@liverpool.ac.uk (E.A. Patterson).

http://dx.doi.org/10.1016/j.pbiomolbio.2016.08.007
0079-6107/© 2016 Published by Elsevier Ltd.

Please cite this article in press as: Patterson, E.A., Whelan, M.P., A framework to establish credibility of computational models in biology, Progress in Biophysics and Molecular Biology (2016), http://dx.doi.org/10.1016/j.pbiomolbio.2016.08.007

apparatus or environment in which simulations or '*in silico*' experiments are performed for the purpose of exploring hypotheses and revealing features of behaviour for which only sparse or no observational data is available (Winsburg, 2010). If the revealing of features is a sufficient outcome, then an adequate process of model validation to underpin credibility could be to simply ensure that the model is useful and functional in providing relevant insights. This approach has been employed, for example, in materials science and termed 'validation of phenomena' (Patterson, 2015).

Biology overlaps with engineering when it is used to create man-made components and products or when engineered products interact with human biology, such as in pharmacology and toxicology. In these circumstances, when models are used, it would be appropriate to adopt the level of rigour employed routinely by the engineering sector to demonstrate their credibility. Engineers use computational models to evaluate and refine the performance, reliability and safety of designs of engineered products. Hence for these models, which might be termed predictive rather than heuristic, the consequence of making a mistake will be typically measured in socioeconomic costs, often significant, such as loss of life or injury. This implies the need for strong evidence that the computational model closely reflects reality, and leads to the definition of validation as 'determining the degree to which a model is an accurate representation of the real world from the perspective of its intended uses' (ASME V&V 10-2006, 2006). The engineering community has developed a series of quantitative validation procedures, (e.g. in solid mechanics (Sebastian et al., 2013)), that allow the evidence to be assembled in a framework that is recognised by modellers and end-users, (e.g. for solid mechanics models (CWA 16799, 2014)), and supports the establishment of credibility and confidence.

In *in silico* biology, when computational models are used to reveal features of behaviour, even the 'validation of phenomena' can be challenging in the absence of reliable data from the real-world, which of course is often the reason for wanting to use a model in the first place. Some computational models of biological systems would appear to be untestable due to their complexity and the difficulty in acquiring reliable data from the biological system. It is tempting at this point, to trust to the judgment of the modeller and accept that the simulation will provide interesting information. However, Hughes (1999) has said that *in silico* experiments reveal information about three types of world: the actual world, possible worlds and impossible worlds; and that it is not possible to know which has been revealed without taking an extra step, such as some form of validation. So, it would be inappropriate to abandon some effort to test the reliability of computational biology models. Thus, our aim is to develop a framework for establishing the credibility of computational biology models that are classified according to our ability to test them and identify their epistemological foundations, to support the work of both modellers and those making decisions based on results from models.

2. Credibility matrix

Untestable models are employed in physics and, to a lesser extent, engineering. Tegmark has drawn an epistemological boundary between physics and metaphysics that is defined by whether or not a theory is experimentally testable (Tegmark, 2014). While for engineering models, Patterson (Patterson, 2015) has gone further and constructed a 2 × 2 diagram that identifies the appropriate approach to establishing the credibility of testable and untestable or meta models based on whether they are principled or unprincipled, i.e. whether the underlying physics is known or unknown. In Fig. 1, we have developed this approach for use in computational biology and *in silico* medicine.

Sober (Sober, 1993) has stated that there are no exceptionless laws in biology. Notwithstanding that some would point to the first law of biology being ‘the tendency for diversity and complexity to increase in evolutionary systems’ (McShea and Brandon, 2010), it is clear that it is difficult to identify universally accepted biological laws. Thus, the use of principled and unprincipled on the horizontal axis is potentially problematic when referring to biology. Instead, in Fig. 1 the more general terms 'known biology' and 'unknown biology' have been used. The allocation of a model between these two categories should be made based on whether or not its knowledge base is founded on one of the three types of scientific reasoning (Osimani and Mignini, 2015), namely (i) inductive reasoning from empirical data to a theory, (ii) hypothesis falsification through *modus tollens*, or (iii) explanatory reasoning. These modes of reasoning are generic, and in biology it would be appropriate to embrace Hill's criteria for causation (Villeneuve et al., 2014). Computational biology models are unlikely to be as readily categorised as implied above, so it is appropriate to include a transition zone between models based on known biology and those based on unknown biology, i.e. between principled and unprincipled. For example, a model of a biological system is usually constructed by combining models of its sub-systems, each perhaps reflecting different scales of biological organisation, and each based on varying degrees of phenomenological understanding. Such models would be located in this transition zone (i.e. the middle column in Fig. 1) especially when the linkages between the subsystems are not understood.

