PLOS ONE

RESEARCH ARTICLE

The InterModel Vigorish (IMV) as a flexible and portable approach for quantifying predictive accuracy with binary outcomes

Benjamin W. Domingue[1*], **Charles Rahal**[2*], **Jessica Faul**[3‡], **Jeremy Freese**[4‡], **Klint Kanopka**[5‡], **Alexandros Rigos**[6,7‡], **Ben Stenhaug**[1‡], **Ajay Shanker Tripathi**[8‡]

1 Graduate School of Education, Stanford University, Stanford, California, United States of America, **2** Demographic Science Unit and Nuffield College, University of Oxford, Oxford, United Kingdom, **3** Michigan Center on the Demography of Aging, University of Michigan, Ann Arbor, Michigan, United States of America, **4** Department of Sociology, Stanford University, Stanford, California, United States of America, **5** Steinhardt School of Culture, Education, and Human Development, New York University, New York, New York, United States of America, **6** Institute for Futures Studies, Stockholm, Sweden,, **7** Department of Economics, Lund University, Lund, Sweden, **8** Department of Electrical Engineering, Stanford University, Stanford, California, United States of America

‡ Alphabetized.
* bdomingue@stanford.edu (BWD); charles.rahal@demography.ox.ac.uk (CR)

OPEN ACCESS

Domingue BW, Rahal C, Faul J, Freese J, Rigos A, Rigos A. et al. (2025) The InterModel Vigorish (IMV) as a flexible and portable approach for quantifying predictive accuracy with binary outcomes. PLoS ONE 20(3): e0316491. https://doi.org/10.1371/journal.pone.0316491

Editor: Leopoldo Trieste, Sant'Anna School of Advanced Studies Institute of Management: Scuola Superiore Sant'Anna Istituto di Management, ITALY

Received: March 26, 2024

Accepted: December 11, 2024

Published: March 21, 2025

Data availability statement: The datasets generated and/or analyzed during the current study are available in the following repositories: HRS data can be accessed at RAND HRS (https://www.rand.org/well-being/social-and-behavioral-policy/centers/aging/dataprod.html), GSS data at NORC GSS (https://gss.norc.org/get-the-data), Titanic data at Kaggle (https://www.kaggle.com/c/titanic),

Abstract

Understanding the "fit" of models designed to predict binary outcomes has been a long-standing problem across the social sciences. We propose a flexible, portable, and intuitive metric for quantifying the change in accuracy between two predictive systems in the case of a binary outcome: the InterModel Vigorish (IMV). The IMV is based on an analogy to weighted coins, well-characterized physical systems with tractable probabilities. The IMV is always a statement about the change in fit relative to some baseline model—which can be as simple as the prevalence—whereas other metrics are stand-alone measures that need to be further manipulated to yield indices related to differences in fit across models. Moreover, the IMV is consistently interpretable independent of baseline prevalence. We contrast this metric with alternatives in numerous simulations. The IMV is more sensitive to estimation error than many alternatives and also shows distinctive sensitivity to prevalence. We consider its performance using examples spanning the social and natural sciences. The IMV allows for precise answers to questions about changes in model fit in a variety of settings in a manner that will be useful for furthering research and the understanding of social outcomes.

1 Introduction

There has been a recent increase in the use of methods focused on prediction and in the 'fit' of models [1], coinciding with calls for a closer integration of explanation and prediction more broadly [2]. An independent long-standing question involves understanding, evaluating and—perhaps most importantly—comparing the quality of predictions from models trained on binary outcomes. An array of (frequently related) techniques have been developed: the

PISA data at OECD (https://www.oecd.org/pisa/data/), Essays data at Harvard Dataverse (https://dataverse.harvard.edu/dataverse/SAT_and_Essays), Football data at Dryad (https://doi.org/10.5061/dryad.8931zcrrs), and FFC data at Princeton POP (https://pop.princeton.edu/). Raw prediction examples are available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/datasets). The COVID-19 data are available from the SAIL databank but restrictions apply to their availability. Data can be obtained from SAIL Databank (https://www.hdruk.ac.uk/organisations/sail-databank/).

Funding: This work was supported by the Jacobs Foundation (BD), the Leverhulme Centre for Demographic Science (CR), The Leverhulme Trust (Grant RC-2018-003; CR) and Nuffield College (CR), and from Handelsbankens forskningsstiftelser (P21-0244; AR). The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

ROC curve [3], the harmonic mean of precision and recall (the F_1 score [4]), other quantities related to the confusion matrix determined by a given decision rule [5], cross-entropy [6], and Information Criteria such as the Akaike and Bayesian Information Criteria [7]. This is in addition to various pseudo-R^2 estimates [8].

The provisioning of accessible, intuitive, and portable metrics is essential to realizing the potential of machine intelligence and other predictive approaches across social science domains [9,10]. To that end, existing approaches have critical shortcomings. First, some metrics do not generalize given that they depend on sample-specific quantities, thus necessitating attempts to both generate sample size-sensitive benchmarks [11] and other attempts to reduce sample size dependency in related contexts [12]. Second, there is a lack of guidance about how to compare predictive gains relative to the base rate (i.e., the problem of "prevalence" or "imbalance") of the outcome [13]. Third, most metrics are absolute statements about the fit of a given model. If interest is in a comparison between models, further manipulation of the metrics is frequently needed (and such manipulations may not be readily interpretable). Collectively, these limitations challenge our ability to make generalizable inferences about the quality of models used across various contexts both within and beyond the social sciences.

Statements about a single model have utility in many settings. For example, the evaluation of whether a black box diagnostic test is of sufficient accuracy to be used in a specific setting. In that case, something like the AUC can be interpreted alongside established benchmarks [14] (e.g., the clinical accuracy of COVID-19 tests [15]). However, such stand-alone approaches have limitations. Suppose there is an outcome y and two predictive systems f and g. A stand-alone metric (e.g., R^2 as defined in the Fragile Families Challenge [16]: $R^2 = 1 - \frac{\sum(y_i - p_i)^2}{\sum(y_i - \bar{y})^2}$) produces an index of fit based on each system's predictive accuracy, which we can denote as m_f and m_g. Predictions from f are better than those from g if $m_f > m_g$; indeed, much existing work stops there and just notes the direction of this inequality. But "how much better?" and "how does this relate to other applications?" are important and challenging questions that require answers if we are to maximize the impact of predictive social science and move towards a more coherent realization of external validity. Having a single numeric summary of the "difference" between m_f and m_g would allow us to better answer those questions, a topic of current interest within the social sciences [16,17]. Many existing metrics either do not allow for a summative comparison of m_f and m_g (e.g., AUC) or produce summaries that can be challenging to interpret given that the values depend on sample size (AIC) or may have an unclear dependence on $\widehat{y}$ (R^2).

This paper introduces a novel metric designed to overcome these challenges for use in predictive systems that generate predictions in the form of probabilities (c.f., class labels). It is based on translating the level of uncertainty for a given predictive system into a canonical physical system—a weighted coin—and then building inference around the well-characterized statistical properties of that physical system. This metric, the InterModel Vigorish (hereafter, IMV), generalizes across multiple predictive systems that may vary in outcome, predictors, and approaches to prediction (so long as the approach generates probabilities rather than classes). Note that this metric is discussed in the specific context of psychometric models for dichotomous item responses elsewhere [61]. In tying notions of profits from gambles to questions of prediction, this work ties into the deep traditions of early statisticians such as Pascal and Huygens [18] who used games of chance as a means to better understand probability.

2 The InterModel Vigorish

2.1 Introducing the IMV

We focus on the problem of constructing a generalizable—in the sense that values of the IMV are comparable across outcomes—metric for comparing the accuracy of two predictive systems for binary outcomes. These are the 'baseline' and 'enhanced' predictions. These names are chosen to increase intuition as one might anticipate the enhanced prediction containing valuable 'side' information not available to the baseline prediction (but the enhanced prediction need not, in fact, be an improvement to the baseline prediction). At first glance, requiring two systems may seem restrictive. However, given that one of the models can be a prediction based on prevalence alone (i.e., the outcome's mean), it is not (alternative approaches such as pseudo-R^2 may use prevalence in a similar capacity). This approach is a multi-step process (see the schematic in Fig 1) described in detail below.

2.2 Defining the IMV

2.2.1 Quantifying randomness Suppose for $i \in \{1, ..., n\}$ we have a vector $y \equiv (y_i)$ of observations of some binary outcome variable Y (so, $y_i \in \{0, 1\}$ for each observation i). A *predictive system* p consists of a probabilistic prediction $p_i \in (0, 1)$ for each observation $i = 1, \ldots, n$. The interpretation is that the system predicts that, for the i-th observation, $y_i = 1$ with probability p_i and $y_i = 0$ with probability $1 - p_i$. The likelihood assigned by the system to an observation i is $L_i \equiv p_i^{y_i}(1-p_i)^{1-y_i}$, while the log-likelihood of that observation is $\ell_i \equiv \log(L_i) = y_i \log p_i + (1 - y_i) \log(1 - p_i)$. To evaluate the system's ability to predict the outcome variable, consider the system's likelihood evaluated on the data set y:

$$L(p; y) \equiv \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i}. \quad (1)$$

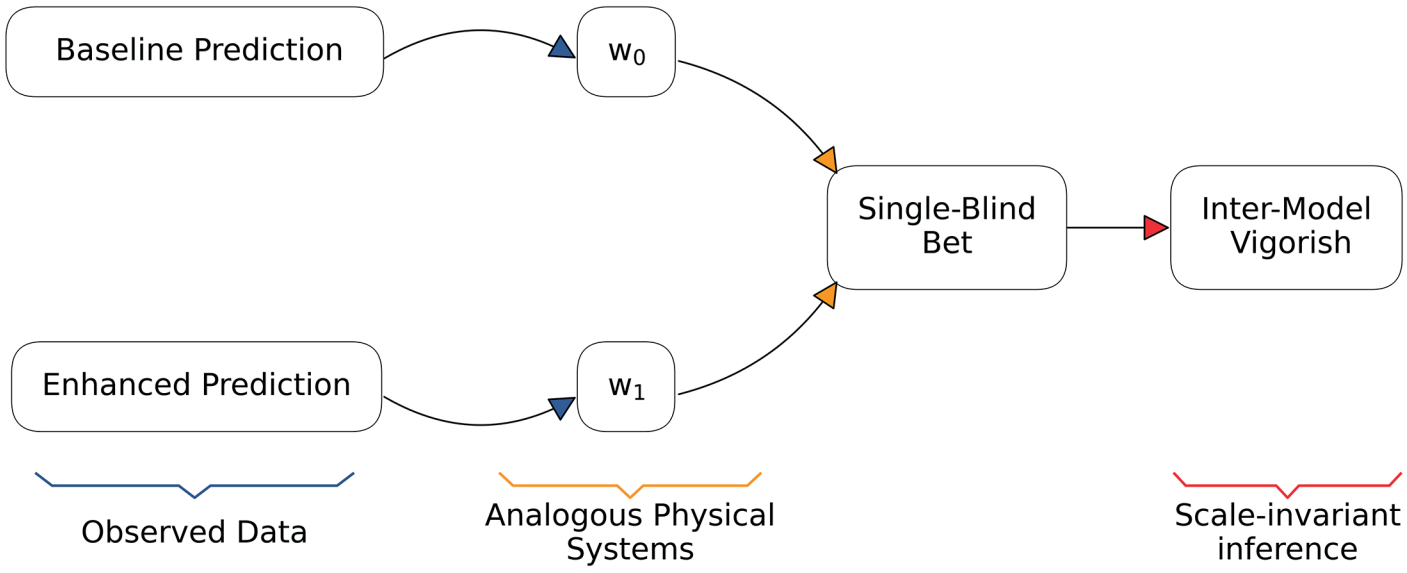

Fig 1. The IMV approach to quantifying prediction. Predictions are translated to an analogous physical system (weighted coins). The single-blind bet is then constructed based on payoff odds generated via w_0 yet one player knows the true probability of success is w_1. The IMV is constructed from the expected winnings associated with the side information contained in w_1. Results can be compared across outcomes given that the fair bet is based on w_0.

