Statistical Science
2023, Vol. 38, No. 4, 543–556
https://doi.org/10.1214/23-STS891

Defining Replicability of Prediction Rules

Giovanni Parmigiani

Abstract. In this article, I propose an approach for defining replicability for prediction rules. Motivated by a recent report by the U.S.A. National Academy of Sciences, I start from the perspective that replicability is obtaining consistent results across studies suitable to address the same prediction question, each of which has obtained its own data. I then discuss concept and issues in defining key elements of this statement. I focus specifically on the meaning of "consistent results" in typical utilization contexts, and propose a multi-agent framework for defining replicability, in which agents are neither allied nor adversaries. I recover some of the prevalent practical approaches as special cases. I hope to provide guidance for a more systematic assessment of replicability in machine learning.

Key words and phrases: Replicability, prediction, decision theory.

1. INTRODUCTION

1.1 Preface

Prediction and machine learning technologies are playing increasingly important roles in science, as many fields leverage rapidly evolving data-generating technologies. Yet replicability, to many an essential element of science, remains inadequately studied in prediction, in part because its definition in relation to prediction remains somewhat elusive. In this article, I discuss concepts and issues in replicability of prediction from a perspective rooted in my experience as a practitioner of prediction approaches in biomedical research, and propose a framework for defining replicability.

1.2 Examples

I begin with examples, to give a concrete sense of application contexts and constituents. All are drawn from medicine, where prediction rules are regularly used to support decision making, and replicability is a practical concern. I hope the framework I propose will also guide investigations in other areas. While illustrative examples may make the concepts more concrete, my intent is not to provide a descriptive account, but rather to encourage discussion about prescriptive theories of replicability quantification.

In medicine, recent years have seen a trend toward regulating models and software using criteria similar to those used for medical devices. For example, the European Commission's Medical Devices Coordination Group has published guidance on the classification of software for regulatory purposes. A critical element of their definition is whether the software is intended to support decision making affecting patients or the public [2]. Regulatory requirements include validation in independent data as well as surveillance of the stability of performance in newly gathered data from practical settings after approval. These general trends, as well as many specific applications, are motivating the definitions in this paper.

1.2.1 *Predicting sepsis*. Models alerting clinicians to the presence of bacterial sepsis are widely used in emergency medicine. In April 2020, one such model, provided by a commercial entity (Epic Systems), was deactivated by one of the hospitals using it (the University of Michigan Hospital) because of the frequency of false positive alerting associated with COVID-19 [13]. To investigate this issue, [50] quantified the performance of sepsis detection models in 24 hospitals before and during the COVID-19 pandemic. Among individuals with COVID-19 virus infections, the relationship between fevers and sepsis differs from what is observed in the majority of the individuals in the dataset originally used for training the model. This leads to a deterioration of the performance of the model, as captured, for example, by an increase in the frequency of alerts, as the prevalence of individuals with COVID-19 increases. In this case, the issue emerged as a result of a data collection specific to the user's context. More broadly, Kelly et al. [21] discuss challenges for clinical implementation of machine learning algorithms across a diverse set of populations and systems.

Giovanni Parmigiani is Professor, Department of Data Science, Dana Farber Cancer Institute & Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: gp@ds.dfci.harvard.edu).

1.2.2 *Predicting survival of ovarian cancer patients*. High throughput measurement of gene transcription offers an opportunity to more accurately predict the survival times of patients diagnosed with cancer. In Waldron et al. [48], focusing on ovarian cancer, we carried out a comprehensive review of published prediction / scoring rules of this kind, identifying 14 from published articles that could be recoded with a high degree of reproducibility. Modelers divulge these algorithms with clinical users in mind, hoping they will serve to inform decision making at the bedside. In parallel, we comprehensively surveyed and collected all available data sets that could be used to assess model performance and systematically evaluated every rule on every dataset. Our work followed, and was in part motivated by, a case involving "premature use of omics-based tests" [56]. Our results, among others such as Chang and Geman [7], systematically document a consistent gap between the performance of prediction rules within the training studies (as measured, say, by cross-validation) and the performance of the same rule in relevant independent datasets. I will revisit this example later as "the ovarian cancer example."

1.2.3 *Evaluating retinal images*. Diabetic retinopathy is diagnosed with the support of imaging techniques. Automated interpretation of images is important for primary care settings. Investigators at the United States' Veteran Administration (VA) Health System [25] carried out a large prospective multi-center validation study to perform a head-to-head comparison of seven algorithms, including one FDA-approved algorithm, evaluating retinal images. I will revisit this example later as "the VA example."

1.2.4 *Screening for tuberculosis*. Chest radiography is used to screen people for pulmonary tuberculosis (TB). Deep learning (DL) neural networks are now available to interpret the images. Qin et al. [34] acquired images from two existing studies in Nepal and Cameroon, and compared three commercially available deep learning neural networks algorithms in both countries. I will revisit this example later as "the TB example."

1.2.5 *External assessment*. In each of these examples, replicability is evaluated via multiple data sets, with either no overlap of individual units with the data used to train the prediction rule, or a clear indication of whether this overlap exists and how it affects the results.

2. GLOSSARY

In this section, I try to clarify the use of the terms "prediction," "prediction rule," "replicability" and "study," also pointing briefly to challenges and issues with these definitions.

2.1 Prediction

A prediction, for the purpose of this discussion, is a statement $p \in \mathcal{P}$ about a future or unknown observable $y \in \mathcal{Y}$ (the label). A prediction rule generates predictions on the basis of observations $x \in \mathcal{X}$ (the predictors), and is thus a mapping $\phi : \mathcal{X} \to \mathcal{P}$. In scoring systems $\mathcal{P} \subseteq \mathbb{R}$; in statistical prediction $\mathcal{P}$ is either a probability space on $\mathcal{Y}$ or a probability space on probability distributions on $\mathcal{Y}$. I will also consider the simple binary case where the prediction rule directly assigns each point to one of two possible estimated labels, in which case $\mathcal{P} = \mathcal{Y}$.

The terms *observable* and *unknown*, which I used in my definition, are far from being self-explanatory. By *observed* (y or x) I mean that there is agreement, within a relevant group of individuals to be discussed further, about the precise value of the labels or predictors. This is not to say that I exclude disagreement altogether. Say radiologists A and B classify the label of the same medical image differently. Radiologist A judges it to reveal a "malignant" condition while B judges it to be "benign." This could be formalized by defining separate dimensions y_A and y_B within y. My discussion, however, is within confines where at some point, the disagreement within the group ceases, for example, because it is at least agreed that the two radiologists' answers are indeed y_A and y_B. By *observable* I mean that, should the observation process be carried out, there will typically be agreement on the value of the result. By "unknown" y, I mean simply that knowledge of the value of y is not part of the making of ϕ.

More broadly, observations are not in general separable from the theories that provided the framework to generate them, and from the contextual values of the prediction tasks. Think of predicting an individual's mental health outcomes, or their subversive political behavior, as examples. Thus my definitions, and by extension any ensuing consideration about replicability, are contextual to the goals, value systems, and theories that underlie the agreement among the individuals in the reference group. The nature and size of the reference group may vary widely in different contexts.

Prediction is regarded by several pioneers of statistical thought as the fundamental problem of statistics [16]. Examples include de Finetti, Pearson and arguably Bayes [43] and Laplace. Predictive approaches are supported by both empiricist and pragmatist considerations. An important motivation when discussing foundations is avoiding the additional degree of abstraction necessary to define concepts such as parameter, hypothesis, representation, latent class and so forth.

2.2 Replicability

Among the fundamental premises of the scientific enterprise is a degree of concordance among experimental

observations made in sufficiently similar circumstances. From this follows the desideratum that scientific predictions also agree well with experimental observations made across sufficiently similar circumstances. Defining this rigorously is not straightforward.