In computational biology, at the boundary between testable and untestable models in Fig. 1, there will be another transition zone that originates from the difficulties in making quantitative observations of real-world biology, which leads to sparse or incomplete data. This is in part due to our inability to control the real-world, as observed by Viceconti (Viceconti, 2015).

The credibility of models that fall into the bottom left corner in Fig. 1 can be established using the type of quantitative validation procedures that are being codified by the engineering community

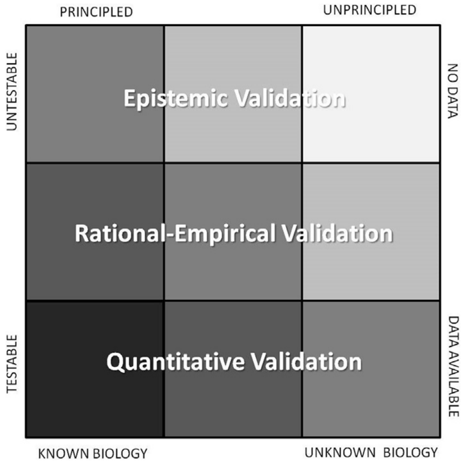


Fig. 1. a schematic diagram illustrating the relationship between testable and untestable models that are either based on known biology (i.e. principled) or unknown biology (i.e. unprincipled) together with the approaches to performing a validation and the resultant level of credibility that can be established indicated by the greyscale. Testable models are those for which it is possible to acquire measured data from real-world experiments, while untestable models are those for which it is not possible to make measurements corresponding to the model's predictions. Epistemic validation is based on the epistemic values of the model including simplicity, consistency and explanatory power; rational-empirical validation involves a series of three ‘tests’ using rationalism, empiricism and demonstration of predictive accuracy; while quantitative validation employs the rigorous methods described in engineering standards.

Please cite this article in press as: Patterson, E.A., Whelan, M.P., A framework to establish credibility of computational models in biology, Progress in Biophysics and Molecular Biology (2016), http://dx.doi.org/10.1016/j.pbiomolbio.2016.08.007

[e.g. ASME V&V 10-2006 (2006), CWA 16799 (2014)]. The fact that the underlying principles of the model are understood has the consequence that the gathering of observational data can be selective and efficient. The bottom right corner is a little more difficult because the principles are unknown. However, when sufficient observational data is available that covers the full parameter space and applicability domain of the model, then it is possible to perform a quantitative validation without demonstrating the legitimacy of assumptions underlying the model. This is reminiscent of the approach taken by Milton Friedmann, the economist, who proposed that the veracity of a model depends not on its assumptions but on its ability to predict the behaviour of the dependent variables treated by the model (Friedmann, 1953).

In the context of climate modelling, where sparse observational data is available in the near-term time domain and reliable detailed predictions in the far-term time domain are required, Biddle and Winsberg (2010) have proposed that models with epistemic properties are more likely to be appropriate than others. Epistemic values include simplicity, explanatory power, and internal and external consistency. This validation approach is viable for the top left corner in Fig. 1, where a model is principled, but becomes less so as one ventures into the top right corner. Here, unprincipled and untestable models do not lend themselves to validation and possess radical uncertainty or unforeseeable outcomes (Roth, 2009) that renders them inappropriate for supporting decisions that have socioeconomic consequences. In summary, the degree of belief in a model tends to increase with the body of evidence available to support it (Audi, 2011) based, in this context, on varying degrees of phenomenological understanding and observational data. Thus the level of credibility for models, which can be established through validation, tends to decrease radially outwards from the bottom left corner of Fig. 1 (as indicated by the greyscale), i.e. it is easier to establish credibility for models that are principled and testable than for those that are unprincipled and untestable.

3. Model validation

Validation is the process of establishing the degree to which a model is an accurate representation of the real-world for its intended purpose. A model does not pass or fail a validation process, because no decision is involved. Instead, a validation process provides information that allows others to make a decision on the acceptability or trustworthiness of the predictions from a model. As discussed earlier, a predictive model whose outputs will inform decisions with consequences requires a rigorous demonstration of its fitness for purpose, often involving a ‘normative-based’ validation that includes the quantification of model uncertainty. On the other hand, an informative model whose purpose is heuristic carries a lower burden of demonstration of its credibility. In general, informative models lie in the top row in Fig. 1, and when modelling in support of decision-making with socioeconomic consequences, it is advisable to avoid models that fall near the top right corner. However, it seems entirely appropriate for computational biologists to deploy such models in scientific endeavor and the pursuit of a greater understanding of biological systems. The use of epistemic properties to establish credibility for models in the top row in Fig. 1 is perhaps the only viable approach to validation and might be termed ‘epistemic validation’. Hence, a more detailed consideration here of the epistemic properties of a biological model in terms of simplicity, explanatory power and consistency is worthwhile and appropriate.

