Position paper: Extending Credibility Assessment of In Silico Medicine Predictors to Machine Learning Predictors

*

Position paper: Extending Credibility Assessment of In Silico Medicine Predictors to Machine Learning Predictors

Marco Viceconti, Filippo Lanubile, Antonella Carbonaro, Sabato Mellone, Cristina Curreli, Alessandra Aldieri, Saverio Ranciati, and Angela Montanari

***Abstract*— There are several situations where it would be convenient if a quantity of interest essential to support a medical or regulatory decision could be predicted as a function of other measurable quantities rather than measured experimentally. To do so, we need to ensure that in all practical cases, the predicted value does not differ from what we would measure experimentally by more than an acceptable threshold, defined by the context in which that quantity of interest is used in the decision-making process. This is called *Credibility Assessment*. Initial work, which guided the elaboration of the first technical standard on the topic (ASME VV-40:2018), focused on predictive models built from available mechanistic knowledge of the phenomenon of interest. For this class of predictive models, sometimes called biophysical models, a credibility assessment practice based on the so-called verification, Validation, Uncertainty, Quantification and Applicability (VVUQA) analysis is accepted. This position paper aims to summarise, through theoretical considerations, a complex debate on whether such an approach can be extended to predictive models built without any mechanistic knowledge (machine learning (ML) predictors). We conclude that the VVUQA can be extended to ML-based predictors; however, since there is no certainty that the features used to predict the quantity of interest are necessary and sufficient, according to the VVUQA framework, such credibility assessment is limited to the test sets used for the validation studies. This calls for a Total Product Life Cycle approach, where periodic retesting of ML-based predictors is part of post-marketing surveillance to ensure that no "unknown bias" may play a role.**

***Index Terms*— In Silico methodologies, Credibility of Predictors, Machine Learning, Total Product Life Cycle.**

I. Introduction

In Silico Medicine, defined as the use of computer modelling and simulation in healthcare, is growing exponentially. But before a predicted quantity can be used to support a clinical decision (Digital Twins in Healthcare) or to assess the safety and/or the efficacy of a new intervention (In Silico Trials), the *credibility* of the predictor must be demonstrated and documented. The process through which the credibility of a predictor is assessed depends on the predictive model form; in particular, it is quite different if the model is built using *explicit knowledge*[1], for example, codified in differential equations (hereinafter referred to as biophysical models) or using implicit knowledge inferred from a data set using some training process (hereinafter referred to as Machine Learning (ML) models). While there are now procedures to assess the credibility of biophysical models for in silico medicine, based on the well-established engineering practice of verification, validation, uncertainty quantification and applicability analysis (hereinafter simply VVUQA), how to extend such practices to assess the credibility of machine learning models remains an open problem.

To date, the only technical standard that codifies a process to

*Manuscript received 24 April 2024.

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, "DARE - DigitAl lifelong pRevEntion" initiative, code PNC0000002, CUP: B53C22006450001. Some concepts exposed here were first discussed in a public workshop organised in March 2024 by the H2020 project "In Silico World: Lowering barriers to ubiquitous adoption of In Silico Trials" (topic SC1-DTH-06-2020, grant ID 101016503).

Marco Viceconti (corresponding author) is with the Department of Industrial Engineering, Alma Mater Studiorum - University of Bologna (IT), and with the Medical Technology Lab, IRCCS Istituto Ortopedico Rizzoli, Bologna (IT) (e-mail: marco.viceconti@unibo.it).

Filippo Lanubile is with the Department of Informatics, University of Bari (IT) (e-mail: filippo.lanubile@uniba.it).

Antonella Carbonaro is with the Department of Computer Science and Engineering, Alma Mater Studiorum - University of Bologna (IT) (e-mail: antonella.carbonaro@unibo.it).

, Sabato Mellone is with the Department of Computer Science and Engineering, Alma Mater Studiorum - University of Bologna (IT), and with the Medical Technology Lab, IRCCS Istituto Ortopedico Rizzoli, Bologna (IT) (e-mail: sabato.mellone@unibo.it).

Cristina Curreli is with the Medical Technology Lab, IRCCS Istituto Ortopedico Rizzoli, Bologna (IT) (e-mail: cristina.curreli@ior.it).

Alessandra Aldieri is with PolitoBIOMedLab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino (IT) (e-mail: alessandra.aldieri@polito.it).

Saverio Ranciati and Angela Montanari are with the Department of Statistical Sciences, Alma Mater Studiorum – University of Bologna (IT) (e-mail: saverio.ranciati2@unibo.it; angela.montanari@unibo.it).

[1] We use the term **explicit knowledge** to indicate the causal knowledge of a physical phenomenon produced through the scientific method (for example, the second law of dynamics). We use the term **implicit knowledge** to indicate the causal knowledge that may be hidden in a dataset and that we can try to infer through machine learning procedures.

assess the credibility of an in silico model used in medicine is the ASME VV-40:2018 [1]. The standard proposes a risk-based selection of the credibility targets for computational models used to support medical device development and evaluation, which are then assessed using the typical VVUQA practice for biophysical models. De facto, while the concept of risk-based credibility assessment is general, the concrete application of the VV-40:2018 is currently limited to biophysical models for which this practice is defined. There is now extensive literature documenting the VV-40 to various types of biophysical in silico medicine predictors [2], [3], [4], [5], [6].

