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

Benjamin W. Domingue[a,1], Charles Rahal[b,1], Jessica Faul[c,2], Jeremy Freese[a,2], Klint Kanopka[a,2], Alexandros Rigos[d,e,2], Ben Stenhaug[a,2], and Ajay Shanker Tripathi[a,2]

[a]Stanford University
[b]University of Oxford
[c]University of Michigan
[d]Institute for Futures Studies, Stockholm
[e]Lund University
[1]bdomingue@stanford.edu & charles.rahal@sociology.ox.ac.uk
[2]Alphabetized

January 12, 2022

Abstract

Understanding the "fit" of models designed to predict binary outcomes has been a long-standing problem. 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 well-characterized physical systems with tractable probabilities: weighted coins. The IMV is always a statement about the change in fit relative to some baseline—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 illustrate the flexible properties of this metric in numerous simulations and showcase its flexibility across examples spanning the social, biomedical, and physical sciences.

1 Introduction

Understanding, evaluating and comparing the quality of predictions from models trained on binary outcomes has been of long-standing interest. Indeed, different fields have developed heterogeneous preferences with regards to quantifying predictive accuracy; in epidemiology and medicine for example, there has been interest in the ROC curve [1]. In contrast, machine learning research has focused on the harmonic mean of precision and recall (the F_1 score [2]), other quantities related to the confusion matrix determined by a given decision rule [3], and cross-entropy [4]. In psychological measurement, there has been interest in summaries across observations of the log-likelihood [5] and other quantities [6, 7]. Information criteria such as the Akaike & Bayesian Information Criteria and pseudo-R^2 estimates such as those of McFadden, Cox and Snell, or Nagelkerke are also in widespread use [8] without any consistent rationale for their application.

Alongside its historical importance in many fields, the provisioning of accessible and intuitive metrics is essential if the utility of machine intelligence throughout the sciences is to be fully realized [9, 10]. Consider the shortcomings of existing approaches. First, some metrics do not generalize given that they depend on sample-specific quantities (e.g., the magnitude of the log likelihood).[1] 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 (see, for example, the discussion in [13]). In combination, these first two problems

[1]Consider attempts to both generate sample size-sensitive benchmarks [11] and other attempts to reduce sample size dependency in related contexts [12].

make it challenging to make comparisons across contexts. 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 are typically difficult to readily interpret). Collectively, these limitations challenge our ability to make generalizable inferences about the quality of models used in various domains.

We desire a metric that is inherently a comparison between two models rather than an absolute statement about a single model. Statements about a single model have utility in many settings. For example, evaluation of whether a black box diagnostic test is of sufficient accuracy to be used in a specific clinical setting. In that case, something like the AUC can be interpreted alongside established benchmarks [14]. This is an important problem as debates about the clinical accuracy of COVID-19 tests have shown [15]. However, we view such stand-alone approaches as having the crucial limitations noted above; we now further describe these limitations. Suppose we have an outcome y and two predictive systems a and b. A stand-alone metric (e.g., AUC, R^2) produces an index of fit based on each system's predictive accuracy, which we can denote as m_a and m_b. We might readily say that a is a better prediction than b if $m_a > m_b$; 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 we need to address if we want to maximize the impact of predictive science and move towards a more coherent utilization of external validity. We could analyze $\frac{m_a}{m_b}$, $m_a - m_b$, or any number of other derived quantities. Even supposing that we have decided upon a means of quantifying improvement in a relative to b, we will also still need be concerned about baseline prevalence (i.e., $\bar{y}$). For a given level of improvement in fit, would we prefer that level if $\bar{y} = 0.5$ or $\bar{y} = 0.75$? Being able to answer such a question is especially relevant if we are going to make meaningful comparisons of predictive gains *across* outcomes, a topic of current interest [16, 17].

We introduce 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). Our approach is based on translating the level of uncertainty for a given predictive system into a canonical physical system—a weighted coin[2]—and then building inference around the well-characterized statistical properties of that physical system. Inferences take the form of "vigorishes"—where a vigorish is the profit made by a bookmaker or casino associated with taking the bet—in the sense of expected gains in bets where one party has additional 'side' information (which could take the form of, for example, additional dimensionality, feature engineering or refinements to estimation). This metric, which we denote as the InterModel Vigorish (hereafter, IMV), generalizes across multiple predictive schemes where the schemes may vary in outcome, predictors, and approaches to prediction (so long as the approach generates probabilities rather than classes). In tying notions of profits from gambles to questions of prediction, this work ties into the deep traditions of early statisticians such as Blaise Pascal and Christiaan Huygens [19] who used gambling as a means to better understand probability (and vice-versa). Below we briefly describe the IMV before showing how it can be used to benchmark the nature of prediction of binary outcomes across a range of social, biomedical, and physical sciences.

2 Introducing the InterModel Vigorish

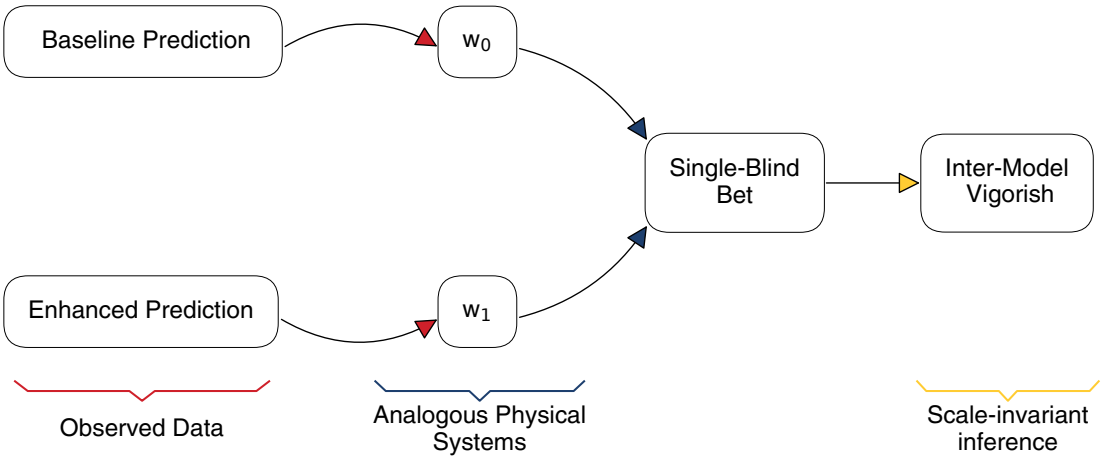
We focus on the problem of constructing a generalizable (in the sense that values of the IMV are comparable across outcomes) metric for comparing the accuracy of two predictive systems for binary outcomes. We refer to these as the 'baseline' and 'enhanced' predictions. In general, we anticipate the enhanced prediction contains 'side' information not available to the baseline prediction but these names are chosen to increase intuition (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 we can always consider one of the models to be a prediction based on prevalence alone (i.e., the training mean), it is not.

Our approach is a 3-step process (see Figure 1):

1. We first use identify a physical system that is analogous to a given predictive system with respect to the level of randomness (i.e., entropy); in particular, we identify the appropriate weight a coin would need to have to be as random as our predictive system. Outcomes based on a coin with a weight of

[2]While we use the notion of a weighted 'coin' because of its canonical use as a stimulant for statistical intuition, biasing the flip of coin via weighting might in fact be physically impossible [18]. Of course there are physical systems (e.g., dice, drawings, roulette wheels) whose outcomes can be dichotomized and can be induced to deviate from 50/50 by different artifices.

Figure 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 where one player knows w_1. The IMV is constructed from the expected winnings associated with knowing w_1. Results can be compared across outcomes given that the fair bet is based on w_0.



0.5 are perfectly random while those based on a coin with a weight of 1 are not random; we gauge the randomness in our system by benchmarking against a coin with equal randomness.

2. We take advantage of the fact that the coins are sufficiently well-characterized to construct a fair bet based on the coin associated with the baseline prediction (i.e., the predicted probability of success). The construction of the fair bet is what ensures that the metric is portable across outcomes.

3. Finally, we update the information held by one player in that bet to construct a "single-blind bet" (the emphasis is on the fact that one player has side information). Expected winnings associated with that bet are our metric for predictive accuracy.

Denominating the increased predictive power of the second predictive system in this way makes it intuitive, generalizable, and in possession of valuable scale properties (we can, for example, reasonably talk about a doubling of predictive value if we observe one IMV being two times another). We further describe each step in Methods.

2.1 Key features of the IMV

Our definition of the IMV—which we denote as ω in equations—in terms of expected winnings offers several important benefits. Increases are quantified on a common scale that accounts for features of the baseline prediction (e.g. where the baseline model is simply imputation based on the sample mean). This accounting is not ad-hoc but rather based on the construction of the fair bet. This is crucial, as appropriate metrics of predictive accuracy need to account for differences in the baseline prevalence if we would like to compare fit across contexts. Further, increases in predictive power are translated into easily interpretable, portable, and consistently meaningful terms; the IMV is the amount of money you'd expect to win for a bet of a single dollar if the coin implied by the baseline model has been replaced by the coin implied by the enhanced model.

We briefly note a number of additional features which are more thoroughly explored in the Supplemental Information (SI). First, we argue that the sensitivity of the IMV to the underlying prevalence is crucial in terms of understanding differences across outcomes and offer an illustration differentiating the behavior of the IMV from stand-alone alternatives such as R^2 as a function of prevalence (S2). Second, we argue (S2) that the

IMV is a proper scoring rule, contrast the IMV to the "Kelly criterion" [20], and also introduce comparisons to vigorishes from a number of common parlor games of chance (i.e., roulette, blackjack, baccarat). Third, we conduct a variety of simulation studies (S3) meant to further contrast the IMV's behavior to commonly used alternatives; based on these studies, we argue that it demonstrates more sensible performance in many conditions. We now turn to a wide variety of empirical illustrations.

3 Empirical Illustrations

We consider a variety of empirical examples. We focus discussion on three in-depth empirical examples using data from existing and canonical studies. In order, these three examples allow us to: (i) probe the utility of different types of side information for predicting health outcomes, (ii) show the change in the value of demographic information while appropriately accounting for changes in prevalence in the prediction of party affiliation, and (iii) reconsider a large 'common task' challenge to which aimed to predict a variety of sociobehavioral outcomes. We then briefly discuss results from a range of other prediction tasks further discussed in the SI. Results are accumulated in Table 1.

3.1 Illustration One: Prediction of health outcomes

We first predict health outcomes within small age windows using data from a population-based survey, the Health and Retirement Study (HRS; [21, 22]). Additional information on data and the full set of predictions is available in S4.1, with key predictions for all illustrations summarized in Table 1. 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., arthritis, stroke, heart disease, death) are predicted more weakly. We then consider prediction based on adding educational attainment. Predictive gains are extremely modest with $\omega < 0.01$ in virtually all cases. This suggests that educational attainment offers fairly limited predictive power relative to prediction based on age and sex.

We next consider predictions based on relatively expensive-to-collect pieces of health data: cognition and physical functioning (as measured by grip and gait). Amongst older respondents, the cognitive score predicts death and proxy-based responding ($\omega \approx 0.02$) at the next wave. Turning to the grip and gait predictors, they are predictors of, for example, heart disease amongst respondents 80 years old ($\omega = 0.019$). However, for both cognition and grip/gait, the predictive power of these specialized variables is fairly low ($\omega < 0.02$) given that such data is expensive to collect.

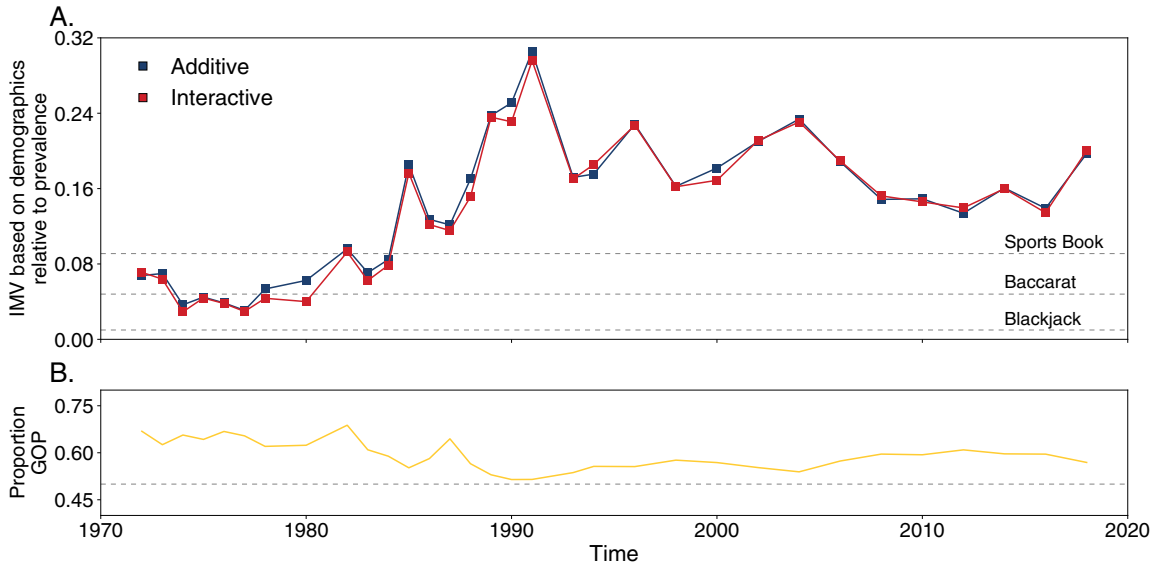
Collectively, these results suggest a relatively high degree of randomness in health outcomes relative to these predictors. Prediction of similar outcomes in clinical samples is far superior (see Table 1): e.g., heart disease 0.12, Breast Cancer 0.53, diabetes 0.62. These differences presumably reflect the value of predictors ascertainable in clinic settings and also show the relatively limited value of similar covariates designed to be informative about individual health in population studies (i.e., grip and gait). As an additional benchmark for interpreting these values, we can consider models including age as a linear predictor of the health indicators considered here (amongst respondents 60-90y); age is maximally predictive of heart disease ($\omega = 0.016$).