https://doi.org/10.1371/journal.pone.0316491.g001

To make comparisons between models of different sample size, we take the geometric mean of an observation's likelihood according to system p:

$$A(p;y) \equiv L(p;y)^{1/n}. \tag{2}$$

Our goal is to identify a probability (or weight) of a weighted coin such that the expected log likelihood of the toss of a coin with that weight is the same as the mean log likelihood of the system p evaluated on data y. This weight $w(p;y)$ is given by:

$$w(p;y) \equiv \left\{ w \in [1/2, 1] : w \log w + (1-w)\log(1-w) = \frac{1}{n}\sum_{i=1}^{n} \ell_i \right\}. \tag{3}$$

For our purposes, a coin with weight $1 - w(p;y)$ is equivalent to a coin with weight $w(p;y)$; we choose $w \geq 1/2$. The weight $w(p;y)$ describes the quality of predictions p in describing the observed y. Note also that the coin with weight w will be equivalent to the predictive system in terms of entropy [19]. A visualization of the curve linking $A(p;y)$ to $w(p;y)$ is shown in Fig 2 Panel A.

2.2.2 A betting analogy Consider a bet involving two parties, the house and a gambler. The parties wager over the outcome of a binary variable, which can be either positive or negative. The house bets \$1 on the positive outcome and the gambler bets \$1/O on the negative outcome. The party whose chosen outcome is realized wins the bet and takes the combined pot of \$(1+1/O). This bet is *fair* if the house has zero expected gains should the gambler take up the wager. So, if the probability of a positive outcome is w, the bet is fair if

$$w \times \frac{1}{O} - (1-w) \times 1 = 0. \tag{4}$$

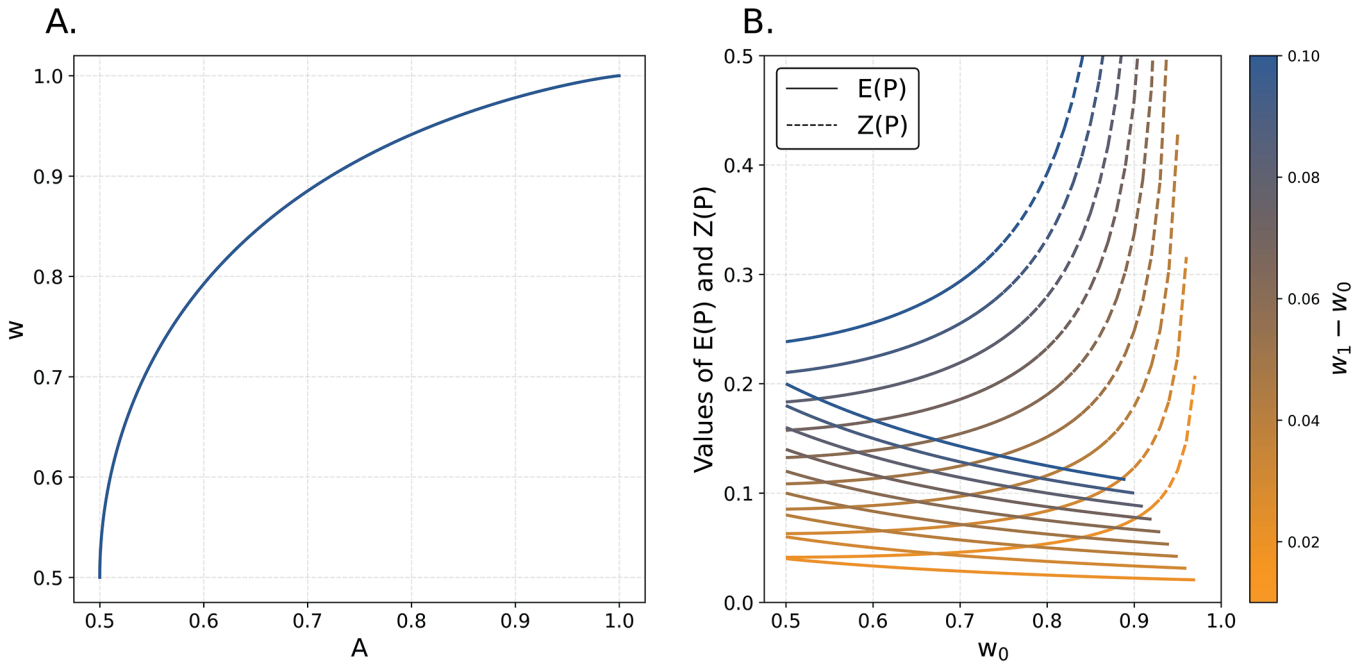

Fig 2. Properties of the IMV. Panel A shows a mapping between A and w. Panel B shows $\mathbb{E}(P)$ as a function of w_0 and $w_1 - w_0$.

https://doi.org/10.1371/journal.pone.0316491.g002

Thus, the bet will be fair if O is equal to the odds ratio, $O(w) \equiv \frac{w}{1-w}$.

Our method aims to compare the predictive accuracy between two different prediction models, the baseline and enhanced predictions. Consider the following scenario. Suppose that the house offers odds of $O(w_0)$ where w_0 is based on the baseline prediction. However, unbeknownst to the gambler, the probability of the positive outcome is actually w_1 where w_1 is the weight of the coin associated with the enhanced prediction. In this scenario, the bet is not fair if $w_0 \neq w_1$. Should the gambler take up the bet, the house expects to win

$$w_1 \times \frac{1}{O(w_0)} - (1 - w_1) \times 1 = \frac{w_1 - w_0}{w_0}. \tag{5}$$

This amount is known as the house's edge or the 'vigorish' of the bet. The IMV (or $\frac{w_1 - w_0}{w_0}$) takes values in [-1,1] as $w_0, w_1 \in [0.5, 1]$.

To define the IMV for predictions p^0 and p^1 for data y, we first identify $w_0 = w(p^0; y)$ and $w_1 = w(p^1; y)$ via Eq 3. The IMV—denoted as $\omega(p^0, p^1; y)$—is defined as

$$\omega(p^0, p^1; y) \equiv \frac{w_1 - w_0}{w_0}. \tag{6}$$

It is the house's vigorish supposing that the probability of the positive outcome was $w(p^1, y)$ yet the house was offering odds based on $w(p^0, y)$.

2.3 Computing the IMV

2.3.1 A toy example in R We can compute the IMV in a simple example so as to develop intuition for this quantity. Code to reproduce this example in `R` is shown below with equivalent examples in `Python` and `MATLAB` available in the Supporting Information (S1-II.1). Suppose we produce outcomes via a combination of 20 tosses of a fair coin (probability of 0.5 for heads) and 20 tosses of a heavily weighted coin (probability of 0.95 for heads). Researchers would be blind to this information about the weights of the coin in general. The fair coin produces 14 heads and the weighted coin produces 19 heads; thus our observed data is 33 heads and 7 tails. We do not attempt to quantify the level of randomness in this data; randomness, for our purposes, is only defined in the context of a specific model for the data generating process. As an illustration, suppose one has a set of heads and tails. If the model is a fair coin, this is pure randomness. If the model is two coins—one that always produces heads and one that always produces tails—there is no randomness. Speaking of the randomness of these outcomes necessitates reference to the data-generating process.

Suppose, arbitrarily, that our baseline prediction is that all outcomes are produced via a coin with probability $p_i^0 = 0.55$ of being heads. We first compute $A(p^0; y) = 0.53$ and then translate into an analogous coin of weight $w_0 = 0.67$ (see also Panel A in Fig 2). So as to forestall confusion, note the distinction between the implied coins, w_0 and w_1, and the coins with weights 0.5 and 0.95 used to generate the data; we make predictions about the latter and use the former to compute the IMV. Now, suppose that our enhanced prediction is $p_i^1 = 0.5$ for the first 20 observations (those produced by the fair coin) and $p_i^1 = 0.9$ for the second 20 observations (those produced by the weighted coin). We compute $A(p^1; y) = 0.63$ and translate that into $w_1 = 0.83$. Note that the coin suggested by w_1 argues for a far less random system than the coin suggested by w_0, this is intuitive given the fact that p^1 is a far-superior approximation of the data-generating process. The improvement in prediction is now $\omega = 0.24$, the IMV. The additional predictive information offered by the enhanced prediction

translates to an expectation of winning nearly a quarter (i.e. 24 cents) for every dollar wagered.

A Toy Example in R

```
set.seed(8675309)

# Combine tosses from fair and weighted coins:
x1<-rbinom(20,1,.5)
x2<-rbinom(20,1,.95)
x<-c(x1,x2)

# Define a function to compute the log-likelihood:
ll<-function(x,p) {
    z<-log(p)*x+log(1-p)*(1-x)
    z<-sum(z)/length(z)
    exp(z)
}

# Create a baseline estimate:
p=.55
a0<-ll(x=x,p=.55)
f<-function(p,a) abs(p*log(p)+(1-p)*log(1-p)-log(a))
p0<-nlminb(.5, f, lower=0.5, upper=.999, a=a0)$par

# Create an improved estimate:
p<-c(rep(.5, 20),rep(.9, 20))
a1<-ll(x=x,p=p)
p1<-nlminb(.5, f, lower=0.001, upper=.999, a=a1)$par

# Calculate the single-blind bet:
imv<-(p1-p0)/p0
```

2.3.2 Computation of the IMV in practice. The IMV can be computed directly with information about the likelihoods for the baseline and enhanced model; computation can be done via a simple application and functionality made available as part of the replication materials hosted on an organizational GitHub account ('InterModelVigorish'). In practice, the quantity can be computed using in-sample or out-of-sample values for the likelihood (or, equivalently, based on training or test sets of estimated probabilities and observed responses). Training-based IMV estimates will be biased in favor of more complex models (see S1-III.4). Here, focus is on the prediction of test data given this problem and the generic issues associated with overfitting [20]. The key ideas related to out-of-sample prediction and cross validation are discussed at length elsewhere [20,21]. A note regarding computational costs: for a given set of predictions, the IMV can be computed very rapidly even for large datasets (i.e., the optimization in Eq 3 is fairly simple). However, producing out-of-sample predictions

may be computationally complex; when used in our preferred out-of-sample framework, computational costs will thus depend on the complexity of implementing the predictive systems in question.