While the VVUQA emerged as an engineering practice, there is now a theoretical framing [7] that can be used to explore extensions of this paradigm to other model forms. For example, such an approach was used to explore how the verification process needs to be revised when the biophysical model is implemented through a rule-based finite state machine (agent-based model) rather than with differential equations [8].

The authors participated in several formal and informal debates organised by the In Silico World Project[2], the DARE project[3], and the EDITH support action[4], where it was discussed the possibility and the difficulties in extending the VVUQA practice and its theoretical framing to ML-based predictors. With this position paper, we aim to condense the consensus from these debates into a theoretically robust form.

II. Definitions

A complete set of definitions used in this manuscript can be found in Annex 1. For a System of Interest (SI), we define the class of quantities that can be measured (Ω), of which one member called Quantity of Interest (ω) is what we want to predict and use the prediction for a specific purpose (*Context of Use*), for which the prediction must have a minimum accuracy (*acceptable threshold*) to be useful. We assume that a subset of Ω (feature set) $\mathbf{X}$ can be identified so that $\omega = f(\mathbf{X}(t), t)$. We expect the error in predicting ω always to be smaller than the acceptable threshold for all cases of interest over the entire region of validity of the predictor.

III Assumptions

To simplify the treatment, in the following we make some assumptions:

- In general, ω is a vector of quantities, but here, to keep the treatment easier, we assume ω to be a scalar quantity; the generalisation to a vector of outputs poses some challenges (especially if the outputs are correlated), but as a starting point for this discussion on the credibility, we make such assumption.
- We observe a physical phenomenon by measuring through controlled experiments the value that several quantities representative of the phenomenon (features) assume as the phenomenon takes place. Inference aims to define the relationship between the values assumed by the various features, for example, if the value of ω can be expressed in terms of the values that a subset $\mathbf{X}$ of features observed experimentally assumes. Prediction is to estimate the value of ω for specific values of the subset $\mathbf{X}$ that have never been observed experimentally. Machine learning models can be used for inference or prediction. Here, we consider only ML models used for prediction.
- We assume ω is a continuous quantity. Thus, we do not consider predictors of binary, categoric, or otherwise discrete quantities here. To use ML terminology, we limit our attention to regression models and do not consider classification models.
- In the following, we will assume that the steps of feature selection and assessment of the predictor's accuracy are done separately. We acknowledge that some methods combine the two steps (e.g., the LASSO method [9]).
- Here, we limit our analysis to supervised learning, where a set of experientially observed values for ω and associated values for the features vector $\mathbf{X}$ is used to train a predictor $\omega = f(\mathbf{X})$. We call this *training set*.

IV VVUQA Process For Biophysical Models

A detailed theoretical framing of verification, validation, uncertainty quantification, and applicability (VVUQA) analysis can be found here [7]. Here, we provide only a high-level description. The credibility assessment of a biophysical model is built on the assumption that the explicit knowledge used to build the model is true. In particular, the explicit knowledge ensures that $\mathbf{X}$ is necessary and sufficient to predict ω.
Still, if we conduct controlled experiments to measure both $\mathbf{X}$ and ω compare the measured ω with the one predicted using $\mathbf{X}$ and the explicit knowledge that links it to ω, we will commit a prediction error. To demonstrate the credibility of a biophysical model, we need to:

- Calculate the prediction error for a sufficiently large number of controlled experiments and confirm that it is below the acceptance threshold.
- Confirm that the observed prediction error can be explained in terms of known sources of error and that their values and distribution match the expectations for that

[2] https://insilico.world/
[3] https://www.fondazionedare.it/
[4] https://www.edith-csa.eu/

source. Biophysical models are affected by three sources of error:

- The *approximation error* one may commit in computing the model; the expectation is that the approximation error is negligible compared to the other sources of error.
- The *aleatoric error*, due to how the uncertainty with which **X** is measured propagates in the prediction of ω; the expectation is that such error has a null mean, so its effect manifests in the variance but cancels in the mean.
- The *epistemic error* is caused by the fact that to compute ω given **X,** we use a predictor *f*()* built by performing some idealisations of the knowledge *f()* necessary to make *f*()* computable or simply efficiently computable.

- Verify that **X** values tested in the controlled experiments are a representative sampling of the region of validity of the biophysical model.

Biophysical models are frequently expressed with systems of differential equations, which ensure the existence, uniqueness, smoothness and non-chaoticity of ω over **X**. These properties are not granted for biophysical models built as discrete systems (e.g., agent-based models) and need to be tested as part of the calculation verification process [8].

V Conceptual Generalisation of Machine Learning Predictors

The complex process of developing a valid ML-based predictor can be simplified, for our purposes, in three steps: feature selection, training, and credibility assessment.