3.2 Illustration Two: Prediction of political party affiliation

We next predict political party using data from the General Social Survey (GSS; [23]) (additional detail shown in S4.2). This application is motivated by recent events involving targeting of political messaging [24]. Figure 2 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 $\omega = 0.3$, the predictive power declines to roughly between 0.15 and 0.2 between 2000 and 2020. Note that we consider models wherein the three demographic predictors are additive (red) and interactive (blue). The interactive model produces nearly identical results in this context (note some small deviations around 1980).

As compared to the health outcomes in HRS, party affiliation is far less random after adjusting for simple demographics; the IMV is roughly an order of magnitude higher in many cases. Our results complement others discussing the changing nature of US political partisanship [25, 26] and suggests that prediction of political affiliation based on fairly simple covariates is in the vicinity of predictions from other scientific disciplines with highly specific predictors.

Figure 2: 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 within the GSS respondents by year.



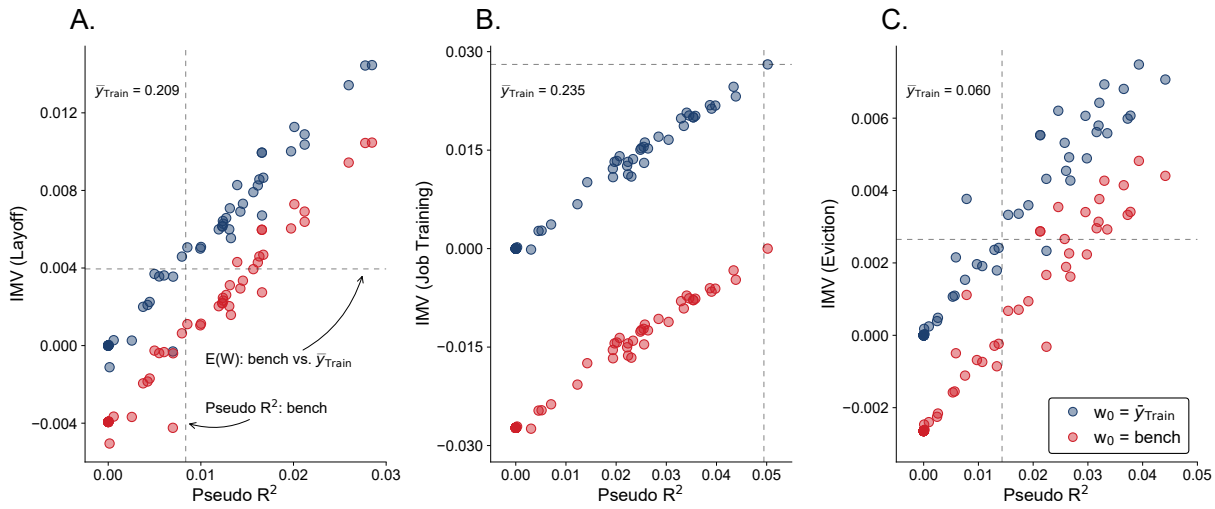
3.3 Illustration Three: Reconsidering the Fragile Families Challenge

The Fragile Families Challenge (FFC; [16]) aimed to quantify the level of predictability in life course outcomes using data from Fragile Families Children and Wellbeing Study (FFCWS; [27]). Widely heralded as a (much-needed and) progressive approach to bringing out-of-sample prediction into the main-stream social science literature [9, 10], it incorporated a 'common task method' [28] where 160 teams of independent researchers submitted predictions for a hold-out set on any of six key outcomes in Wave Six of the FFCWS. Here we focus on the three binary outcomes, with additional details shown in S4.4. We postulate that the IMV might have more cross-outcome comparative value for evaluation of submissions against the four variable benchmark model as compared to the pseudo R^2 metric (i.e. the 'Brier Skill Score') used in the original study.

The outcomes have different prevalences in the training data (21%, 23%, and 6% for layoff, job training, and eviction respectively), and these differences make comparisons between the pseudo R^2 values used in the original paper challenging to interpret across outcomes. In contrast, the IMV can be readily interpreted given that it is conditioned on the baseline prevalence (or, alternatively, the predictions of a baseline model). In Figure 3, the blue markers show the predictive value of the submitted models relative to predictions based on prevalence alone. Gains in the IMV metric relative to prevalence alone are 0.015, 0.028, and 0.007 for the maximally predictive submissions. Thus, the submitted predictions generally show gains against the baseline prevalence of $\bar{y}_{\text{Train}}$ (i.e. $\omega > 0$), although we note that the increase in prediction for eviction is meager.

However, improvement in prediction compared to prevalence alone is a low bar. The FFC was primarily interested in whether more sophisticated modeling approaches improve upon four variable benchmark models determined by a domain expert. For layoff and eviction, the more sophisticated models submitted to the FFC do indeed yield improvements. These improvements are fairly modest however (e.g., $\omega \leqslant 0.01$). For job training, the submitted models do not yield any improvements, other than in one exception ($\omega = 1.38 \times 10^{-6}$). This is presumably due to the fact that, for job training, the benchmark model is itself quite predictive ($\omega = 0.028$) relative to prediction based on prevalence alone. Our results echo those of the FFC in suggesting that alternate modeling approaches lead to fairly meager gains in our ability to predict life course outcomes with such a small sample size, but the IMV ensures comparability across outcomes.

Figure 3: Re-evaluating the Fragile Families Challenge. IMV metrics plotted against the pseudo R^2 for all FFC submissions when evaluated against either $\bar{y}_{\text{Train}}$ (blue markers, the mean of the training data) or against the 4-variable benchmark (red markers). Panels A-C respectively evaluate Layoff, Job Training and Eviction.



3.4 Additional empirical examples

We consider several additional examples to provide insight and further emphasize the flexibility of the IMV; they are briefly described here with additional detail in S5. We consider prediction of item responses to cognitive tasks using item response theory models using data from the OECD's Programme for International Student Assessment (PISA; [29]). We build off of recent work using text data from college application essays [30] to illustrate how our metric can be used with natural language processing.

We also show how the IMV can be used to make straightforward comparisons of logistic regression coefficients—a vexing problem [31]—by comparing the role of sex in prediction using data from the GSS and the Titanic disaster (S5.3). In brief, we show that the inclusion of gender is over 50 times as valuable a predictor in prediting death amongst Titanic passengers as compared to predicting political affiliation in the GSS. Note that the outcomes are of relatively similar prevalence in both cases—so it is not simply the case that these are driven by such differences—and the baseline model for deaths in the Titanic absent gender is more predictive than the baseline model in the GSS. We also use this data to outline the utility of the IMV in Kaggle-type predictive competitions (S5.4), where we consider a range of commonly used algorithms and metrics to show a potential lack of agreement in the ordinal ranking of algorithms when considering the IMV and–specifically–label based metrics. While the rankings are highly correlated, this further emphasizes the importance of metric choice in competition design. We advocate for consideration of the IMV, given that it is designed to facilitate comparisons of prediction across multiple competitions and outcomes.

To the above, we also consider the prediction benchmarks from a variety of scientific disciplines (e.g., biology, physics, medicine). We aim to evaluate the IMVs associated with prediction of out-of-sample cases using standard (e.g., logistic regression) approaches to combine information about a variety of predictive factors across a variety of scientific processes that lead to variation in the stochasticity of the outcome.

3.5 Summary of Empirical Results

Table 1 compares the IMVs observed across a range of outcomes and includes vigorish benchmarks from games of chance (based on calculations in S2.1). To the predictions based on the HRS, GSS, and FFC, we add the additional examples discussion in Section 3.4. Individual item responses collected from administering cognitive tasks to adolescents are the outcomes wherein we have lowest predictive accuracy as measured by IMV; this should not be surprising as these individual item responses reflect relatively little information about the respondents compared to, for example, their health or political affiliation. However, health outcomes

are also predicted with relatively low accuracy with data from social surveys; the maximum gains from models for predicting health outcomes from demographics rarely exceed 0.02. Relative to prediction based on demographics alone, addition of other predictors yields even smaller values.

In contrast, prediction of health outcome in clinical settings resulted in much higher IMVs (presumably due to the higher quality information available in those clinical studies). For the FFC outcomes, the IMV ranges substantially but are generally consistent with the previous conclusion that these life course outcomes are largely stochastic [16]. In another example of a highly predictive outcome, using basic demographics to predict political party affiliation yields fairly large improvements compared to prevalence alone. At a minimum, the IMV was 0.03 but increased by an order of magnitude at its largest (0.307 in 1990 for the additive model). Examples from the physical and biological sciences (e.g., prediction of abalone rings or glass type) benchmark the levels of predictions associated with well-understood scientific processes; such predictions are, in many cases, orders of magnitudes more valuable than those based on, for example, predictions of health problems in population-based surveys. This summary across a variety of outcomes illustrates the portability of the IMV approach and its capacity to compare either a given model to baseline prevalence or the gains generated by moving from one model to another.

4 Discussion

As computational power increases, dimensions of data expand, and new empirical methods are developed, the applicability and relevance of prediction increases. Therefore, so does the researcher's need to be able to evaluate prediction in a consistent and tractable fashion. We introduce a flexible and portable metric for evaluating predictive accuracy with binary outcomes. Our approach focuses on anchoring a given predictive system to a physical system with readily understood statistical properties: weighted coins. We use these coins to 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. We emphasize the flexibility of our approach in that it only requires predicted probabilities generated *in any way* for each outcome, and thus can be used with a large class of predictive models (and inputs), contrasting its simplicity and portability with the interpretive challenges of, for example, log-likelihood based pseudo-R^2 values [11].

Our empirical illustrations point to a wide range of potential uses. The study of party affiliation in the GSS shows how the metric can be used to track the level of predictability of a fundamental social science outcome. 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 clearly it is here. We argue that IMV has desirable properties—clear intuition, clear dependence on prevalence—relative to many of these alternative metrics. Our analysis of HRS health outcomes suggests, first and foremost, that these outcomes are less well predicted given demographics than is party affiliation. We might suspect that the addition of targeted information—cognitive functioning, grip, and gait—would provide substantial increases in prediction of health outcomes relative to demographics alone. While we do see some predictive power from these more specialized variables, the gains are modest (especially considering how expensive these data are to procure). We also found the limited predictive power of educational attainment to be noteworthy given substantial interest in educational disparities in health conditions [42].

In our final empirical example, we reconsider the results of the recent FFC. Using the IMV metric, we show that predictive gains from the maximally predictive submitted models for layoff are roughly twice as large as the predictive gains for eviction. Note that this assessment can be made despite the disparate prevalences of these two outcomes in the training data. In contrast, submitted models for job training offered no predictive gains relative to the baseline model; this is presumably due to the fact that the baseline model was a relatively strong predictor of this outcome already (e.g., an IMV relative to prediction based on the prevalence of 0.028 as compared to smaller quantities—0.015 and 0.007—for layoff and eviction).

As illustrated by these various use cases, our metric could help future work better understand the relative degree of randomness in wide-ranging binary outcomes of scientific interest.[3] Table 1 shows predictions that

[3]Consider the way that different outcomes can be placed on a skill/luck continuum [43]. In the sense of our paper, we would posit that "skill" is just a loss of randomness due to the utility of certain predictors. The IMV may be a useful tool in helping to operationalize further explorations of such ideas.

Table 1: Summary of select results. IMV for various empirical illustrations (alongside gambling comparators) plus prevalences; results ordered by IMV.

Binary Outcome	Data	Model 1	Model 2	Prevalence	IMV
Job Training	FFC	Benchmark Model	Top predictor	0.23	< 0.001
Math item responses	PISA	2PL	3PL	0.47	0.002
Eviction	FFC	Benchmark Model	Top predictor	0.06	0.005
Eviction	FFC	$\bar{y}_{\text{Train}}$	Top predictor	0.06	0.007
Blackjack					*0.010*
Math item responses	PISA	Rasch	2PL	0.47	0.010
Layoff	FFC	Benchmark Model	Top predictor	0.21	0.011
Layoff	FFC	$\bar{y}_{\text{Train}}$	Top predictor	0.21	0.015
High blood pressure (age 63)	HRS	Age and sex	+ education	0.52	0.018
Heart problems (age 80)	HRS	Age, sex, and education	+ grip and gait	0.39	0.019
Death (age 90)	HRS	Age, sex, and education	+ cognition	0.29	0.025
Job training	FFC	$\bar{y}_{\text{Train}}$	Top predictor	0.23	0.028
Political Party affiliation (1977)	GSS	Prevalence	GLM based on age and sex	0.66	0.030
Baccarat					*0.048*
High family income	[30]	SAT scores	+ topics	0.50	0.073
High blood pressure (age 63)	HRS	Prevalence	Age and sex	0.52	0.082
Sports book					*0.091*
Heart disease	[32]	Prevalence	GLM	0.28	0.123
Death on Titanic	Titanic	Passenger and ticket features	+Sex	0.62	0.136
Nonmarine coarse siltstone	[33]	Prevalence	GLM	0.23	0.163
Skin?	[34]	Prevalence	GLM	0.79	0.196
Hospital readmissions in DM patients	[35]	Prevalence	GLM	0.46	0.196
Home team in European football	[30]	Prevalence	[36]	0.61	0.244
Excess alcohol consumption	[37]	Prevalence	GLM	0.51	0.245
Political Party affiliation (1991)	GSS	Prevalence	GLM based on age and sex	0.51	0.307
Death on Titanic	Titanic	Prevalence	LightGBM	0.62	0.410
Glass Manufacturing process	[38]	Prevalence	GLM	0.41	0.420
Marine siltstone and shale (v. Mudstone)	[33]	Prevalence	GLM	0.46	0.446
Breast Cancer	[39]	Prevalence	GLM	0.37	0.526
Early dection of diabetes	[40]	Prevalence	GLM	0.62	0.617
Abalone rings	[41]	Prevalence	GLM	0.50	0.667

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.

differ in value by over 500%. We emphasize one additional point about these various use cases shown in Table 1; these IMV values are based on several types of models. While we primarily consider straightforward logistic regression-based approaches, this is not always the case. In some cases we use latent variable models (i.e. IRT in the context of PISA), machine learning approaches (in the FFC and Titanic examples), and natural language processing (essays and income). The IMV is ultimately flexible in terms of its ability to allow for comparisons of different specifications and estimators given that it requires minimal information (i.e., just the model-based predictions), and poses a minimal computational burden whilst being portable across domains.

