ASSESSING THE CREDIBILITY OF QUANTITATIVE INFORMATION: A GENERAL FRAMEWORK

Marco Viceconti[1]

[1] Department of Industrial Engineering, Alma Mater Studiorum - University of Bologna (IT)

ORCID IDs

Marco Viceconti https://orcid.org/0000-0002-2293-1530

SUBMITTED TO NPJ Digital Medicine

ON May 2025

CORRESPONDING AUTHOR:
Prof Marco Viceconti
Department of Industrial Engineering
Via Terracini 24, 40131 Bologna (IT)
Email: marco.viceconti@unibo.it

ASSESSING THE CREDIBILITY OF QUANTITATIVE INFORMATION: A GENERAL FRAMEWORK

Abstract

Quantitative information can be measured, inferred, or predicted, and the process used to assess the credibility of such information varies depending on how it is produced. In this paper, we propose a general process called the 7S Framework, which can be used to assess the credibility of information, whether measured, inferred or predicted. The 7S Framework integrates and generalises the credibility assessment approaches used for measured information in metrology, inferred information in statistics, and predicted information in computational science and engineering. In this study, when applied to seven fairly different use cases, the proposed framework was effective, sufficiently general, and capable of capturing all the subtle differences that the concept of credibility implies for these different kinds of information. We propose the 7S Framework as a generalisation useful in the credibility assessments of complex *in silico* medicine scenarios such as in-silico augmented clinical trials, physics-informed machine learning predictors, or the use of synthetic datasets to overcome privacy limitations, train machine learning predictors, and run large-scale In Silico Trials.

Keywords

Credibility analysis; information; measurements; inferences; predictions.

Introduction

Quantitative information can be **measured**, **inferred** from measured data (e.g., through interpolation) or **predicted** by assuming a quantitative causal relationship between the quantity of interest and other measurable (or inferable) quantities.

This is linked to a process sometimes known as "The Pyramid of Knowledge", where we uplift raw data (for example, obtained as experimental quantifications) into *information* by annotating the data with metadata about who, what, how, and when the data were produced. Then, we uplift the information into tentative causal *knowledge* by observing and modelling the correlation with the information space. Lastly, we use other information to challenge such tentative knowledge, and in doing so, we uplift knowledge into something some authors call *wisdom*, which is when tentative knowledge is credible enough to be used operationally, for example, to support decisions.

Traditionally, the credibility of quantitative information is discussed separately in these cases. The metrology community debated the credibility of measured information [1], the statistics community of inferred information [2], and the computational science and engineering research community of predicted information [3]. The uplift of data into information is seen as an information science issue; that of information into knowledge is a statistical sciences issue; the uplift of knowledge into wisdom is seen as an epistemology issue.

But these separations are getting blurred. Machine learning, synthetic data, and physics-informed machine learning mix all three sources of quantitative information. The credibility assessment of a machine learning predictor (which can be seen as the materialisation of

wisdom) is not independent of the quality of data, metadata (information) and correlative knowledge used to build it.

There is a growing need to frame the problem of information credibility in a more general way, encompassing all three sources. To the author's knowledge, such a general framework has not been proposed in the literature so far. However, starting points for the considerations presented in this paper are the "Consensus statement on the credibility assessment of ML predictors" produced by the In Silico World Community of Practice [4] and a position paper on the credibility of machine learning predictors[5] published recently.

This study proposes a seven-step general framework (7S Framework) to assess the credibility of every quantitative information, whether measured, inferred, or predicted. After the presentation of the framework, we will apply it to three exemplary information sets, each obtained from a different source, to show the applicability and generality of such a framework.

Definitions and assumptions

We define *quantitative information* as an annotated set of values where the annotation metadata provides the "domain of information", i.e. the details that both the sender and the receiver of that information need to know for that set of values to be informative. These typically include answers to the questions "who", "what", "where", and "when"; it does not include the answers to the "why" question, which uplifts information to knowledge [6].

We define the *System of Interest as* the target of our investigation. Let us assume that $\mathfrak{I}$ is the set of all types of quantitative information that could be observed for the system of interest. We are interested in obtaining one in particular that we call *Quantity of Interest* (QoI); hereinafter, we will assume it to be a single scalar value for simplicity of exposition. While the treatment can be considerably complicated if the QoI is multiscalar, conceptually, the process of defining the credibility of information remains unchanged. For the same reasons, we will exclude multi-level QoIs from our analysis, such as nominal, ordinal, interval, and ratio scales.

