Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Phenotyping part of the target family improves within-family genomic prediction in rice nested association mapping families

Jixiao Li

yuriljx@gmail.com

Hokkaido University

Research Article

Keywords: Oryza sativa, genomic prediction, genomic selection, within-family prediction, training population design, partial phenotyping, target-family phenotyping, nested association mapping, ridge regression, Gaussian kernel

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11533831/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Genomic prediction uses DNA markers to help rice breeders choose promising lines before costly yield trials. A new breeding family usually lacks its own yield records, so its lines are predicted from other families, often less accurately. With limited field plots, breeders can measure yield in some lines of the family being selected, called the target family, or adopt a more flexible model, which needs no extra plots. It is unclear which helps more, whether they combine, and whether better prediction selects higher-yielding lines. Here, two-site data for 15 nested association mapping families sharing one parent were reanalyzed with 2000 training records per target family, so gains reflect which lines were measured, not how many. In a four-strategy comparison added after the original analysis, replacing 100 other-family records with records from 5–15% of the target family raised average within-family predictive ability, the correlation between predicted and recorded yield of held-out lines, from 0.16–0.20 to 0.19–0.26 for both linear ridge regression and a nonlinear Gaussian-kernel model, with gains in 11–14 of 15 families. Switching to the Gaussian model alone, tuned for selected-line yield, changed it by only 0.005–0.011, although the two gains did not differ clearly after multiple-comparison correction; combining both gave the highest predictive ability. At 10% selection, lines chosen with target-family records yielded more on average, with 95% intervals above zero only at Shanghai; the originally planned comparison of records versus model was positive but not conclusive. In a multiparent population predicted from these families, its own records lowered selected-line yield at one site, with mixed effects at the other. In an elite International Rice Research Institute cohort, adding 25 current-season records raised average predictive ability of both models in every season, while selected-line yield varied. In these related families, phenotyping part of a new family before ranking the rest improved within-family prediction for both models. The results support using ridge regression trained with target-family records as the benchmark for flexible models, with strategies compared by both predictive ability and selected-line yield; predictions and code are publicly released.

Introduction

Each breeding cycle, rice breeders must decide which lines from new crosses to advance, often before those lines can be tested for yield. Better decisions matter: rice is the staple food of more than half of the world's population, and continued yield gains through breeding remain essential for food security (Seck et al. 2012). Because yield trials are slow and expensive, only a limited number of lines can be measured (Akdemir and Isidro-Sánchez 2019). Genomic prediction helps to ease this bottleneck. A statistical model learns how DNA markers relate to yield in lines that have both marker data and yield records, known as the training population, and then predicts yield for lines that have only been genotyped. Selecting lines on the basis of such predictions is called genomic selection (Meuwissen et al. 2001; Spindel et al. 2015).

Predictions are usually most reliable when the training lines are close relatives of the lines being ranked, ideally full siblings from the same cross. Different crosses segregate for different gene variants, so marker–yield associations learned in other families may not hold in a new one, and the genetic

relationship between training and target lines strongly affects prediction (Habier et al. 2007; Riedelsheimer et al. 2013). Yet when the lines of a new family are first ranked, most available yield records usually come from other families. Breeders then have two broad options for improving predictions for the new family, here called the target family. They can measure yield in some of its lines and add these target-family records to the training data; if the number of field plots is fixed, this means measuring fewer lines from other families. Alternatively, they can keep training on other families and switch to a more flexible model, such as a Gaussian-kernel model, which predicts a line from its overall genetic similarity to tested lines and can capture interactions among genes (Gianola and van Kaam 2008; Jiang and Reif 2015); this requires no extra field work. The two options can also be combined.

The value of records from close relatives is well established. Full-sib records predict a family better than records from half-sib or unrelated families (Brauner et al. 2020; Zhu et al. 2021), and small training populations of closely related lines have outperformed larger general ones in some maize crosses (Brandariz and Bernardo 2019). Phenotyping part of each new population to predict the rest is already practiced, for example in the test-half-predict-half approach of maize breeding, where spreading phenotyping across several connected families, a few lines each, improved prediction over within-family training (Atanda et al. 2021). Sparse-testing designs similarly grow only some genotypes in each environment and predict the rest (Jarquin et al. 2020). How to choose or compose the training lines has also been studied in maize, wheat, rice and other crops (Rincent et al. 2012; Lehermeier et al. 2014; Isidro et al. 2015; Akdemir and Isidro-Sánchez 2019; de Verdal et al. 2023; Fernández-González et al. 2023; He et al. 2023). Evidence for more flexible models is less consistent. Linear methods such as ridge regression and the closely related genomic best linear unbiased prediction (GBLUP) remain the standard benchmarks (VanRaden 2008; de los Campos et al. 2009; Endelman 2011). Neural-network and transfer-learning approaches continue to be developed (Li et al. 2024; Montesinos-López et al. 2025a, b; Li et al. 2026), but greater complexity does not consistently improve prediction (Azodi et al. 2019; Nguyen et al. 2023; Montiel et al. 2025; Kusmec et al. 2026). Some studies have varied both the training data and the model in the same dataset (Spindel et al. 2015; He et al. 2023).

