# Replicability in Science:
# 3: Replicability and Decision Theory

giovanni_parmigiani@dfci.harvard.edu

Padova, September 19, 2023

# nas report

An **ad hoc committee of the National Academies** of Sciences, Engineering, and Medicine explored the issues of reproducibility and replication in scientific and engineering research, focusing on defining reproducibility and replicability, and examining the extent of non-reproducibility and non-replicability.

# nas replicability

NAS: We define ... replicability to mean obtaining **consistent** results across **studies** aimed at **answering** the same scientific question, each of which has obtained its own **data.**

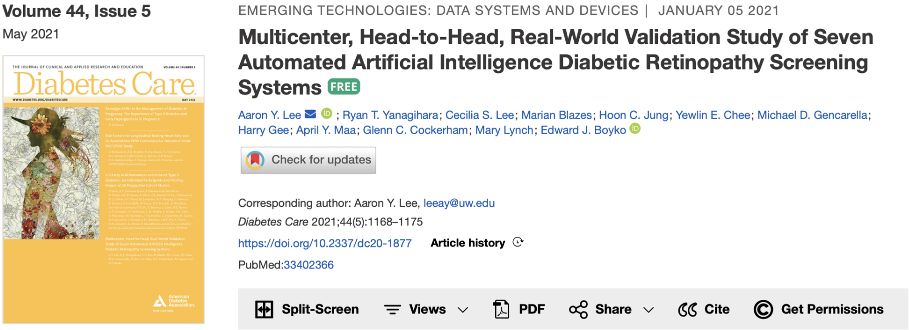
# nas multiple criteria

NAS: One challenge in assessing the extent of non-replicability across science is that different types of scientific studies lead to different or multiple criteria for determining a successful replication.

# nas no standard

NAS: ... there is no standard across science for assessing replication between two results

# retinopathy example, reference



**Volume 44, Issue 5**

May 2021

THE JOURNAL OF CLINICAL AND APPLIED RESEARCH AND EDUCATION
Diabetes Care
WWW.DIABETES.ORG/DIABETESCARE
MAY 2021
American Diabetes Association.

EMERGING TECHNOLOGIES: DATA SYSTEMS AND DEVICES | JANUARY 05 2021

## Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems FREE

Aaron Y. Lee; Ryan T. Yanagihara; Cecilia S. Lee; Marian Blazes; Hoon C. Jung; Yewlin E. Chee; Michael D. Gencarella; Harry Gee; April Y. Maa; Glenn C. Cockerham; Mary Lynch; Edward J. Boyko

Corresponding author: Aaron Y. Lee, leeay@uw.edu

*Diabetes Care* 2021;44(5):1168–1175

https://doi.org/10.2337/dc20-1877 **Article history**

PubMed:33402366

# retinopathy example, context

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 carried out a large prospective **multi-center study** to perform a head-to-head comparison of **seven algorithms**, including one FDA-approved algorithm, evaluating retinal images.

# notation, predictions

A prediction, 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 $\phi : \mathcal{X} \to \mathcal{P}$.

E.G. 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 $\mathcal{P} = \mathcal{Y}$.

# replicability of prediction rules, nas revised

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

### Definition (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 compared to NAS 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.

# "suitable"

NAS report: replication is “the act of repeating an entire study, independently of the original investigator without the use of original data.” **(replication by design)**

Here: both replication by design and **observational replication**, defined in reference to a specific prediction task and a collection of relevant datasets

# study

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 $n_k$.

# retinopathy example, studies

311,604 retinal images from 23,724 veterans who presented for teleretinal DR screening at the Veterans Affairs (VA) Puget Sound Health Care System (HCS) or Atlanta VA HCS from 2006 to 2018.

$K = 2$

Assessing whether discrimination ability is comparable in Seattle and Atlanta was among the goals.

# agents

- Modeler, generates prediction or scoring rule $\phi$
- Agent, needs to solve a decision problem
- Assessor, identifies relevant directions of variations and gold standard studies $S_1, \ldots, S_K$, and quantifies replicability

Replicability: Assessor(s) agree that the modeler's tool,
in the context of the user's decision problem,
is providing similar average utility across studies.

1, 2 or 3?