In 2019, the U.S.A.'s National Academy of Sciences (NAS) established a Committee on Reproducibility and Replicability in Science. Their report [32] is essential reading for those interested in this topic. Though their focus is primarily on scientific hypotheses, their definition is a good starting point for this discussion. Conclusion 3-1 on page 36 states: "Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." This definition is consistent with that given by the American Statistical Association [6] and with other work in statistics [18].

A distinct concept is that of repeatability: a repeatable prediction approach produces predictions without variation across independent tests carried out by repeating the entire process, including data collection, on the same individual or sampling unit [28]. This is important but is not examined here.

Replicability is also used in contrast to reproducibility, defined as "obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis" [32]. Usage of these terms is often inconsistent and is plainly reversed in computer science—a sarcastic twist in the parallel evolution of the same concepts in siloed fields. A thread of literature debates and documents related terminologies and their usage [22, 17, 1].

For predictions, I propose to modify the NAS definition to say:

DEFINITION 2.1 (Replicability of prediction rules). Replicability is obtaining consistent results across studies **suitable** to address the same scientific **prediction** question, each of which has obtained its own data.

Key edits are in bold. I narrowed the scientific question to prediction, but I broadened the definition to the consideration of suitable studies or datasets irrespective of the original aim of the data collection or design.

Determining whether any two studies are suitable for answering the same scientific prediction question is a matter of judgment. As was the case for observations in Section 2.1, my view is that it is useful to frame this determination in terms of inter-subjective agreement. Given the challenges of objectively defining both experimental results, and suitability of studies to a specific prediction question, I do not think it is useful to attempt an objective definition of prediction replicability. A more realistic goal is to aim for a broad consensus on observables and data, so that data can be used to descriptively quantify replicability in a way that will be found to be convincing by many.

The NAS report proposes to think of replication as "the act of repeating an entire study, independently of the original investigator without the use of original data." A strength of such replication activity is that similarities and differences between the original study and its replica are themselves part of the experimental design. In my opinion, this type of activity can in principle lead to the most compelling evidence about replicability. I will call it replication *by design*.

Definition 2.1, however, allows for a broader empirical scope, including data generation activities that may or may not originate to answer the same prediction question, or any prediction question. I will call it *observational* replication. While replication by design is defined in reference to a specific study and the activity of replicating it, observational replication is defined in reference to a specific prediction task and a collection of relevant datasets. Interest in forming collections of datasets for the purpose of understanding prediction rule replicability and study heterogeneity "in the wild" is growing—see, for example, the WILDS data collections [23].

External validation studies of prediction rules gather evidence about the applicability of a prediction rule beyond the conditions wherein it was trained, using available independent data [42]. This can be implemented by design or observationally or both depending on the circumstances. There is an element of replicability in these analyses, insofar as they compare properties of prediction rules across datasets. Many study designs and analytic techniques are relevant for both tasks. An important distinction is in the questions asked. Simplifying, validation asks weather a model's prediction ability is adequate for a certain set of tasks, while replicability asks whether prediction ability varies across multiple independent studies.

2.3 Studies

Formally, a study S is a collection of units, where a unit is a point in $(\mathcal{X} \times \mathcal{Y})$. So a study of size n is a point in $(\mathcal{X} \times \mathcal{Y})^n$. It is useful to frame discussions of replicability of predictions around a collection of relevant studies $S_1, \ldots, S_K$. The size of study k is n_k.

In an example of replicability by design, focus may be on decision rule ϕ, associated with a specific publication or software tool. Investigators may prospectively perform replication studies $S_1, \ldots, S_K$, (as in the VA Example 1.2.3 where $K = 2$) not necessarily from identically distributed populations. In this case, we have a sharp pre-existing definition of ϕ, $\mathcal{X}$ and $\mathcal{Y}$.

In an example of observational replicability one may gather evidence about the applicability of ϕ beyond the

conditions where it was trained, using existing data. $S_1, \ldots, S_K$ are chosen based on a set of inclusion criteria which could include: sufficient similarity of $\mathcal{X}$ and $\mathcal{Y}$ to those used in ϕ; sufficient relevance of the units sampled; sufficient quality, and so forth. Specifics will be heavily dependent on the context so it is difficult to provide general guidance. In Section 2.5, I will revisit the examples of Section 1.2 to help fix ideas.

Generally speaking, replicability by design is implemented prospectively, while observational replicability can also occur via retrospective data collections. The Medical AI Evaluation Database [51] catalogues medical artificial intelligence devices recently approved by the United States' FDA, and systematically reports on how they were evaluated before approval. Included are the three algorithms for the analysis of retinal images covered in Example 1.2.3, where prospective validation studies were carried out after the training of the algorithm was finalized. In two cases in the database the prospective study was multi-site. Prospective validation now accounts for a small minority of the approval processes reported by [51], just 4 of 130. However, there is interest in a more systematic use of prospective study design [12] for both validation and replicability assessment.

2.4 Study-to-Study Variability

A useful way to think about replicability is to identify interesting sources of variation across which it would be desirable for ϕ to be replicable, and define studies accordingly. For example, these can include variation in the technologies used for data collection, or in the selection criteria for including study units.

Ideally, identification of these sources of variation may begin as part of the initial study, through substantive insight as well as formal statistical analysis. Guidance on how to assess and report potential sources of variation exists in various application niches, such as the analysis of batch effects in high throughput biology [27, 26, 54].

While replicability can be evaluated across any collection of studies, the utility of this assessment is far greater if the study collection is defined and gathered in a systematic and comprehensive way, and based on criteria defined prior to the replicability analysis. Considerations are similar to those relevant in meta-analysis.

One way to conceptualize $S_1, \ldots, S_K$ is to think of it as a draw from a multi-level probability model, composed by a set of $q_k(x, y)$'s that generate units within each study, and a $q(1, \ldots, K)$ drawing study indices from a hypothetical population of studies. Much of the relevant variability discussed so far will translate into variation in the joint distributions q_k.

In machine learning, cross-study heterogeneity is described as "dataset shift." More specifically, "concept shift" refers to changes in the conditional probability of labels given predictors, while "covariate shift," refers to changes in the joint distribution of the predictors [24, 55] and "label shift" refers to changes in the marginal distribtion of labels. Moreno-Torres et al. [31] review and compare terminology and concepts.

2.5 Examples

This section revisits three of the earlier examples to illustrate inclusion criteria and sources of variation.

2.5.1 *Ovarian cancer example.* In Waldron *et al.* [48], we discuss in detail a case study where we form a collection of studies deemed suitable for the replicability analysis of a family of prognostic rules. We carried out a comprehensive review of available data, with predefined inclusion criteria. Sources of variation across studies include different microarray analysis technologies, differences in laboratory utilization of these technologies, differences in patient populations, including variation in stage and tumor size, and differences in clinical annotations (e.g., surgical outcomes). Nonetheless, studies are sufficiently comparable that meta-analytic biomarker discovery and model training provide robust results [38, 15].

When data collection technologies vary across studies, a nontrivial step, both practically and conceptually, is to map variables across studies. In this example, Ganzfried [15] illustrates the challenges of mapping transcriptomics data across high throughput technologies.

2.5.2 *VA example.* [25] prospectively collected data within the VA system at two separate locations, which constitute the studies in this case. Studies are homogeneous in important ways, including a shared IT infrastructure and data dictionaries, but vary in the populations served and some of the clinical workflows.

2.5.3 *TB example.* [34] retrospectively identified two relevant existing studies with sufficiently similar chest radiography images and clinical annotations. Variation arises from differences in populations and in referral patterns, among other sources.