The accuracy of measurements is defined in a metrological sense; it is expressed by a pair of values, a measure of *Trueness* and one of *Precision*. Trueness expresses the systematic errors affecting the quantification; Precision expresses the casual errors. For simplicity of exposition, we will define the *Class of Accuracy* of a quantity as the average trueness over repeated estimations, normalised by the maximum value that quantity can assume in the system of interest.

In the following, we will generalise a set of terms used in statistics to serve our purpose. We will refer to the general process of assigning a value to the QoI using the term *estimation*. Consistently, we will refer to the QoI as an *estimand*, the value assigned as an *estimate* and the method for producing such an estimate as an *estimator.* So, an estimator produces an estimate of the estimand. However, contrary to the classic statistical terminology focusing on inference, an estimator can produce an estimate through measurement, inference, or prediction.

Generally, the quantities we can observe for the system of interest are not mutually independent. Thus, the particular values we can observe may correlate with the values assumed by other observable quantities $\boldsymbol{I} \in \mathfrak{I}$. Ideally, we would like to solve the *causation problem* and determine which observable quantities are necessary and sufficient to define the value of the QoI. But hereinafter, we will only assume we know $\boldsymbol{C} \in \mathfrak{I}$ the observable quantities that are sufficient but not necessarily necessary to define the value of the QoI. On the contrary, we will not discuss here the case where some of the necessary quantities are not part of the observable

set $\mathfrak{I}$. In other words, we will assume that every quantity necessary to cause the QoI is observable, although not necessarily known.

The QoI usually correlates with each observable quantity only for a finite range of possible values of such a quantity; we call it the *limits of validity*. Hereinafter, we assume that we know the limits of validity for the QoI and for each observable quantity that correlates with the QoI. If $\boldsymbol{C}$ includes only two quantities, we could, in principle, plot all true values O that the QoI can assume as a 3D surface (2-manifold), but because all quantities in $\boldsymbol{C}$ are limited, that would only be a *patch*, a portion of that surface. If $\boldsymbol{C}$ includes n quantities, that entire *information space* can be represented as a bounded n-manifold.

Last, we define the *credibility* of an estimator as the lowest accuracy with which the estimator provides an estimate of the estimand over the entire information space.

Credibility can be categorised into three levels, depending on the expectation for accuracy [7]:

- Level 1 (L1) requires that information estimation falls within a pre-defined range of uncertainty of the true value. A good example is the credibility assessment of measured information. The measurement chain used to estimate the information should have given a precision class (say, for example, 0.5, which means an absolute intrinsic error never exceeding 0.5% of its largest reading). As far as the measured information deviates from the true value in the validation experiments for less than this limit, the measured information is considered L1-credible (in class 0.5).
- Level 2 (L2) requires that the average value of the information over repeated validation experiments is not significantly different to the average of the true value; in other words, the estimate and the true value are close in a statistical sense. A good example is the credibility assessment of inferred information when using the quadratic norm to calculate the average difference between the inferred and measured values.
- Level 3 (L3) requires a local accuracy at any validation point to be below or equal to an acceptability threshold. A good example is the credibility assessment of predicted information. Modelling how the prediction error varies as a function of the input values (and in particular if there is a correlation between the input values and the prediction error) can help us to estimate the upper boundary of the prediction error within the limits of validity and of applicability of the model.

Definition of a general credibility framework

Our definition of credibility immediately makes the difficulties in assessing it evident. First, to calculate the accuracy, you need the true value of the estimand; in the best-case scenario, we can obtain an accurate measurement of the estimand only for a finite number of instances of $\boldsymbol{C}$. In principle, we could estimate credibility by brute force, pushing the induction process to large numbers. Estimating the error distribution over $\boldsymbol{C}$ requires an infinite number of experimental measurements (estimating credibility by induction). Still, for the Law of Large Numbers (LLN), a finite but sufficiently large number of experiments could provide us with a reliable estimate of such error distribution. Here, we will assume producing such a large number of true values is practically impossible. Furthermore, the LLN postulate certain regularity properties that not all time of estimators considered here can ensure. Thus, we assume that estimating credibility by induction is not a viable option.