For values shown here, computation is typically done via a standard cross-validation procedure. We calculate the mean IMV based on 10-fold cross-validation: observations are randomly assigned into one of ten mutually exclusive folds with uniform probability (in a few cases, we use different numbers of folds or folding techniques where indicated in order to highlight the versatility of our approach). For a given fold, we first estimate model parameters using data assigned to other folds and then treat the given fold as an out-of-sample test dataset by computing ω using observations from this fold and predictions from the model constructed in the other folds. Uncertainty in the IMV can be assessed using standard deviations across the folds (see S1-III.3 for discussion regarding this as a way of describing uncertainty).

S1-II.2 describes an applied use case of the IMV; the Supporting Information accompanying this paper contains this example in `R`, but our online supporting repository of code also calculates this in `Python` and `MATLAB`. This example of how to use the IMV in practice takes the standard 'Titanic' dataset and predicts survival based on two simple logistic models: one that includes only a constant and one with sex and passenger class. In this context, these additional predictors are greatly predictive of survival (mean IMV is 0.352 with a standard deviation of 0.143) relative to prediction based on prevalence alone.

2.4 Properties of the IMV

Note several properties of the IMV. First, given that A is the geometric average likelihood for an observation, the IMV captures the expected winnings for the prediction of a single outcome generated by the enhanced coin. Second, note that the IMV is not symmetric. If interest is in the IMV associated with the side information in the baseline model relative to the enhanced model, the relevant quantity would be $\omega = \frac{w_0 - w_1}{w_1}$. Third, observe that the IMV decreases as w_0 increases for a fixed $w_1 - w_0$; this behavior is shown in Panel B of Fig 2.

Is this desirable? Consider the house's profit

$$P(O) = \begin{cases} \frac{1}{O} = \frac{1-w_0}{w_0} & \text{if the outcome is positive (with probability } w_1) \\ -1 & \text{if the outcome is negative (with probability } 1 - w_1) \end{cases} \quad (7)$$

for a gamble based on a coin with $w_1 = w(p^1, y)$. P is a Bernoulli random variable whose support depends upon O, but with parameter w_1. If $O = \frac{w_0}{1-w_0}$, then $\mathbb{E}(P) = \omega(p^0, p^1; y) = \frac{w_1 - w_0}{w_0}$ (see Eq 5). Consider a potential alternative, $Z(P) = \frac{\mathbb{E}(P)}{\sqrt{\mathbb{V}(P)}}$ (where $\mathbb{V}$ is the variance operator). Given that $\mathbb{V}(P)$ decreases as w_0 nears unity, this would have the effect of our preferring gains, in terms of $w_1 - w_0$, when w_0 is near one (i.e. where the dashed curves in Panel B of Fig 2 are upward sloping). Doing so merely informs us that we are comparing changes in $w_1 - w_0$ to very small levels of uncertainty. The goal in betting is to make money—to maximize ω—not to make smaller amounts of money in games with relatively little randomness; this logic applies to prediction of stochastic outcomes of scientific interest as well. Our metric produces values consistent with this logic. However, other approaches may require the prioritization of relatively small chanegs to $w_1 - w_0$ when the underlying uncertainty is relatively low (e.g., something more akin to $Z(P)$); our approach would be inappropriate in such cases.

S1-I contains additional arguments about the IMV. The IMV is a proper scoring rule [22] and is related to the "Kelly criterion" [23] which is an optimal approach to betting in certain

scenarios that are related to the two bets considered in the construction of the IMV. Building on this connection to betting, we introduce comparisons to vigorishes from a number of common parlor games of chance (i.e., roulette, blackjack, and baccarat).

Finally, note the emphasis on predictions of unobserved ('test'/'out-of-sample') data to compute the IMV. This is practically consequential given that, when used with training data, the IMV will always be biased in favor of more complex models. Our utilization of test data removes this bias—a fact which can be observed with a simple logistic regression example (S1-III.4)—but which is also consistent with the broader conceptual turn towards a focus on prediction [1]. In empirical cases, variation in the IMV across folds can be used as an index for sampling-related uncertainty in the IMV (see illustration in S1-III.3).

3 Simulation studies

We conduct several simulation studies pertaining to the IMV. These studies build on the simulations in earlier work which focused on a specific application of the IMV to the modeling of item responses [61]. We first describe specialized use cases that showcase the flexibility of the IMV given that it only relies upon predictions and outcomes and then describe a variety of simulation studies meant to contrast the IMV with common alternatives. We focus on a set of metrics—the AIC and BIC, pseudo-R^2 approaches, the AUC, and the F_1 score—that are meant to be representative of the large set of options available for use in this context. These metrics and the IMV behave similarly in many settings. In terms of the yes/no question "Does model A fit better than model B?", we anticipate roughly similar answers from all the metrics; interest here is on scenarios that depict qualitative differences in the behavior of the magnitudes of these metrics.

3.1 Two specialized versions: The oracle and the overfit

The fact that the IMV is quite flexible and requires only fairly generic inputs—outcomes and two predictions—can be used to introduce two specialized versions of the IMV: the Oracle and Overfit. These are meant to help us better understand parameter recovery and overfitting in simulation studies. These metrics quantify the value of truth (when it is known) relative to estimates based on test data (Oracle) and training data (Overfit). We illustrate their use via a simple univariate logistic regression problem. We simulate data y_i from a Bernoulli distribution based on

$$p_i \equiv \Pr(y_i = 1) = \sigma(\beta_1 x_i) \tag{8}$$

where $x_i \sim \text{Normal}(0, 1)$ for $i \in \{1, ..., N\}$ and σ is the logistic sigmoid, $\sigma(x) = (1 + \exp(-x))^{-1}$. We vary N by sampling $n \sim \text{Unif}(\log_{10} 50, \log_{10} 10000)$ and setting $N = 10^n$ and let $\beta_1 \in \{0.01, 0.1, 0.5\}$. We estimate the logistic regression model using (x_i, y_i) from which we can create fitted values $\widehat{p_i}$. We then generate a second set of outcomes, $y^\star$, for the same values of β_1 and x; that is, we use the same p_i values to generate a second set of outcomes that will be used as out-of-sample test data.

Note that the models are trained on y while the $y^\star$ outcomes are hypothetical test data (given that we use the same x to generate $y^\star$ and y, the $\widehat{p_i}$ are germane for each). For each simulated set of data, we then consider (using the order of arguments as in Eq 6):

- ω_0: $\text{IMV}(\bar{y}, \widehat{p_i}; y^\star)$ where $\bar{y} = \frac{1}{N}\sum_i y_i$,
- Overfit: $\text{IMV}(\widehat{p_i}, p_i; y)$,
- Oracle: $\text{IMV}(\widehat{p_i}, p_i; y^\star)$.

The oracle and overfit values are only available due to the fact that the data generating mechanism is known (i.e., we observe the true p_i); while limited in scope they can be valuable in benchmarking the performance of estimates $\widehat{p_i}$. The oracle and overfit values are computed based on the same probabilities, but with different data. Crucially, overfitting is indicated by negative values of the IMV. Negative overfit values imply that the estimates $\widehat{p_i}$ are more valuable (because they are overfit to observed data) than the true p_i. In contrast, the ω_0 value shows that the IMV can function as a stand–alone metric since it relies only on $\widehat{p_i}$ and the mean $\bar{y}$ within the training data (i.e., the prevalence).

Results for 5,000 choices of N for each value of β_1 are shown in Fig 3.1. Several key points emerge. When $\beta_1 = 0.01$, there is (unsurprisingly) little value in using estimates to predict $y^\star$ as opposed to just the mean (ω_0 is near zero). However, there is a substantial cost paid due to overfitting for small N. For $\beta_1 = 0.1$, note two key facts. First, the oracle IMV declines as sample size increases due to declines in $|p_i - \widehat{p_i}|$ (i.e., estimates improve for large N). Second, the value associated with ω_0 increases as a function of sample size for the same reason. For $\beta_1 = 0.5$, a clearer influence of sample size on all three values is apparent. Values of ω_0 increase as a function of sample size while both the oracle and overfit IMV values decline towards zero. This first simulation demonstrates that the IMV behaves sensibly in this simple context, and its flexibility allows us to utilize the Oracle and Overfit variants which may allow for future studies about the performance of various estimators.

3.2 The IMV versus alternatives

3.2.1 Comparisons to R^2, the AUC, and the F_1 score We begin to contrast the IMV with alternatives by focusing on key measures of prediction accuracy: the R^2, the F_1 score, and the AUC. The below study is designed to emphasize the fundamental difference between the IMV and these alternatives (S1-III.1 contains a straightforward scenario wherein behavior across

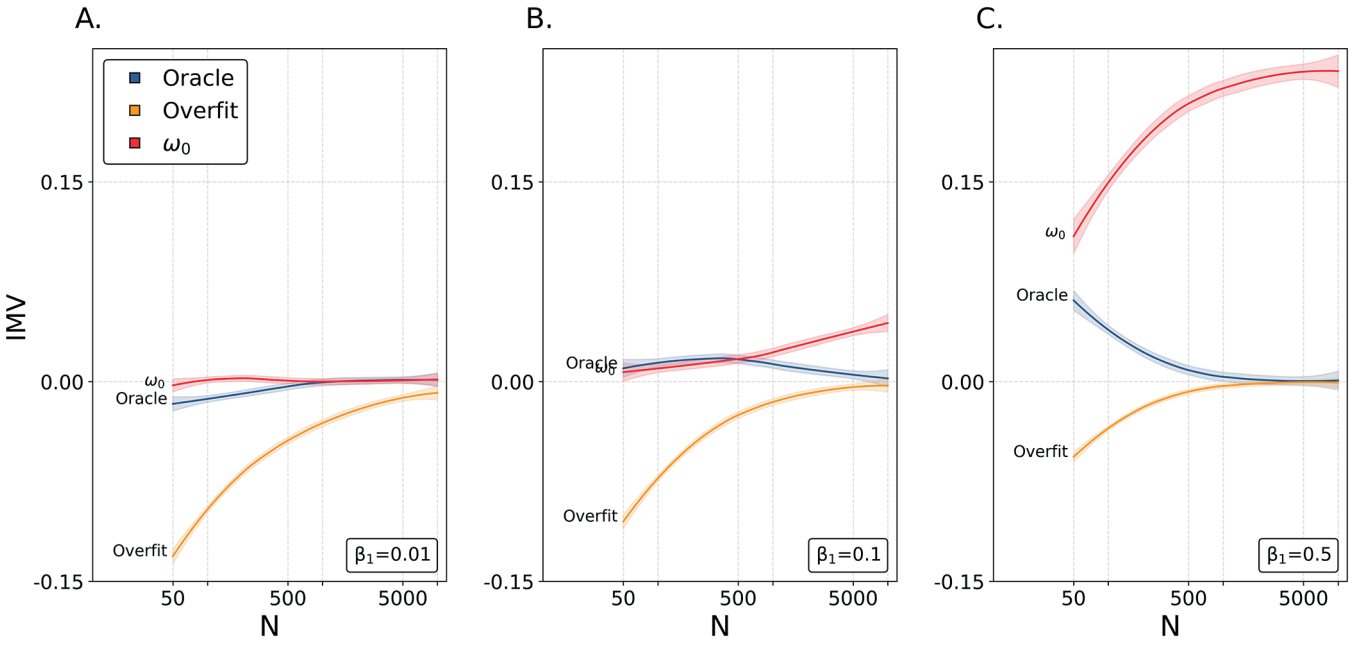

Fig 3. The Oracle, Overfit, and Omega. A comparison of ω_0, oracle, and overfit values of IMV in the logistic regression context as a function of β_1 and N, fit to 99.9% confidence intervals.

https://doi.org/10.1371/journal.pone.0316491.g003

the indices is quite similar). We simulate data based on parameters (N, a, b, Ψ, ψ). We generate N values from Beta (a, b) where we use the following density for the Beta distribution:

$$f(x) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))x^{(a-1)}(1-x)^{(b-1)}. \quad (9)$$

These are then linearly rescaled so as to lie on the interval $(\Psi, 1-\Psi)$; call these rescaled values p. Based on these true probabilities we generate two sets of outcomes (where the i–th outcome is generated via Bernoulli(p_i)); the first set is used to derive a training mean $\bar{y}$, while the second set is treated as the test data. The ψ quantity is effectively error; we use ψ to generate an 'estimate' of p, denoted p_1, where $p_1 = p \pm \psi$ where we randomly choose to add or subtract. The parameter ψ serves as a proxy for the quality of estimates as $p_1 \to p$ as $\psi \to 0$. When $\psi = 0.2$, the difference $|p_1 - p| = 0.2$ is large relative to the difference $|p - \bar{y}|$ (i.e., for $a = b = 1$ we have $\mathbb{E}(y) = 0.5$ and thus if $\Psi = 0.2$ we have $|p - \bar{y}| \leq 0.3$) meaning that p_1 are low-quality estimates. We thus anticipate strong sensitivity in our metrics to the value of ψ.