Simplicity appears to be an intuitively appropriate property of a reliable model based on Occam's razor and likely to support a willingness in others to trust a model, i.e. to support credibility. Wimsatt (1981) has proposed that a false assumption in a model does not matter with respect to the phenomenon of interest, if it can be shown that replacing it with a more realistic one does not change the answer provided by the model to the question of interest. This allows models to be simplified and perhaps justifies the epistemic value of simplicity. The property of explanatory power echoes the Friedmann approach described above, in the sense it implies that if a model predicts a phenomenon then it is more likely to be correct than if it does not; but, this is not particularly helpful since a model that does not have explanatory power is not useful and would be rejected anyway. The requirement for consistency is a better test and ideally should be applied both internally and externally. External consistency with observations is not viable for untestable models, due to the absence of data, but consistency with other models is a feasible test and is the first of a series of strategies which Franklin (Franklin, 1986) has identified that experimentalists use to ascertain the reliability of data from experiments. It is suggested that it is good practice to adopt his remaining strategies for *in silico* experiments, which are to demonstrate that (a) the experiment produces an already known result; (b) when perturbed the experiment produces the expected result; (c) the experiment is able to detect artifacts known to exist; (d) effects disappear when expected to do so and (e) when all plausible sources of error are eliminated the remaining observations must be real.

The nature of untestable and unprincipled models renders the application of some of Franklin's strategies problematic due to the lack of physical observations and makes it necessary to distinguish between cases in which there is no quantitative data, i.e. measurement is impossible, and no observation of any type is feasible, in other words even qualitative data cannot be acquired. In these latter conditions, Winsberg (Winsburg, 2010) has proposed that the theoretical ancestry of a simulation carries a heavy burden in providing credibility. Laying out the theory is the first stage in modelling, a process that Hacking terms 'speculation', and then the model needs to be built, which Hacking calls 'calculation' (Hacking, 1983). Model building usually involves discretization of the problem, possibly in many domains including time and space, in order to make it tractable with the computing resources available. Discretization converts what might be an analytical, or at least mathematically, continuous theory into an approximate estimation in a computational model and hence can be a major source of error. In addition, in complex multi-scale models, several layers of discretization may be necessary to describe the structure and function of the complete system. The connections between these discrete sub-models are often cobbled-together fictions or ‘kluge’ that add substantially to the uncertainty in the predictions. In these circumstances, model credibility must come not only from the 'ancestry' of the theory, which will not be available for unprincipled models, but also from the established credentials of the model-building techniques, leading to what are known as ‘self-vindicating’ models (Winsberg, 2003). The process of demonstrating that a model has epistemic properties will not constitute a quantitative validation because the degree to which the model represents the real world cannot be quantified in this process. Instead, qualitative evidence is accumulated and can be used to establish credibility for the model as a heuristic tool for exploring and extending our understanding of phenomena.

The quantitative approaches to validation associated with principled testable models can be applied to models across the bottom of the diagram in Fig. 1. The reliability of the predictions from unprincipled models can be demonstrated in the same manner as for principled ones, that is, through quantitative comparisons of comprehensive datasets. The comparative analysis of measured and predicted datasets will allow the degree to which the model represents reality for its intended uses to be established in terms of probability distributions. In this region of the matrix the

Please cite this article in press as: Patterson, E.A., Whelan, M.P., A framework to establish credibility of computational models in biology, Progress in Biophysics and Molecular Biology (2016), http://dx.doi.org/10.1016/j.pbiomolbio.2016.08.007

approach mimics that employed in engineering though it may be more complicated due to variability and complexity of biological systems. These approaches are described in detail elsewhere, for instance by Roy and Oberkampf (Roy and Oberkampf, 2011), and so are not discussed further here. Overall, the level of credibility of a model will increase with the body of evidence supporting it (Audi, 2011), and even though a lack of principled-knowledge can be compensated to some extent with more observational data, and vice versa, the most credible models will be underpinned by an optimal balance of both principled knowledge and observational data.

The validation of models falling in the zone between the testable and untestable models is more problematic and more likely to be required by computational biologists because in many cases the quantity of measured data will be small or sparse. In circumstances in which we have incomplete information about the real-world behaviour of a system, Naylor et al. (Naylor et al., 1967) proposed a three-step approach based on rationalism, empiricism and demonstration of predictive accuracy. Rationalism is the belief that a model is simply a system of logical deductions from a series of synthetic premises of unquestionable truth, where a synthetic premise is a proposition based on observation rather than analysis. So the first step involves identifying the relevant deductions and premises, which is equivalent to identifying whether or not the model lies on the principled or unprincipled side of the matrix in Fig. 1. In this context, Viceconti has provided a useful taxonomy of predictive models (Viceconti, 2011). Empiricism refutes any postulates or assumptions that cannot be independently confirmed, hence step two involves the testing of individual assumptions used in constructing the model. These assumptions should be tested against real-world observations in order to establish their validity. The third step is to confirm the predictive ability of the model by making quantitative comparisons to the available data from the real-world regardless of whether the first and second step gave completely positive outcomes, i.e. following the approach championed by Friedmann. We would recommend that at each step, or as a fourth step, the uncertainties associated with the model are identified and characterised. If this is performed as a fourth step then the six sources of model uncertainty identified by Roy and Oberkampf (Roy and Oberkampf, 2011) form a suitable framework for estimating total uncertainty.