To keep the parallel with the VVUQA for biophysical models, we will assume that the feature selection process has produced a set of measurable quantities **X**. However, here, nothing guarantees that **X** is necessary and sufficient to predict ω. While this is not a hard requirement, we exclude methods that combine feature selection and training for simplicity of exposition so that we do not entangle the different sources of error and approximation between the two steps of the analysis.

For the same reason, we focus on supervised learning methods here. Nothing conceptual forces us to restrict ourselves to this class of learners; however, if we use labelled data, some concepts, such as training residuals, are easier to explain.

The main methods for supervised learning are regression analysis, artificial neural networks, decision trees (including random forests), and support vector machines. Classification methods that can be adapted as regression methods (for example, Rand Forests Regression). Native regression methods provide a law that links the output ω to the inputs **X**. In contrast, the classification methods adapted to operate as regression methods define such a law implicitly by searching for some partitioning of the data space that provides such regression.

In both cases, the training process can be abstracted and generalised as the search for the combination of parameters of the predictor that minimises the norm of the differences between the predicted values $\widehat{\Omega}_j$ and the observed values Ω_j. This generic ML-based predictor can be written as

$$f(\boldsymbol{X}_j) = \widehat{\Omega}_j, where\ (\boldsymbol{X}_j, \Omega_j) \in \hat{Q}_t, and\ \|\widehat{\Omega}_j - \Omega_j\| < \varepsilon \qquad (1)$$

$\hat{Q}_t$ is the *training set*, the set of $(\boldsymbol{X}_j, \Omega_j)$ pair that we were able to measure experimentally.
The predictor *f*() is one of the set of predictors *F* that we defined by assigning specific values to its parameters $\boldsymbol{\theta}$ (whose specific form depends on the machine learning algorithm):

$$f(\boldsymbol{\theta}, \boldsymbol{X}_j) \in F(\boldsymbol{\theta}) \qquad (2)$$

The training of a machine learning model can be generalised as the search for the specific predictor *f*(); in other words, the search for the values $\boldsymbol{\theta}$ that minimise the error function *E*():

$$E(\boldsymbol{\theta}) = \frac{1}{n_{tr}} \sum_{i=0}^{n_{tr}} \|\widehat{\omega}_i - \omega_i\|^2 \qquad (3)$$

where $E(\boldsymbol{\theta})$is the error function, $n_{tr}$is the number of measured values in the training set, ω_i is the measured value, and $\widehat{\omega}_i$ is the corresponding predicted value.

VI Extension of the VVUQA to ML-based predictors

Given this generalisation of ML-based predictors, we can now see if the VVQUA approach can be extended also to ML-based predictors.

First, we need to perform the feature selection process. Once we have identified the relevant components to the CoU for vector **X**, we need to define the region of validity. Biophysical models usually derive the region of validity from the explicit knowledge used to build the model. In ML-based predictors, this is not the case; thus, here, we define the region of validity by identifying the maximum and minimum values that can be observed for each quantity in **X**. We have an expectation that our ML-based predictor will have a prediction error lower than the acceptable threshold, for any possible combinations of **X** within these limits.

Once **X** is identified, we can generate and use the training set to develop our ML-based predictor.

The process of determining if the computational model can be trusted to make predictions of the system behaviour for its context of use includes performing VVUQA analyses.

Verification. Verification is the process that establishes if the model is implemented correctly and solved accurately. It is typically composed of two main activities: code and solution verification. The first relies on having a robust software quality assurance program to minimise the occurrence of bugs in the software. The second aims to check if the errors due to numerical approximation are negligible compared to those due to other sources of error [1]. The main source of numerical error

for an ML predictor is the uncertainty affecting the hyperparameters. For ML-based predictors, properties such as existence, uniqueness, smoothness, and non-chaoticity of ω over **X** must also be demonstrated. The methods can be similar to those used for discrete biophysical systems [2].

Uncertainty Quantification: Uncertainty quantification examines the degree to which uncertainty in the model inputs propagates into uncertainty in the model outputs. As for biophysical models, the aleatoric error in ML models is due to the quantification uncertainties affecting the feature set **X**.

Validation: Validation addresses the question of the adequacy of the selected models for representing the reality of interest. The predictions of the ML model must be compared to the reference values of one or more test sets to calculate the predictive error. To calculate the prediction error for a sufficiently large number of controlled experiments and confirm that it is below the acceptance threshold, we can distinguish between internal and external validation. Internal validation occurs when a single data collection, performed under homogeneous conditions, is divided into training and test sets. One or more test sets with the same experimental setting used to generate the training set are considered to calculate how closely the predicted output matches observations of the physical system [3]. External validation involves using new data collected from cohorts that differ from those used to build the model. As with biophysical models, the discussion on how large and diverse the set of validation experimental data used for this step should be is complex. However, the risk-based approach to define the severity of the credibility assessment plan described in the ASME VV40:2018 also remains valid for ML predictors.