We set $N = 1000, \Psi = 0.2, a = 1, b = 1$ and consider 2,000 samples of $\psi \sim \text{Unif}(0, \Psi)$. Results are shown in Fig 4 wherein we consider the metrics as a function of ψ (focusing on smoothed curves estimated via LOWESS regression of metrics on ψ). For each metric, we first compute the value when true p is known (e.g., for R^2 we consider $1 - \frac{\sum(y_i - p_i)^2}{\sum(y_i - \bar{y})^2}$; we choose this approach given that it is used in [16]) as compared to prediction based on $\bar{y}$; these values are shown in the solid lines. Note that they are not dependent on ψ. We then compute each metric for p_1 compared to $\bar{y}$; these are shown as dashed lines lines. As expected, we observe declines in the dashed lines as ψ increases. When $\psi = 0.2 = \Psi$, the p_1 predictions are in fact worse than $\bar{y}$ in predicting observations. We can illustrate this via a consideration of squared errors. We have $(p_1 - p)^2 = \psi^2 = 0.04$. Alternatively, we can compute $\mathbb{E}((\bar{y} - p)^2) = \mathbb{V}(p)$. We can use the fact that p is a re-scaled Beta random variable to compute that

$$\mathbb{V}(p) = \frac{1}{12}(1 - 2\Psi)^2 = 0.03. \quad (10)$$

Further, observe that the prediction errors should be equal in expectation when $\psi^2 = \frac{(1-2\Psi)^2}{12}$ which has a solution for approximately ψ=0.17 which is where the dashed line crosses the origin for both R^2 and the IMV. This fact is apparent for both the R^2 and IMV which are below 0.

The AUC and F_1 metrics are both fairly insensitive to ψ; the dashed lines decline from the solid lines, but the declines are relatively slight. From our perspective, these relatively slight changes—especially as p_1 estimates become extremely low quality—make these metrics poor tools for a generalized understanding of the predictions from different models. While the R^2 and IMV show sensitivity to *psi*, we turn now to a simulation study meant to illustrate their differences.

3.2.2 A key distinction between the IMV and R^2. In Fig 4 both the IMV and R^2 show relatively strong sensitivity to increased estimation error. However, they will not always be similar. We consider an analysis that contrasts the sensitivity of R^2 to the strength of a predictor in a scenario wherein the prevalence of the outcome is being adjusted (via a logistic regression model) so as to make the IMV constant. We define $p_i = \Pr(y_i = 1) = \sigma(\beta_0 + \beta_1 x_i)$ for $x \sim \text{Normal}(0, 1)$. We then choose β_0 values over a grid between 0 and 0.5 and we set $N = 100,000$ to make clear that our results aren't driven by sample size. We specify a value for $\omega \in \{0.01, 0.1\}$ that determines the desired IMV for prediction relative to the mean (i.e., ω_0

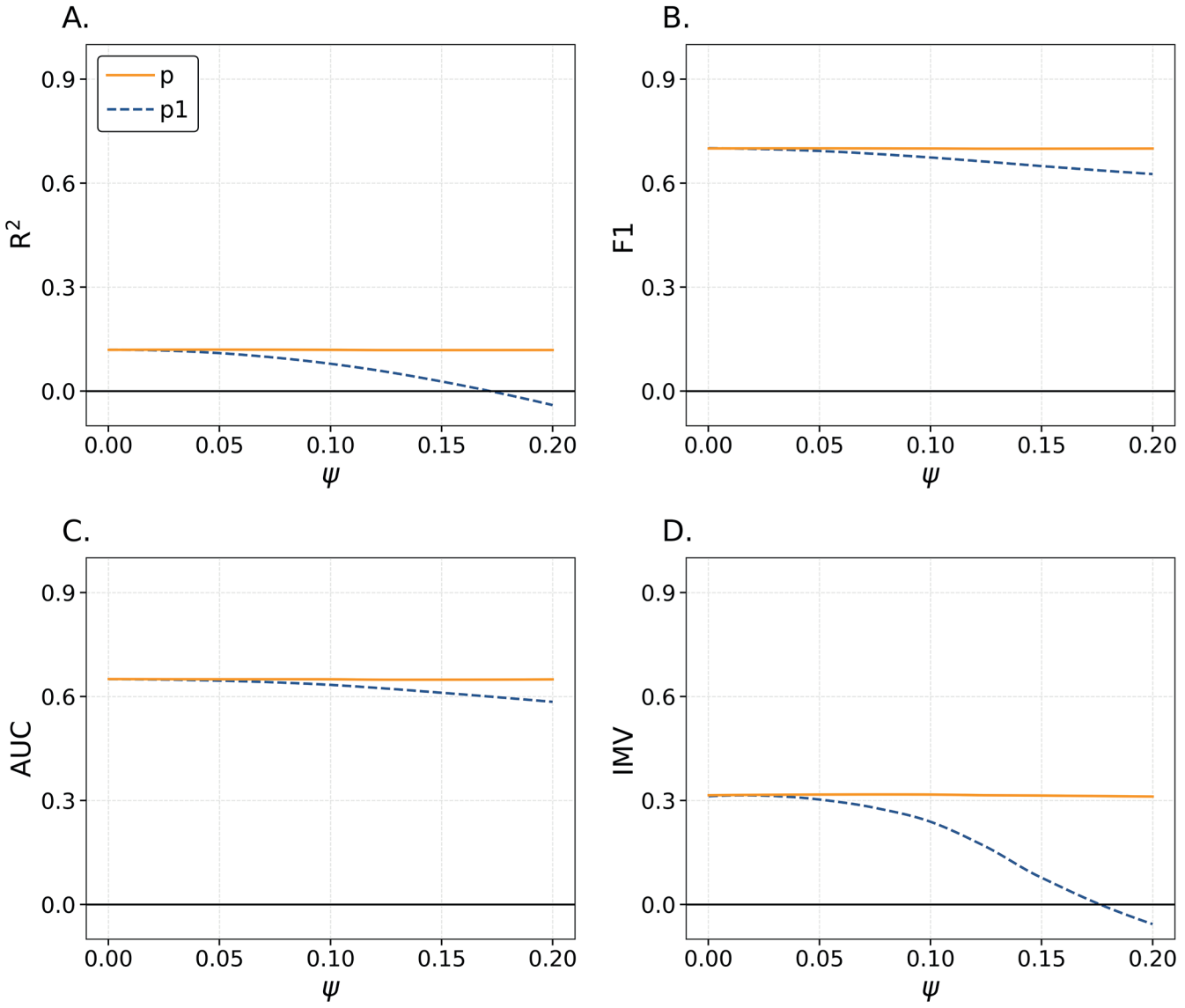

Fig 4. Metrics and ψ. Smoothed values of various metrics as a function of ψ.

https://doi.org/10.1371/journal.pone.0316491.g004

as in the discussion of the oracle and the overfit indices). Our goal is to then find the necessary β_1 to generate the specified ω value; we are effectively querying the trade-off in β_1 necessary to maintain a constant IMV when we increase the prevalence via β_0. As β_0 increases, a larger β_1 value is required to hold ω constant. We identify the appropriate β_1 (using an optimizer) and then use these values to simulate data (x, y) from which we construct $R^2 = 1 - \sum(y_i - p_i)^2 / \sum(y_i - \sigma(\beta_0))^2$ (we use the true p_i here so as to make clear that this is not a point about estimation error).

Fig 5 examines the (β_0, β_1) curves while also providing information about the R^2 values for extreme choices of β_0 (left panel). Each point on a given line has the same IMV. For a given value of β_0, the solid line is above the dashed line. This is due to the fact that from the

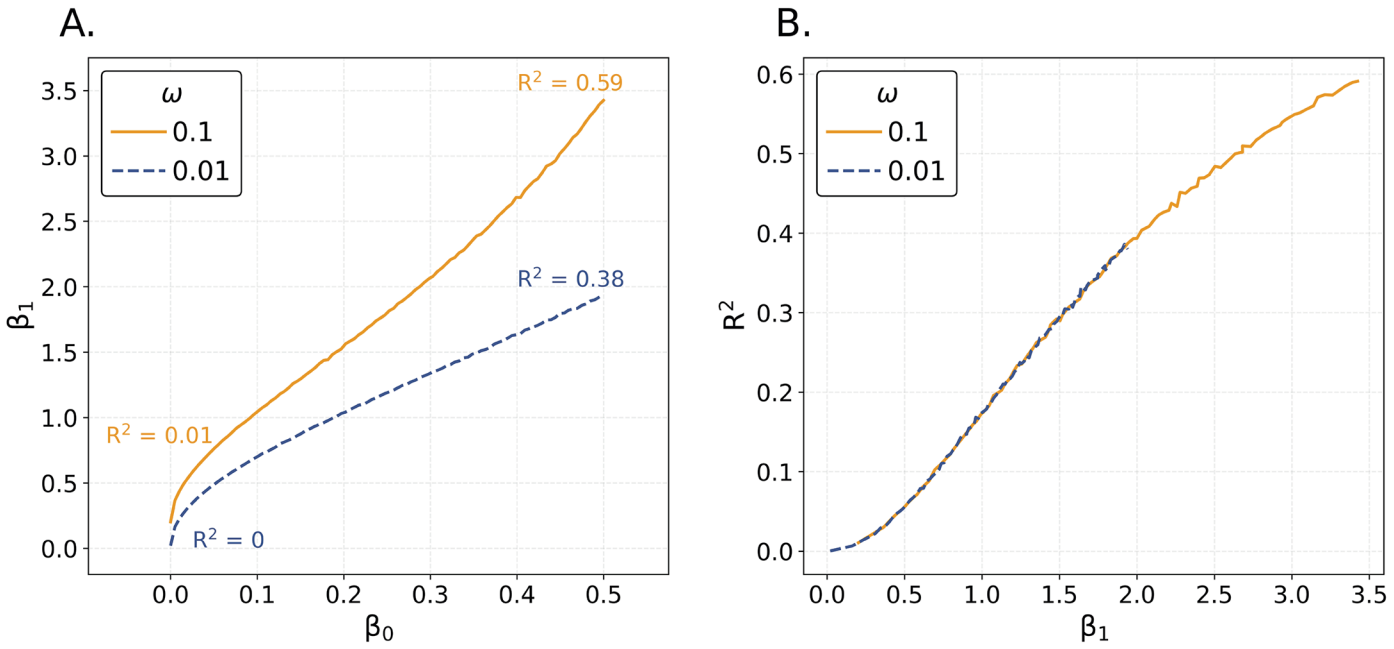

Fig 5. Comparisons of prevalence. Comparisons of prevalence versus β_0 and β_1 for constant values of ω.

https://doi.org/10.1371/journal.pone.0316491.g005

perspective of the IMV, an increase in β_0 decreases randomness in the outcomes and therefore larger values of β_1 are necessary to generate the same predictive value. We then compare β_1 and the resulting R^2 (right panel). As one would expect, an increase in β_1 leads to an increase in R^2. However, the two lines representing the different choices of ω are overlapping. The R^2 value is highly sensitive to β_1 while the IMV is sensitive to both β_1 and β_0. This divergence between the meaning of ω and R^2 is crucial in explicating the novel information provided by the IMV (similar evidence regarding the relationship between these parameters and prevalence can also be found in S1-III.2).