4. Acceptance

The credibility of a model is in the gift of the decision-maker and not the modeller. This is because credibility is about the willingness of others to make decisions based on the outcome from the model, i.e. its predictions. The degree of belief in a model increases with the body of evidence supporting its claim for being an acceptable representation of reality for the purposes of the decision. The modeller is rarely the decision-maker and indeed Jeffrey (Jeffrey, 1956) argued that the scientist's proper role is to provide rationale agents in society with the probabilities related to a hypothesis, and the rationale agents, or decision-makers, then make a decision. This is an approach that Biddle and Winsberg (2010) have applied to climate change modelling and it seems appropriate to use in the context of *in silico* biology and medicine, where in the latter case regulators, clinicians and patients are likely to be the rationale agents.

It is relevant to consider the process by which a decision-maker should review the information provided by the modeller since this should influence the behaviour of the modeller. In some areas of engineering and applied science a simulation review is conducted and Kaizer et al. (Kaizer et al., 2015) have described the process in the nuclear industry, which is sufficiently generic to be applied in computational biology. The supporting evidence is analysed to determine (a) the trustworthiness of the results of the simulation and (b) the level of trustworthiness required for the intended purpose. Then, based on these two pieces of information, a decision is made on whether to trust the specific simulation for the intended purpose. Clearly this process becomes easier as the body of evidence grows so that reviews of simulations based on predictive models that lie in the bottom left corner of Fig. 1 are more likely to lead to positive outcomes than for models in other areas of the figure.

The above discussion implies that it is unlikely, for example, that an initial and novel computational biology model of a virtual tissue would be accepted by a regulator making decisions on licensing a pharmaceutical. It is probable that the model would be unprincipled because we do not understand all of the mechanisms involved in the underlying pathways and processes, and it might also be untestable in the real-world, i.e. in humans. Such a model would likely lie towards the top right corner in Fig. 1 with a very low prospect of achieving credibility and acceptance, though it might enhance our understanding of important phenomena. It is likely that this exemplar model would consist of a number of sub-elements or sub-models in order to represent the complexity of the process and we can exploit this network of sub-elements to create a knowledge base that improves the prospect of establishing credibility and acceptance. Recently, Villeneuve et al. (2014) have described strategies for developing Adverse Outcome Pathways (AOPs) that can be considered as models of toxicological processes. AOPs are built from knowledge about biologically plausible, and empirically supported, links between a molecular-level perturbation of a biological system and an adverse outcome at an organ or organism level. AOPs consist of a molecular-initiating event (MIE) and an adverse outcome connected by a series of key events (KE) and key event relationships (KER). Key events are measurements of change in a biological state that are indicative of progress towards the adverse outcome, while key event relationships are causal associations between the key events. Putative AOPs usually start out towards the top right corner in Fig. 1, but move to the bottom right corner as the identification of the key events proceeds because, by definition, the key events have to be measurable. The identification of the key event relationships that are biologically plausible and supported by empirical data transfers the AOP to the bottom left corner because the model(s) embedded in an AOP can be considered principled, i.e. the biology is known. Hence, the process of developing an AOP or a similar knowledge-based model, such as a disease AOP (Langley et al., 2015) or therapeutic outcome pathway, offers an approach with the potential to substantially increase the prospects of establishing credibility. Of course, translations across the diagram towards the bottom left corner in Fig. 1 bring with them an implied requirement to adjust the validation approach appropriately.