Few studies, however, have measured how much within-family prediction benefits from devoting part of a fixed number of yield records to the target family, compared with switching to a more flexible model, whether the two gains add up, and whether either improves the yield of the lines actually selected. The answer matters because it tells breeders whether a new family is better served by some of their limited field plots or by a different model.

How predictions are judged also matters. Correlation describes how closely predictions track differences in recorded yield: zero indicates no linear association, and one indicates a perfect positive linear relationship. When lines from several families are pooled, much of this correlation can come from differences between families, which inflates apparent accuracy without improving the ranking of siblings within a family (Werner et al. 2020). Within-family correlations are therefore typically modest, but they measure the breeder's actual task of choosing among siblings. Moreover, breeders keep only the best few lines, and a model can follow yield differences across all candidates more closely without selecting

a higher-yielding group. Ranking-based objectives and training-set designs aimed at identifying superior genotypes have therefore been proposed (Blondel et al. 2015; Chen et al. 2024; Liu and Liao 2026), and measures that score the selected lines complement correlation rather than replace it (Wang and Yan 2026). This study therefore reports two measures. Predictive ability (r) is the Pearson correlation between predicted and recorded yield among the evaluated lines, computed within each family in the main analysis. The selected-line yield advantage (U) is the amount by which the average recorded yield of the top-ranked lines, such as the top 10%, exceeds the average of all lines evaluated with them. Neither measure is an estimate of genetic gain in the next generation, which also depends on breeding-value accuracy, genetic variation, selection intensity and cycle length (Cobb et al. 2019).

This retrospective study used the public 18K-rice nested association mapping (NAM) population, in which each of 15 families of recombinant inbred lines descends from a cross between the elite indica cultivar Huanghuazhan and a different, genetically diverse accession (Wei et al. 2024). Lines within a family are full siblings, and the families are related as half-sib families through their common parent. They therefore serve as a model of related families in a breeding program, although they are larger and more diverse than many elite crosses; each family has roughly 650–2100 lines with Shanghai yield records. Each family was treated in turn as a new family: its yield records were used neither to train the models nor to tune their settings, except that strategies with target-family records added the records of 100 randomly chosen lines to the final model fit. This differs from designs in which every family contributes training lines (Kitony et al. 2026). With the total held at 2000 yield records, four strategies were compared at two field sites: ridge regression or a Gaussian-kernel model, each trained with or without 100 target-family records. All four were scored on the same evaluation lines, about half of each family, which were never used for training or tuning. Three strategies belonged to the original analysis, which was designed around the yield of selected lines; the Gaussian model with target-family records and all correlation summaries were added later.

To test how far the pattern extends, the same question was examined when NAM records were used to predict a different, multiparent advanced generation intercross (MAGIC) population (Wei et al. 2024), and in an elite breeding cohort from the International Rice Research Institute (IRRI; Spindel et al. 2015), where 25 current-season records were added to 626 historical records rather than replacing any. The comparison concerns the yield information available for prediction, not the monetary cost or time needed to obtain it.

The objectives were to test, with the number of yield records held fixed, whether a small number of target-family records improves within-family prediction more than switching to a more flexible model, whether the two approaches combine, and whether better prediction leads to higher-yielding selections. In brief, target-family records improved within-family prediction for both models, whereas switching to the Gaussian model alone changed it little. Combining target-family records with the Gaussian model gave the highest predictive ability, and the gain from these records did not clearly differ between the two models. Lines selected with target-family records yielded more on average at both sites, with 95% intervals above zero at Shanghai but not at Hangzhou. The study makes three contributions. First, it

combines the two options in a two-by-two design at a fixed number of yield records, so the value of a few target-family records is measured against, and together with, a more flexible model on the same evaluation lines. This extends maize evidence that phenotyping a few lines in each connected family improves prediction (Atanda et al. 2021) and soybean evidence that full-sib records outperform half-sib records (Zhu et al. 2021). Second, it judges both options by the recorded yield of the lines actually selected as well as by predictive ability. Third, it releases archived predictions, evaluation-line identities and scoring code for the public 18K-rice and IRRI data, so that others can re-score these comparisons and compare new models or phenotyping designs on the same evaluation lines (Li 2026).