3.2.3 The IMV versus information criteria We now consider the AIC and BIC [7] vis-a-vis the IMV in a simulation study (while there are similarities between the AIC and BIC, note that there are also important differences between them [24]). These quantities are based on adjustments to the likelihood given the number of estimated parameters. These adjustments are meant to minimize overfitting; i.e., the AIC is asymptotically equivalent to leave-one-out cross validation [25] (although conventional usage of the AIC may be sub-optimal if the underlying statistical model is mis-specified [26]). Given modern computational power, such adjustments can perhaps be replaced in favor of the testing of different models in novel data not used for training (i.e., model estimation). While adjustments are available to remove dependency on sample size [27], information criteria values are also sample-size dependent; this dependence makes generalizations challenging, a point made in initial simulation studies with the IMV [61].

To illustrate the substantive differences between the AIC/BIC and the IMV, we consider the simulation study summarized in Fig 6. For a choice of (N, β), we generate N values (y) based on $\Pr(y = 1|x) = \sigma(\beta x)$. We then fit the correct model and contrast it with three alternatives: (1) an overfit model with quadratic terms in x, (2) a model where we observe

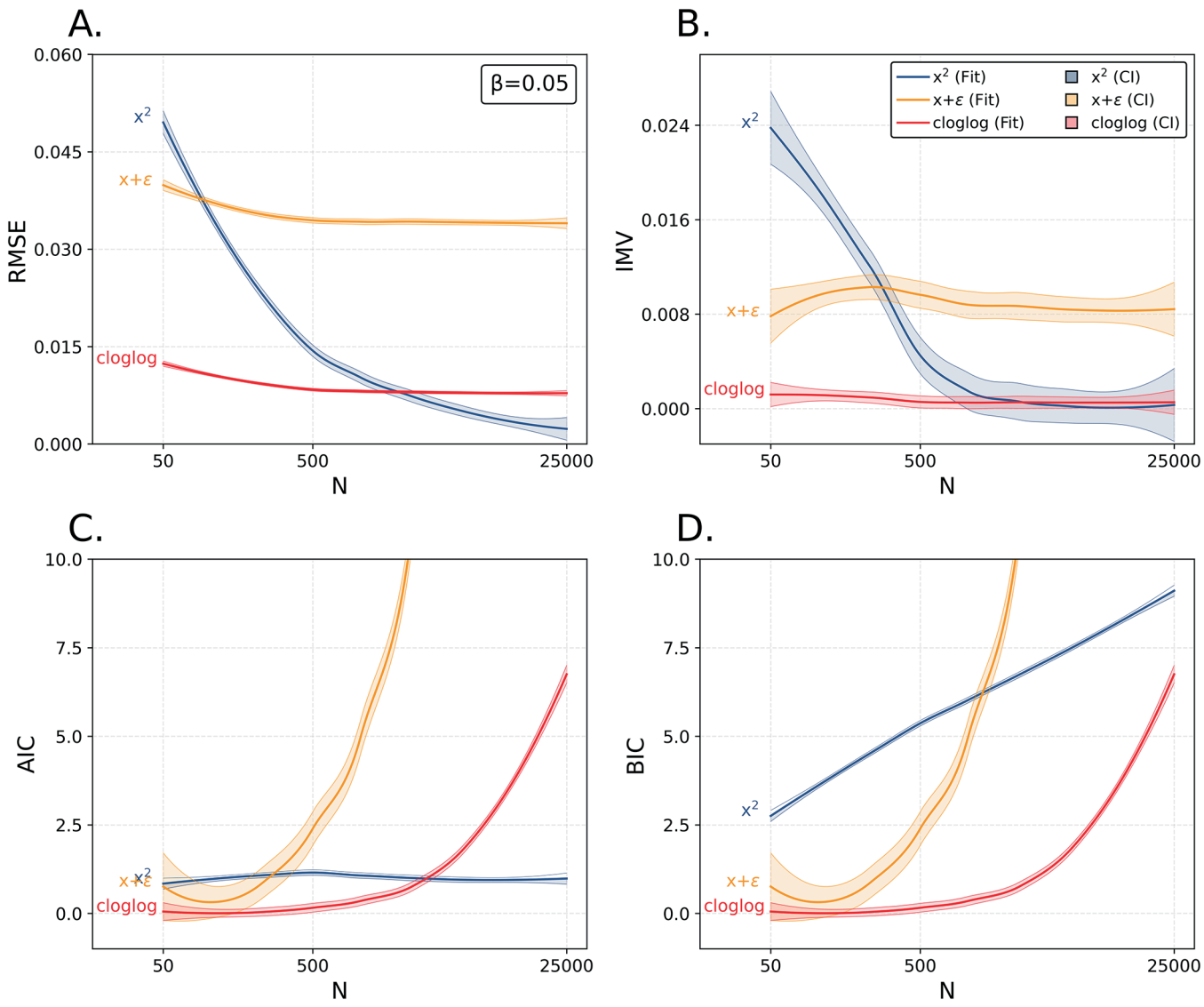

Fig 6. Oracle, Overfit, and Omega. Comparison of IMV and information criteria for different choices of N. Columns represent IMV, AIC, and BIC from left to right respectively. Y-axes are truncated so as to emhasize area of key variation.

https://doi.org/10.1371/journal.pone.0316491.g006

$x + \epsilon$ rather than x (where $\epsilon \sim \text{Normal}(0, 0.3^2)$), and (3) a model where the complementary log-log link is used instead of the logistic. We fix $\beta = 0.5$ and vary N (we sample $n \sim \text{Unif}(\log_{10} 50, \log_{10} 25000)$ and let $N = 10^n$) so as to focus on the behavior of these curves in the context of changing sample size. We generate two sets of outcomes; the first is used for estimation, and the second for computation of out-of-sample IMV values. Each curve is based on a comparison of the correct model to one of the incorrectly specified alternatives. To supplement our interpretation of these curves, we also add the root-mean-square-error of the estimated probabilities as compared to true probabilities.

Fig 6 readily captures the difference in behavior across the IMV and the information criteria. Consider the comparison of the overfit model that includes the quadratic term of the true model (the blue curve). The difference in prediction error (panel A) declines to zero as N

increases; the IMV (panel B) behaves similarly. The IMV curve captures the critical information: with small *N*, an overfit model can be costly but an overfit model with large *N* behaves functionally equivalently to the true model. In contrast, the AIC (panel C) is constant as a function of sample size; in fact, the AIC is roughly unity for all *N* (as expected given that the difference between the models is the estimation of a single additional parameter). The BIC (panel D) increases as a function of sample size.

For the yellow curves—which compare the prediction based on a noisy covariate compared to the true model—there is a relative insensitivity in the IMV value to sample size; this is to be expected given that the attenuation bias does not depend on sample size (see Eq 26.8 in [28]) and is confirmed by the RMSE. The information criteria are, on the other hand, strongly increasing as a function of sample size with respect to this mis-specification. Finally, the mis-specified link (red curve) has a consistently small IMV with some decline as *N* increases; again, the RMSE behaves similarly. In contrast, the information criteria sharply differentiates between the two models for large *N*. Our aim is not to critique the information criteria; these metrics behave appropriately given their designed purpose. But, clearly, the IMV displays distinctive behavior that quantifies the difference between predictions (rather than attempt to authoritatively arbitrate in favor of one set of predictions, as with the information criteria) and this behavior matches that of the prediction error as a function of sample size.

3.3 Summary of simulation results

The simulation studies described here indicate some unique features of the IMV relative to alternatives. It is clearly distinguished from metrics like the AIC and BIC in that its variation as a function of sample size only matters to the extent that sample size improves prediction (rather than allowing for very precise adjudication between fairly similar approaches). This evidence cumulatively suggests that the IMV offers novel perspectives on fit when considering binary outcomes. Again, note that the IMV behaves similar to other metrics if interest hinges on a choice between two models. However, the IMV is designed to be portable such that values can be compared in a straightforward way across settings. That is, the IMV can be used to not just choose between models, but to also understand how much better one model is than another in a consistent manner. Further, the flexibility of the IMV allows for easy computation of the Oracle and Overfit quantities. These quantities can be used to, for example, study the implications of estimation error.

4 Empirical illustrations

We now consider the IMV's potential use in a range of canonical examples across the social sciences. These examples are meant to offer useful benchmarks in the form of IMV values against which future work can be compared, and to illustrate the fact that the IMV can be interpreted as a meangingful quantity even when there are variations in prevalence. Focus is on the IMV rather than alternative metrics. In our view, comparisons to alternatives are best made via controlled simulation settings (as above); this omission is not meant to imply that alternative metrics fail to offer complementary information.

4.1 Illustration One: Prediction of health outcomes

Predictive models are being used to study a variety of health-related phenomena including the social determinants of health [29] and age-related social care [30]. To illustrate how the IMV can be used to index such predictions, we build models related to health outcomes using data from a population-based survey; the Health and Retirement Study (HRS [31,32]; see

S1-IV.1 for further details). We illustrate how the IMV can be used to quantify prediction of health outcomes. The risk of these health outcomes varies substantially as respondents age (see S1–S6 Figs). We thus focus on prediction within age bins. However, given the change in prevalence as a function of age it would be challenging to examine changes in age-related predictability of these health risks using metrics that do not appropriately account for changes in prevalence; the IMV is thus a useful tool for this exercise.

Consider first predictions of health outcomes based solely on demographic information. Certain outcomes—high blood pressure and arthritis in relatively young respondents, and heart disease in relatively old respondents—are predicted with $\omega > 0.02$ using race and sex relative to prevalence alone while others (e.g., stroke, death) are predicted more weakly (results shown in S1–S7 Figs). While we focus on predictions of health problems in age bins, we can utilize predictions of health as a function of age as a benchmark; age is maximally predictive of heart disease ($\omega = 0.016$) compared to prevalence alone. We next consider prediction based on adding educational attainment. Relative to prediction using demographics, gains from adding information on education are extremely modest with $\omega < 0.01$ in virtually all cases. This limited increase in prediction associated with inclusion of information about education is noteworthy given substantial interest in educational disparities in health conditions [33].