5. Discussion

Most biological systems exhibit behaviour that is characterised by feedback loops and non-linearity. The solutions of the resultant non-linear equations describing such systems are not single values but series of values that are often represented in state (or phase) space as patterns or portraits, known as attractors. There are three classes of attractor: point, periodic and strange. Point attractors describe systems in stable equilibrium and periodic attractors those oscillating in a periodic equilibrium. Strange attractors correspond to systems exhibiting chaotic behaviour and may have bifurcation points in the phase portrait at which system behaviour changes and a new order is established (Capri and Luisi, 2014). Examples of the use of non-linear systems models (e.g. (Lachowicz, 2011; Ghergu and Radulescu, 2012)) are not commonplace in computational biology and the use of phase portraits or similar approaches (e.g. Gross et al. (2011)) to represent system behaviour is rare. However, the implications for validation protocols are serious because the model is no longer deterministic but will produce a set of solutions described by the phase portrait and it is likely that our real-world measurements will only represent a tiny fraction of the possible locations in the phase portrait. Indeed, without a precise knowledge of the initial starting conditions we will not even know where to look in the phase portrait. As a consequence, the accuracy of a non-linear dynamics model is likely to be underestimated and the apparent variability in the experimental data cause it to be discarded or down-graded when actually it simply represents different locations in state space. O'Leary et al. (2015) have identified the need to develop alternative strategies for fitting data accounting for dynamic state estimation but we need to go a step further and develop data comparison methods with similar capabilities.

Biological systems are complex in the sense that they exhibit non-trivial emergence and self-organising behaviours (Mitchell, 2009). In 1925 C.D. Broad used the term 'emergent property' to describe behaviour that emerges at a certain level of complexity but does not exist at lower levels (Broad, 1925). This has implications in hierarchical and multi-scale modelling and DeLanda (2011) has discussed how emergent properties allow simulations to decompose reality and replicate phenomena at one scale without the need for high fidelity representation at all scales. This implies that when we are interested in the behaviour of a complex system we do not need to model all of its components nor do we need to validate the performance of the models of the sub-systems. It is clear that cognizance of emergent behaviour is always necessary to produce a reliable model. However, when high fidelity modelling is not undertaken at all scales it might mask, intentionally or otherwise, a lack of knowledge of the relevant biology which would place the model in the right half of Fig. 1 with the associated credibility issues.

The underlying issue is that the traditional approach to modelling complex systems is based on Cartesian reductionism, which is the concept that everything about a complex system can be understood by reducing it to the smallest constituent part. In the context of natural systems, a move away from reductionism towards a systems view has been championed by Capri and Luisi who have stated that a reductionist approach has revealed much about the subunits within the cell but nothing about the balancing of their independent metabolic pathways and cycles (Capri and Luisi, 2014). This shift away from a reductionist approach presents new challenges to establishing model credibility because in a reductionist approach the emphasis is usually on measuring and quantifying behaviour, whereas with a systems view the focus is on the mapping of relationships and behavioural patterns. The complexity of biological systems implies that an exact knowledge of relationship and patterns will usually not be attainable and we must be satisfied with an approximate knowledge. In other words, we cannot be certain about the behaviour of the system of interest. There will always be a degree of uncertainty or a probability that the system will behave in a particular way. Experiments *in vivo*, *in vitro* and *in silico* can increase our understanding and Bayesian techniques can be used to update our probabilistic knowledge of a system's behaviour as new information becomes available (Osimani and Mignini, 2015). Bayesian networks (e.g. Needham et al. (2007)) and computation (e.g. Liepe et al. (2014); Gasche, Mahevas, Marchal) have found applications in computational biology and a recent review (Price et al., 2014) by a working group of the Drug Information Association supported the use of Bayesian methods for the design and analysis of safety trials for new pharmaceutics. However, further work is necessary to establish viable methods to handle these issues.

Recently, Carusi (2014) has discussed the impact of inherent variability in biological systems on validation which she identifies as arising at three levels: sub-cellular, cellular variability between collocated cells of the same type and cellular variability between non-collocated cells of the same type. At a more fundamental level, the randomness of chemical reactions based on chance collisions of molecules ensures that some randomness or variability will always be present. This generates an irreducible or aleatory uncertainty that can be characterised by a probability density distribution, whereas our lack of knowledge about the system generates an epistemic uncertainty that can be reduced by acquiring more knowledge. The implication for model validation and credibility is that these uncertainties need to be identified and characterised, and then utilised in decision-making, which emphasises the importance of reporting uncertainties as a key output of the validation process.

Viceconti et al. (2005) discussed the verification and validation of finite element models intended to be used to extract clinically relevant data. They proposed that verification, i.e. ensuring that the mathematics of the model is implemented correctly, was the responsibility of individual scientists while validation involves the whole community. In the engineering community, verification is performed by the commercial suppliers of the software, usually using internationally accepted benchmarks and this would seem to be an appropriate approach to adopt in computational biology. However, a community-owned approach to validation, including the sharing of collections of models and validation-quality data would lead to more robust validations, particular when diverse data sets are available. In turn, such validation processes would possess a high degree of transparency and traceability, and hence would contribute considerably to credibility and acceptance of models. Going further, such validated models combined with their associated real-world datasets could evolve to become digital twins of biological systems particular to specific sub-populations or even individuals. The concept of digital twins probably originated in the aerospace industry (Glaessgen and Stargel, 2012) and its deployment in the civil nuclear industry has been proposed recently (Patterson et al., 2016). In these industry sectors, the advantages of digital twins have been identified as shortening development times, reducing costs and improving safety, reliability and usability. It is reasonable to expect similar advantages to accrue in biomedical science and engineering.