3. A MULTI-AGENT FRAMEWORK FOR REPLICABILITY

Definition 2.1 refers to obtaining "consistent results." In this section, I propose a framework for quantifying consistency of results.

3.1 Roles

In the applications of Section 1.2, obtaining "consistent results" depends on modeling strategies of the developer, utilization patterns by the user, as well as the nature and variety of the collection of studies used for assessment. The process is complex. Multiple constituencies are involved and their goals are only partly overlapping. I propose to model it by defining three essential roles.

* *Modeler.* This is the entity developing ϕ through statistical / machine learning techniques.
* *User.* This is the entity applying ϕ to execute policy, commercial, legal, medical, or other decisions in practice.
* *Assessor.* This is the entity or group defining the relevant collection of studies $S_1, \ldots, S_K$, including definition of variables, and criteria for data quality. In Section 2.1, I mentioned a reference group who needs to agree on observables. The assessor groups needs to be a subset of this group.

My premise is that it is helpful to distinguish these three roles to arrive at definitions that address important societal uses of prediction algorithms. In some applications, roles may overlap. For example, the user and assessor may be the same entity. In others, the developer may also be the user. I find it hard, however, to frame replicability as a traditional single-agent decision problem in the vein of, say Savage [40]. In none of the examples of Section 1.2 are the interests of all entities involved fully aligned, although broadly speaking all agents may have an interest in replicability to occur.

In one version of these roles, the modeler, solely concerned about the construction of a useful ϕ, is *Algorithmic* in the sense of Breiman's two cultures [5]; the user, immersed in a specific medical or commercial reality with clearly defined goals, is a *Rational Bayesian* in the tradition of Ramsey [35]; and the assessors, in an effort towards neutrality, limit their scope to *Descriptive* statistics, a practice as old as the field. Alternatively, assessors could take a Fisherian perspective [14] and test for significance of departures from replicability. Further discussion on this can be found in Section 6.2.

Throughout, I assume that the modeler has not used any of studies $S_1, \ldots, S_K$ in the training of ϕ.

3.2 Single User

Consider first the scenario where ϕ is used by a single rational agent for supporting a specific decision, defined as the choice of a point a in a decision space $\mathcal{A}$ with the goal of maximizing the expectation of a utility function

$$U(a, x, y) : (\mathcal{A} \times \mathcal{X} \times \mathcal{Y}) \to \mathbb{R}.$$

A decision function is a mapping $\delta(\phi) : \mathcal{P} \to \mathcal{A}$ from predictions to actions.

I assume that the user is an expected utility maximizer and holds a personal probability distribution $\pi(x, y)$ on the observables relevant for their decision problem. π will affect the replicability analysis via the choice of the optimal decision function. Also, π may or may not reflect information arising in studies $S_1, \ldots, S_K$, but, to begin, will be independent of k. While it is critical that studies $S_1, \ldots, S_K$ are not used by the modeler in the development of ϕ, the same does not necessarily hold, in my view, for the user, although the nature of the replicability evaluation does change depending on whether π reflects these studies.

An optimal decision function δ^* satisfies

$$\delta^*(\phi) = \max_{\delta \in \Delta} E_\pi \{U(\delta(\phi(x)), x, y)\}.$$

I assume the user, as a rational agent, will utilize ϕ solely via δ^*.

The assessor, for each study in turn, will describe the user's utility through the vectors

$$(U(\delta^*(\phi(x_{1k})), x_{1k}, y_{1k}), \ldots, U(\delta^*(\phi(x_{n_k k})), x_{n_k k}, y_{n_k k}))$$

for $k = 1, \ldots, K$. For study k, the user's utility is, on average,

$$\mathcal{U}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} U(\delta^*(\phi(x_{ik})), x_{ik}, y_{ik}). \tag{1}$$

From here, I propose to define the prediction rule ϕ to be replicable if its optimal application to the same decision problem in different data sets leads to approximately the same average utility to the user. The degree of approximation can be formalized in many ways, and could itself be viewed as a decision problem if the assessor's role can be modeled in those terms.

A summary of deviations among $\mathcal{U}_k$'s is a useful departure point. For example, the $K \times K$ matrix $\mathcal{U}$ with generic element $\mathcal{U}_k - \mathcal{U}_{k'}$ could be examined or visualized. Binary summarizations of the $\mathcal{U}_k$'s can be used to define replicability. Two examples are in the following definitions:

DEFINITION 3.1 (Absolute ϵ-replicability). ϕ is ϵ-replicable in absolute utility over $S_1, \ldots, S_K$ if

$$\max_{k,k'} |\mathcal{U}_k - \mathcal{U}_{k'}| \leq \epsilon.$$

DEFINITION 3.2 (Relative ϵ-replicability). ϕ is ϵ-replicable in relative utility over $S_1, \ldots, S_K$ if

$$\max_{k,k'} \frac{2|\mathcal{U}_k - \mathcal{U}_{k'}|}{\mathcal{U}_k + \mathcal{U}_{k'}} \leq \epsilon.$$

Definitions 3.1 and 3.2 can be applied both to replication by design and observational replication, depending on how $S_1, \ldots, S_K$ is formed.

Definitions 3.1 and 3.2 are conditional on observed data. Agreeing on the conclusions only requires agreeing on the choice of studies and data integrity.

This descriptive, empirical, definition is in contrast to potential definitions that may require additional theoretical constructs, such as collections of hypothetical datasets defined by a data generating model, or families of such data generating models. In such constructs, the summation in Equation (1) would be replaced by the expectation with respect to a joint predictive distribution reflecting the

assessor's knowledge and beliefs. This distribution would not necessarily coincide with $\pi(x, y)$.

The ideas of this section are the basic building blocks for replicability assessment across multiple users and utility specifications. For example, our user could have multiple applications for the same prediction rule in different decision problems, each requiring a separate replicability analysis. More generally, the process could be repeated for various users separately, as illustrated in some of the examples below, each with potentially different decision spaces, priors, and utility functions.

3.3 Examples

3.3.1 *VA example*. In the VA study [25], the modelers are 5 participating companies which commercialize the algorithms considered. The replicability evaluation is carried out in parallel for 7 algorithms provided by the 5 companies. To the extent that any of these algorithms focus on the same clinical detection task, the companies are in direct competition. The users are physicians in the VA Health System. It is unknown, but possible, that individual physicians may have dual interest with some of the companies. The assessors are scientists working within the VA system. The utility of the algorithms is defined for the VA as an entity. User and assessor in this case are somewhat aligned, but are not necessarily in complete agreement. Assessors report not to have duality of interest with the companies [25]. This exemplifies a complex overlap of interests in the three roles.

The study collection consists of $K = 2$ VA hospitals, one in Seattle and the other in Atlanta. These two studies are used (a) together, to produce a replicability-by-design analysis of previous claims, and (b) separately, to examine replicability across components of the VA system. Labels are abstracted from medical records. A subset was regraded by a second expert, and differing grades were arbitrated by a retina specialist who did not know the identities of the graders. This illustrates the strengths gained from inter-subjective agreement on data. The clinical decision varies with the algorithm. For simplicity one may approximate it as whether or not additional follow-up is needed based on the retinal scan. Though multiple metrics are examined, the closest to a utility is the "value per encounter," defined by the authors as "the estimated pricing of each algorithm to make a normal profit (i.e., revenue and costs = 0) if deployed at the VA." This calculation was based on a two-stage scenario in which an AI algorithm would be used initially and then the images that screened negative would not need additional review by an optometrist or ophthalmologist" [25], page 1170. In the replicability analysis $|\mathcal{U}_k - \mathcal{U}_{k'}|$ is the difference in value of encounter between Atlanta and Seattle. This turned out to be nontrivial, owing, according to the authors, both to differences in the populations served, and to the quality of the images. One of the centers did not perform a useful preliminary dilation as often as the other before collecting the images.