We next consider predictions based on relatively expensive-to-collect pieces of health data: cognition and physical functioning (as measured by grip and gait). These expensive data are, for certain outcomes, as predictive as information about respondent age. Amongst older respondents, the cognitive score predicts death and proxy-based responding ($\omega \approx 0.02$) at the next wave. Turning to grip and gait, they predict, for example, heart disease amongst respondents aged 80 ($\omega = 0.019$). As a contrast, we can compare these predictions with those from clinical samples. Prediction of health outcomes in clinical samples is far superior (see Table 1): e.g., heart disease $\omega = 0.12$, Breast Cancer $\omega = 0.53$, diabetes $\omega = 0.62$. These differences presumably reflect the value of predictors ascertainable in clinic settings and show the relatively limited value of similar covariates designed to be informative about individual health in population studies (i.e., grip and gait).

4.2 Illustration Two: Prediction of political party affiliation

Predicting political orientation has recently become a mainstay within the field of social data science [34], with voter-based microtargeting (for the purpose of political messaging) allegedly occurring regularly and in high-profile, consequential circumstances [35]. We examine the relative information content held within simple demographic variables by predicting political party affiliation using data from the General Social Survey (GSS, [36]; detail in S1-IV.2). The relative popularity of the political parties has also changed over time (see Fig 7 Panel B) thus making it challenging to quantify temporal changes in the degree to which party affiliation is structured by these demographic features using many common metrics. The IMV is well-suited to this problem.

We predict affiliation using age, sex, and race. Fig 7 Panel A shows the IMV of demographics in predicting party affiliation beginning in 1970. There is a sharp increase across the 1980s in the predictive power of demographics. After a peak in the 1990s near $\omega = 0.25$, the predictive power declines to roughly between 0.15 and 0.2 between 2000 and 2020. There are only minor differences between additive and interactive models. Especially in the 1990s, party affiliation is relatively strongly predicted by demographic features; the IMV is roughly an order of magnitude higher than those observed with health outcomes in the HRS. Our results complement others discussing the changing nature of US political partisanship [37,38] and suggest a potentially strikingly high level of predictability of partisan affiliation as a function of

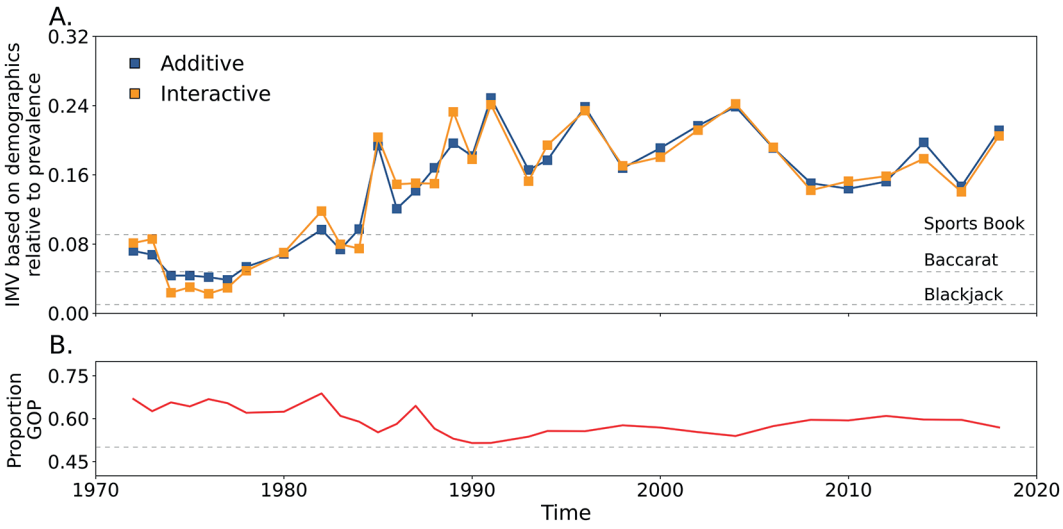

Fig 7. IMV and the GSS. Panel A (top) shows the IMV for prediction of political party affiliation across GSS survey years and Panel B (bottom) the proportion of GOP ("Grand Old Party") within the GSS respondents by year.

https://doi.org/10.1371/journal.pone.0316491.g007

demographics, not least before 'big' data [39] or psychological profiles are integrated into the domain.

4.3 Illustration Three: The Fragile Families Challenge

The Fragile Families Challenge (FFC; [16]; detail in S1-IV.3) aimed to quantify the level of predictability in sociological and behavioural life course outcomes using data from Fragile Families and Child and Wellbeing Study (FFCWS; [40]). Widely heralded as a (much-needed and) progressive approach to bringing out-of-sample prediction into the main-stream social science literature [9,10,41], it incorporated a 'common task method' [42] where 160 teams of independent researchers submitted predictions based on a reserved, unseen hold-out set of six key outcomes in Wave Six of the FFCWS (three of which were binary). Teams differed substantially in the methodological sophistication of their submitted approaches (see also [43]). Some built models based on theory and prior research [44], whereas others began with many variables [45] or took a 'human in the loop' based approach [46].

The binary outcomes had different prevalences in the training data (21%, 23%, and 6% for layoff, job training, and eviction respectively); these differences are a first challenge in making comparisons between the R^2 values used in the original paper (i.e., Brier Skill Scores) which are difficult to interpret across outcomes. Further, the baseline models were differentially successful in predicting outcomes, thus furthering the challenge of making comparisons regarding the degree to which sophisticated approaches led to improved prediction. The IMV also allows us to unequivocally state that the benefits of the 'enhanced' models submitted by challenge participants were unilaterally small. To emphasize the IMV's utility, consider a simple question: for which outcome were predictive gains the largest? Mirroring the FFC, we can consider this using R^2 values (See Fig 8). We first compute the R^2 value based on the benchmark model (0.009, 0.049, 0.014 for layoffs, job training, and evictions respectively) and then compute it for each candidate model and look at the difference in these quantities. For the three outcomes, the maximal R^2 values were 0.028, 0.050, and 0.044 respectively. Taking differences yields 0.019, 0.0004, and 0.030; from this perspective, the modeling innovations were

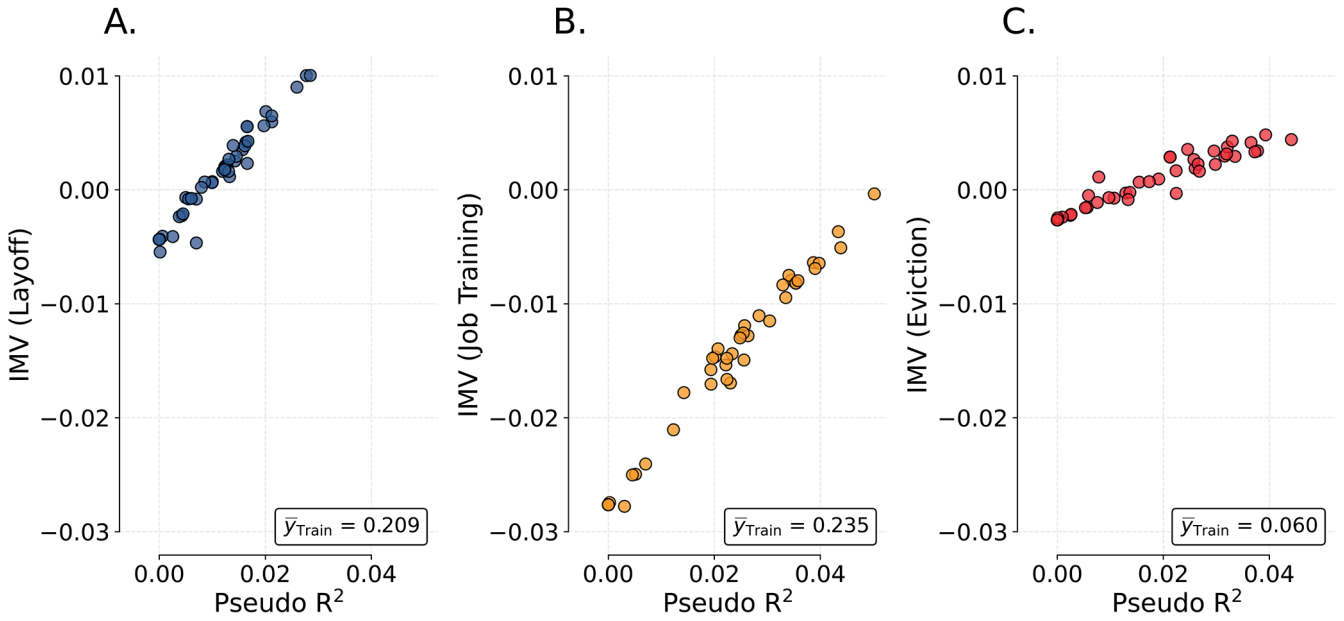

Fig 8. Re-evaluating the Fragile Families Challenge. IMV metrics (from each submission with a Pseudo $R^2 > 0$) plotted against the Pseudo R^2 for all FFC submissions when evaluated against a four variable benchmark [16]. Panels A-C evaluate Layoff, Job Training and Eviction respectively.

https://doi.org/10.1371/journal.pone.0316491.g008

most useful in improving predictions of evictions and layoffs with the biggest gain seeming to come from predicting evictions.

Alternatively, we can use the IMV to analyze this question. In terms of rank-ordering, the IMV and the R^2 values provide similar information (the Spearman correlations between these are greater than 0.97 for all three outcomes); there is little difference across the two metrics as to the rank-ordering of the modeling approaches in terms of the relative improvements. We again observe that the IMV and a metric such as the R^2 will produce similar information if there is interest in making a simple yes/no decision related to any single outcome. However, the IMV offers a different perspective on the question of where the gains are the most impressive across outcomes. Across layoffs and eviction, maximal IMVs (between FFC submissions and the benchmark model) are 0.010 and 0.005 (excluding job training, which was zero to four decimal places). The IMV indicates that the reduction in uncertainty provided by the best FFC model is twice as valuable in removing uncertainty when predicting layoffs as the best FFC model for predicting evictions. The ordering as compared to the R^2 approach is reversed and, more critically, we can meaningfully compare the magnitudes of the IMV given that they are designed to be used in such a manner.

4.4 Empirical benchmarks

Alongside the core empirical illustrations, additional examples further demonstrate the range of IMV values. The first is the prediction of item responses to cognitive tasks using item response theory models using data from the OECD's Programme for International Student Assessment (PISA; [59]). Another is the prediction of social class using natural language processing, based on recent work using text data from college application essays [47]. We also consider whether a home team wins in European football [50], and the prediction of a positive COVID diagnosis based on symptomology [48] (see S1-V.2). Finally, we consider the

prediction of outcomes from a variety of scientific disciplines (e.g., biology, physics, medicine; [55]) that serve as interesting contrasts of more highly predictive systems. See S1-V for additional detail on these data.

Table 1. Summary of results. IMV for various empirical illustrations (alongside gambling comparators) plus prevalences; results ordered by IMV. The IMV is the mean ω across many folds (along with the SD of the ω so as to indicate uncertainty). In some examples (e.g. FFC re-examination) there is only one fold. In such cases (and in casino games), no SD is reported.