It is important to appreciate validation is a process that, in some ways, is never completed. This is because the outcome of the validation process is a statement about the degree to which a model is an accurate representation of the real-world for the intended purposes of the model, and the level of representation can nearly always be improved. The outcome from a quantitative validation process includes a statement about the uncertainties associated with the predicted and measured data used in the process, which together allow the accuracy of the representation to be expressed in terms of probabilities. Subsequently, the process of establishing credibility in a model is two-way exchange between the decision-maker, or end-user, and the modeller, which can be described as a process of social epistemology, i.e. arriving at a collective understanding of the value of the model. When the predictions of the model are to be used to support decisions with socioeconomic consequences then it is likely that a high level of confidence and low levels of uncertainty will be demanded by the decision-maker. However, increases in confidence levels and reductions in uncertainty will usually have resource implications in terms of updating and refinement of the model and, or the acquisition of better quality data from the real-world. Hence a cost-benefit analysis will often need to be performed. In practice, biological data is rarely acquired from the real-world but instead from an experiment that is itself a representation of reality and so consideration needs to given to the extent to which the experiment is an accurate representation of the real-world for the intended purposes of the model. Here, it is important that the experiments are designed in collaboration with the modeller in another process of social epistemology, i.e. a two-way exchange leading to a shared understanding of the model and reality. Some people have described this process as a cycle, or as part of model-simulation-experiment system (Carusi et al., 2012), which is sometimes portrayed as circular. We would suggest that it is more akin to cycles of oscillation as shown schematically in Fig. 2. The interaction between the design of the model and experiment starts with the identification of initial conceptual ideas leading to an initial or first model that can be used to design an appropriate experiment and leads to the acquisition of measured data, i.e. experience of reality. This data can be used to evaluate the performance of the model, including its validation, which permits the interpretation of the predicted data that can be used to revise the conceptual ideas. The cycle is repeated so that progress is made towards both an increased conceptual understanding and knowledge of reality. This parallel progress towards conceptual understanding and knowledge of reality can be the outcome of both heuristic and predictive modelling. This social epistemological approach can lead to concerns about so-called 'double-counting' of measured data when it is used for both calibration or updating of models and in the validation process, particularly when data is limited as for models lying in the middle row of Fig. 1. Carusi and her co-workers (Carusi, 2014; Carusi et al., 2012) avoid this issue by performing validation using experimental datasets that are independent of those used in the construction of the model. However, Steele and Werndl, (2013, 2016) have argued that these practices are justifiable within a Bayesian framework and with appropriate reference to the logic of confirmation or validation, as we have discussed.

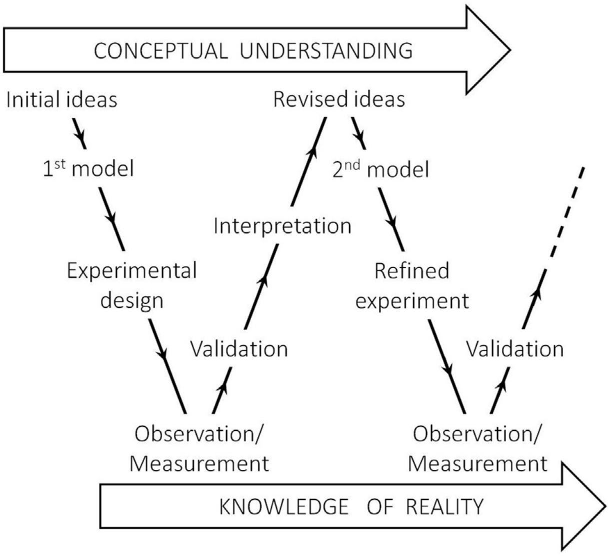


Fig. 2. a schematic representation of a social epistemological approach to advancing conceptual understanding and knowledge of reality through the integrated use of modelling and experiments. Conceptual understanding and knowledge of reality are developed in parallel through using conceptual understanding to create ideas that form the basis of models which inform experiment design and are validated using observations and measurements of reality that also enhance our knowledge of reality.

6. Conclusions

Model credibility is about the willingness of people to make decisions based on the predictions from the model. It is the responsibility of the modeller to provide evidence of the trustworthiness of the model which should include the results of a validation process. The decision-maker should review both the evidence provided in support of the model's trustworthiness and the level of trustworthiness required for the intended purpose, and then using these two pieces of information, make a decision on the acceptability of the predictions.