While we argue that our metric's ability to readily compare outcomes in lieu of difference prevalences is valuable, we also note that such comparisons will 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. 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 probabalistic loss functions—with care in settings wherein there is interest in minimizing one of those two quantities.

Scientists have long prioritized knowledge about *what* predicts an outcome. Interest, however, is turning towards *how* predictable an outcome is, given a broad set of predictors and approaches to model-building. 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 this has several desirable properties. The scientific community has accumulated great insights about 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.

5 Methods

Additional detail on the process of constructing the IMV is shown below followed by a discussion of the materials used herein.

5.1 The IMV

5.1.1 Identifying an analogous coin

Suppose that we have two predictive models (i.e., the baseline and enhanced models) for some binary outcome. We are interested in the predictive accuracy of the enhanced model relative to the baseline model (again emphasizing that the baseline model could be quite simple and potentially invariant across observations; e.g., the outcome's prevalence). Suppose $y_i \in \{0, 1\}$ are the individual outcomes and $p_i \in [0, 1]$ are the probabilities implied by one of our models (when we need to specify, we denote $p_i^{(0)}$ for the baseline and $p_i^{(1)}$ for the enhanced). The likelihood assigned by the model to each observation is

$$L_i = p_i^{y_i}(1 - p_i)^{1-y_i}, \tag{1}$$

and from that the log likelihood:

$$\ell_i = \log(L_i) = y_i \log p_i + (1 - y_i) \log(1 - p_i). \tag{2}$$

We summarize these quantities via the geometric mean of the likelihoods

$$A = \left(\prod_{i=1}^{n} L_i \right)^{\frac{1}{n}}. \tag{3}$$

A is a natural quantity to consider as it is a transformation of the sample log-likelihood $\bar{\ell}$; specifically, it is the mean log-likelihood transferred back to the likelihood scale. It can be used as a representation of the in-sample likelihood that does not depend on sample size. Note as well that $\log A = \bar{\ell}$. We use A for identification of an analogous weighted coin.

We now aim to identify the Bernoulli random variable with this expected log likelihood. That is, we want to identify the coin with weight w that would be expected to produce $\log A$ if we consider its likelihood. We thus solve

$$w \log(w) + (1 - w) \log(1 - w) = \log A \tag{4}$$

for w.[4] A visualization of this is shown in Figure S1 Panel A where we show the curve linking A to w. In this transformation, we translate our level of predictive control, A, from our model-based predictions ($p^{(0)}$ or $p^{(1)}$) to the necessary weight for a coin to be similarly random.[5] Note also that there is a unique coin associated with a given value of $\log A$; that is, a given system of predictions translates to a singular equivalent weighted coin. Conceptually, these weights represent coins whose outcomes contain results with a similar level of uncertainty as those of our predictive system; for example, a weight of $w = 0.99$ would represent a highly accuracy predictive system while a weight of $w = 0.51$ would represent a system whose predictions are only marginally better than chance. We denote the weights for our (baseline and enhanced) coins as w_0 and w_1; they play a central role in what follows.

We additionally emphasize that Eqn 4 implicitly uses notions of entropy as the mechanism for identifying an analogous physical system. Entropy is a measure of the level of uncertainty [44]. For a coin with associated probability of w, entropy is defined as

$$-1(w \log w + (1 - w) \log(1 - w)).^{6} \tag{5}$$

Thus, in Figure S1 Panel A we are identifying a coin with equivalent uncertainty, in terms of entropy, as our predictive system. Note that entropy is both symmetric around and maximized at $w = 0.5$. Further, note the clear parallel between Eqn 2 and Eqn 5.

5.1.2 The fair bet

We now consider a thought experiment composed of two bets based on our hypothetical coins in a casino. First, a fair bet based on w_0. Suppose that the house flips a coin with weight w_0. The house bets one dollar on an outcome of heads and a bettor puts in $1/O_0$ dollars where O_0 is the odds, equal to $w_0/(1-w_0)$. What can the house expect to win (in net)? Denoting this quantity as W_0, if the coin shows heads, the house wins the pot which translates to a net winnings of $W_0 = 1/O_0$ (having subtracted the initial stakes) while if the coin shows tails, the house lose a dollar: $W_0 = -1$. This is a fair bet in the sense that expected net winnings are zero. That is

$$\mathbb{E}(W_0) = \left(\frac{1}{O_0}\right) w_0 + (-1)\,(1 - w_0) \tag{6}$$

$$= \left(\frac{1 - w_0}{w_0}\right) w_0 + -1 + w_0 \tag{7}$$

$$= 1 - w_0 - 1 + w_0 = 0. \tag{8}$$

There is no advantage to being either the house or the bettor (i.e., the game is fair). Note one important benefit of having translated our predictive system into the coin with weight w_0: for a given toss of the coin, there are only two outcomes and only a single bet to be made about this scenario (i.e., was the coins heads or tails)? We rely on this fact in our assertion that the IMV metric as used here is not the byproduct of ambiguous choices.

5.1.3 The single-blind bet

Now we turn to our single-blind bet. Suppose that the coin used in the bet has been replaced with one of weight w_1. Crucially, the house knows this, but the bettor does not (hence, the single-blind nature of the

[4] Computationally, our minimization problem is $\text{argmin}_{w \in [0.5,1]} |w \log(w) + (1 - w) \log(1 - w) - \log(A)|$.

[5] Note that values of A below 0.5 are possible (for example, suppose that an outcome is 1 with probability 0.99 but where the prediction is 0.4). Such a model is inconsistent with the entropy of any weighted coin (hence we only consider values of $A > 0.5$). These models get $w = 0.5$ for the relevant computations.

[6] Note that, given our use of the natural logarithm, our units of entropy are "natural" units of information, denoted "nats" [45].

bet). In this game, we can again calculate expected winnings for the house:

$$\mathbb{E}(W_1) = \left(\frac{1}{O_0}\right) w_1 + (-1)(1 - w_1) \tag{9}$$

$$= \frac{w_1 - w_0 w_1 - w_0 + w_0 w_1}{w_0} = \frac{w_1 - w_0}{w_0}. \tag{10}$$

Given its presence in the numerator, we write $\delta = w_1 - w_0$. We finally define the IMV, denoted as ω in equations, as

$$\omega \equiv \frac{\delta}{w_0}. \tag{11}$$

For every flip of the coin, the house expects to win ω.

This quantity is intuitive and, crucially, portable across outcomes with different prevalences given the way w_0 was used to construct the fair bet. The IMV name is motivated by the fact that $\mathbb{E}(W_1) \equiv \omega$ is effectively the profit that accrues as a function of the intermodel difference in prediction quality.[7] That is, prediction quality is assessed as the level of profit that the house expects to make based on the side information offered by the enhanced model relative to the baseline (i.e., knowing the w_1 coin is being flipped rather than the w_0 coin). Recall that w_0 and w_1 are the unique coins associated with the two predictive systems, the fair bet is the unique fair bet that can be made about a coin of weight w_0, and the single-blind bet is similarly the unique profit associated with exchanging w_1 for w_0 while blinding one player. These are crucial facts as they ensure that the IMV is uniquely identified for two predictive systems.

5.1.4 Computation of ω

We make a few notes regarding computation of ω. We include computation of ω for a toy example in S1. This quantity can be computed directly with information about the likelihoods for the baseline and enhanced model. We also provide a simple website for computing this quantity.[8] In practice, the quantity can be computed using in-sample or out-of-sample values for the likelihood (or, equivalently, based on in-sample or out-of-sample sets of estimated probabilities and observed responses). We focus here largely on out-of-sample prediction given the issues associated with overfitting (see [46] or Figure S2 in SI) with in-sample work but emphasize that the IMV is a versatile metric that can be readily utilized in either computational framework.

5.2 Materials

We consider three core empirical examples.[9] In the first two cases, we focus on mean IMV based on 10-fold cross-validation. That is, for each calculation of IMV, we split the data into 10 folds. We treat each fold as an out-of-sample set in which we compute ω based on a model trained on the remaining 9 folds of the data. In the final illustration, we re-evaluate the FFC submissions. Additional detail on these data is available in the SI. We also make use of a range of illustrations of prediction in various datasets so as to benchmark the levels of prediction observed in different settings; all data are discussed in the SI.

Acknowledgements

This work was supported by the Jacobs Foundation (B.D.), the Leverhulme Centre for Demographic Science (The Leverhulme Trust; C.R.) and Nuffield College (C.R.). The authors would like to acknowledge Dan Bolt, 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. 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.

[7] In this discussion, we have presumed that $w_1 > w_0$; that is, we assume the enhanced prediction is better than the baseline prediction. This need not be the case; in such cases, as we shall see below, we end up with $\omega < 0$.

[8] `https://klint-kanopka.shinyapps.io/imv-app/`

[9] Code for these analyses is available at `https://github.com/crahal/InterModel_Vigorish`.

References

[1] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[2] Alex P Zijdenbos, Benoit M Dawant, Richard A Margolin, and Andrew C Palmer. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE transactions on medical imaging*, 13(4):716–724, 1994.

[3] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[4] Daniel Ramos, Javier Franco-Pedroso, Alicia Lozano-Diez, and Joaquin Gonzalez-Rodriguez. Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, 20(3):208, 2018.

[5] Taehoon Kang and Allan S Cohen. Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, 31(4):331–358, 2007.

[6] John B DiTrapani. *Assessing the Absolute and Relative Performance of IRTrees Using Cross-Validation and the RORME Index*. PhD thesis, The Ohio State University, 2019.

[7] Yoav Bergner, Stefan Droschler, Gerd Kortemeyer, Saif Rayyan, Daniel Seaton, and David E Pritchard. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. *International Educational Data Mining Society*, 2012.

[8] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[9] Mario Molina and Filiz Garip. Machine learning for sociology. *Annual Review of Sociology*, 45(1):27–45, 2019.

[10] Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, May 2017.

[11] Giselmar AJ Hemmert, Laura M Schons, Jan Wieseke, and Heiko Schimmelpfennig. Log-likelihood-based pseudo-r 2 in logistic regression: deriving sample-sensitive benchmarks. *Sociological Methods & Research*, 47(3):507–531, 2018.

[12] Giovanni Nattino, Michael L Pennell, and Stanley Lemeshow. Assessing the goodness of fit of logistic regression models in large samples: A modification of the hosmer-lemeshow test. *Biometrics*, 76(2):549–560, 2020.

[13] Christopher KI Williams. The effect of class imbalance on precision-recall curves. *Neural Computation*, 33(4):853–857, 2021.

[14] Ana-Maria Šimundić. Measures of diagnostic accuracy: basic definitions. *Ejifcc*, 19(4):203, 2009.

[15] Beatriz Böger, Mariana M Fachi, Raquel O Vilhena, Alexandre de Fátima Cobre, Fernanda S Tonin, and Roberto Pontarolo. Systematic review with meta-analysis of the accuracy of diagnostic tests for covid-19. *American journal of infection control*, 2020.

[16] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.

[17] Eli Puterman, Jordan Weiss, Benjamin A Hives, Alison Gemmill, Deborah Karasek, Wendy Berry Mendes, and David H Rehkopf. Predicting mortality from 57 economic, behavioral, social, and psychological factors. *Proceedings of the National Academy of Sciences*, 2020.

[18] Andrew Gelman and Deborah Nolan. You can load a die, but you can't bias a coin. *The American Statistician*, 56(4):308–311, 2002.

[19] AWF Edwards. Pascal's problem: The'gambler's ruin'. *International Statistical Review/Revue Internationale de Statistique*, pages 73–79, 1983.

[20] John L Kelly Jr. A new interpretation of information rate. *The Bell System Technical Journal*, 34(4), 1956.

[21] F Thomas Juster and Richard Suzman. An overview of the health and retirement study. *Journal of Human Resources*, pages S7–S56, 1995.

[22] Amanda Sonnega, Jessica D Faul, Mary Beth Ofstedal, Kenneth M Langa, John WR Phillips, and David R Weir. Cohort profile: the health and retirement study (hrs). *International journal of epidemiology*, 43(2):576–585, 2014.