The other important activity considered part of validation is the decomposition of the prediction error in its components and the confirmation that each component behaves as expected. The numerical component of the predictive error is estimated through verification, and the aleatoric component is estimated through uncertainty quantification. By eliminating these two, what is left of the prediction error is due to the *epistemic error*. In biophysical models, this is the error due to imprecisions with which the prior knowledge used to build the model describes the reality of interest. However, ML predictors do not involve prior knowledge. Are ML predictors also affected by an epistemic error? In the past, some of the authors of this paper suggested that the answer was no [4]. However, a closer look at the problem may suggest otherwise. An essential step in building a model, which is also true for ML predictors, is the choice of the model form. Such a choice carries the potential for epistemic error.

Applicability: Applicability is the relevance of the evidence from the verification, validation and uncertainty quantification analyses to support the use of the model in a specific context of use. This involves *verifying that **X** values tested in the controlled experiments are a representative sampling of the region of validity of the biophysical model.* As with biophysical models, applicability is also possible and necessary for ML-based predictors, but with one major caveat. In biophysical models, we expect **X** to be necessary and sufficient to predict ω. From this, we can deduce that if we repeat the same validation experiment in different conditions for similar values of **X**, we should observe similar values ω. This is because we are sure that there is no unknown factor not included in the feature set **X** that could affect the quantity of interest ω. With ML models built using implicit knowledge, such certainty does not exist. Nothing ensures that if tomorrow we repeat the validation experiments in a different setting (different hospital, different ethnicity of the patients, etc.), we will observe for similar **X** equally similar prediction errors over ω.

ML-based predictors have sources of error similar to the biophysical ones: approximation, aleatoric and epistemic. The primary differences lie in how these errors are generated and the processes used to calculate them. As for biophysical models, the aleatoric error is due to the uncertainties affecting **X**. The epistemic error for an ML-based predictor is expected to decrease as the training data size increases. Following this logic, the approximation error is the residual error $E(\boldsymbol{\theta})$ left at the end of the optimisation process. The expectations for these sources of errors are also similar: we would expect the approximation error (average prediction error over the training set) to be negligible compared to the prediction error (average prediction error over the test sets). Similarly, we would expect that if the measurements of **X** are not affected by significant systematic error, also the component of the prediction error due to this source has a null mean. If this is the case, the error $E(\boldsymbol{\theta})$ calculated as quadratic mean error over the test sets, it should represent the epistemic error (since the approximation error is negligible and the aleatoric error has a null mean).

VII An example of VVUQA for an ML predictor

To provide an example of extending the VVUQA process to an ML predictor, we use BBCT-ML. BBCT-ML is a surrogate ML predictor of BBCT-hip, a biophysical digital twin that predicts the risk of hip fracture upon falling based on CT scan data of the subject at risk [5]. The biophysical predictor has been the subject of an extensive VVUQA assessment [6].

The BBCT surrogate ML predictor is a regression ML predictor so configured:

$$[\varepsilon_{11}, \varepsilon_{33}] = f([PC1 - PC34], \alpha_{AP}, \beta_{ML}, \rho)$$

where *f* predicts the peak values for the major (ε_{11}) and minor (ε_{33}) principal components of the strain tensor induced in the femur for a side fall impact force with an intensity of 1000 N. The strain values can be predicted if one knows the 34 principal component values that describe morpho-densitometry of the patient's femur (PC1-PC34), the two angles that define the orientation of the femur at the impact (α_{AP}, β_{ML}), and the change of the mineral density over time due to the progression of the disease (ρ).

In our example, since BBCT-ML is a surrogate model, we would expect it to have an average predictive accuracy equal to or better than the model's predictive accuracy. When validated in ex vivo experiments, where strain can be directly measured, BBCT-Hip showed an average (Root Mean Square, RMS) error

of 7% of the maximum measured strain [7]; we expect BBCT-ML to have a better average predictive accuracy, say 5%. However, since the predicted strain is used to decide whether a certain impact force is sufficient to fracture the subject's femur, defining an acceptance threshold for the maximum predictive error is necessary. Here, we refer to the clinical stratification accuracy of the original biophysical model, which is around 87% [5]. Thus, the acceptance threshold for the peak predictive error should be 1300 microstrains.

There are some ML approaches where feature selection and hyperparameter optimisation are done simultaneously. This approach is not ideal from a credibility assessment perspective, so it is advisable to separate feature selection from the VVUQ process. In our example, a preliminary feature selection analysis with the Random Forest showed that only seven of the 34 principal component modes were necessary in this training set.

At the end of the development process, we selected an XGBoost regression [8] as the best regression model, in which hyperparameters were optimised using the GridSearch algorithm with a 5-fold cross-validation within the training set.

As an illustration of the process, we now apply the VVUQA process to the BBCT-ML predictor. For simplicity, the reference values (that we assume are true) are provided by the biophysical model and not by a controlled experiment. However, to illustrate the process, this makes no difference.