3.3.2 *Ovarian cancer example*. In the Ovarian Cancer Study [48], the modelers are 14 research groups who published prognostic algorithms meeting a set of criteria in terms of clinical goals and reproducibility of code. The replicability evaluation is carried out in parallel. The assessors are scientists funded by the NIH. There is some overlap between the assessors and the modelers for at least two of the models. The potential conflict is addressed by "freezing" the algorithms to the version originally published and by providing a transparent and reproducible analysis workflow for the replicability work. The potential users are physicians, although none of the 14 algorithm was in broad clinical use at the time of the evaluation. In fact the validation and replicability analyses included among their goals to assess whether this family of rules was ready for clinical application. No formal decision framework is considered in [48].

3.4 Dominance

Consider now comparing ϕ to other classifiers. Beginning with two studies, a useful perspective is to partition the average utility space as in Figure 1. The dot is positioned at coordinates $(\mathcal{U}_1, \mathcal{U}_2)$ for ϕ. Compared to ϕ, an alternative classifier in region B would display a better average utility in both studies, as well as better replicability, because the empirical average utilities would be closer to each other. An alternative in region C displays better average utility in both studies, but worse replicability; the reverse is true in region A.

This reasoning suggests that, if we are interested in both high utility and replicability, we can define dominance in this context as follows. Consider classifiers ϕ and ϕ', the latter with average utilities $\mathcal{U}'_k$, $k = 1, \ldots, K$.

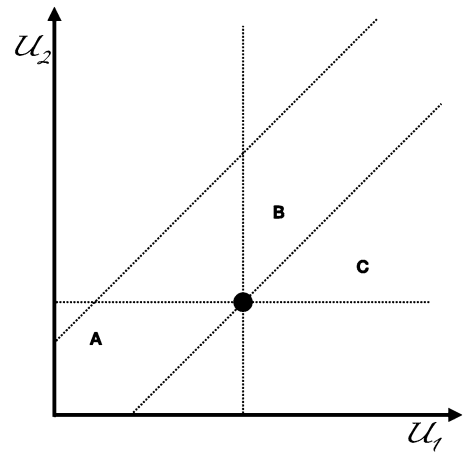

FIG. 1. *Regions of Average Utility for comparison between ϕ and alternative classifiers. The dot is positioned at coordinates $(\mathcal{U}_1, \mathcal{U}_2)$ for ϕ. The two 45° lines are equidistant from the main diagonal, not shown. Letters denote regions of interest considered in the text.*

DEFINITION 3.3 (Dominance). ϕ dominates ϕ' in absolute utility and replicability over $S_1, \ldots, S_K$ if both

$$\max_{k,k'} |\mathcal{U}_k - \mathcal{U}_{k'}| \leq \max_{k,k'} |\mathcal{U}'_k - \mathcal{U}'_{k'}|$$

and

$$\mathcal{U}'_k \leq \mathcal{U}_k, \quad k = 1, \ldots, K.$$

Formally, this definition could be applied to the context of Section 3.2. However, a decision maker with their own probability model π may not decide based on the empirical average utilities, but rather update π in the light of the studies and evaluate expected utilities accordingly. On the other hand, the assessor would remain interested in the empirical average utilities, and, to the extent that the utility function captures objectives of common interest, may value comparisons such as those of Figure 1. For example, assessors can gain insight about whether replicability is achieved at the cost of a worsened performance in some studies.

Figure 1 suggests the possibility of a formalization where the assessors hold their own utility function, capturing replicability. We could then leverage multi-agent decision theory approaches [20] to understand the trade offs between the goals of accuracy and replicability. This might be interesting in some applications but too restrictive in others. For example, if ϕ is of general public health utility, both the user and the assessor would have an interest in high values of $\mathcal{U}$.

4. THE BINARY CASE

4.1 Replicability of Binary Prediction Rules

In many applications, the product of the prediction rule are binary class labels rather that probabilities or risk scores. I will denote this special class of algorithms by the script variant of the letter phi, as in $\varphi : \mathcal{X} \to \mathcal{Y}$.

In the user-centered approach of Section 3.2 this is a special case obtained setting $\mathcal{A} = \mathcal{Y}$ and considering the rule φ constructed from prediction/scoring rule ϕ via

$$\varphi(x) = \delta^*(\phi(x)) = \max_{\delta \in \Delta} E_\pi\{U(\delta(\phi(x)), x, y)\}.$$

The definition of $\mathcal{U}_k$ specializes to:

$$\mathcal{U}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} U(\varphi(x_{ik})), x_{ik}, y_{ik}) \tag{2}$$

based on which we can apply Definitions 3.1 and 3.2.

A formal tie between φ and the trio (δ, U, π) is not always made explicitly. Nonetheless it is often the case that binary prediction rules are built with some consideration of the modeler's expectation of x and y, and desired properties of δ for users. A common example occurs when modelers first build a prediction rule ϕ and then apply thresholds to dichotomize the results. Another interesting scenario covered by this case is one where modeler and user are the same entity and the action space is binary.

When δ and π are not explicitly specified, one can still evaluate replicability of φ by positing a utility function directly as

$$U(\varphi, y) : (\mathcal{Y} \times \mathcal{X} \times \mathcal{Y}) \to \mathbb{R}.$$

For example, if $U(\varphi, x, y) = I_{\varphi=y}$ for every x, then $\mathcal{U}_k$ is the empirical proportion of cases where prediction and outcomes coincide in study k and ϵ-replicability obtains when this proportion does not vary by more than ϵ in any two-study comparison. Here replicability can be characterized without reference to a user's subjective probability distribution π. The utilities still refer to a specific, albeit hypothetical decision problem. In practice utilities like $U(\varphi, x, y) = I_{\varphi=y}$ are widely used for simplicity and in the hope that they may capture the quality of a classifier well enough across several potential applications.

Another commonly used approach in binary prediction is to separately penalize the two possible errors. This is particularly important in medical applications where it is rarely the case that the error of assigning a low risk individual to the high risk class is as severe as the opposite. This holds across a wide range of clinical applications.

A typical generalization of $U(\varphi, x, y) = I_{\varphi=y}$ is

$$U(\varphi, x, y) = u_{01} I_{\varphi<y} + u_{10} I_{\varphi>y}, \tag{3}$$

where u_{01} and u_{10} are negative numbers quantifying the consequences of each of the two error types. Defining the relative frequencies of the study-specific errors as

$$f^k_{01} = \frac{1}{n_k} \sum_{i=1}^{n_k} I_{\varphi(x_{ik})<y_{ik}}$$

and

$$f^k_{10} = \frac{1}{n_k} \sum_{i=1}^{n_k} I_{\varphi(x_{ik})>y_{ik}}$$

leads to

$$\mathcal{U}_k = u_{01} f^k_{01} + u_{10} f^k_{10}. \tag{4}$$

In the canonical two-by-two table of labels versus predictions, f^k_{01} and f_{10} represent the off-diagonal joint probabilities. Any, but not necessarily all, aspects of the K joint distributions of ϕ and y may be relevant for assessing replicability. The utility function reflects information on which are relevant in a specific problem. For example, the two frequencies of correct predictions are not distinguished in this utility specification, as they are both assigned the maximum utility, that is, 0. The function U defines a three-set partition of the four-set sample space for ϕ and y into equivalence classes relevant for the user's decision. In this case, cross-study variation in the frequencies of correct predictions would be irrelevant for user-based replicability as long as the overall proportion remained constant.

Expression (4) can be further rewritten in terms of the empirical sensitivity and specificity of detection of $y = 1$, and the empirical proportions of $y = 1$ cases, as long as all the latter are nonzero. Viewed this way, replicability requires sufficient stability of these three quantities across studies, so that their combination according to (4) may not vary by more than ϵ. The empirical proportions of $y = 1$ cases remains necessary for the evaluation of (4). As a corollary, metrics that depend solely on sensitivity and specificity cannot be viewed as special cases of (4).