In computational biology and biomedicine these processes are often not straightforward because of the difficulties associated with acquiring experimental data to support the validation process and establishing the veracity of the premises on which the models are constructed. A framework based on a 3 × 3 matrix has been presented that allows models to be categorised according to their testability and basis in known biology, or epistemic foundations. The ease with which credibility can be established increases with both testability and the strength of the epistemic foundations. Strategies for demonstrating the fitness for purpose of a model have been proposed according to whether the available data from reality is potentially unlimited, sparse or unobtainable. These strategies have been termed: quantitative, rationale-empirical and epistemic respectively and provide decreasing levels of quantitative information about the probability that the model is an accurate representation of reality for its intended purpose or application.

In order to increase the prospects for establishing credibility and acceptance of a model, it is recommended that modellers should adopt strategies to increase the testability and strengthen its epistemic foundation. It has been illustrated how the use of a knowledge driven approach to model building, such as the Adverse Outcome Pathway (AOP) can support this strategy by assisting in translating a model of a complex system from being untestable and unprincipled, with a low prospect of establishing credibility or being accepted, into a set of testable, principled models with a good prospects for credibility and acceptance.

Acknowledgements

EAP is a recipient of a Royal Society Wolfson Research Merit Award.

References

ASME V&V 10-2006, 2006. Guide for verification & validation in computational solid mechanics. American Society of Mech. Engineers, New York.
Audi, R., 2011. Epistemology: a Contemporary Introduction to Theory of Knowledge, third ed. Routledge, New York.
Biddle, J., Winsberg, E., 2010. Value judgments and the estimation of uncertainty in climate modeling. In: Magnus, P.D., Busch, J. (Eds.), New Waves in Philosophy of Science. Palgrave MacMillan, Basingstoke.
Broad, C.D., 1925. Mind and its Place in Nature. Routledge & Kegan Paul, London.
Capri, F., Luisi, P.L., 2014. The Systems View of Life: a Unifying Vision. Cambridge University Press, , Cambridge.
Carusi, A., 2014. Validation and variability: dual challenges on the path from system biology to systems medicine. Stud. Hist. Philos. Biol. Biomed. Sci. 48, 28–37.
Carusi, A., Burrage, K., Rodriguez, B., 2012. Bridging experiments, models and simulations: an integrative approach to validation in computational cardiac electrophysiology. Am. J. Physiol. Heart Circ. Physiol. 303, H144–H155.
CWA 16799, 2014. Validation of Computational Solid Mechanics Models. Comité Européen de Normalisation (CEN). CEN Workshop Agreement, 2014.
DeLanda, M., 2011. Philosophy and Simulation: the Emergence of Synthetic Reason. Continuum International Publishing Group, London.
Franklin, A., 1986. The Neglect of Experiment. Cambridge University Press, Cambridge.
Friedmann, M., 1953. The methodology of positive economics. In: Essays in Positive Economics. Chicago University Press, Chicago.
Gasche, L., Mahevas, S., Marchal, P., 2013. Supporting fisheries management by means of complex models: can we point out isles of robust on a sea of uncertainty? PLoS One 8 (10), e77566.

Ghergu, M., Radulescu, V.D., 2012. Pattern formation and the Gierer-Meinhardt model in molecular biology. In: Nonlinear PDEs: Mathematical Models in Biology, Chemistry and Population Genetics. Springer, Berlin.

Glaessgen, E.H., Stargel, D.S., 2012. The digital twin paradigm for future NASA and US Air Force vehicles. In: Proc 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference. AIAA paper 2012-2018, NF1676L–13293.

Gross, F., Metzner, G., Behn, U., 2011. Mathematical modelling of allergy and specific immunotherapy Th1-Th2-Treg interactions. J. Theor. Biol. 269, 70–78.

Hacking, I., 1983. Representing and Intervening: Introductory Topics in Philosophy of Science. Free Press, New York.

Hughes, R., 1999. The Ising model, computer simulation, and universal physics. In: Morgan, M.S., Morrison, M. (Eds.), Models as Mediators: Perspectives on Natural and Social Science. Cambridge University Press, Cambridge.

Jeffrey, R.C., 1956. Valuation and acceptance of scientific hypotheses. Philos. Sci. 22, 237–246.

Kaizer, J.S., Heller, A.K., Oberkampf, W.L., 2015. Scientific computer simulation review. Reliab. Eng. Syst. Saf. 138, 210–218.

Lachowicz, M., 2011. Individually-based Markov processes modelling nonlinear systems in mathematical biology. Nonlinear Anal. Real World Appl. 12, 2396–2407.