Verification. As previously mentioned, the main source of numerical error for an ML predictor is related to the nature of optimisation algorithms used during training. In our example, we considered the Grid Search algorithm [9]. A key parameter in this algorithm is the grid spacing; to obtain the numerical approximation, one could first repeat the hyperparameters optimisation with a denser grid and then see how the average or peak predictive error changes over the training set when these new set of hyperparameters is used. Also, for ML predictors, we recommend extending the verification to include tests for existence, uniqueness, smoothness, and non-chaoticity conditions; such conditions are not tested for biophysical models formulated with differential equations because that mathematical form ensures their satisfaction. However, they are required for other biophysical models, such as agent-based predictors [2].

To assess the existence condition, we run a Monte Carlo with a Latin Hypercube over the input space, assigning equal probability for any value between the maximum and minimum admissible values. We kept sampling until the mean and variance of e11 and e33 that did not change significantly, adding more samples. For this type of data, convergence is observed for 1000 samples or less. We simply check that there is no input value for which the predictor fails to predict an output value. This test can be replaced by analytical considerations on the mathematical form of the predictor that ensures a solution always exists in that input space.

To test uniqueness, for a deterministic predictor like this one, we need to estimate the round-off numerical error by looking at the variability of the least significant digit provided by the calculus and confirm that it is one or more orders of magnitude smaller than the least significant digit with a physical meaning. For e11 and e33, this least significant digit is 10^{-5} strain (or 10^{1} microstrains). As the reproducibility of the simulation may be affected by the input values, the test should be repeated for multiple input sets uniformly sampling the solution space.

To test smoothness and non-chaoticity, we expand the Monte Carlo run to assess the existence conditions by adding to each input set variations where one of the input quantities is change ± 1% and calculate the partial derivative that these small variations induce in the predicted values. The average value of these derivatives over the entire range of inputs explored should remain small (smoothness). The maximum value of the derivative for any input tested should also remain small (non-chaoticity).

Validation. Internal and external validation should be considered for ML predictors. Here, for simplicity, we used only an internal validation based on 20% of the experimental data, randomly selected, while the remaining 80% was used as the training set. The average (RMS) error of the BBCT-ML XGBoost regression over this test set is 194 microstrains, while the peak error is 1282 microstrains. Both error indicators are below the respective acceptability thresholds. The error histogram for ε_{11} (Fig. 1) shows how the average error is very close to zero (-0.2293 microstrains), suggesting no significant systematic error by the predictor.

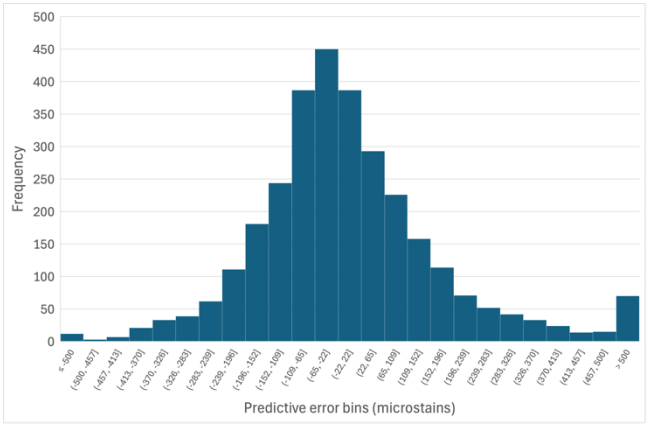


Figure 1. Histogram of the prediction error for BBCT-ML.

Uncertainty quantification. Using the same Monte Carlo analysis run for the verification test, we were also able to quantify the aleatoric uncertainty caused by the propagation through the predictive model of the uncertainty associated with the input parameters. When measured experimentally in the cadaver studies originally used to validate the biophysical BBCT-Hip predictor, the two impact angles α_{AP} *and* β_{ML} had a reproducibility of ± 1 degree. The uncertainty affecting ρ is obtained by CT images through densitometric calibration. The resulting uncertainty is ±0.01 gr/year. The estimation of uncertainty for the PCA modes is more complex, as it is related not to the reproducibility of experimental measurement but to the fitting of a statistical atlas to a training set. Here, we assume that ±1% variation for each principal mode used in the smoothness test is also representative of the aleatoric

uncertainty affecting those inputs. Using the Monte Carlo simulation, we calculated how all these uncertainties propagated into the prediction of the two principal strain components.

Applicability analysis

The last step is to ensure that the test sets we used for the validation are a representative sampling of the region of validity of the biophysical model. Here, we are in a special case, as we are not testing an ML predictor trained with experimental data, which are inherently scarce and patchy. Being our ML predictor a surrogate model, we produced the dataset to train and test the ML predictor running the biophysical model over the entire admissible range of values for each input feature.

Given the numerical error was negligible compared to the sum of the other two, and the aleatoric error is also quite small compared to the prediction error found in this validation study, we can confirm that the primary source of error is the error caused by the form of the ML predictor and the identification of its hyperparameters. We can confirm that the observed prediction error can be explained in terms of known sources of error and that their values and distribution match the expectations for that source. In particular, the prediction error for ε_{11} (which we know is dominated by the epistemic error) shows a very weak correlation (CC = 0.57) with the "true" value provided by the biophysical model. The linear regression slope is only 0.15, and the $R^2 = 0.32$. Such a weak correlation can be easily explained, considering that the biophysical model assumed the behaviour of the bone tissue to be linear elastic, something which becomes less and less true as the strain increases.