Importantly, the same collection of studies may meet the user-centered replicability definition for one user and not another, depending on of their specification of u_{01} and u_{10}. The same classifier may be sufficiently replicable for a user that requires high sensitivity, but not for another requiring high specificity.

Definition 3.2 can be used to build extensions to multiple utility functions. These could arise either from the need for the agent to use ϕ in multiple applications, or from the desire to define replicability across a set of agents, each with their own utility. Defining user-centered replicability over a class of users requires either scaling the utility functions so they are comparable or using a vector of bounds instead of a single ϵ in the definition. This variant is not pursued here in any detail.

4.2 Examples

4.2.1 *TB example*. In [34], the modelers are companies developing algorithms for image analysis. The assessors are supported by philanthropic funding, and no conflicts with companies are reported. Assessors selected companies based on a comprehensive literature review, further strengthening the independence of the assessment. The user is a hypothetical TB clinician in one of the two countries considered, Cameroon and Nepal.

The algorithms generate "abnormality scores" ϕ. Labels y encode bacteriologically confirmed TB status. Assignment of images to predicted classes $\delta^*(\phi)$ is based on a threshold on the abnormality score. According to the authors, there are no generally recommended threshold scores to use, which motivates them to consider ROC curves.

Each specification of u_{01} and u_{10} implies an optimal threshold, which depends on these quantities as well as empirical specificity, sensitivity and prevalence [30]. Perhaps the ROC could be viewed as a step towards assessing performance and replicability across a range of specifications for u_{01} and u_{10}, each corresponding to a different hypothetical user. However, explicit consideration of prevalence is required for the computation of any instance of (2), and is lacking from ROC analysis and related summaries. A similar consideration applies to the C-index for time-to-event outcomes used in [48] in Example 1.2.2.

4.2.2 *Sepsis*. In Example 1.2.1, ϕ produces a risk score used for classifying patients according to their risk of sepsis. The developer is a company, providing IT services to many hospitals. Again we have multiple users, each aiming to optimally use the score to inform clinical decisions in their emergency rooms. Assessors are academic researchers, some of whom work at one of the $K = 24$ hospitals studied. Replicability is evaluated across hospitals as well as over time within each hospital. Replicability is evaluated with respect to the proportion of patients generating sepsis alerts per day, or $\sum_i I_{\varphi(x_{ik})=1}$.

4.3 The Case of K = 1

In replicability by design, as well as in many model validation efforts [42, 8] one may begin the assessment with a single study S_1. A user-based replicability analysis can still be carried out if a benchmark value $\mathcal{U}_0$ is available, typically from the modeler. In this case, replicability analysis reduces to the comparison of the vector

$$\big(U\big(\varphi(x_{11}), x_{11}, y_{11}\big), \ldots, U\big(\varphi(x_{n_1 1}), x_{n_1 1}, y_{n_1 1}\big) \big)$$

to the benchmark $\mathcal{U}_0$, followed by appropriate summarizations. When knowledge of the user's δ^* is available to the modelers, they can then evaluate $\mathcal{U}_0$ using their own validation data, or earlier published evidence, or set-aside data from the training set. In another scenario, modeler and user cooperate in evaluating $\mathcal{U}_0$ before external assessment.

5. DISTANCE REPLICABILITY

In this section, I consider a scenario with a modeler and an assessor, possibly coinciding, and no user. This may be relevant in early stages of development of a model, before individual users are identified specifically, or in the case of models that target a very wide range of users with distinct goals, as do models embedded in smartphone apps.

Irrespective of the decision problem at hand and of the user's utility, the realized average utilities $\mathcal{U}_1, \ldots, \mathcal{U}_K$ only depend on the data through the triplets

$$\big(\phi(x_{ik})\big), x_{ik}, y_{ik}) \quad i = 1, \ldots n_k, k = 1, \ldots, K.$$

Let F_k be the empirical joint cumulative distribution of the points $(\phi(x_{ik})), x_{ik}, y_{ik})$ $i = 1, \ldots n_k$. A driver of replicability across different utility functions is the similarity among the distributions $F_1, \ldots, F_K$. If ϕ is a classifier for a discrete label, and the utility only depends on x through ϕ, then F is a simple bivariate distribution on a contingency table, or confusion matrix. If ϕ generates a probability distribution, more general spaces are required, but the concepts are similar.

Mirroring the development in Section 3.2, define $D(F_k, F_{k'})$ to be a distance between the c.d.f.'s F_k and $F_{k'}$, such as the total variation distance on the appropriate space. Then we can posit the following definition.

DEFINITION 5.1 (Distance ϵ-replicability). ϕ is ϵ-replicable in distance over $S_1, \ldots, S_K$ if

$$\max_{k,k'} D(F_k, F_{k'}) \leq \epsilon.$$

A predictor ϕ defines a partition of $\mathcal{X}$ into sets with equal ϕ. Generally, for given $S_1, \ldots, S_K$, the coarser the partition, the smaller $\max_{k,k'} D(F_k, F_{k'})$ will be. Even for the degenerate case in which ϕ is constant over $\mathcal{X}$, Definition 5.1 may not hold owing to differences in the distribution of y's.

A contrast with Definitions 3.1 and 3.2 is provided by the following observation. For given π and U, $\mathcal{U}_k$ is a functional of F_k. The difference

$$D_{\pi,U}(F_k, F_{k'}) \equiv |\mathcal{U}_k - \mathcal{U}_{k'}|$$

fails to satisfy the definition of distance among empirical c.d.f's, because it is possible to have $D_{\pi,U} = 0$ with $F_k \neq F_{k'}$.

Analyses that consider solely properties of the distributions of prediction rules conditional on class labels, such as the ROC curve or the C-index are also not covered by this definition, again because equality of conditional distributions alone does not imply equality of the joint distributions.

When $\epsilon = 0$, we can refer to distance replicability and user-based replicability as exact. This case is not of much practical interest as long as replicability is defined descriptively, as sampling variation will generally be present and will generate some variation across studies. Nonetheless it is conceptually interesting to note that exact distance replicability is a more strict requirement than exact user-based replicability. In other words if ϕ is exactly replicably by distance than it must be exactly replicable for any user in the sense of Section 3.2.

To see this, consider that equality of the empirical c.d.f's requires equality of the support points and associated point masses. This in turn occurs only if one of the two studies is formed by collating b copies of the other, $b = 1, 2, 3, \ldots$. From this, follows that each element in the sum (1) for the study with smaller sample size has b identical terms in the sum (1) for the other. As the sample size of the larger studies is b times that of the smaller one, b cancels in the averaging and the result follows.

For an example, return to the setting of Expression (4) and define f_{00}^k and f_{11}^k to be the frequencies of the two possible correct classifications. A pair of studies such that $f_{00}^k + f_{11}^k = f_{00}^{k'} + f_{11}^{k'}$, $f_{01}^k = f_{01}^{k'}$, and $f_{10}^k = f_{10}^{k'}$ but $f_{00}^k \neq f_{00}^{k'}$ will have exact replicability for users with utility (3) but will fail to achieve exact distance replicability.

6. INFERENCE

6.1 Uncertainty Quantification for Replicability

All the descriptive statements presented so far could be complemented by uncertainty quantification. Inevitably, this would require an additional layer of assumptions on the part of the assessor. I will mention here approaches that require a minimal amount of modeling and therefore a modest degree of additional stipulations.