Qualitatively speaking, the approach all credibility frameworks use (metrological credibility of measurements, statistical credibility of inference, Verification, Validation and Uncertainty

Quantification of knowledge-based predictors) is similar. Based on the nature of the estimator, we identify the possible sources of error and how, for each source, the associated error is expected to be distributed in a well-behaved estimator. Then, we try to decompose the estimation error we calculate over a finite number of true values into its components and check how each component distributes. Suppose the estimation error is found sufficiently low for that finite number of true values available, and the contribution to that error for each source is distributed as expected. In that case, we can assume the estimator is well-behaved and accepts the induction risk.

We can generalise this process in the following seven steps.

S1. Context of use and error threshold

The first step of the proposed credibility framework is the extension of the domain of information to include the *Context of Use* for that information. In particular, it is necessary to define the maximum error $\bar{\varepsilon}$ affecting the information that still makes it useful for that use. The first requirement is that any estimator of the QoI must satisfy to be considered credible is that in all points of the solution space where the estimator accuracy is tested, the error is $\varepsilon_e < \bar{\varepsilon}$.

S2. Source of true values

The second step is to define the source of the true values. We postulate that true values can be obtained only through measurement, using a measurement chain that ensures both for the QoI and the correlated quantities $\boldsymbol{C}$ a class of accuracy that is at least one order of magnitude smaller than the maximum error defined for a context of use.

S3. Quantification of estimation error

The estimator's error can be quantified by sampling the solution space through controlled experiments where the correlated quantities $\boldsymbol{C}$ are imposed or measured, and the true values for the QoI and the corresponding correlated quantities are quantified.

S4. Identification of the sources of error

Depending on the nature of the estimator, we now identify the possible sources of such estimation error. This is the most delicate step and requires a profound understanding of how each particular class of estimators operates.

S5. Decomposition of the estimation error

This step is even more challenging: once the various sources of estimation error are identified, we must find ways to estimate how the overall estimation error can be divided among these sources. In some cases, this might require generating true values under special conditions where all but one source of error is excluded.

S6. Critical review of error distributions

As we identify the various potential sources of error affecting the estimator, we can also define some expectations of how that component of the estimation error should be distributed over

repeated estimands across the solution space. Once the estimation error is decomposed over its sources, we check if these expectations are confirmed.

S7. Robustness to biases and applicability

Once the estimator is considered credible for a context of use, it will be used routinely; the last step in the credibility assessment is to infer if there might be in this routine use additional biases, which may alter the accuracy of the estimator, that do not appear in the controlled experiments used to quantify the estimator error. A special case is that of applicability: we need to ensure that when the estimator is used routinely, the QoI and its correlates never exceed the limits of validity we defined during the credibility assessment, and if they do, how the credibility of the estimator may be compromised.

Application – case descriptions

We applied this approach to seven cases from our group's research to make this methodological proposal more concrete. However, this exposition is lengthy, so we provide it as supplementary material[1]. Here, we summarise only three of those cases for measured, inferred, and predicted information.

Credibility of strain gauge measurements

The Context of Use (CoU) is that of measurements of bone tissue deformation to validate a finite element model to predict bone fracture. The error threshold is calculated from the difference in strain used to decide fracture, assuming two orders of magnitude (the strain gauge validates the FE model, and the FE model predicts the fracture) (S1). The validation experiment is done on aluminium alloy specimens machined as slender rectangular beams, so the calculations based on the beams' theory, corrected for curvature error, provide the true value (S2). We calculate the trueness of the measurement as the root mean square error (S3). Measurements are affected by systematic and aleatoric errors. We confirm that the errors are normally distributed using a normality test, with a mean value close to zero (S4-S6). Validation experiments are repeated for bone specimens; if the errors are comparable to the ones found with the aluminium specimens, we can assume sufficient applicability (S7).