[23] James A Davis and Tom W Smith. *The NORC general social survey: A user's guide*, volume 1. SAGE publications, 1991.

[24] Carole Cadwalladr and Emma Graham-Harrison. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian*, 17:22, 2018.

[25] Joshua N Zingher. Polarization, demographic change, and white flight from the democratic party. *The Journal of Politics*, 80(3):860–872, 2018.

[26] Joseph Bafumi and Robert Y Shapiro. A new partisan voter. *The Journal of Politics*, 71(1):1–24, 2009.

[27] Nancy E Reichman, Julien O Teitler, Irwin Garfinkel, and Sara S McLanahan. Fragile families: Sample and design. *Children and Youth Services Review*, 23(4-5):303–326, 2001.

[28] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[29] OECD Pisa. Pisa: Results in focus. *Organisation for Economic Co-operation and Development: OECD*, 2015.

[30] AJ Alvero, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L Stevens, and Benjamin W Domingue. Essay content and style are strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. *Science advances*, 7(42):eabi9031, 2021.

[31] Richard Breen, Kristian Bernt Karlson, and Anders Holm. Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, 44:39–54, 2018.

[32] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

[33] Brendon Hall. Facies classification using machine learning. *The Leading Edge*, 35(10):906–909, 2016.

[34] Rajen B Bhatt, Gaurav Sharma, Abhinav Dhall, and Santanu Chaudhury. Efficient skin region segmentation using low complexity fuzzy decision tree model. In *2009 Annual IEEE India Conference*, pages 1–4. IEEE, 2009.

[35] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[36] Victor Martins Maimone and Taha Yasseri. Football is becoming more predictable; network analysis of 88 thousand matches in 11 major leagues. *Royal Society Open Science*, 8(12):210617, 2021.

[37] Peter D Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of artificial intelligence research*, 2:369–409, 1994.

[38] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[39] Kristin P Bennett and Olvi L Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1(1):23–34, 1992.

[40] MM Faniqul Islam, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*, pages 113–125. Springer, 2020.

[41] Samuel George Waugh. *Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks*. PhD thesis, University of Tasmania, 1995.

[42] Paula A Braveman, Catherine Cubbin, Susan Egerter, David R Williams, and Elsie Pamuk. Socioeconomic disparities in health in the united states: what the patterns tell us. *American journal of public health*, 100(S1):S186–S196, 2010.

[43] Daniel Getty, Hao Li, Masayuki Yano, Charles Gao, and AE Hosoi. Luck and the law: quantifying chance in fantasy sports and other contests. *SIAM Review*, 60(4):869–887, 2018.

[44] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[45] Fazlollah M Reza. *An introduction to information theory*. Courier Corporation, 1994.

[46] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

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

Supplemental Information (SI)

S1 A toy example

We use a simple simulation to generate outcomes based on the combination of 20 tosses of a fair coin ($w = 0.5$) and 20 tosses of a heavily weighted coin ($w = .95$). Code to reproduce this example in `R` is shown below. We 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 emphasize that we don't 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.[1]

Suppose, arbitrarily, that our baseline prediction is that all outcomes are produced via a coin with probability $p^{(0)} = 0.55$ of being heads. We first compute $A_{p^{(0)}} = 0.53$ (where we subscript A to indicate the model upon which it is based). We translate this into an analogous coin of weight $w_0 = 0.67$ (see also Panel A in Figure S1). So as to forestall confusion, we emphasize 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. 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 $A_{p^{(1)}} = 0.63$ 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. When predicting the result of a coin flip generated by this same data-generating process, we would expect to win nearly a quarter (i.e. 24 cents) for every dollar wagered.

S1. A toy example

```
set.seed(8675309)
# Combine tosses from fair coin with a heavily weighted coin:
x1<-rbinom(10,1,.5)
x2<-rbinom(10,1,.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 approximation:
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=0.001,upper=.999,a=a0)$par
# Create an improved approximation:
p<-c(rep(.5,10),rep(.9,10))
a1<-ll(x=x,p=p)
p1<-nlminb(.5,f,lower=0.001,upper=.999,a=a1)$par
# Calculate the single-blind bet:
imv<-(p1-p0)/p0
```

[1] As an illustration, suppose we are shown 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 H and one that always produces T—there is no randomness. Speaking of the randomness of these outcomes necessitates reference to the data-generating process.

S2 Key features of the IMV

We emphasize a few important features of IMV. First, given that A is the geometric average likelihood for an observation, the IMV captures the expected winnings for prediction of a single outcome in the test data. Second, note that the IMV decreases as w_0 increases for a fixed δ; this behavior is shown in Panel B of Figure S1. Is this desirable? Consider a potential alternative, $Z(W) = \frac{\mathbb{E}(W)}{\sqrt{\mathbb{V}(W)}}$ (where $\mathbb{V}$ is the variance operator). Given that $\mathbb{V}(W)$ decreases as w_0 nears unity, this would have the effect of our preferring gains, in terms of δ, when w_0 is near one (i.e. the dashed curves in Panel B of Figure S1 are upward sloping).[2] However, doing so would just tell us that we are making big gains relative to very small levels of uncertainty; in terms of their impacts on the reductions in uncertainty as quantified by our single-blind bet, they are less meaningful than comparable changes of size δ relative to smaller values of w_0. The goal in betting is to make money—to maximize $\mathbb{E}(W)(= \omega)$—not to make smaller amounts of money in games with relatively little randomness; we argue that 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 prioritization of δ values when the underlying entropy is relatively low (e.g., something more akin to $Z(W)$); our approach would be inappropriate in such cases.

S2.1 The IMV as a proper scoring rule

In the realm of probabilistic forecasting, scoring rules are used to evaluate the quality of predictions [1]. Because the IMV requires probabilities as inputs, it makes sense to consider its use as a potential scoring rule. Recall that the IMV is meant to quantify the predictive gain of one model relative to another. Despite this difference in application, if we fix the baseline model, the IMV constitutes a scoring rule for the evaluation of the predictive quality of enhanced models. A desirable property of scoring rules is that they are *strictly proper*, or they produce a maximum expected score if and only if the forecasted probabilities are equal to the true probabilities [1]. As the IMV with its fixed baseline model is a re-scaling of the log likelihood of the enhanced model, and as the log likelihood itself is a proper scoring rule (the *logarithmic score*), we see that the IMV with fixed baseline model is itself a strictly proper scoring rule. As such, it prioritizes accurately describing the underlying probability distribution which generates the observed data over attributing extreme confidence to predictions.

S2.2 Connections to betting and the Kelly Criterion

The interpretation of the IMV in non-statistical terms (given that it calculates the expected payoff from an uncertain investment) can be immediately connected to expected returns in other domains outside of formal statistical modelling. For example, our approach shares some of the logic in previous derivations of key principles of decision analysis from the basis of coin-tossing [2]. An even more useful comparison is to gambling. The vigorish (the "profit" associated with taking the bet that is generated by assignment of asymmetric odds to competing outcomes) at most sports books is 10% [3] implying that $\omega = 0.091$. This is relatively high; vigorishes for baccarat and blackjack (under certain assumptions, such as play being strategically optimal) are 5% [4] and 1% [5] respectively, which translate into expected winnings of 0.048 and 0.0099 cents per dollar respectively for the casino.[3] We use these as benchmarks in empirical illustrations below.

Our construction of the IMV is also related to one approach to betting: the "Kelly Criterion" [6]. This relates to an optimal strategy for uncertain investment. Suppose that in an (infinite) series of instances one is faced with a bet that pays out V dollars for each dollar wagered in the event of a win, while one loses the whole wager in the event of a loss. If the probability of winning is p, what proportion of one's wealth should

[2] As an alternative view of this issue, consider $\Pr(y = 1) = \sigma(\beta_0 + \beta_1 x)$ where $\sigma(\cdot)$ is the standard logistic sigmoid. As β_0 increases, one needs increasingly large values of β_1 to maintain a constant $\mathbb{E}(W)$ (i.e., see discussion of Figure S1 Panel D below).

[3] Consider a hypothetical two-outcome betting example where a book prices the ('European'/decimal) odds of an event happening at 1.5 (i.e. net 50 cents profit on a dollar successfully wagered), and offers odds on the same event not happening at 2.5. The 'over-round' (O) on this book is calculated as the sum of the implied probabilities: $(\frac{1}{1.5} + \frac{1}{2.5}) = 106.67\%$, and the 'vig' as $\omega = \frac{O-100}{O} = 0.0625$.

they bet at each instance? The Kelly criterion proposes that one can maximize long-run wealth via optimal wagering. In particular, the optimal proportion of one's wealth one should wager (the Kelly bet) is

$$K \equiv p - \frac{1-p}{V}. \tag{S1}$$

In our case, the amount won (per dollar) in the event of a win is the one implied by the benchmark model, i.e., $V = 1/O_0 = (1-w_0)/w_0$, while the probability of winning (for the player who uses the enhanced model) is $p = w_1$. The Kelly bet is, then,

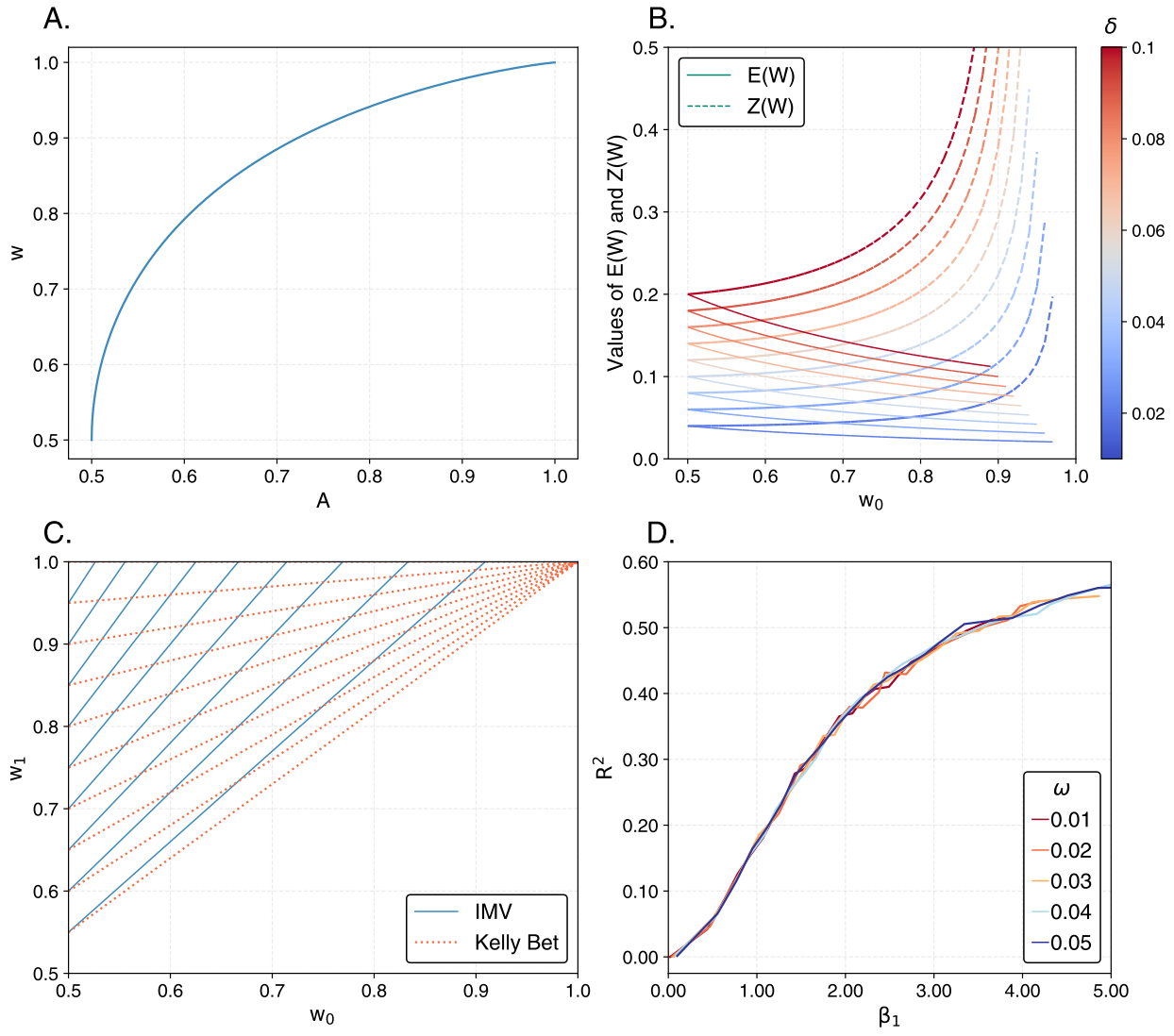
$$K = w_1 - \frac{1-w_1}{\frac{1-w_0}{w_0}} = \frac{w_1(1-w_0) - w_0(1-w_1)}{1-w_0} = \frac{w_1 - w_0}{1-w_0}. \tag{S2}$$

Notice the similarity between the Kelly bet (Eqn. S2) and the IMV (Eqn. 10). The numerator in both formulas is $w_1 - w_0$, but the denominator differs. The values of both metrics increase in w_1 and decrease in w_0 if $w_1 < 1$. Suppose, however, that the better model is perfect at predicting the outcomes (i.e., $w_1 = 1$). This implies $K = 1$ which is is intuitive: if one is certain to win the wager, one should bet their entire bankroll. This can be seen in the contour plots for IMV and the Kelly bet in Panel C of Figure S1; note the flat line associated with $w_1 = 1$. From our perspective, the lack of discrimination by K between values of w_0 when $w_1 = 1$ suggests a limited utility of K for model comparisons (admittedly in an edge case) given that there is much more to be gained when w_0 is relatively small. If $w_0 = 0.999$, one should bet their bankroll, but accruing wealth will be a slow process.