Langley, G., Austin, C.P., Balapure, A.K., Birnbaum, L.S., Bucher, J.R., Fentem, J., Fitzpatrick, S.C., Fowle, J.R.I.I.I., Kavlock, R.J., Kitano, H., Lidbury, B.A., Muotri, A.R., Peng, S.-Q., Sakharov, D., Seidle, T., Trez, T., Tonevitsky, A., van de Stolpe, A., Whelan, M., Willett, C., 2015. Lessons from toxicology: developing a 21st-Century paradigm for medical research. Environ. Health Perspect. 123 (11), A268–A272.

Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C.P., Stumpf, M.P.H., 2014. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nat. Protoc. 9 (2), 439–456.

McShea, D.W., Brandon, R., 2010. Biology's First Law: the Tendency for Diversity and Complexity to Increase in Evolutionary Systems. Chicago University Press, Chicago.

Mitchell, M., 2009. Complexity: a Guided Tour. Oxford University Press, Oxford.

Naylor, T.H., Finger, J.M., McKenney, J.L., Schrank, W.E., Holt, C.C., 1967. Verification of computer simulation models. Manag. Sci. 14 (2), B92–B106.

Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R., 2007. A primer on learning in Bayesian networks for computational biology. PLoS Comput. Biol. 3 (8), 1409–1416.

Oreskes, N., Shrader-Frechette, Belitz, K., 1994. Verification, validation and confirmation of numerical models in Earth Sciences. Science 263, 641–646.

Osimani, B., Mignini, F., 2015. Causal assessment of pharmaceutical treatments: why standards of evidence should not be the same for benefits and harms? Drug Saf. 38, 1–11.

O'Leary, T., Sutton, A.C., Marder, E., 2015. Computational models in the age of large datasets. Curr. Opin. Neurobiol. 32, 87–94.

Patterson, E.A., 2015. On the credibility of engineering models and meta-models. J. Strain Anal. 50 (4), 218–220.

Patterson, E.A., Taylor, R.J., Bankhead, M., 2016. A framework for an integrated nuclear digital environment. Prog. Nucl. Energy 87, 97–103.

Price, K.L., Xia, H.A., Lakshminarayanan, M., Madigan, D., Manner, D., Scott, J., Stamey, J.D., Thompson, L., 2014. Bayesian methods for design and analysis of safety trails. Pharm. Stat. 13, 13–24.

Roth, W.-M., 2009. Radical uncertainty in scientific discovery work, Science. Technol. Hum. Values 34 (3), 313–336.

Roy, C.J., Oberkampf, W.L., 2011. A comprehensive framework for verification, validation and uncertainty quantification in scientific computing. Comput. Methods Appl. Mech. Eng. 200, 2131–2144.

Rudner, R., 1953. The scientist qua scientist makes value judgements. Philos. Sci. 20, 1–6.

Schruben, L.W., 1980. Establishing the credibility of simulations. Simulation 34, 101–105.

Sebastian, C., Hack, E., Patterson, E.A., 2013. An approach to validation of computational solid mechanics models for strain analysis. J. Strain Anal. 48 (1), 36–47.

Sober, E., 1993. Philosophy of Biology. Westfield Press, Boulder, CO.

Steele, K., Werndl, C., 2013. Climate models, confirmation and calibration. Br. J. Philos. Sci. 64, 609–635.

Steele, K., Werndl, C., 2016. Model tuning in engineering: uncovering the logic. J. Strain Anal. 51 (1), 63–71.

Tegmark, M., 2014. Our Mathematic Universe: My Quest for the Ultimate Nature of Reality. Allen Lane, London.

Viceconti, M., 2011. A tentative taxonomy for predictive models in relation to their falsifiability. Phil. Trans. R. Soc. A 369, 4149–4161.

Viceconti, M., 2015. Biomechanics-based in silico medicine: the manifesto of a new science. J. Biomech. 48, 193–194.

Viceconti, M., Olsen, S., Nolte, L.-P., Burton, K., 2005. Extracting clinically relevant data from finite element simulations. Clin. Biomech. 20, 451–454.

Villeneuve, D.L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T.H., LaLone, C.A., Landesmann, B., Lettieri, T., Munn, S., Nepelska, M., et al., 2014. Adverse outcome pathway (AOP) development II: best practices. Toxicol. Sci. 142 (2), 321–330.

Wimsatt, W., 1981. Robustness, reliability and over-determination. In: Brewer, M., Collins, B. (Eds.), Scientific Inquiry and the Social Sciences. Jossey-Bass, San Francisco.

Winsberg, E., 2003. Simulated experiments: methodology for a virtual world. Philos. Sci. 70, 105–125.

Winsburg, E., 2010. Science in the Age of Computer Simulation. The University of Chicago Press, Chicago.

Please cite this article in press as: Patterson, E.A., Whelan, M.P., A framework to establish credibility of computational models in biology, Progress in Biophysics and Molecular Biology (2016), http://dx.doi.org/10.1016/j.pbiomolbio.2016.08.007