Binary Outcome	Data	Model 1	Model 2	Prevalence	Mean ω_k	SD $\omega_k^{\dagger}$
Job Training	FFC	Benchmark Model	Top predictor	0.23	0.0000	–
Math item responses	PISA	2PL	3PL	0.47	0.002	3e-4
Eviction	FFC	Benchmark Model	Top predictor	0.06	0.005	–
Eviction	FFC	$\bar{y}_{\text{Train}}$	Top predictor	0.06	0.007	–
High blood pressure (age 63)	HRS	Age and sex	+ education	0.52	0.008	0.017
Blackjack					*0.010*	
Survival on Titanic	Titanic	Logistic Regression	LightGBM	0.38	0.010	0.048
Math item responses	PISA	Rasch	2PL	0.47	0.010	0.001
Layoff	FFC	Benchmark Model	Top predictor	0.21	0.010	–
Layoff	FFC	$\bar{y}_{\text{Train}}$	Top predictor	0.21	0.014	–
Knowledge of initial coin state					*0.019*	
Death (age 90)	HRS	Age, sex, and education	+ cognition	0.29	0.024	0.022
Heart problems (age 80)	HRS	Age, sex, and education	+ grip and gait	0.39	0.026	0.063
Political Party affiliation (1976)	GSS	Prevalence	GLM based on age, sex and race	0.66	0.026	0.037
Job training	FFC	$\bar{y}_{\text{Train}}$	Top predictor	0.23	0.028	–
Baccarat					*0.048*	
High blood pressure (age 63)	HRS	Prevalence	Age and sex	0.52	0.067	0.108
High family income	[47]	SAT scores	+ topics	0.50	0.073	0.006
Sports book					*0.091*	
COVID Infection	[48]	Prevalence	First month, small spec GLM	0.31	0.092	0.098
Heart disease	[49]	Prevalence	GLM	0.28	0.123	0.287
Survival on Titanic	Titanic	Prevalence	+Sex + ticket class	0.38	0.352	0.143
Home win in European football (average)	[50]	Prevalence	Network model	0.64	0.159	–
Nonmarine coarse siltstone	[51]	Prevalence	GLM	0.23	0.163	0.033
Skin ID	[52]	Prevalence	GLM	0.79	0.196	0.004
Hospital re-admissions in DM patients	[53]	Prevalence	GLM	0.46	0.196	0.015
Excess alcohol consumption	[54]	Prevalence	GLM	0.51	0.245	0.170
Political Party affiliation (1991)	GSS	Prevalence	GLM based on age, sex, and race	0.51	0.256	0.094
Glass Manufacturing process	[55]	Prevalence	GLM	0.41	0.420	0.078
Marine siltstone and shale (v. Mudstone)	[51]	Prevalence	GLM	0.46	0.446	0.157
Breast Cancer	[56]	Prevalence	GLM	0.37	0.526	0.157
Early detection of diabetes	[57]	Prevalence	GLM	0.62	0.617	0.230
Abalone rings	[58]	Prevalence	GLM	0.50	0.667	0.031

FFC results based on the top-performing model. Predictions of health status from HRS selected by identifying the maximum IMV for each pair of model contrasts. For the GSS application, max and min values for additive models across survey years are shown. For games of chance, the house vigorish is shown.

† Values omitted when the IMV is based on out-of-sample prediction from a single test dataset.

https://doi.org/10.1371/journal.pone.0316491.t001

Table 1 contains IMV values observed from prediction exercises across all aforementioned domains. We can compare these results to the vigorish benchmarks from popular games of chance as well as the profit associated with knowledge of the coin's initial state (i.e., head up?) pre-toss [60]. As one example of a relatively weak improvement in predictive value, increasing model complexity for item responses (S1-V.3) on cognitive assessments for adolescents adds limited predictive value, $\omega \leq 0.01$ (these values are similar to those observed for the FFC predictions). Concerning an NLP example: the text used in college admissions essays predicted whether an applicant's family income was above or below the median with $\omega = 0.073$ (S1-V.4). We also replicated previous observations [50] of a changing patterns of predictability of a home team victory in European football over time (See S1-V.5), which are surprisingly predictable ($\omega = 0.159$). Turning to examples from the physical and biological sciences, there are numerous cases (e.g., prediction of abalone rings, $\omega = 0.667$, or glass type, $\omega = 0.420$, see S1-V.6) that serve to benchmark the high levels of predictive value associated with simple models for outcomes determined by well-understood scientific processes. These predictions are, in many cases, orders of magnitudes more valuable than those based on, for example, predictions of health problems in population-based surveys.

Table 1 emphasizes the flexibility of the IMV in allowing for straightforward comparisons across outcomes irrespective of prevalence or modeling strategy. It also provides a range of values against which future studies can be benchmarked. However, we deliberately avoid suggesting specific values as benchmarks. The fact that the IMV can be interpreted in monetary terms suggests that the answer to "What is a large IMV?" is as context-dependent as the answer to "What is a large amount of money?". Future work can place IMVs for a given scenario in a range of contexts using results from Table 1 which includes a large number of outcomes that are both highly stochastic (in the context of predictions made here), such as evictions and layoffs, as well as relatively strongly determined, such as the relationship between age and rings amongst abalones.

5 Discussion

As our capacities for computation and data collection expand, the applicability and relevance of prediction increases. Therefore, so does our need to evaluate prediction in a consistent and tractable fashion. The IMV is a flexible and portable metric for evaluating predictive accuracy with binary outcomes. Our approach focuses on anchoring a given predictive system to a physical analogue with readily understood statistical properties: weighted coins. The coins establish a system that informs us about the expected winnings associated with an improvement in prediction. We compare this approach via simulation to alternative metrics of predictive accuracy and then undertake various simulated and empirical illustrations. Note that the IMV is portable—values can be consistently interpreted across outcomes—and thus can be used across a large range of scientific outcomes and predictive models.

We emphasize a few select facts about the IMV that are intriguing arguments for its use in future settings. First, simulation studies suggest that the IMV is quite sensitive to error in estimated probabilities (i.e., Fig 4) and is also distinct from other metrics that are similarly sensitive in terms of how they respond to changes in prevalence (i.e., Fig 5). Second, the fact that the IMV is inherently a metric of change allows it to be used in interesting ways; in particular, the Oracle and Overfit metrics might be used in simulation work to further our understanding of estimation error, sample size, and the problem of overfitting in many scenarios. Third, the IMV can be used to clarify the meaning of logistic regression coefficients (see S1-V.1.1). Understanding of those coefficients is frequently challenging given that they require

discussion of odds ratios; the IMV can be used to straightforwardly state the relative predictive value of a given covariate in such models in ways that might help ease understanding in future work. Note also that conventional metrics are not helpful for clarifying the relative predictive contribution of a predictor in a portable fashion.

One clear distinguishing feature of the IMV is its sensitivity to prevalence. As an outcome's prevalence moves away from 0.5, this leads to an increase in w_0; thus, a given value of $w_1 - w_0$ will be associated with a smaller IMV as prevalence increases. When discussing the proprties of the IMV, we discussed this behavior based on a consideration of profit maximization in gambling. This defense does not require one to think of prediction as gambling but merely emphasizes that highly prevalent outcomes have less uncertainty as compared to less prevalent ones. This fact is critical in understanding why the IMV behaves as it does. Here we offer an alternative rationale. Increases in w_0 can occur due to either increases in prevalence or increases in the predictive capacity of the baseline model. Focusing on the latter, the IMV's behavior is consistent with the logic that increases in predictive power (i.e., values of $w_1 - w_0$) are more meaningful in terms of resolving uncertainty when we have less predictive power from the baseline model (i.e., a smaller w_0). In our view, this logic is persuasive given that predictive innovations for outcomes that are poorly understood are harder to come by relative to further increasing clarity about relatively well-understood outcomes.

We offer a wide range of empirical illustrations to showcase the IMV. These examples illustrate a wide range of predictability—over two orders of magnitude—of outcomes across a range of scientific disciplines. While we refrain from offering specific values to which future IMVs can be compared, the range shown in Table 1 will allow for rapid contextualization of future results. We also use the IMV to illustrate change in prediction over time. The study of party affiliation in the GSS shows how the metric can be used to index changes in the predictability of outcomes over time. Past approaches may have documented changes in the level of a covariate's estimated magnitude over time but interpretation of such estimates is compromised if the prevalence is fluctuating as it clearly is here; the IMV resolves this issue. The IMV can also be used to clarify interpretation and comparison of logistic regression coefficients (e.g., see discussion of the role of sex in predicting death in the Titanic disaster in S1-V.1).

Alongside the different empirical settings, the illustrations also make use of a wide range of modeling approaches. Alongside logistic regression examples, we also make use of latent variable models (i.e. IRT in the context of PISA; see also [61]), machine learning approaches (in the FFC), and natural language processing (essays and income). The IMV represents a reasonable evaluation metric for endeavours such as Kaggle-like competitions (S1-V.1.2) and is flexible in terms of its ability to allow for comparisons of different specifications and estimators given that it only requires outcomes and associated predictions.

The IMV can be used to compare a multitude of outcomes, but such comparisons need to allow for the role of context. For example, a small increase in an already highly predictive medical diagnostic test may have major implications in terms of time, money, and human lives that render such gains much more important than similar increases in other settings. The portability of the IMV values allows for ready comparisons, but these comparisons will need to be informed by other concerns. A related limitation of the IMV is that it does not differentially weight false positives and negatives. It may need to be used—as with other probabilistic loss functions—with care in settings wherein there is interest in minimizing one of those two quantities. Future work could focus on extending the IMV to incorporate the broader notion of cost/loss or utility functions used in decision theory [62].

Scientists have long prioritized knowledge about *what* predicts an outcome. Interest, however, is turning towards *how* predictable an outcome is (e.g., [16]) and specifically to decomposing 'predictability' in social systems and life prediction tasks [63]. Having metrics that

can be readily used to understand the degree of randomness in a given predictive system is thus highly desirable. The metric introduced here is relatively easy to compute—based on an intuitive analogy to a physical system—and has a range of desirable properties. The scientific community has accumulated a multitude of insights regarding what factors may be relevant for predicting certain outcomes; our work is meant to offer a tool for further advancing our understanding of the stochastic nature of those outcomes.

Supporting information

S1 File. Supporting information for The InterModel Vigorish (IMV) as a flexible and portable approach for quantifying predictive accuracy with binary outcomes.
(PDF)

Acknowledgments

The authors would like to acknowledge Dan Bolt, Richard Breen, Kyla Chasalow, Davide Chicco, Per Engzell, Giuseppe Jurman, David Rehkopf, Mike Sklar, Niklas Tötsch, Mark Verhagen, Shixuan Wang, and Tobias Wolfram for helpful feedback on early drafts of this manuscript, and Taha Yasseri and Victor Maimone for assistance with the football data.

Author contributions

Conceptualization: Benjamin W. Domingue, Charles Rahal, Jeremy Freese, Klint Kanopka, Ben Stenhaug, Ajay Shanker Tripathi.

Data curation: Benjamin W. Domingue, Charles Rahal.

Formal analysis: Benjamin W. Domingue, Charles Rahal, Klint Kanopka, Alexandros Rigos, Ajay Shanker Tripathi.