S2.3 Distinguishing IMV from R^2

To emphasize the distinction between IMV and alternatives, we focus here on a key discrepancy between the IMV and pseudo-R^2 (the latter used as a stand-in for similar metrics) via simulation. Based on the logistic regression model which we introduce in the simulation study (Eqn S4), we fix β_0 and a value for ω. We then identify a β_1 value consistent with these choices (for $x \sim \mathrm{N}(0,1)$). We use the β_0 and β_1 values (and set $\beta_2 = 0$) to simulate data and then compute the R^2 value associated with application of a logistic regression model to the simulated data. Results are shown in Figure S1 Panel D wherein we compare β_1 and R^2. We emphasize that the lines are all based on common values of ω. While R^2 increases as a function of β_1, by construction ω is constant. This is due to the fact that ω is sensitive to prevalence in a way that other metrics may not be. As β_0 increases, a larger β_1 value is required to hold ω constant (recall Figure S1 Panel B). From the IMV perspective, this is as it should be given the loss of entropy that comes from an increase in β_0. This divergence between the meaning of ω and R^2 is crucial in indicating the novel intuition provided by the IMV.

Figure S1: Properties of the IMV. Panel A shows a mapping between A and w. Panel B shows $\mathbb{E}(W)$ as a function of p and δ. Panel C displays Contour plots for the IMV and the Kelly bet given w_0 and w_1. Panel D plots values of R^2 and β_1 relative to a fixed ω and β_0.



S3 Simulation studies

We now describe a variety of simulation studies to illuminate behavior of the IMV. First, we illustrate how the IMV performs in a simple univariate logistic regression context. Second, we then contrast performance of the IMV with alternative metrics using a simulation strategy proposed elsewhere [7]. Third, we continue this comparison with alternative metrics with an emphasis on the role of prevalence. Fourth, we emphasize differential behavior when metrics are being used to index change between predictive models.

S3.1 Three uses of the IMV: ω_0, the oracle, and overfitting

We use a univariate logistic regression problem to emphasize some key facts about the IMV. We simulate data as $y_i \sim \delta(\beta_1 x_i)$ (where $x_i \sim \mathrm{N}(0,1)$) for $i \in \{1, ..., N\}$ and consider variation in N and β_1. We estimate the logistic regressio model using (x_i, y_i) and use this to generate fitted value for y_i, $\widehat{p}_i$. We then generate a second set of outcomes, $y^\star$, for the same values of β_1 and x. For each simulated set of data, we then consider:

- ω_0: IMV($\bar{y}$, $\widehat{p}_i$; $y^\star$),
- Overfit: IMV($\widehat{p}_i$, p_i; y),
- Oracle: IMV($\widehat{p}_i$, p_i; $y^\star$)

where p_i is the true probability ($\sigma(\beta_1 x_i)$) and we use the notation IMV$(b, e; o)$ to reference the IMV computed with baseline model b, enhanced model e, and outcomes o. The oracle values are only available due to the fact that we know the data generating mechanism (i.e., we know p_i). The overfit values will illustrate the cost of overfitting in-sample by focusing on y instead of the novel $y^\star$. Crucially, overfitting will be indicated by negative values of the IMV. Here, these suggest that the estimates are more valuable than the true p_i, a clear indication of a problem.

Results for 5000 simulations for each value of β_1 are shown in Figure S2. Several key points emerge. When $\beta_1 = 0.01$, there is (unsurpisingly) little value in estimates to predict $y^\star$. However, note that overfitting leads to a substantial cost associated with the use of the true p_i. For $\beta_1 = 0.1$, we can note two key facts. First, the oracle IMV declines as sample size increases due to declines in $|p_i - \widehat{p}_i|$. Second, the value associated with ω_0 increases as a function in sample size for the same reason. For $\beta_1 = 0.5$, we see a clear role of sample size. Values of ω_0 increase as a function of sample size while both the oracle and overfit IMV values decline towards zero.

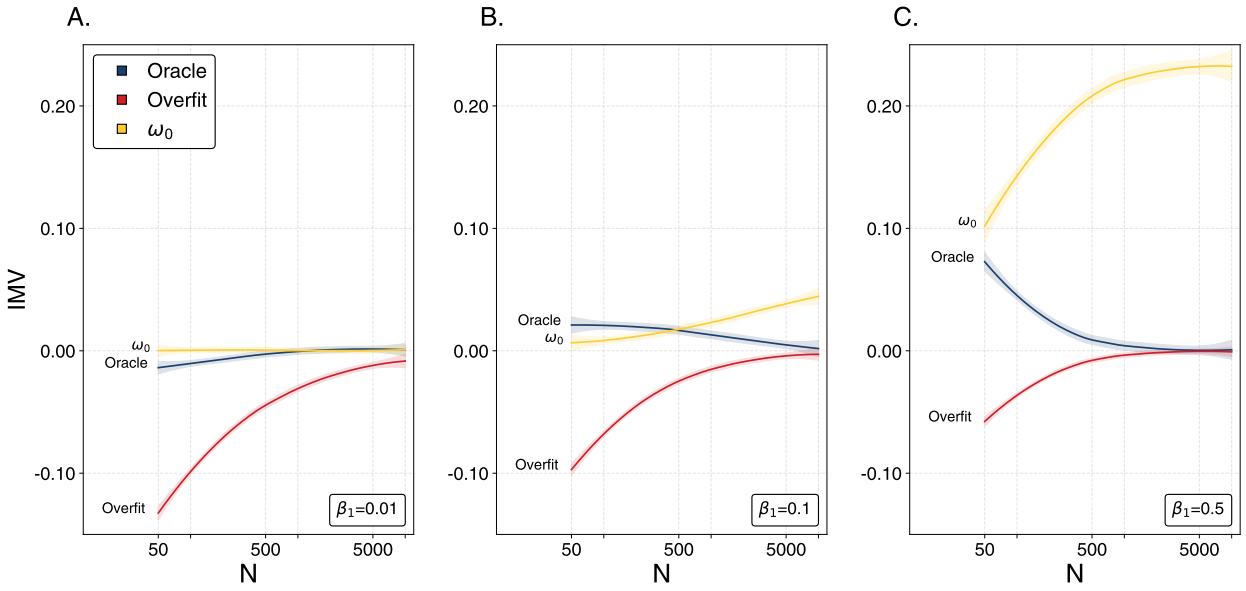
S3.2 Alternative Metrics

So as to compare the performance of the IMV, we introduce three alternative metrics. Using $y_i \in \{0, 1\}$ to denote the i-th observation of some binary outcome and $\widehat{p}_i \in (0, 1)$ as a prediction for that observation, we consider:

1. Pseudo R^2: Motivated by recent work [8], we consider a metric inspired by the traditional R^2 metric of regression:
$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{p}_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{S3}$$
This metric is closely related to the Brier score [9] which typically focuses on $\frac{1}{N}\sum_i (y_i - \widehat{p}_i)^2$ where $i \in \{1, ..., N\}$.
2. AUC: The AUC (area under curve, [10]) metric has long been used to study performance of classifiers based on a comparison of the true and false positive rates.
3. F_1: The F_1 score [11] is the harmonic mean of recall (percentage of true cases that are correctly identified) and precision.

This is, of course, only a subset of the large number of potential metrics we could consider; we view these as being fairly popular and representative of the larger set.

Figure S2: 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.



We begin by contrasting these fit statistics using the framework from [12] wherein we generate probabilities from two Beta distributions. Specifically, we sample four numbers from Unif$(1, 15)$. Denoting these as a_k, we generate 5000 probabilities from Beta(a_1, a_2) and 5000 probabilities from Beta(a_3, a_4). Ten thousand outcomes are then generated based on these probabilities and we compute metrics using the known probabilities. In Figure S3 we show these metrics for 1000 draws of a_k parameters alongside information about the difference between the expectations of the Beta distributions (where we've ordered things such that this difference is always positive) and the prevalence (based on $|0.5 - \hat{y}|$).

Note first that all of the metrics are positively correlated with the difference in the means of the Beta distributions (i.e., $\frac{a_1}{a_1+a_2} - \frac{a_3}{a_3+a_4}$), thus indicating that they are sensitive to changes in the differences in probabilities for responses. The indices are sensitive to prevalence in the sense that there is less variation in the metrics when $|0.5 - \hat{y}|$ is nearer to 0.5. The metrics are all positively correlated amongst themselves, where the lowest correlation between the IMV and another metric is with the F_1 score ($r = 0.38$). This is driven by the fact that the F_1 score shows peculiar behavior as a function of prevalence (note the "spike" in that scatterplot); a potential problem with usage of this statistic. Similarly, usage of the R^2 and AUC statistics may be compromised by the potential inflation in these statistics for large values of $|0.5 - \hat{y}|$ (note the departure of the points from the x-axis in those scatterplots).

S3.3 Fit statistics as a function of prevalence

Below we make use of the following equation to simulate data:

$$\Pr(y_i = 1) = \sigma(\beta_0 + \beta_1 x_i + \beta_2 z_i) \tag{S4}$$

where $\sigma(\cdot)$ is the logistic sigmoid ($\sigma(x) = (1 + \exp(-x))^{-1}$) and x_i, z_i are independently samples from the standard normal distribution. For each iteration of the simulation, we first generate triplets $(\beta_0, \beta_1, \beta_2)$. The predictors x and z are then independently drawn from the standard normal distribution, with y being drawn from a Bernoulli distribution with probability specified by Equation S4. This process is repeated to generate observations (x_i, z_i, y_i) for $i \in \{1, ..., 4000\}$.

We begin with an examination of the fit statistics under different assumptions about the prevalence of the underlying binary indicator. We consider results based on 1000 simulations wherein β_0, β_1 are independently sampled from Unif$[0, 1]$ and $\beta_2 = 0.3$. Here, we blind ourselves to z and predict based on x alone. For IMV, the comparison is to prediction based on prevalence alone. Results in the form of a third order polynomial

fit (with 99.9% confidence intervals) to the 1000 points, showing patterning as a function of β_0 and β_1 are shown in Figure S4. As expected, all metrics are increasing as a function of β_1. However, note that there is varying sensitivity to values of β_0. AUC and R^2, for example, are unassociated with β_0. Whether this is desirable is a crucial question; from our perspective, the IMV provides a clear answer. Larger values of β_0 are consistent with less stochastic outcomes and consequently there is less to be gained (in terms of our single blind bet) from knowledge of x. This is apparent in the decrease of IMV as a function of β_0.

S3.4 Fit statistics as indicators of differences in model quality

We turn now to a second simulation wherein we introduce information about z so as to probe the performance of fit statistics when they are meant to index change between two predictive models. In these simulations, we hold β_0 and β_1 constant but sample β_2 from Unif$[0, 1]$. Results are shown in Figure S5. In the top panel, we are considering raw versions of the metrics (where the IMV values are computed relative to prediction based on prevalence alone). The metrics all increase as a function of β_2 as expected. Note that even when $\beta_2 = 0$, there is still predictive value due to x and thus the IMV (and all other metrics) are greater than 0.

We now focus on the key question: how does inclusion of z increase prediction relative to prediction based on x alone? This analysis is addressed in the bottom panel. Consider first the IMV values. For the IMV, this is a straightforward question. When $\beta_2 = 0$, there is no gain from inclusion of z and the IMV is 0. It increases to different maximum values—the precise value depending on the values of β_0 and β_1—when β_2 is near 1. For the other metrics, assessing the increases is less clear. These metrics need to first be transformed to address such a question. Let us call m_x the value of a given metric when we predict with just x and $m_{x,z}$ the value when we predict with x and z. We consider differences, $m_{x,z} - m_x$, in the bottom panels of Figure S5 but note that this choice is arbitrary; alternative formulations (e.g., $\frac{m_{x,z}}{m_x}$, $\frac{m_{x,z}-m_x}{m_x}$) are possible. We emphasize that the relative differences vary substantially across the different columns. For example, the increase in prevalence between first two columns leads to a large change in the slope of the F_1 curve. This curve gets still steeper when we reduce β_1 in the final column. The variation across the columns suggests that inferences about the relative change in going from prediction based on x to prediction based on x and z will be potentially sensitive to parameters in a way that is poorly understood. In contrast, the IMV metric has a clear definition and interpretation when the goal is comparing two models; when such comparisons are of interest, we view the IMV as being uniquely tailored for the task.

Figure S3: Comparison of four metrics, along with the difference ($\frac{a_1}{a_1+a_2} - \frac{a_3}{a_3+a_4}$) and prevalence, based on data simulated from Beta distributions.