To begin, one can consider resampling. Bootstrap of units within each study in turn would provide variance estimates for each $\mathcal{U}_k$ and, assuming independence of the studies, of each element of the matrix $\mathcal{U}$. Variance estimates obtained in this way would condition on the selection of studies $S_1, \ldots, S_K$. For a simple extension, Davison and Hinkley [10] describe a randomized cluster bootstrap procedure where both clusters (in this case studies) and observations within a cluster are sampled with replacement.

In some cases, uncertainty may extend to study membership of individual units. In one example, data are extracted from a single encompassing data collection infrastructure and partitioned into k studies based on geographical or administrative criteria, which could reasonably be specified at different levels of resolution. In another, individuals may be assigned to studies based on ethnicity, a trait that may not be known with certainty in some cases. The study strap approach of Loewinger et al. [29], is a resampling technique that generates a collection of "pseudo-studies," generalizing the randomized cluster bootstrap. The study strap is controlled by a tuning parameter that determines the proportion of observations to draw from each study, and can be used to dial the amount of study heterogeneity in the synthetic data throughout a range going from what is empirically observed to the case of complete exchangeability of units. The latter extreme is not a useful setting for a replicability analysis, but choosing tuning parameters that generate collections of studies close to the empirical distribution could provide a useful sensitivity analysis.

For a given ϵ, and for any of the definitions in the preceding sections, resampling procedures would produce a proportion of cases that satisfy ϵ-replicability. In turn, these could serve as an uncertainty quantification of whether ϕ meets the definition.

6.2 Rejecting Replicability

Another basic inferential question is whether replicability can be rejected via a significance testing approach. For an example with distance replicability, consider $\max_{k,k'} D(F_k, F_{k'})$ from Definition (5.1) to be the test statistic of interest. A simple procedure for producing significance statements in this context is to generate a permutation null for the vector $(\mathcal{U}_1, \ldots, \mathcal{U}_K)$ and its functions by permutations of the study labels, and compute a p-value based on the permutation distribution. Multiple testing methods can also be relevant if one wishes to separately assess replicability for each of the pairwise comparisons. Elements of $\mathcal{U}$ are not independent, which requires additional care.

7. DISCUSSION

7.1 Sampling Frame

In my definition of a study, the sampling frame is the space $(\mathcal{X} \times \mathcal{Y})$. Before units are sampled into a study, both x and y are unknown. All the descriptive measures of replicability proposed in this paper consider joint variation of both x and y. Any heterogeneity of this joint variation across studies can and should challenge replicability. In machine learning terminology, replicability should be challenged by any of label shift, covariate shift or concept shift [24, 55]. This applies to both the utility-based definitions and the distance-based definitions.

The theory outlined in this paper does not cover efforts aiming at the useful but more limited goal of assessing the discrimination ability of prediction rules. These efforts, in analysis and often in design, condition on the class labels y. Examples include the ubiquitous ROC analysis which is often used as a criterion, or the sole criterion, for evaluating prediction rules. In scenarios with label shift, failures of replicability with important practical consequences can elude class-conditional analyses.

7.2 Local Replicability

My discussion considered properties of algorithms when applied to entire datasets. In this sense they are all global properties. In many applications, it may be very interesting to consider groups within these studies. Replicability could be differentially evaluated within each group. If we define $\mathcal{X}^* \subset \mathcal{X}$ to be any subset of the feature space, we can revisit every definition given in the preceding sections, upon restricting the analysis to $x \in \mathcal{X}^*$, provided the set is not empty. In general, the variation of $\mathcal{U}$ across studies will depend on the $\mathcal{X}^*$ chosen, and it may be the case that replicability is achieved in some groups but not others. When $\mathcal{X}$ is coarse and cells are sufficiently populated, this logic can be pushed to the level of considering each cell separately to serve as $\mathcal{X}^*$.

7.3 Algorithmic Fairness

It is interesting to think about replicability across collections of studies where individuals in different studies have the same rights (applicants for credit or for educational opportunities), or users of predictions have the same ethical responsibilities (medical providers). In this type of circumstance, it may be possible to construct meaningful collections of studies around notions of algorithmic fairness of the prediction rule studied [11, 53]. Algorithmic fairness in classification is concerned with preventing discrimination against individuals based on their membership in some group. A connection arises with (global) distance replicability if one chooses the labels $1, \ldots, K$ to represent these groups. If ϕ satisfies distance replicability, it will be difficult for any user to discriminate among groups, on average, using ϕ.

It is more difficult to tie fairness to user-based replicability, as fair algorithms could produce different user's utilities in different groups, as a result of "benign" variation in the F_k's, that is, variation that is not associated with a discriminative use of ϕ.

Alternatively, a protected groups assignment could be used to investigate local replicability by appropriately choosing $\mathcal{X}^*$. Achieving local replicability within a protected group would not protect from discrimination if it exists, but may quantify its replicability.

These considerations apply to groups. Dwork et al. [11] define a seminal framework for fair classification at the individual level. In their words, it comprises:

1. *a (hypothetical) task-specific metric for determining the degree to which individuals are similar with respect to the classification task at hand;*
2. *an algorithm for maximizing utility subject to the fairness constraint, that similar individuals are treated similarly.*

Here the definition of similarity among individuals should not include the group membership we intend to protect. Their approach, like the one I describe here, is also multi-agent, and also considers the modeler separately from the user. In contrast, it also explicitly considers the rights of the individuals being classified, which I did not consider. Extending the framework of this paper to include individuals as a fourth role could be interesting. Depending on the application context, individuals may have an interest in any subset of fairness, replicability, and prediction accuracy.

7.4 Retraining

I discussed how to define and quantify replicability of an algorithm ϕ that was previously trained and remains fixed throughout the analysis. My goal is to capture the implementation stage of a machine learning algorithm, once the development is completed. From a methodological perspective, however, the question of replicability can and should also be asked of model fitting techniques.

Given a collection of k studies, and analogously to what I described in Section 3.2, replicability of training techniques can be explored using designs that consider every pair of studies. For an example, in Bernau et al. [3] we defined a cross-study validation matrix whose generic element measures predictive performance when one trains a predictor in study k and evaluates it externally in k'. Our goal was to investigate properties of methodologies for training classifiers when external replicability is a goal. In contrast to how $\mathcal{U}$ is defined in Section 3.2, ϕ_k is different in every row and trained *de novo* using study k. Both $\mathcal{U}$ and the Bernau et al. version, offer the opportunity to learn about study heterogeneity and outlying studies. See also [44].

Another useful design for investigating replicability of training techniques is the leave-one-study-out design [39], which applies the jackknife logic at the study level. When training on $K-1$ studies, however, a challenge is to properly incorporate potential study-to-study heterogeneity, an issue considered in Section 7.6.

In the social sciences, Vijayakumar and Cheung [46] investigated the replication success of R^2 in both cross-validation and cross-study validation. They focus on three replication aims: (1) tests of inconsistency to examine whether single replications reject the originally reported study-specific R^2; (2) tests of consistency based on a region of equivalence, and (3) meta-analytic intervals for accuracy measures—a goal also pursued by Waldron et al. [48].

In addition to prediction performance and the utility thereof, it is interesting from a methodological standpoint to investigate replicability of various aspects of model construction, such as dimensionality, smoothness, variable selection and variable importance. Examples include [47]. Yu and Kumbier [52] emphasized model stability as a guiding principle. They define it as *acceptable consistency of a data result relative to appropriate perturbations of the data or model*. As examples of perturbation they suggest jackknife, bootstrap, and cross validation. Multi-study extensions would be interesting from a replicability standpoint.

7.5 Testing Replicability and Replicability of Testing

In a testing approach to replicability of prediction rules, such as that sketched in Section 6.2, the null hypothesis is the equality of the expected utility of a prediction rule across studies for a specific user or users. This has not to my knowledge been explored in depth.