Credibility of a biophysical predictor of human bone deformation

The context of use of the BBCT-Hip predictor is to predict the mechanical strains induced in a subject's bone under a given loading when a properly calibrated Computed Tomography (CT) of the bone is provided [8]. The error threshold is defined as 2% of the cortical bone failure strain in compression, i.e., 146 με (S1). The source of true value is the strain gauge measurements we perform on the human cadaver femurs, which are properly preserved to retain their mechanical properties [9] (S2). The predictions of biophysical models are affected by numerical, aleatoric and epistemic errors; the verification, validation and uncertainty quantification (VVUQ) procedure separates them [10] (S4-S5). In all cases, prediction errors are expected not to exceed the acceptability threshold. The numerical error component must be negligible compared to the other two; the aleatoric component must be normally distributed with a mean close to zero, and the epistemic component should have a root mean square error close to zero (S6).

[1] Supplementary material: https://doi.org/10.5281/zenodo.15340551

Applicability analysis, in this case, is done by brute force. Experiments are designed to explore inter-subject variability and all possible loading conditions, the two primary sources of bias [11,12] (S7).

Credibility of a synthetic dataset inferred from a clinical data collection

Another use of the BBCT-Hip predictor is to estimate the efficacy of a treatment in reducing the incidence of hip fractures in a cohort [13]. We have access to a clinical cohort, a group of patients selected according to specified inclusion and exclusion criteria. To assess the credibility of BBCT-Hip, we derived a *Virtual Cohort* by measuring all the quantities necessary as input for the BBCT-Hip digital twin (feature set) in each patient of the clinical cohort. However, we need a much larger cohort to estimate a treatment's efficacy. So, we generate a *synthetic cohort*, a virtual cohort obtained by inferring the feature set values of other patients who belong to the same sub-population of the clinical cohort.

The context of use we are considering is to use, in place of an experimentally collected virtual cohort, a synthetic cohort of feature sets of the BBCT-Hip predictor to evaluate *in silico* the efficacy of different interventions aimed at reducing the risk of hip fracture in fragile elderlies on a much large virtual cohort that it would be practically possible using only experimental data. In this CoU, we plan to compare some central properties of two probability distributions (mean, median), so the error should be expressed in terms of the maximum acceptable difference for such properties. This should apply to both distributions of values for each feature in the set and for the model's predictions when such values are provided as inputs. In terms of error threshold, the difference in the median values for each feature between the virtual cohort and the synthetic cohort, as well as those of the average quantity of interest predicted by the model between the virtual and synthetic cohort, should never exceed the accuracy with which each of those quantities is measured or predicted (S1). The source of true values is the virtual cohort, the set of feature set values measured experimentally on the subjects forming the clinical cohort (S2). Of course, no one-to-one comparison is possible, so each feature's estimation error is defined as the difference between the median values of the two distributions (S3). Synthetic datasets are produced by inference over the experimental data, so no epistemic error is involved. The aleatoric error is due to the uncertainty of the experimental measurements. The numerical errors are those produced for each feature by the interpolation function that generates new synthetic values (S4). The aleatoric error can be estimated by conducting a sensitivity analysis. Once the aleatoric error is defined, we need to verify that the numerical error of the interpolators is negligible compared to that (S5, S6). The applicability analysis of information produced by inference is simple: the resulting information can be used only in cases that fall in the portion of the information space defined by the feature set sampled in the virtual cohort (S7).

DISCUSSION

The aim of this paper was to present the 7S Framework, a general process that can be used to assess the credibility of information, whether measured, inferred or predicted.

The 7S Framework integrates and generalises the credibility assessment used in metrology for measured information, which is used in statistics for inferred information, and the so-called VVUQ used in computational science and engineering (CS&E) for predicted information.

When applied to seven fairly different cases (credibility of bone tissue deformation measurements; credibility of a biophysical model used to predict bone deformation, strength, and clinical risk of fracture; credibility of a synthetic dataset to perform In Silico Trials on a

cohort larger than that experimentally available; credibility of In Silico Trials to estimate the efficacy of treatments to reduce the incidence of hip fractures; credibility of a machine learning predictor to be used as a surrogate of a biophysical model to improve computational efficiency), the proposed framework showed to be effective, sufficiently general, and capable of capturing all the subtle difference that the concept of credibility implies for these different kinds of information.