Investigation: Benjamin W. Domingue, Charles Rahal.

Methodology: Benjamin W. Domingue, Charles Rahal, Klint Kanopka.

Project administration: Benjamin W. Domingue.

Resources: Benjamin W. Domingue, Charles Rahal.

Software: Benjamin W. Domingue, Charles Rahal, Klint Kanopka.

Supervision: Benjamin W. Domingue.

Validation: Benjamin W. Domingue, Charles Rahal.

Visualization: Benjamin W. Domingue, Charles Rahal.

Writing – original draft: Benjamin W. Domingue, Charles Rahal, Jessica Faul, Jeremy Freese, Klint Kanopka, Alexandros Rigos, Ben Stenhaug, Ajay Shanker Tripathi.

Writing – review & editing: Benjamin W. Domingue, Charles Rahal, Jessica Faul, Jeremy Freese, Klint Kanopka, Alexandros Rigos, Ben Stenhaug, Ajay Shanker Tripathi.

References

1. Rahal C, Verhagen MD, Kirk D. The rise of machine learning in the academic social sciences. AI & Society; p. 1–4.

2. Hofman JM, Watts DJ, Athey S, Garip F, Griffiths TL, Kleinberg J, et al. Integrating explanation and prediction in computational social science. Nature. 2021;595(7866):181–8. https://doi.org/10.1038/s41586-021-03659-0 PMID: 34194044
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747 PMID: 7063747
4. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging. 1994;13(4):716–24. https://doi.org/10.1109/42.363096 PMID: 18218550
5. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. https://doi.org/10.1186/s12864-019-6413-7 PMID: 31898477
6. Ramos D, Franco-Pedroso J, Lozano-Diez A, Gonzalez-Rodriguez J. Deconstructing cross-entropy for probabilistic binary classifiers. Entropy (Basel). 2018;20(3):208. https://doi.org/10.3390/e20030208 PMID: 33265299
7. Burnham K, Anderson D. Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res. 2004;33(2):261–304.
8. Hossin M, Sulaiman M. A review on evaluation metrics for data classification evaluations. Int J Data Mining Knowl Manag Process. 2015;5(2):1.
9. Molina M, Garip F. Machine learning for sociology. Annu Rev Sociol. 2019;45(1):27–45. https://doi.org/10.1146/annurev-soc-073117-041106
10. Mullainathan S, Spiess J. Machine learning: An applied econometric approach. J Econ Perspect. 2017;31(2):87–106. https://doi.org/10.1257/jep.31.2.87
11. Hemmert G, Schons L, Wieseke J, Schimmelpfennig H. Log-likelihood-based pseudo-R^2 in logistic regression: deriving sample-sensitive benchmarks. Sociol Methods Res. 2018;47(3):507–31.
12. Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: a modification of the Hosmer-Lemeshow test. Biometrics. 2020;76(2):549–60. https://doi.org/10.1111/biom.13249 PMID: 32134502
13. Williams CKI. The effect of class imbalance on precision-recall curves. Neural Comput. 2021;33(4):853–7. https://doi.org/10.1162/neco_a_01362 PMID: 33513323
14. Šimundić A-M. Measures of diagnostic accuracy: basic definitions. EJIFCC. 2009;19(4):203–11. PMID: 27683318
15. Böger B, Fachi M, Vilhena R, de Fátima Cobre A, Tonin F, Pontarolo R. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. Am J Infection Control. 2020.
16. Salganik M, Lundberg I, Kindel A, Ahearn C, Al-Ghoneim K, Almaatouq A. Measuring the predictability of life outcomes with a scientific mass collaboration: Correction. Proc Natl Acad Sci USA. 2021;118(50):e2023706118. https://doi.org/10.1073/pnas.2023706118
17. Puterman E, Weiss J, Hives B, Gemmill A, Karasek D, Mendes W, et al. Predicting mortality from 57 economic, behavioral, social, and psychological factors. Proc Natl Acad Sci USA. 2020.
18. Edwards A. Pascal's problem: The'gambler's ruin'. International Statistical Review/Revue Internationale de Statistique. 1983; p. 73–9.
19. Cover TM. Elements of information theory. Wiley; 1999.
20. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning, vol. 112. Springer; 2013.
21. Song Q, Tang C, Wee S. Making sense of model generalizability: a tutorial on cross-validation in R and shiny. Adv Methods Pract Psychol Sci. 2021;4(1):2515245920947067.
22. Gneiting T, Raftery A. Strictly proper scoring rules, prediction, and estimation. J Am Statist Assoc. 2007;102(477):359–78.
23. Kelly Jr JL. A new interpretation of information rate. Bell Syst Tech J. 1956;34(4):917–26.
24. Kuha J. AIC and BIC: comparisons of assumptions and performance. Sociol Methods Res. 2004;33(2):188–229.
25. Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J Roy Statist Soc Ser B: Statist Methodol. 1977;39(1):44–7. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x
26. Lv J, Liu J. Model selection principles in misspecified models. J Roy Statist Soc: Ser B (Statist Methodol). 2014;76(1):141–67.
27. Wagenmakers E-J, Farrell S. AIC model selection using Akaike weights. Psychon Bull Rev. 2004;11(1):192–6. https://doi.org/10.3758/bf03206482 PMID: 15117008

28. Cameron A, Trivedi P. Microeconometrics: methods and applications. Cambridge University Press; 2005.
29. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. SSM Popul Health. 2017;4:95–9. https://doi.org/10.1016/j.ssmph.2017.11.008 PMID: 29349278
30. Bardsley M, Billings J, Dixon J, Georghiou T, Lewis GH, Steventon A. Predicting who will use intensive social care: case finding tools based on linked health and social care data. Age Ageing. 2011;40(2):265–70. https://doi.org/10.1093/ageing/afq181 PMID: 21252036
31. Juster F, Suzman R. An overview of the health and retirement study. J Human Resour. 1995:S7–56.
32. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JWR, Weir DR. Cohort profile: the Health and Retirement Study (HRS). Int J Epidemiol. 2014;43(2):576–85. https://doi.org/10.1093/ije/dyu067 PMID: 24671021
33. Braveman PA, Cubbin C, Egerter S, Williams DR, Pamuk E. Socioeconomic disparities in health in the United States: what the patterns tell us. Am J Public Health. 2010;100 Suppl 1(Suppl 1):S186-96. https://doi.org/10.2105/AJPH.2009.166082 PMID: 20147693
34. Blaemire R. The evolution of microtargeting. Campaigns and Elections American Style. 2018;19.
35. Cadwalladr C, Graham-Harrison E. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. 2018;17:22.
36. Davis J, Smith T. The NORC general social survey: a user's guide. vol. 1, Sage; 1991.
37. Zingher J. Polarization, demographic change, and white flight from the democratic party. J Politics. 2018;80(3):860–72.
38. Bafumi J, Shapiro R. A new partisan voter. J Politics. 2009;71(1):1–24.
39. Colleoni E, Rozza A, Arvidsson A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. J Commun. 2014;64(2):317–32.
40. Reichman N, Teitler J, Garfinkel I, McLanahan S. Fragile families: sample and design. Child Youth Serv Rev. 2001;23(4–5):303–26.
41. Wolfram T, Tropf F, Rahal C. Short essays written during childhood predict cognition and educational attainment close to or better than expert assessment. 2022. Available from: htttp://osf.io/preprints/socarxiv/a8ht9.
42. Donoho D. 50 Years of data science. J Comput Graph Statist. 2017;26(4):745–66. https://doi.org/10.1080/10618600.2017.1384734
43. Salganik MJ, Lundberg I, Kindel AT, McLanahan S. Introduction to the special collection on the fragile families challenge. Socius. 2019;5:10.1177/2378023119871580. https://doi.org/10.1177/2378023119871580 PMID: 37309412
44. McKay S. When 4≈10000: The power of social science knowledge in predictive performance. Socius. 2019;5:2378023118811774. https://doi.org/10.1177/2378023118811774
45. Rigobon DE, Jahani E, Suhara Y, AlGhoneim K, Alghunaim A, Pentland A, et al. Winning models for grade point average, grit, and layoff in the Fragile Families Challenge. Socius. 2019;5:2378023118820418.
46. Filippova A, Gilroy C, Kashyap R, Kirchner A, Morgan AC, Polimis K, et al. Humans in the loop: incorporating expert and crowd-sourced knowledge for predictions using survey data. Socius. 2019;5:10.1177/2378023118820157. https://doi.org/10.1177/2378023118820157 PMID: 33981842
47. Alvero AJ, Giebel S, Gebre-Medhin B, Antonio AL, Stevens ML, Domingue BW. Essay content and style are strongly related to household income and SAT scores: evidence from 60,000 undergraduate applications. Sci Adv. 2021;7(42):eabi9031. https://doi.org/10.1126/sciadv.abi9031 PMID: 34644119
48. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med. 2020;26(7):1037–40. https://doi.org/10.1038/s41591-020-0916-2 PMID: 32393804
49. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989;64(5):304–10. https://doi.org/10.1016/0002-9149(89)90524-9 PMID: 2756873
50. Maimone VM, Yasseri T. Football is becoming more predictable; network analysis of 88 thousand matches in 11 major leagues. R Soc Open Sci. 2021;8(12):210617. https://doi.org/10.1098/rsos.210617 PMID: 34925866
51. Hall B. Facies classification using machine learning. The Leading Edge. 2016;35(10):906–9.
52. Bhatt R, Sharma G, Dhall A, Chaudhury S. Efficient skin region segmentation using low complexity fuzzy decision tree model. In: 2009 Annual IEEE India Conference. IEEE; 2009. p. 1–4.

53. Strack B, DeShazo J, Gennings C, Olmo J, Ventura S, Cios K, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int. 2014;2014.

54. Turney PD. Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm. J Artif Intell Res. 1994;2:369–409.

55. Dua D, Graff C. UCI machine learning repository. 2017. Available from: http://archive.ics.uci.edu/ml

56. Bennett KP, Mangasarian OL. Robust linear programming discrimination of two linearly inseparable sets. Optimiz Methods Softw. 1992;1(1):23–34. https://doi.org/10.1080/10556789208805504

57. Islam M, Ferdousi R, Rahman S, Bushra H. Likelihood prediction of diabetes at early stage using data mining techniques. Comput Vision Mach Intell Med Image Anal. Springer; 2020. p. 113–25.

58. Waugh S. Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks. University of Tasmania; 1995.

59. Pisa O. Pisa: results in focus. Organisation for Economic Co-operation and Development: OECD. 2015.

60. Bartos F, Sarafoglou A, Godmann HR, Sahrani A, Leunk DK, Gui PY, et al. Fair coins tend to land on the same side they started: evidence from 350,757 flips; 2023.

61. Domingue BW, Kanopka K, Kapoor R, Pohl S, Chalmers RP, Rahal C, et al. The InterModel Vigorish as a lens for understanding (and quantifying) the value of item response models for dichotomously coded items. Psychometrika 2024;1–21.

62. Vehtari A, Ojanen J. A survey of Bayesian predictive methods for model assessment, selection and comparison. Statist Surv. 2012;6:142–228.

63. Lundberg I, Brown-Weinstock R, Clampet-Lundquist S, Pachman S, Nelson T, Yang V. The origins of unpredictability in life trajectory prediction tasks. arXiv preprint. 2023. http://arxiv.org/abs/2310.12871