On the other hand, there is a robust and useful literature on assessing the replicability of tests of hypotheses, which considers whether a hypothesis about the data generating mechanism is replicably rejected in multiple studies. Often this is framed in the context of multiple testing. An important foundational paper in this area is Heller and Benjamini [18] who introduce the r-value, defined as the lowest false discovery rate at which a given finding can be called replicated.

The context of meta-analysis, the systematic combination of results from studies investigating the same hypotheses, offers an interesting contrast to replicability analyses in terms of how the two approaches relate to study-to-study variation. For an anecdote tied to Example 1.2.2, in [15] we identified expression of the gene CXCL12 as prognostic of overall survival in patients with ovarian cancer, via the combined analysis of 14 studies, in only two of which CXCL12 expression is a significant predictor. [19] offers additional and more systematic exploration of the discordance of goals and results between meta-analysis and replicability.

Since the publication of [15], substantial biological evidence has accumulated on the prognostic role of this gene (see [9] and references therein), thanks to progress in cancer immunology research. This supports the conclusion of the meta-analysis. On the other hand, the data of [15] would most likely provide evidence against the replicability of the hypothesis that CXCL12 expression is associated with survival, as measured by, say, an r-value. And although such analysis was not carried out, the data would likely have questioned replicability of prediction rules based solely on CXCL12 expression as well. Replicability reaches a different conclusion compared to meta-analysis because it asks a different question. Replicability is intentionally sensitive to the heterogeneity of study designs, the challenges in the normalization across technologies of the measured expression of a gene with only moderate transcriptional activity, and, if implemented via significance, the study sample sizes. In contrast, meta-analysis hopes to find signal in the midst of this variation.

7.6 Learning Replicability

After discovering failures of replicability, whether in inference or prediction, a reasonable next step is to move beyond a single study analysis, and tackle the study-to-study variation as part of the learning process. In inference, meta-analysis exemplifies this.

In prediction, availability of multiple datasets suitable to address the same or similar prediction question offer the opportunity to train algorithms that can incorporate knowledge of cross-study heterogeneity and produce predictions that are more likely be replicable in future data from the same or other studies. The domain generalization literature is particularly germane here as it focuses on leveraging multiple datasets in model training to improve prediction performance on an unseen, but related, domain [49].

In statistics, interest in drawing upon multiple data sets in prediction is also emerging. Approaches include meta-analyzing model coefficients (e.g., [37, 36, 45]) and ensembling models with weights that reward replicability. Specifically, in Patil and Parmigiani [33] we propose a multi-study generalization of stacking [4] to achieve this goal. Our approach comprises two stages: (1) training models on each study separately, and (2) ensembling them via a stacking regression on the merged data. This structure rewards cross-study prediction performance as ensemble weights are primarily driven by how well each model predicts across studies different from the one where it was trained.

7.7 Prediction Tasks

An interesting direction for generalization is to characterize replicability of a broadly defined prediction task, such as response to a drug treatment based on information

on a patient's genome. In this case, we would not necessarily have a specific ϕ or a methodology of interest. We would need to establish a class of measurements of x and y that constitute a sufficiently homogeneous collection to be worth studying from a replicability viewpoint, and potentially extend the definition of replicability to classes of ϕ's or optimally selected ϕ's within a class.

7.8 Broader Perspectives

An open question is whether or not to approach the definition of replicability from a game theoretic perspective. I imagine there would be many ways of meaningfully doing so depending on the context in which the prediction is developed, used, and assessed. My attempt here is to take a slightly more general perspective where the goal is to provide definitions that address trustworthiness of predictions across a relevant scientific community. Of course this is predicated on trust in the assessor and data. But even taking this trust for granted, there remains a gap before we arrive at statements about a specific prediction rule for a specific application. This is the gap my definitions try to fill.

While I discussed illustrative examples to make the concepts more concrete, I did not intend to provide a descriptive theory but rather to encourage discussion about prescriptive theories of replicability quantification. As a first step, my sense is that an explicit and transparent statement about who are the actors in the roles of modeler, user and assessor is a foundational step that should be encouraged in these analysis. The next is to explicitly connect the metrics used to quantify replicability to users' decisions.

I will close by noting that ultimately, as a field, we would benefit from examining the development, validation and trustworthiness of prediction rules in their historical, cultural, and social contexts. Such efforts would be close in scope to the field of science and technology studies [41] which has already contributed important perspectives to epistemology.

ACKNOWLEDGMENTS

I presented a preliminary version of Section 3 at a 2022 symposium on "Statistical methods and models for complex data," held in Padova. I am grateful to my discussants Marco Alfò and Gianmarco Altoè for very thoughtful comments, and to Aldo Solari for encouraging me to think about falsifiability in the context of replication. Mike Daniels, Lorenzo Trippa, Michael Lavine and two insightful reviewers helped with comments on earlier drafts.

FUNDING

Work supported by NSF Grant DMS-2113707.

REFERENCES

[1] Barba, L. A. Terminologies for reproducible research. Available at arXiv:1802.03311.

[2] Beckers, R., Kwade, Z. and Zanca, F. (2021). The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. *Phys. Med.* **83** 1–8. https://doi.org/10.1016/j.ejmp.2021.02.011

[3] Bernau, C., Riester, M., Boulesteix, A., Parmigiani, G., Huttenhower, C., Waldron, L. and Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112. https://doi.org/10.1093/bioinformatics/btu279.

[4] Breiman, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64. https://doi.org/10.1007/BF00117832

[5] Breiman, L. (2001). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–231. https://doi.org/10.1214/ss/1009213726

[6] Broman, K., Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D. and Wickham, H. (2017). *Recommendations to funding agencies for supporting reproducible research*. Amer. Statist. Assoc., Alexandria, VA.

[7] Chang, L.-B. and Geman, D. (2015). Tracking cross-validated estimates of prediction error as studies accumulate. *J. Amer. Statist. Assoc.* **110** 1239–1247. https://doi.org/10.1080/01621459.2014.1002926

[8] Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M. et al. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med. Res. Methodol.* **14** 40. https://doi.org/10.1186/1471-2288-14-40

[9] D'Alterio, C., Spina, A., Arenare, L. and Chiodini, P. (2022). Biological role of tumor/stromal CXCR4-CXCL12-CXCR7 in MITO16A/MaNGO-OV2 advanced ovarian cancer patients. *Cancers* **14** 1849.

[10] Davison, C. A. and Hinkley, D. V. (1997). *Boostrap Methods and Their Applications*. Cambridge Univ. Press, New York.

[11] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* 214–226. ACM, New York. MR3388391

[12] Ebrahimian, S., Kalra, M. K., Agarwal, S., Bizzo, B. C., Elkholy, M., Wald, C., Allen, B. and Dreyer, K. J. FDA-regulated AI algorithms: Trends, strengths, and gaps of validation studies. *Acad. Radiol.* **29** 559–566. https://doi.org/10.1016/j.acra.2021.09.002

[13] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S. and Saria, S. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385** 283–286. https://doi.org/10.1056/NEJMc2104626

[14] Fisher, R. A. (1925). *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh.

[15] Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J. et al. (2013). curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database* (*Oxford*) **2013** bat013. https://doi.org/10.1093/database/bat013.

[16] Geisser, S. (1993). *Predictive Inference*: *An Introduction*, Chapman & Hall, New York.

[17] Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci. Transl. Med.* **8** 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

[18] HELLER, R., BOGOMOLOV, M. and BENJAMINI, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc. Natl. Acad. Sci. USA* **111** 16262–16267. https://doi.org/10.1073/pnas.1314814111

[19] JALJULI, I., BENJAMINI, Y., SHENHAV, L., PANAGIOTOU, O. A. and HELLER, R., Quantifying replicability and consistency in systematic reviews. *Stat. Biopharm. Res.* **15** 372–385. https://doi.org/10.1080/19466315.2022.2050291

[20] KEENEY, R. L., RAIFFA, H. and MEYER, R. F. (1976). *Decisions with Multiple Objectives*: *Preferences and Value Tradeoffs*, Wiley & Sons, New York.