While we do not expect the 7S Framework to replace well-established practices in metrology, statistics, and CS&E, in silico medicine pushes the envelope of what can be done with computational methods, and the separation between measurements, inferences, and predictions is becoming increasingly blurred. However, there is an evident need for epistemologically robust credibility assessment strategies for medical applications. We hope the approach proposed here can help assess the credibility of complex scenarios such as in-silico augmented clinical trials [14], physics-informed machine learning predictors [15], or synthetic datasets to overcome privacy limitations [16], train machine learning predictors [17], and run large-scale In Silico Trials [13].

ACKNOWLEDGEMENTS

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: B53C22006460001.

Several use cases and datasets used to explore them were produced in the frame of the European Commission-funded H2020 project "In Silico World: Lowering barriers to ubiquitous adoption of In Silico Trials" (topic SC1-DTH-06-2020, grant ID 101016503).

CONFLICT OF INTEREST

The authors declare that they do not have any financial or personal relationships with other people or organisations that could have inappropriately influenced this study.

OPEN ACCESS DATA

A detailed application of the 7S Framework to seven relevant use cases is provided as open-access supplementary material here: https://doi.org/10.5281/zenodo.15340551.

REFERENCES

1. Prenesti, E. & Gosmaro, F. Trueness, precision and accuracy: a critical overview of the concepts as well as proposals for revision. *Accred Qual Assur* **20**, 33–40 (2015).
2. Aitkin, M. & Liu, C. Confidence, credibility and prediction. *METRON* **76**, 251–268 (2018).
3. Viceconti, M. *et al.* In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. *Methods* **185**, 120–127 (2021).

4. Aldieri, A. *et al.* Consensus statement on the credibility assessment of machine learning predictors. *Briefings in Bioinformatics* **26**, bbaf100 (2025).
5. Viceconti, M. *et al.* Position paper: Extending Credibility Assessment of In Silico Medicine Predictors to Machine Learning Predictors. *IEEE J Biomed Health Inform* **PP**, (2025).
6. Rowley, J. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science* **33**, 163–180 (2007).
7. Viceconti, M. *et al.* Theoretical Foundations of Good Simulation Practice. in *Toward Good Simulation Practice: Best Practices for the Use of Computational Modelling and Simulation in the Regulatory Process of Biomedical Products* (eds. Viceconti, M. & Emili, L.) 9–23 (Springer Nature Switzerland, Cham, 2024). doi:10.1007/978-3-031-48284-7_2.
8. Aldieri, A., Curreli, C., Szyszko, J. A., La Mattina, A. A. & Viceconti, M. Credibility assessment of computational models according to ASME V&V40: Application to the Bologna Biomechanical Computed Tomography solution. *Comput Methods Programs Biomed* **240**, 107727 (2023).
9. Ohman, C., Dall'Ara, E., Baleani, M., Van Sint Jan, S. & Viceconti, M. The effects of embalming using a 4% formalin solution on the compressive mechanical properties of human cortical bone. *Clin Biomech (Bristol, Avon)* **23**, 1294–1298 (2008).
10. Viceconti, M. *et al.* Credibility of In Silico Trial Technologies-A Theoretical Framing. *IEEE J Biomed Health Inform* **24**, 4–13 (2020).
11. Schileo, E. *et al.* An accurate estimation of bone density improves the accuracy of subject-specific finite element models. *J Biomech* **41**, 2483–2491 (2008).
12. Grassi, L. *et al.* Accuracy of finite element predictions in sideways load configurations for the proximal human femur. *J Biomech* **45**, 394–9 (2012).
13. Oliviero, S., La Mattina, A. A., Savelli, G. & Viceconti, M. In Silico clinical trial to predict the efficacy of hip protectors for preventing hip fractures. *J Biomech* **176**, 112335 (2024).
14. Haddad, T. *et al.* Incorporation of stochastic engineering models as prior information in Bayesian medical device trials. *J Biopharm Stat* **27**, 1089–1103 (2017).
15. Shi, J., Manjunatha, K., Behr, M., Vogt, F. & Reese, S. A physics-informed deep learning framework for modeling of coronary in-stent restenosis. *Biomech Model Mechanobiol* **23**, 615–629 (2024).
16. Thambawita, V. *et al.* DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Sci Rep* **11**, 21896 (2021).
17. Rashidi, H. H. *et al.* Prediction of Tuberculosis Using an Automated Machine Learning Platform for Models Trained on Synthetic Data. *J Pathol Inform* **13**, 10 (2022).