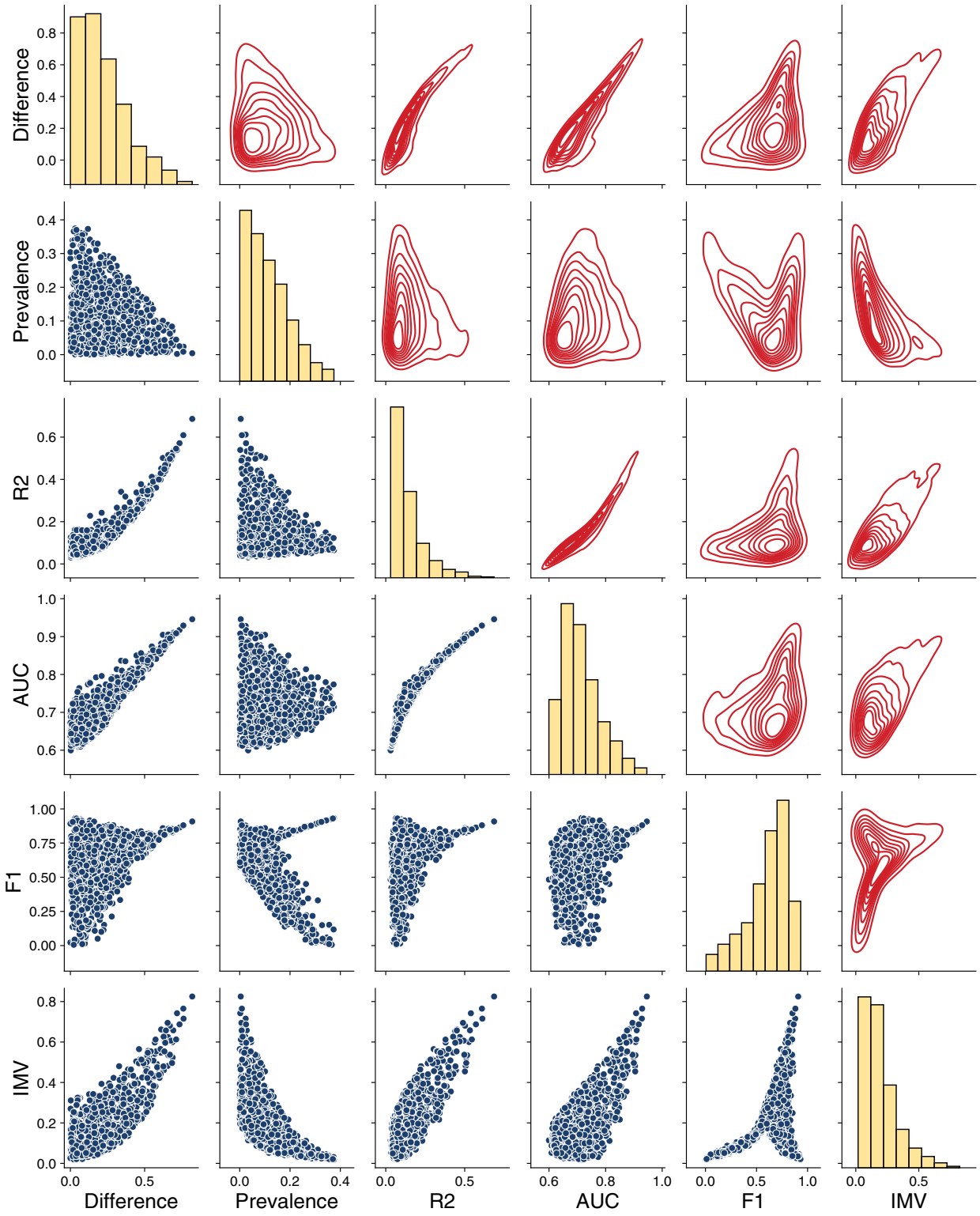


Figure S4: Patterning of fit metrics based on x alone. Metrics are fit as a function of β_0 and β_1 ($\beta_2 = 0.3$). Curves are third order polynomials fit to 1000 iterations of the simulation with 99% confidence intervals.

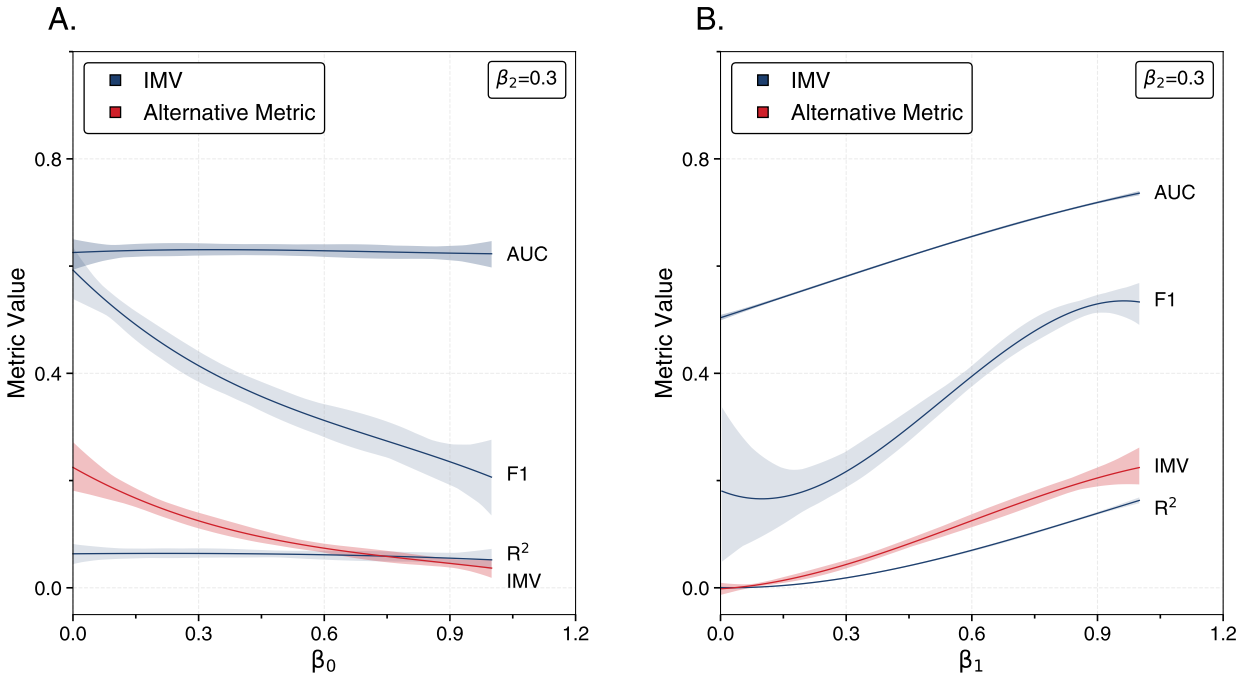
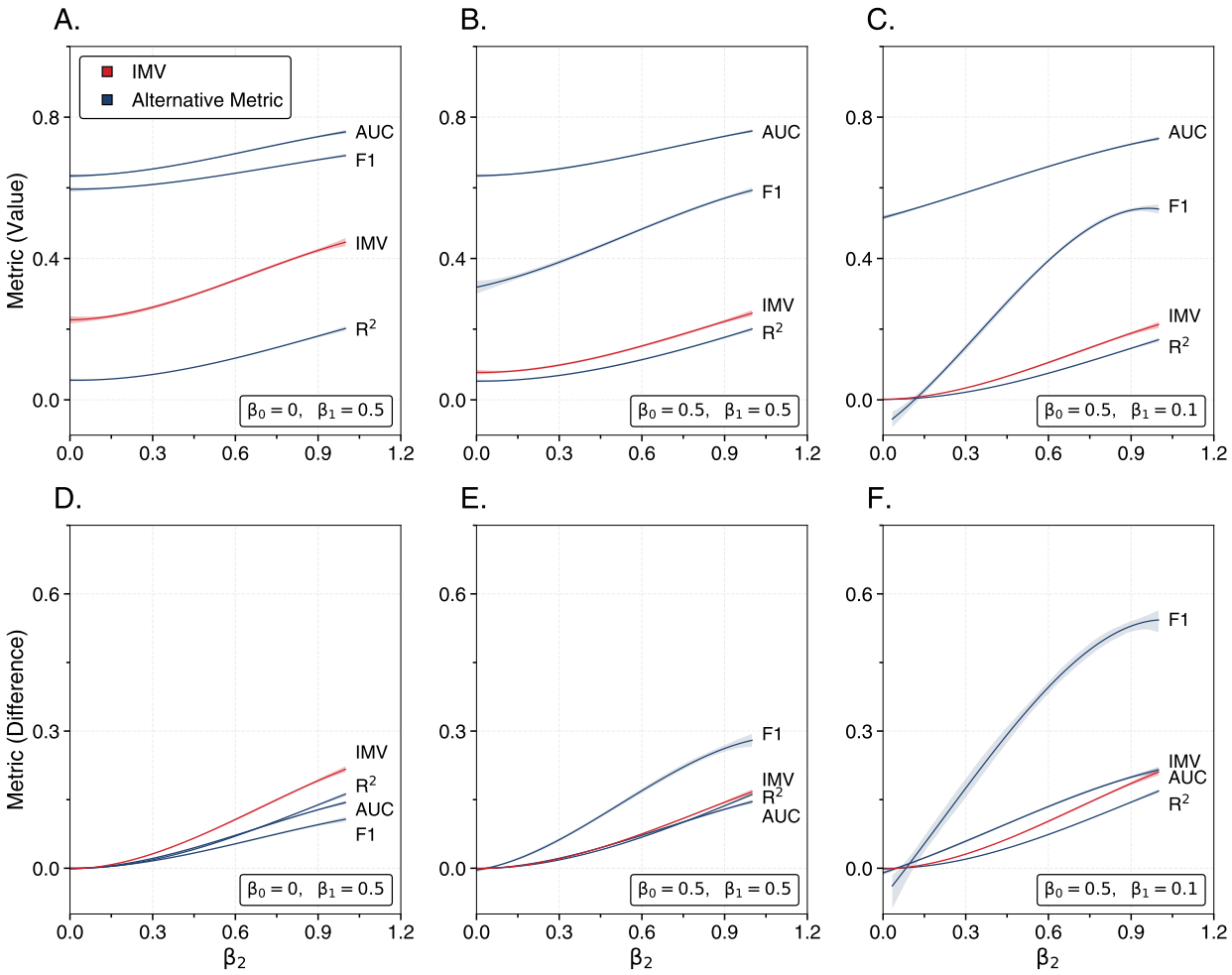


Figure S5: Patterning of fit metrics as a function of β_2 for prediction based on x and z. Curves are third order polynomials fit to 1000 iterations of the simulation with 99% confidence intervals. Columns represent different values of β_0 and β_1. Top: Raw metrics based on prediction via x and z (IMV value compared to prediction based on prevalence). Bottom: Relative difference comparing prediction based on x and z to prediction based on x alone.



S4 Additional details on core empirical illustrations

S4.1 HRS

The US-based Health and Retirement Study (HRS, [13, 14]) is a biennial longitudinal study of US adults over 50. The HRS collects data on both social status as well as health. We use data from the RAND Corporation's 2016 HRS data release.[4] We focus on observations on those aged 60 and over. Across all waves, there are 100,243 such observations. Sample sizes varied by age and outcome. We only compute data in cases with at least 1000 respondents. Sample sizes ranged from 1188 to 15864 (mean of 7582). We use logistic regression to predict outcomes based on linear combinations of the indicated predictors. We focus on the prediction of health outcomes *at the next wave* within an age bin (those within 1.5 years of a focal age) based on predictors measured at the current wave (or that are time-invariant).

We focus on prediction of self-reported physical health measures. At each wave, respondents are asked if they have been diagnosed with any of the following: high blood pressure, diabetes, cancer, lung disease, heart problems, stroke, psychological problems, arthritis. We also ask whether the respondent died prior to the next wave or had information completed by a proxy respondent (which typically occurred due to substantial impairment of the respondent's physical or mental functioning). To predict these outcomes, we rely on demographic information (age, sex, race/ethnicity), years of education reported by the respondent (a time-invariant predictor given the age of the respondent), and two specialized measures that the HRS collects.

- Cognition: The HRS measure of cognitive functioning [15] is based on the TICS survey phone-based assessment of cognitive status [16]. We used the `cogtot` measure constructed by RAND [16] for respondents who did not respond via proxy. The measure summarize word recall (via immediate and delayed recall of a 10 word list) and mental status (serial 7s, backward counting, correct naming of objects, memory, and date).
- Grip strength & Gait speed: The HRS collects measures of grip strength and gait speed (details on measurement can be found in [17]). These are only available at every other wave for each respondent given the structure of the HRS interview (which alternates between in-person and phone-based data collection). Gait speed was only collected in those 65 and older.

Prevalences as a function of age are shown in Figure S6. The outcomes, Panel A, generally increase as a function of age although such increases may be relatively slight. The predictors, Panel B, show more dramatic changes. Gait speed decreases dramatically across age (resulting in increased time) while grip strength and the score on the cognitive test both decline across age. We first predict health outcomes based on demographics (sex and race) relative to prevalence alone (panel A), then add educational attainment as a predictor (panel B). In Panels C and D, we turn to predictions based on relatively expensive-to-collect pieces of health data: cognition and physical functioning (as proxied for by grip and gait).