VIII Generalisation of validity

The mechanistic knowledge we use to build biophysical models should ensure that the features vector $\mathbf{X}$ is necessary and sufficient to predict ω. An important implication of this fact is that if we repeat the validation experiments in a different setting, there is an expectation that similar $\mathbf{X}$ values will be associated with similar ω values. In other words, if we repeat the validation experiment, we should get the same results in terms of predictive accuracy. This is because we are sure the biophysical predictor has no "unknown bias" that may alter the quantity of interest, even if the feature values are the same. However, this is not the case for any predictor not entirely built using reliable mechanistic knowledge. Indeed, ML-based predictors are exposed to the risk of concept drift [10].

This simple fact has a profound implication from a regulatory point of view. Since validation experiments may yield different results when repeated in different conditions, any ML-based predictor (and any hybrid predictor, for that matter) can only be considered credible through a VVUQA process for the conditions in which the validation experiments have been done. This changes entirely the perspective of the credibility assessment. While for biophysical predictors, the credibility can be assessed once the applicability to the context of use has been demonstrated, ML predictors must be re-assessed every time the predictor is used in conditions different from those under which the validation experiments have been conducted. This yields the idea of periodic retesting: the additional test sets should be collected in widely different conditions and settings to allow the effect of eventual "unknown bias" to manifest, building trust in the credibility of the ML-based predictor.

As already suggested in a public proposal for a regulatory framework for Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) the FDA published in 2019[5], one possible approach to this continuous retesting needs is what the FDA calls Total Product Life Cycle (TPLC). According to the FDA, "throughout the total product lifecycle, device developers and regulators take steps to ensure that a new device can be safe and effective when used by the patient population for which it has been designed". The general idea is to frame the credibility assessment in the more general concepts of continuous risk management and of quality assurance.

We support this approach, as far as a significant portion of it is done as post-marketing surveillance. An ML predictor should be assessed, following the approach of the VV-40 standard, with a credibility assessment plan selected based on the risk analysis of the context of use for that ML predictor (how important is the result of the predictor in the clinical decision? How severe are the consequences if that decision is wrong?). This will require, in addition to the training set, the validation against one or more test sets collected with appropriate criteria, again inspired by the context of use. In this sense, it might be wise to narrowly define the "context of use" of the ML predictor. After marketing authorisation, the developer may pursue a so-called "re-labelling change" to broaden it. With this evidence of credibility over a specific context of use well represented by the test sets used for the credibility assessment, the ML predictor should receive marketing authorisation, conditioned to its periodic re-testing.

In our opinion, this approach offers a good balance between safety and innovation. The developers can obtain a first marketing authorisation with a relatively modest amount of test data, even if for a narrowly defined context of use. They can then organise post-marketing studies to collect additional test sets either under the same narrow context of use (to demonstrate no concept drift) or for a broader context of use in preparation for a relabelling request.

IX Discussion

This position paper explored the possibility and difficulties in extending the VVUQA practice and its theoretical framing to machine learning models.

[5] https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

The generalisation of the ML learners we proposed makes it possible to identify ML-based predictors' sources of error similar to those we have in biophysical models, as well as similar verification, validation and uncertainty quantification procedures to separate them and to check they behave as expected. The applicability analysis also retains a similar meaning and thus can be extended to ML-based predictors. So, the answer to the question is yes, the VVUQA practice can be extended to ML-based predictors.

Since ML-based predictors are not based on differential equations, the calculation verification must include evidence of existence, uniqueness, smoothness, and non-chaoticity of ω over **X**. The methods adopted for calculation verification can still be similar to those used for discrete biophysical systems [8].

However, feature identification is the real difference between biophysical and ML-based predictors. The mechanistic knowledge we use to build biophysical models should ensure that the features vector **X** is necessary and sufficient to predict ω. An important implication is that if we repeat the validation experiments in a different setting, there is an expectation that similar **X** values will be associated with similar ω values. In other words, if we repeat the validation experiment, we should get the same results. This is because we are sure the experiment has no "unknown bias" that may alter the quantity of interest, even if the feature values are the same. However, this is not the case for any predictor not entirely built using reliable mechanistic knowledge. Indeed, ML-based predictors are exposed to the risk of concept drift [10]. This calls for periodic retesting of ML predictors. We recommend that this is wisely implemented in the regulatory pathways, following a Total Product Life Cycle in which part of the testing can be done as part of a conditional marketing authorisation as part of the periodic post-marketing surveillance.

It should also be noted that the certainty that no "unknown bias" exists is rarely the case for biophysical models. Many of these models result from strong reductionist idealisations that simply ignore a big portion of the human pathophysiology, assuming they do not affect the predicted interest. However, this might not be true, and in these cases, post-marketing surveillance based on a Total Product Life Cycle approach might also be advisable.