[21] KELLY, C. J., KARTHIKESALINGAM, A., SULEYMAN, M., CORRADO, G. and KING, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17** 195. https://doi.org/10.1186/s12916-019-1426-2

[22] KENETT, R. S. and SHMUELI, G. (2015). Clarifying the terminology that describes scientific reproducibility. *Nat. Methods* **12** 699–699. https://doi.org/10.1038/nmeth.3489

[23] KOH, P. W., SAGAWA, S., MARKLUND, H., XIE, S. M., ZHANG, M., BALSUBRAMANI, A., HU, W., YASUNAGA, M., LANAS PHILLIPS, R. et al. WILDS: A benchmark of in-the-wild distribution shifts. Available at arXiv:2012.07421.

[24] KOUW, W. and LOOG, M. (2019). An introduction to domain adaptation and transfer learning. Available at arXiv:1812.11806.

[25] LEE, A. Y., YANAGIHARA, R. T., LEE, C. S., BLAZES, M., JUNG, H. C., CHEE, Y. E., GENCARELLA, M. D., GEE, H., MAA, A. Y. et al. (2021). Head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care* **44** 1168–1175. https://doi.org/10.2337/dc20-1877

[26] LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11** 733–739. https://doi.org/10.1038/nrg2825

[27] LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** e161. https://doi.org/10.1371/journal.pgen.0030161

[28] LEMAY, A., HOEBEL, K., BRIDGE, C. P., BEFANO, B., SANJOSÉ, S. D., EGEMEN, D., RODRIGUEZ, A. C., SCHIFFMAN, M., CAMPBELL, J. P. et al. (2022). Improving the repeatability of deep learning models with Monte Carlo dropout. *npj Digit. Med.* **5** 174. https://doi.org/10.1038/s41746-022-00709-3

[29] LOEWINGER, G., PATIL, P. KISHIDA, K. T. and PARMIGIANI, G. (2022). Hierarchical resampling for bagging in multistudy prediction with applications to human neurochemical sensing. *Ann. Appl. Stat.* **16** 2145–2165. https://doi.org/10.1214/21-AOAS1574

[30] METZ, C. E. Basic principles of ROC analysis. *Semin. Nucl. Med* **8** 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

[31] MORENO-TORRES, J. G., RAEDER, T., ALAIZ-RODRÍGUEZ, R. and CHAWLA, N. V. (2012). A unifying view on dataset shift in classification. *Pattern Recognit.* **45** 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[32] COMMITTEE ON REPRODUCIBILITY AND REPLICABILITY IN SCIENCE (2019). *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C. https://doi.org/10.17226/25303

[33] PATIL, P. and PARMIGIANI, G. (2018). Training replicable predictors in multiple studies. *Proc. Natl. Acad. Sci. USA* **115** 2578–2583.

[34] QIN, Z. Z., SANDER, M. S., RAI, B., TITAHONG, C., SUDRUNGROT, S., LAAH, S. N., ADHIKARI, L. M., CARTER, E. J., PURI, L. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci. Rep.* **9** 15000. https://doi.org/10.1038/s41598-019-51503-3

[35] RAMSEY, F. (1926). *The Foundations of Mathematics*, Oxford University Press Oxford.

[36] RASHID, N. U., LI QUEFENG, Y., JEN, J. and IBRAHIM, J. G. (2020). Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *J. Amer. Statist. Assoc.* **115** 1125–1138. https://doi.org/10.1080/01621459.2019.1671197

[37] RIESTER, M., TAYLOR, J. M., FEIFER, A., KOPPIE, T., ROSENBERG, J. E., DOWNEY, R. J., BOCHNER, B. H. and MICHOR, F. (2012). Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clin. Cancer Res.* **18** 1323–1333. https://doi.org/10.1158/1078-0432.CCR-11-2271

[38] RIESTER, M., WEI, W., WALDRON, L., CULHANE, A. C., TRIPPA, L., OLIVA, E., KIM, S.-H., MICHOR, F., HUTTENHOWER, C. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl. Cancer Inst.* **106** dju048–dju048. https://doi.org/10.1093/jnci/dju048

[39] RIESTER, M., WEI, W., WALDRON, L., CULHANE, A. C., TRIPPA, L., OLIVA, E., KIM, S.-H., MICHOR, F., HUTTENHOWER, C. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl. Cancer Inst.* https://doi.org/10.1093/jnci/dju048

[40] SAVAGE, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

[41] SISMONDO, S. (2004). *An Introduction to Science and Technology Studies*. Blackwell, Malden, MA.

[42] STEYERBERG, E. W. and VERGOUWE, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35** 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

[43] STIGLER, S. M. (1982). Thomas Bayes's Bayesian inference. *J. Roy. Statist. Soc. Ser. A* **145** 250–258. MR0669120 https://doi.org/10.2307/2981538

[44] TRIPPA, L., WALDRON, L., HUTTENHOWER, C. and PARMIGIANI, G. (2015). Bayesian nonparametric cross-study validation of prediction methods. *Ann. Appl. Stat.* **9** 402–428.

[45] VENTZ, S., MAZUMDER, R. and TRIPPA, L. (2022). Integration of survival data from multiple studies. *Biometrics* **78** 1365–1376.

[46] VIJAYAKUMAR, R. and CHEUNG, M. W. L. Assessing replicability of machine learning results: An introduction to methods on predictive accuracy in social sciences. *Soc. Sci. Comput. Rev.* **39** 768–801.

[47] VIJAYAKUMAR, R. and CHEUNG, M. W. L. (2018). Replicability of machine learning models in the social sciences: A case study in variable selection. *Z. Psychol.* **226** 259–273.

[48] WALDRON, L., HAIBE-KAINS, B., CULHANE, A. C., RIESTER, M., DING, J., WANG, X. V., AHMADIFAR, M., TYEKUCHEVA, S., BERNAU, C. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J. Natl. Cancer Inst.* **106** dju049. https://doi.org/10.1093/jnci/dju049

[49] WANG, J., LAN, C., LIU, C., OUYANG, Y. and QIN, T. (2021). Generalizing to unseen domains: A survey on domain generalization. Available at arXiv:2103.03097.

[50] WONG, A., JIE, C., LYONS, P. G., DUTTA, S., MAJOR, V. J., ÖTLEŞ, E. and SINGH, K. (2021). Quantification of sepsis model alerts in 24 US hospitals before and during the

Covid-19 pandemic. *JAMA Netw. Open* **4** e2135286–e2135286. https://doi.org/10.1001/jamanetworkopen.2021.35286

[51] WU, E., WU, K., DANESHJOU, R., OUYANG, D., HO, D. E. and ZOU, J. (2021). How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27** 582–584. https://doi.org/10.1038/s41591-021-01312-x

[52] YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

[53] ZEMEL, R., SWERSKY, K. and PITASSI, T. (2013). Learning fair representations. In *Proceedings of the* 30*th International Conference on Machine Learning*.

[54] ZHANG, Y., PATIL PRASAD, J., EVAN, W. and PARMIGIANI, G. (2021). Robustifying genomic classifiers to batch effects via ensemble learning. *Bioinformatics* **37** 1521–1527. https://doi.org/10.1093/bioinformatics/btaa986

[55] ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H. and HE, Q. (2020). A comprehensive survey on transfer learning. Available at arXiv:02685.

[56] INSTITUTE OF MEDICINE (2012). *Evolution of Translational Omics*. The National Academies Press, Washington, D.C.