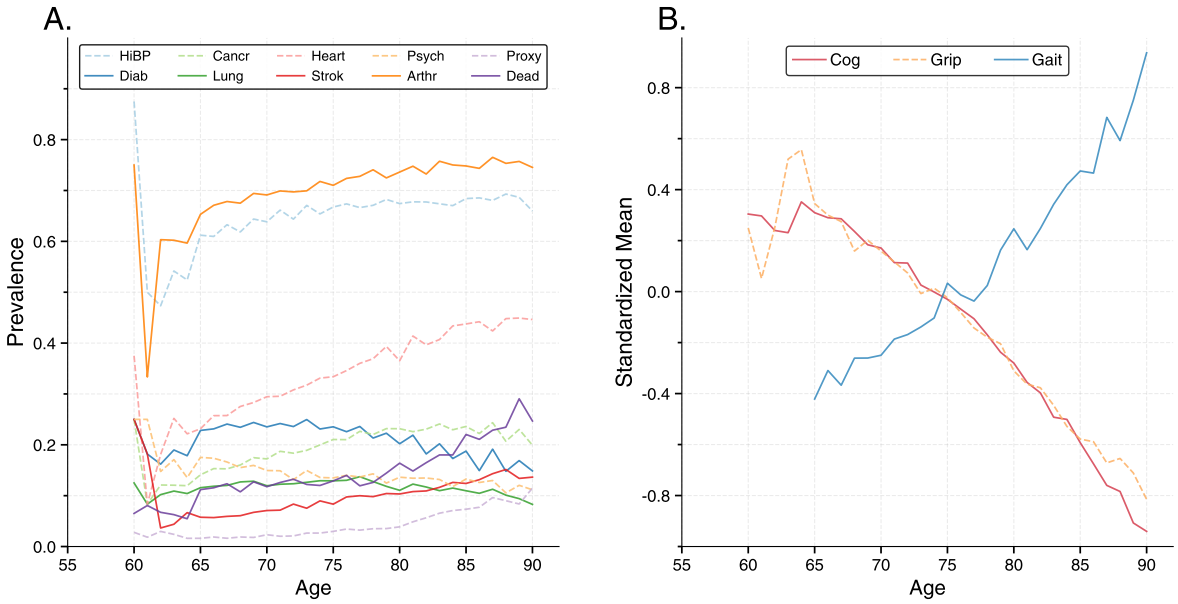
S4.2 The General Social Survey

The General Social Survey (GSS, [18]) is a survey of the social attitudes of Americans that has run since 1972.[5] In particular, it is widely used to understand the changing nature of political ideologies in the US. In total, there are 64,816 observations. Sample sizes varied across years; the smallest sample was 1193 in 1990 and the largest was 3385 in 2006. We focus on prediction of political party affiliation for partisans (the US has two major political parties). We consider independents who report leaning in one direction as partisans in that direction at that point in time (but exclude those who report no lean). Predictions are based on logistic regression models based on respondent demographics: age, sex, and whether the respondent identifies as white (as opposed to black or other) versus prevalence alone. The proportion of the sample classified as Republican is shown in the associated figure of the main text (bottom panel); note that comparisons of predictive accuracy over time will need to account for variation in this proportion.

[4] The data is publicly available, `https://www.rand.org/well-being/social-and-behavioral-policy/centers/aging/dataprod.html`; note that the grip and gait data come from the public release HRS files.

[5] The GSS data is publicly available, `https://gss.norc.org/get-the-data`. Data for this analysis can be found at `https://gssdataexplorer.norc.org/projects/84591`.

Figure S6: Age trends in HRS measures



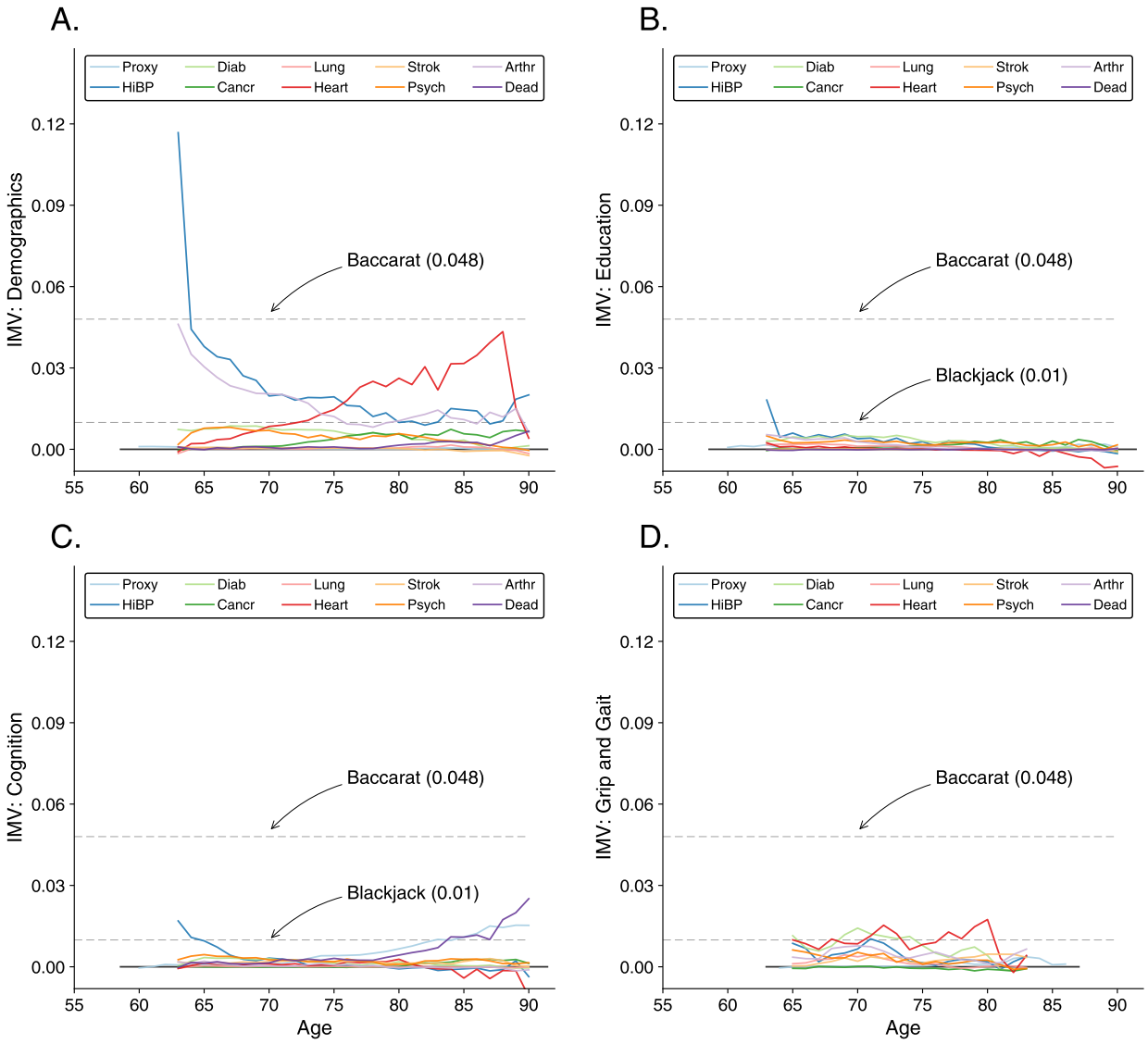
S4.3 The Fragile Families Children and Wellbeing Study

We utilize the replication materials—which include individual team's submissions—from the FFC itself [19] in conjunction with the freely available FFC data.[6] We then re-estimate their baseline models (as logistic regressions, as opposed to linear probability models), drop contributed submissions which took the form of class labels, and keep only those submissions which have a pseudo R^2 >0 (i.e. outperform the prediction made by a simple training mean),[7]. We then apply our IMV metric.

[6] We use the `FFChallenge_v5.zip` data, available from the from Princeton's Office of Population Research. Replicating our application of the IMV to the FFC data requires a simple and accessible registration at `https://pop.princeton.edu/`.

[7] This results in 95 submissions for 'layoff', 92 for 'jobTraining', and 88 for 'eviction'.

Figure S7: IMV for predicting health at next wave as a function of different sets of predictors. Panel A shows prediction based on sex and race relative to prevalence alone while Panel B shows prediction based on years of education added to race and sex. Panel C shows prediction based on cognition relative to race, sex, and education and Panel D the prediction based on grip and gait added to race, sex, and education.



S5 Additional empirical examples

S5.1 OECD's Programme for International Student Assessment (PISA)

PISA is an international assessment conducted every three years and aims to measure "the extent to which 15-year-old students, near the end of the compulsory education, have acquired key knowledge and skills"[20]. We use data from the students who took the 2015 PISA math exam.[8] We use a random sample of 10,000 15 year olds from the 2018 math assessment. We estimate item response theory (IRT) models via the EM algorithm and then obtain EAP ability estimates (both from [21]; we focus on EAP ability estimates and place priors on discrimination and guessing parameters). Predictive fit is based on the "missing response" paradigm [22]; that is, we hold item responses out-of-sample and attempt to predict them based on item- and person-level features derived from analysis of in-sample data.

PISA uses the two parameter logistic model (2PL) based on previous considerations of fit [23]; we compare the fit of the 2PL to fit based on the 3PL and the 2 factor 2PL. We first consider gains from more flexible models relative to the 2PL. For the 3PL, the IMV based on 10-fold CV was 0.0018. For the two factor 2PL, $\omega = -0.0007$ indicating that the two factor 2PL does a worse job with out-of-sample prediction than does the 2PL. Gains from these more flexible approaches as compared to the currently used 2PL are thus quite small. As one way of contextualizing this, going from the previously used 1PL model to the 2PL generated $\omega = 0.01$; an order of magnitude larger than going from the 2PL to the 3PL.

S5.2 Predicting family income from text data

To further illustrate the flexibility of our metric, we build off of recent work examining interplay between income and the kinds of themes identified by topic modeling in a large corpus of essays submitted as part of college applications [24]. In particular, we consider the IMV associated with prediction of whether a reported family income is above or below the median; in these data, the median family income across the 59723 respondents is \$70,000. Given the longstanding observation regarding associations between income and the SAT [25], we inquire about the degree to which the 70 themes predict this indicator net of the applicant's total SAT score. In these data, the correlation between the SAT score and the indicator is 0.43. Across 10-fold CV, the IMV associated with the topics is 0.073.

S5.3 Using the IMV to compare logistic regression coefficients

Here we offer an example of how the IMV can be used to generate additional insight into coefficients from logistic regression. Interpretation of logistic regression coefficients is complicated by the fact that coefficients (when exponentiated) describe changes in the odds ratio, rather than the probabilities, associated with a change in the covariate. This makes it relatively challenging to directly interpret or compare such regression coefficients [26, 27].

When interest is in the coefficient for a given covariate, the basic approach is to fit logistic regression models with and without that covariate—playing the role of the enhanced and baseline model, respectiely—and to then compute the IMV based on the resulting log-likelihoods.[9] To illustrate this point, we compare the role of sex in predicting two outcomes. We first consider the degree to which sex predicts death amongst Titanic passengers.[10] The majority of passengers (65%) were male; of the 891 passengers, the majority (N=468) of the 549 who died were male; descriptive statistics shown in Table S1. We next consider the role of sex in predicting political affiliation, specifically a tilt towards the GOP in the GSS [18]. The GSS sample is far larger and split more evenly by sex. Note that the outcomes occur at roughly similar rates (62% for the Titanic and 59% for the GSS) in both cases thus suggesting that the outcomes do not differ dramatically in the degree to which we might expect models to further explain these outcomes.

We first consider logistic regression models for each outcome after including other predictors. For the Titanic data, we use information on the ticket class, age, number of siblings/spouses onboard, number of parents/children onboard, and the port of embarkation alongside sex. For the GSS, we add age, marital status and race to sex. The baseline model without sex for death in the Titanic data may seem more compelling;

[8] PISA data is publicly available, `https://www.oecd.org/pisa/data/`.

[9] This can be done readily at `https://klint-kanopka.shinyapps.io/imv-app/`.

[10] Data from `https://www.kaggle.com/c/titanic`.

Table S1: Illustration of IMV as a means of understanding logistic regression coefficients.

	N	Proportion Female	Outcome Prevalence	β	ω_0	ω
Titanic	891	0.35	0.62	-2.64	0.29	1.36e-1
GSS	64816	0.56	0.59	0.21	0.14	2.39e-3

Table S2: Illustration of the IMV relative to other metrics in a Kaggle-style competition

	Logistic	SVC	Naive Bayes	LightGBM	Random Forest	Baseline
Log Loss	0.449	0.449	*0.625*	**0.434**	0.505	0.667
Accuracy	0.807	0.809	*0.768*	**0.837**	0.819	0.616
Brier Score Loss	0.142	0.141	*0.179*	**0.132**	0.138	0.237
F1	0.741	0.735	*0.721*	**0.772**	0.749	0
ROC-AUC	0.86	0.849	*0.843*	**0.865**	0.856	0.5
Pseudo-R^2	0.4	0.405	*0.243*	**0.442**	0.414	0
IMV	0.395	0.392	*0.146*	**0.41**	0.302	0

Bold entries indicate the best, and italicized indicates the worst performing models as assessed by each metric (excluding the 'Baseline' entry). The 'Baseline' predictions are simply a vector of the mean of the dependent variable in the training set (i.e. $\hat{y}_i = \bar{y}_{\text{train}} \forall i$) for reference. For the Log Loss and Brier Score Loss metrics, a lower score indicates superior performance.

we comment more on that below. Estimates (β) of association being female and the outcome are -2.64 (SE=0.22) in the Titanic data and 0.21 (SE=0.02) in the GSS. How to directly compare these coefficients in terms of the key question—how much does information about sex resolve uncertainty—is non-obvious.

We next consider IMV estimates based on 10-fold cross-validation. We first compute ω_0 which is the IMV for the model without sex compared to prediction based on outcome prevalence. Note that the baseline mode used to predict Titanic deaths is roughly double as effective as the model for predicting being in the GOP in the GSS; again, differences in the predictive value of sex aren't due to the fact that the baseline model for Titanic deaths is qualitatively weaker than the baseline model in the GSS. We then compute ω based on the logistic regression models with and without sex as a predictor. We can see that the ω value for sex in the Titanic case is roughly 57x the magnitude of sex in the GSS case. In our view, this quantity offers an unambiguous interpretation: in terms of literal value, having access to sex in the prediction of individual outcome (i.e., death) in the Titanic data is 57x as valuable as this same data on the GSS respondents for prediction of those outcomes (i.e., political affiliation).