X Conclusion

The VVUQA process can also be used for ML-based and hybrid predictors; however, the credibility of these predictors is demonstrated only concerning the conditions in which the test sets have been formed. Thus, a Total Product Life Cycle approach is advisable for these predictors, where the predictive accuracy is periodically re-tested on new test sets. We recommend that this re-testing take place as part of the post-marketing surveillance.

While this position paper states the authors' current understanding of the matter, we expect the debate on evaluating the credibility of data-driven predictors to continue. For this purpose, a new public debate channel called #credibility_machine_learning was created on the In Silico World Community of Practice, a Slack forum operated by the VPH Institute[6]. This community is open and free; if you are not a member, follow the instructions here[7] to join.

Acknowledgment

The authors would like to thank the In Silico World Consortium and all the participants of the workshop on regulatory barriers for the In Silico World project in Catania (IT) on March 14th, 2024. The debate on the credibility of ML predictors was essential in forming the key concepts of this manuscript.

Annex 1: definitions

- *B* is a class of systems that satisfy certain inclusion-exclusion criteria. Any instance of *B* is called a System of Interest (SI).
- Ω is the class of all the quantities that can be measured on the SI.
- The goal is to obtain an accurate quantification of one member of Ω, which we call Quantity of Interest (ω), without measuring it.
- We want to use this quantification for a specific purpose (*Context of Use*). In particular, the context of use defines the maximum affordable error for such quantification to remain useful to that purpose (*acceptable threshold*).
- This is possible if the value that ω is assumed in the SI at each instant is in a causal relationship with other measurable quantities of the SI. In that case, knowing the values of that vector of quantities **X** (*feature set*) and the causal relation that links them to ω we can calculate:
  $\omega = f(\mathbf{X}(t), t)$
- Of course, *f()* must be valid for every instance of *B*; in other words, the causal relationship must be sufficiently universal to cover all systems that fit the inclusion-exclusion criteria. We call this property *Universality*.
- The accuracy of predictors built from the causal knowledge *f()* in predicting ω also depends on the internal status of the SI; the *limits of validity* for *f()* are the two feature sets $\mathbf{X}_1$ and $\mathbf{X}_2$. So, *f()* is expected to predict ω with the expected accuracy for any **X** within the region delimited by $\boldsymbol{X}_1 \times \boldsymbol{X}_2$ (*region of validity*).

[6] https://www.vph-institute.org/

[7] https://insilico.world/community/join-the-community-of-practice-channels/

REFERENCES

[1] ASME, 'ASME V&V 40 - Verification & Validation of Medical Devices'. 2018. Accessed: Apr. 02, 2024. [Online]. Available: https://www.asme.org/codes-standards/find-codes-standards/v-v-40-assessing-credibility-computational-modeling-verification-validation-application-medical-devices

[2] T. M. Morrison, P. Hariharan, C. M. Funkhouser, P. Afshari, M. Goodin, and M. Horner, 'Assessing Computational Model Credibility Using a Risk-Based Framework: Application to Hemolysis in Centrifugal Blood Pumps', *ASAIO J.*, vol. 65, no. 4, p. 349, Jun. 2019, doi: 10.1097/MAT.0000000000000996.

[3] M. A. Dharia, S. Snyder, and J. E. Bischoff, 'Computational Model Validation of Contact Mechanics in Total Ankle Arthroplasty', *J. Orthop. Res.*, vol. 38, no. 5, pp. 1063–1069, 2020, doi: 10.1002/jor.24551.

[4] G. Bideault, A. Scaccia, T. Zahel, R. W. Landertinger, and C. Daluwatte, 'Verification and Validation of Computational Models Used in Biopharmaceutical Manufacturing: Potential Application of the ASME Verification and Validation 40 Standard and FDA Proposed AI/ML Model Life Cycle Management Framework', *J. Pharm. Sci.*, vol. 110, no. 4, pp. 1540–1544, Apr. 2021, doi: 10.1016/j.xphs.2021.01.016.

[5] G. Luraghi, S. Bridio, C. Miller, A. Hoekstra, J. F. Rodriguez Matas, and F. Migliavacca, 'Applicability analysis to evaluate credibility of an in silico thrombectomy procedure', *J. Biomech.*, vol. 126, p. 110631, Sep. 2021, doi: 10.1016/j.jbiomech.2021.110631.

[6] A. Aldieri, C. Curreli, J. A. Szyszko, A. A. La Mattina, and M. Viceconti, 'Credibility assessment of computational models according to ASME V&V40: Application to the Bologna Biomechanical Computed Tomography solution', *Comput. Methods Programs Biomed.*, vol. 240, p. 107727, Oct. 2023, doi: 10.1016/j.cmpb.2023.107727.

[7] M. Viceconti, M. A. Juárez, C. Curreli, M. Pennisi, G. Russo, and F. Pappalardo, 'Credibility of In Silico Trial Technologies—A Theoretical Framing', *IEEE J. Biomed. Health Inform.*, vol. 24, no. 1, pp. 4–13, Jan. 2020, doi: 10.1109/JBHI.2019.2949888.