# retinopathy example, agents

| | |
|---|---|
| Modelers: | Industry Providers of AI |
| Users | Physicians within the VA |
| Assessor | Lee et al. |

# the modeler/agent case

The Modeler/Agent holds:

prediction or scoring rule $\phi$

model $\pi$ on $\mathcal{X}$ and $\mathcal{Y}$

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

An optimal decision function $\delta^*$ satisfies

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

# assessor

The prediction rule $\phi$ is replicable if its optimal application to the same decision problem in different data sets leads to approximately the same average utility to the decision maker. Formally:

### Definition (Absolute $\epsilon$-replicability)

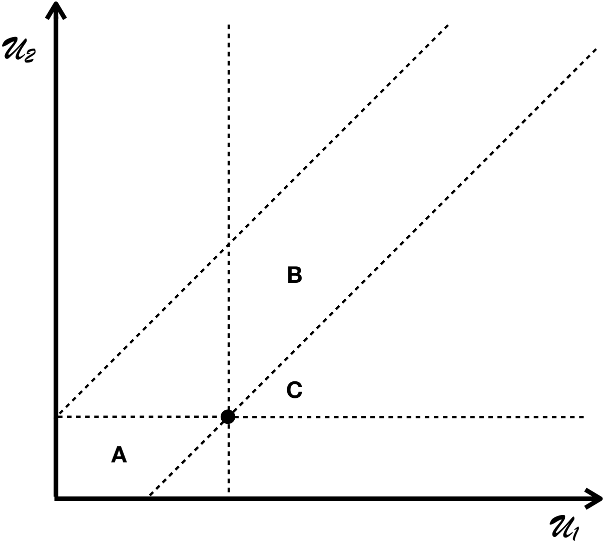
$\phi$ is $\epsilon$-replicable in absolute utility over $S_1, \ldots, S_K$ if

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

where, for study $k$, the agent's utility is, on average,

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

# dominance



# special case: binary *p*

A classification algorithm $\varphi : \mathcal{X} \to \mathcal{Y}$
e.g. $\varphi = \delta^*(\phi(x)) = max_{\delta \in \Delta} E_\pi \left\{ U(\delta(\phi(x)), y) \right\}.$

Utility function defined directly as

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

$\mathcal{U}_k$ defined as

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

and apply Definition 2.

e.g if $U(\varphi, y) = I_{\varphi = y}$ then $\mathcal{U}_k$ is the empirical correct classification proportion in study *k*
and $\epsilon$-replicability obtains when this proportion does not vary by more than $\epsilon$ in any two-study comparison.

# no-user replicability

**Definition (Distance $\epsilon$-replicability)**

$\phi$ is $\epsilon$-replicable in distance over $S_1, \ldots, S_K$ if

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

where $F_k$ is the empirical joint cumulative distribution of the points $(\phi(x_{ik})), x_{ik}, y_{ik})$ $i = 1, \ldots n_k$

$D$ is a metric, or more generally a useful summary

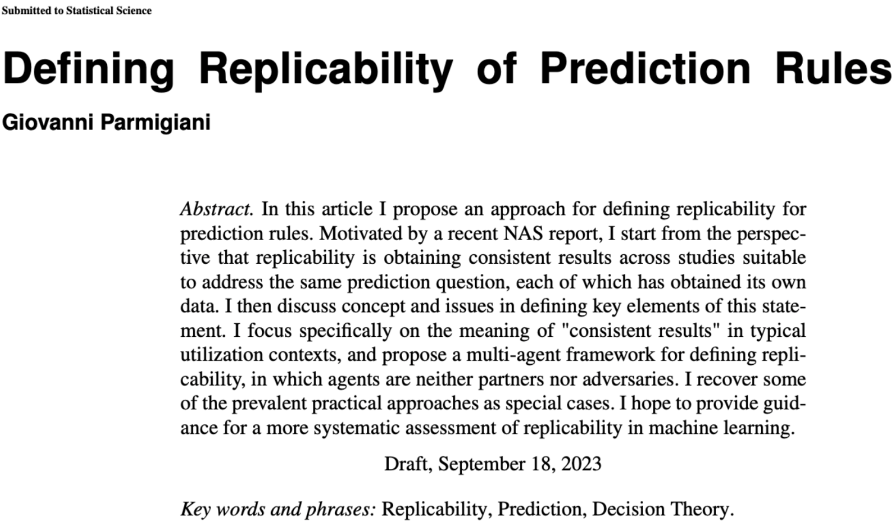
# statistical decisions

Work in progress. Two challenges.

1. Decisions are not applied to individual units but to studies.

2. The utility of a decision is not ultimately observable.

# reference



Submitted to Statistical Science

## 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 NAS report, 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 partners 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.

Draft, September 18, 2023

*Key words and phrases:* Replicability, Prediction, Decision Theory.