S5.4 The IMV versus other metrics when used to benchmark performance of a variety of ML algorithms for predicting survival in the Titanic data

We also use this same Titanic dataset to show the potential for re-ordering of the rankings of different predictive systems when using the IMV in something akin to a Kaggle competition. Historically, the key choice of evaluation metrics has been a point of significant contention. Using the same dataset as in S5.3, we undertake some light additional feature engineering (such as creating age bins, cleaning honorifics, and so forth). We then evaluate the performance of four commonly utilized machine learning tools (a Support Vector Classifier, Guassian Naive Bayes, Light Gradient Boost, and the Random Forest) in addition to calculating the same metrics for probabilistic estimates of the training mean. We utilize stratified 10-fold cross validation, taking the mean of the folds for each metric. We evaluate our models with a range of both probabilistic and class based metrics; alongside the pseudo-R^2, ROC-AUC, and F1 metrics from above we also consider Accuracy, the Log Loss (which is simply $-(y_i log(\hat{y_i}) + (1-y_i) log(1-\hat{y_i}))$), and the Brier Score [9]. For the IMV, we focus on the value computed based on predictions from the model in question versus a baseline model which is simply the mean in the training data.

Results are shown in Table S2.[11] Across all metrics, the LightGBM approach consistently ranks as the best while the Guassian Naive Bayes ranks as the worst. Focusing on the IMV, note that the IMV suggests the random forest approach offers less predictive value than either logistic regression or the SVC (as does the log loss). In contrast, the Accuracy, F1, and the AUC-ROC all rank the random forest as superior to those two. The difference originates from the nature of *label* based evaluatory metrics, as opposed to *probabalistic* ones. The Random Forest predicts a comparatively larger proportion in tails; while these more often represent accurate class labels (when converted under any reasonable threshold), their distance from an incorrect prediction in what is an essentially well-balanced ($\bar{y} = 0.3838$) problem. While we do not formally enter the debate around which metrics should (or should not) be used for evaluating predictive competitions, we do imagine that the cross-problem portability of the IMV makes it a strong and compelling contender for consistent use in problems of this type. This is due to its ability to directly and readily compare performance across all similar types of (binary prediction) problems on one platform, across multiple different platforms, or even beyond.

S5.5 Predicting whether home team wins football in professional European competitions

We consider the year wherein the IMV is highest using data and predictions from [28]. Across all leagues, the highest IMV ($\omega = 0.244$) was observed in 2017 when 61% of home teams won their matches. The general increase in the model's power for predicting the winner is seen in the increase in the IMV over time. As points of contrast, we also consider results for England and the Netherlands. The results for England show a similar pattern while, as in [28], the results for the Netherlands show less change over the period observed here.

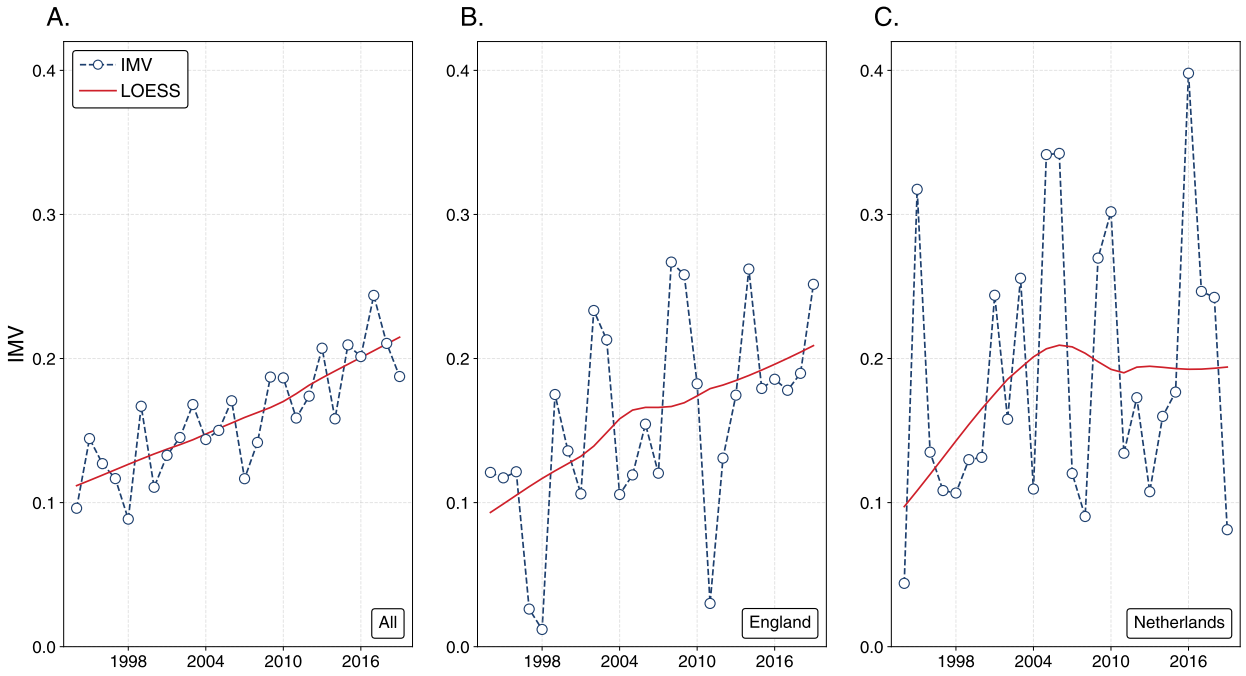
S5.6 Raw prediction examples

To offer additional benchmarks for interpretation of IMVs, we consider a range of raw predictions (meaning we are merely interested in the quality of prediction—which the IMV facilitates—rather than relative improvement in prediction) using a variety of different empirical datasets. Unless otherwise specified, we are computing the IMV based on 10-fold CV using logistic regressions (where covariates are included additively) as compared to prediction based on prevalence in the out-of-fold data. Additional documentation can be found at `https://github.com/ben-domingue/binary-prediction`.

- Prediction of whether a given sample of glass was produced via a "float" process (useful for forensic identification of samples) in 214 samples [29]. Prediction is based on the sample's refractive index and results of chemical assays and produce $\omega = 0.420$ compared to prediction based on prevalence alone.
- Prediction of whether the number of rings in a sample of 4176 abalone is above or below the median [30]. Prediction is based on a variety of anthropometric characteristics (e.g., diameter, weight) and sex. The IMV was $\omega = 0.667$.
- Early prediction of diabetes using patient reported symptoms (e.g., itching, age, sex, obesity, etc.) from a survey of 520 respondents [31]. The IMV was $\omega = 0.617$.
- Diagnosis of breast tissue as malignant or benign based on a subset of imaging characteristics in 569 tissue samples [32]. The IMV was $\omega = 0.526$.
- Excess alcohol consumption (dichotmization as per the rule in [33]) based on 6 blood tests from 345 blood samples [34]. The IMV was $\omega = 0.245$.
- Identification of whether a pixel is a skin tone or not based on B,G,R values [35]; prediction based on both main effects and all possible interactions between these three predictors in 245057 pixels. The IMV was $\omega = 0.196$.

[11] Note that if these accuracy scores were obtained for the truly unobserved test set in the competition, several of our algorithms would generate predictions in the top 5% of submitted predictions.

Figure S8: Prediction of whether home team wins using model from [28]. Gray dots represent points from individual years while the red lines represent LOESS fits.



- Diagnosis of heart disease (based on angiographic disease status) in a clinical sample of 302 patients based on 14 predictors [36]. The IMV was $\omega = 0.123$.
- Hospital readmissions for patients with Diabetes Mellitus (DM) based on a subset of their medical record from a previous hospital visit (e.g., number lab procedures, number medications) in 101766 hospital visits [37]. The IMV was $\omega = 0.196$.
- Lithofacies class (out of 9 classes) amongst 3232 well samples using well-log data [38]. We focused on predictions of each class versus all others and each class versus adjacent classes (see Table 2 in [38]). When predicting classes versus all alternatives, the values ranged from $\omega = 0.163$ for predicting Nonmarine coarse siltstone and $\omega = 0.016$ for predicting Dolomite (which was the rarest of the classes, only occurring in 3% of the well samples). Predicting Dolomite versus the adjacent facies classes of Wackestone or Packstone-grainstone was similarly the weakest of these predictions with $\omega = 0.049$. Predicting Marine siltstone and shale versus the adjacent class of Mudstone was the largest IMV with $\omega = 0.446$.

References

[1] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[2] Ronald A Howard. Decision analysis: Perspectives on inference, decision, and experimentation. *Proceedings of the IEEE*, 58(5):632–643, 1970.

[3] Corey A Shank. Is the nfl betting market still inefficient? *Journal of Economics and Finance*, 42(4):818–827, 2018.

[4] Jon P Morosky. Method of administering and playing a baccarat type card game, May 23 2000. US Patent 6,065,753.

[5] Kurt Lofink and Richard Lofink. Blackjack game with modifiable vigorish, April 17 2001. US Patent 6,217,024.

[6] John L Kelly Jr. A new interpretation of information rate. *The Bell System Technical Journal*, 34(4), 1956.

[7] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[8] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.

[9] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[10] Jin Huang and C.X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[11] Alex P Zijdenbos, Benoit M Dawant, Richard A Margolin, and Andrew C Palmer. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE transactions on medical imaging*, 13(4):716–724, 1994.

[12] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *IEEE Access*, 2021.

[13] F Thomas Juster and Richard Suzman. An overview of the health and retirement study. *Journal of Human Resources*, pages S7–S56, 1995.

[14] Amanda Sonnega, Jessica D Faul, Mary Beth Ofstedal, Kenneth M Langa, John WR Phillips, and David R Weir. Cohort profile: the health and retirement study (hrs). *International journal of epidemiology*, 43(2):576–585, 2014.

[15] Mary Beth Ofstedal, Gwenith G Fisher, A Regula Herzog, et al. Documentation of cognitive functioning measures in the health and retirement study. *Ann Arbor, MI: University of Michigan*, 10, 2005.

[16] Delia Bugliari, Nancy Campbell, Chris Chan, Orla Hayden, Michael Hurd, Regan Main, Joshua Mallett, Colleen McCullough, Erik Meijer, Michael Moldoff, et al. Rand hrs data documentation, version p. *RAND Center for the Study of Aging*, 2016.

[17] Eileen Crimmins, Heidi Guyer, K Langa, Mary Beth Ofstedal, Robert Wallace, and D Weir. Documentation of physical measures, anthropometrics and blood pressure in the health and retirement study. *HRS Documentation Report DR-011*, 14(1-2):47–59, 2008.

[18] James A Davis and Tom W Smith. *The NORC general social survey: A user's guide*, volume 1. SAGE publications, 1991.

[19] Matthew Salganik, Ian Lundberg, Alex Kindel, and Sara McLanahan. Replication materials for "Measuring the predictability of life outcomes using a scientific mass collaboration", 2020.

[20] OECD Pisa. Pisa: Results in focus. *Organisation for Economic Co-operation and Development: OECD*, 2015.

[21] R Philip Chalmers et al. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48(6):1–29, 2012.

[22] Ben Stenhaug. *MODEL SELECTION METHODS FOR ITEM RESPONSE MODELS*. PhD thesis, Stanford University, 2021.

[23] Maria Elena Oliveri and Matthias von Davier. Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, 53(3):315, 2011.

[24] AJ Alvero, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L Stevens, and Benjamin W Domingue. Essay content and style are strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. *Science advances*, 7(42):eabi9031, 2021.

[25] Rebecca Zwick and Jennifer Greif Green. New perspectives on the correlation of sat scores, high school grades, and socioeconomic factors. *Journal of Educational Measurement*, 44(1):23–45, 2007.

[26] Richard Breen, Kristian Bernt Karlson, and Anders Holm. Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, 44:39–54, 2018.

[27] Max A Halvorson, Connor J McCabe, Dale S Kim, Xiaolin Cao, and Kevin M King. Making sense of some odd ratios: A tutorial and improvements to present practices in reporting and visualizing quantities of interest for binary and count outcome models. *Psychology of Addictive Behaviors*, 2021.

[28] Victor Martins Maimone and Taha Yasseri. Football is becoming more predictable; network analysis of 88 thousand matches in 11 major leagues. *Royal Society Open Science*, 8(12):210617, 2021.

[29] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[30] Samuel George Waugh. *Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks*. PhD thesis, University of Tasmania, 1995.

[31] MM Faniqul Islam, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*, pages 113–125. Springer, 2020.

[32] Kristin P Bennett and Olvi L Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1(1):23–34, 1992.

[33] Peter D Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of artificial intelligence research*, 2:369–409, 1994.

[34] James McDermott and Richard S Forsyth. Diagnosing a disorder in a classification benchmark. *Pattern Recognition Letters*, 73:41–43, 2016.

[35] Rajen B Bhatt, Gaurav Sharma, Abhinav Dhall, and Santanu Chaudhury. Efficient skin region segmentation using low complexity fuzzy decision tree model. In *2009 Annual IEEE India Conference*, pages 1–4. IEEE, 2009.

[36] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

[37] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[38] Brendon Hall. Facies classification using machine learning. *The Leading Edge*, 35(10):906–909, 2016.