[8] C. Curreli et al., 'Verification of an agent-based disease model of human Mycobacterium tuberculosis infection', *Int. J. Numer. Methods Biomed. Eng.*, vol. 37, no. 7, p. e3470, Jul. 2021, doi: 10.1002/cnm.3470.

[9] R. Tibshirani, 'Regression Shrinkage and Selection Via the Lasso', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[10] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, 'Learning under Concept Drift: A Review', *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 12, pp. 2346–2363, Dec. 2019, doi: 10.1109/TKDE.2018.2876857.

Marco Viceconti is a professor of industrial bioengineering in the Department of Industrial Engineering at the University of Bologna. He founded the VPH Institute, an international non-profit organisation coordinating this research community. He drove the creation of the Avicenna Alliance, representing the biomedical industry interests in this domain. He served as President of the European Society of Biomechanics and the European Alliance for Medical and Biological Engineering and Science. He is currently one of 25 members of the World Council of Biomechanics. In 2018, he became a Fellow of the UK Royal Academy of Engineering; in 2021, he received the Huiskes Medal for Biomechanics. According to SCOPUS, he published 399 papers (H-index = 58).

Filippo Lanubile is a professor of computer science and Head of the Department of Informatics at the University of Bari, Italy, where he also leads the Collaborative Development Research Group. His research interests include human factors in software engineering, AI-augmented software development, engineering AI-enabled software systems, and health informatics. He is a recipient of a NASA Group Achievement Award, an IBM Eclipse Innovation Award, an IBM Faculty Award, and the Software Engineering Innovation Foundation Award from Microsoft Research. From 2020 to 2023, he was the IEEE Software Advisory Board Chair. In 2024, he served as Deputy Chair of the ACM SIGSOFT Award Committee for Research, Education, and Service Awards.

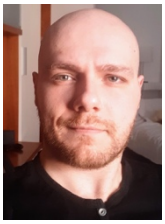
This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3552320

Antonella Carbonaro received a PhD in Intelligent Artificial Systems from the Faculty of Engineering of the University of Ancona, Italy. She won post-doc research grants on Artificial Intelligence. Since 2000, she has been a researcher and then an associate professor at the Department of Computer Science - Science and Engineering - of the University of Bologna, Italy. Her research is on data and knowledge modelling for representing entity semantics and relations, with applications also in healthcare. She is the leader of the WP Technology and Analytics project, ‘‘DARE - digital lifelong prevention,’’ and is supported by the Italian Ministry of University and Research. She leads the WP Knowledge Graph for the Cancer Virtual Lab project, International Foundation Big Data, and Artificial Intelligence for Human Development.

Sabato Mellone is a senior assistant professor at the University of Bologna. He received a PhD in Biomedical Engineering from the same university. His research activities include data and signal processing, medical informatics, ICT in clinical practice, wearable and embedded sensors, personal health systems design and validation, and eHealth / mHealth applications. He was a work package leader responsible for developing technology in multiple European and national projects. He co-founded mHealth Technologies s.r.l., a University of Bologna spin-off company. I am the author of more than 60 papers published in international journals.

Cristina Curreli received a PhD in Mechanical Engineering from the University of Pisa and worked as a post-doctoral researcher at the Biomechanics Lab of the University of Bologna. She joined the Medical Technology Lab at the Rizzoli Orthopaedic Institute as a visiting researcher in April 2019 and as a healthcare researcher in September 2023. Her research activities mainly focus on developing and validating in silico methodologies to support clinical decisions for the prevention, diagnosis, and treatment of musculoskeletal diseases and to provide additional insights into assessing new medical products' safety and/or efficacy.

Alessandra Aldieri is an Assistant Professor with a time contract at Politecnico di Torino, Italy. In 2020, she got a PhD in Bioengineering and medical-surgical sciences from Politecnico di Torino. Later, she worked as a post-doctoral fellow at the University of Bologna. Her main research focus is computational biomechanics, specifically digital twins in healthcare.



Saverio Ranciati is a senior assistant professor in statistics at the Department of Statistical Sciences “Paolo Fortunati” in Bologna, Italy. He graduated from a joint double-PhD programme organised by the University of Bologna and the University of Groningen. His research interests include penalised regression and inference, graphical models, model-based clustering, computational statistics and Bayesian techniques.
He has been involved in past and ongoing national and international projects: COSTNET (COST Action CA15109); Mobilise-D (Innovative Medicines Initiative 2); AlmaHealthDB; PNC (PNRR) project “DARE: Digital Lifelong Prevention”.

Angela Montanari is a full professor of Statistics at the Department of Statistical Sciences at the University of Bologna. She has been the Head of the Department of Statistical Sciences from 2015 to 2021. She represents the University of Bologna in the Self Steering Committee on Data Science and AI of the University alliance UNA Europa. She is the past president of IFCS (the International Federation of Classification Societies). Her main research interests are multivariate analysis, variable selection, and supervised and unsupervised statistical learning.