Materials and methods

Study design overview

Each of the 15 NAM families was treated in turn as the target family (Fig. 1; Table 1); its yield records were never used for tuning and entered model fitting only as the 25 or 100 target-family records of the strategies that used them. About half of each target family's lines formed a fixed scoring set (S) used only for evaluation. The remaining lines formed the adaptation pool (A); at the Hangzhou (HZ) site, this pool comprised the HZ-eligible lines outside the Shanghai (SH) scoring sets. Target-family records came from 100 lines of A, taken in a pseudo-random order fixed in advance by a hash of target family, draw and line identifier, and these lines formed the calibration set (C). Each main strategy used 2000 yield records: 2000 from the other 14 families, or 1900 such records plus the 100 target-family records; in both cases, 700 of the other-family records came from the two families used to tune settings. Ridge regression and a Gaussian-kernel model were each fitted without and with target-family records, giving four strategies: source ridge0, source Gaussian0, pooled ridge100 and calibrated Gaussian100. Their archive codes are ridge0, RKHS0, pool100 and RKHS100, where RKHS denotes reproducing kernel Hilbert space, and Fig. 2 labels them R0, G0, R100 and G100. The number is the count of target-family records, and all training records were weighted equally. Every strategy was tuned on other families only. Within each site, all four were compared on the same scoring lines by r and U over five repeated draws of calibration lines and source records; at HZ, the scoring lines were the SH scoring lines with HZ records. The original analysis comprised the first three strategies, target-mass-weighted ridge with 25 or 100 target-family records and an exploratory target-only comparator; its primary endpoint was U at 10% selection for weighted ridge100 versus source Gaussian0. Calibrated Gaussian100, all r summaries, the 5% and 20% selection fractions and the direct comparison D were added post hoc. MAGIC and IRRI followed the same logic with the differences described below; in IRRI, 25 current-season records were added to 626 historical records.

Table 1
Deployment designs and information accounting

Deployment	Source and C yield records	Scoring domain	Development criterion	Analysis role
NAM	Source2000 + C0, source1975 + C25 or source1900 + C100; the 2000-record total includes 700 development yield records	15 families × 5 draws; SH 7793, HZ 7465 total S	Each arm: equal source pseudo-family U	Original comparisons; later four-arm completion
MAGIC	Six principal strategies with 2000 records; C0/25/100 replaces source records	A = S=1097; q = 110; five overlapping draws	Each arm: fixed NAM pseudo-family U	Original transfer; later matched-λ/domain checks
IRRI original	626 source observations + 25 current-season yield records	One C25 allocation; S = 96/96/94/96; q = 10	Source-only r; calibrated arms inherit	2011DS primary; correlated follow-ups
IRRI fixed20	Same source+current-season C25	20 overlapping C/S configurations per season	Original settings retained	Post-hoc partition sensitivity
IRRI adaptive20	Same source+current-season C25	Same 20 C/S configurations	Each deployment: historical pseudo-target U	Separate post-hoc complete-policy sensitivity

Budgets count material-site/season yield records, not monetary costs. Target-only100 in NAM/MAGIC has a separate 800-record information budget. Actual S yield records do not select parameters. No earlier target season is accumulated in IRRI. NAM names and analysis codes are source ridge0 (ridge0), source Gaussian0 (RKHS0), pooled ridge100 (pool100) and calibrated Gaussian100 (RKHS100); corresponding IRRI calibrated arms use 25 target yield records. Restricted SH denotes SH predictions scored on the HZ-valid intersection.

Resources and information budgets

The NAM families are recombinant inbred lines derived from Huanghuazhan crossed with 15 other accessions and advanced by seven rounds of single-seed descent (Wei et al. 2024). The source experiment used single-row plots of seven plants in an augmented block design, with each family blocked separately and Huanghuazhan included as a check (Wei et al. 2024). The Wei resource contains 16019 NAM materials in 15 families and 2402 MAGIC crosswalk identities (Chen 2024; Wei et al. 2024). Family identities were obtained from Rice Recombination Site Information version 1 (Chen 2026); family origins and record counts are given in Additional file 1: Table S1. Finite grain-yield counts were 15592 at SH and 15336 at HZ; MAGIC had 2194 jointly finite records. The historical 2048-marker panel was retained. Its phenotype-independent quality control used all NAM genotypes, including later target

families; each fitted task subsequently transformed markers using its source core. Grain yield is interpreted as yield per plant, as supported by the primary paper; the column headers of Wei et al. (2024) Data S4 state grams (g). The field year and any aggregation or adjustment behind its Shanghai and Hangzhou yield columns, K and Z, remain unresolved (Additional file 1: Section S1).

IRRI data derive from the Spindel archive (Spindel et al. 2015, 2016a, b). Source-defined qualification retained 201 source identifiers (IDs) and 121 genotype-eligible new-entry target IDs. Source filtering retained 64208 markers, with 64205 active features after transformation. Each IRRI yield record is the mean of at least two observations with distinct replicate identifiers in one season, following the qualification rules in Additional file 1: Section S1; these means, expressed in kg/ha, are not best linear unbiased estimates (BLUEs). The archive is separate from Wei and also underlies the TrG2P Rice299 dataset (Li et al. 2024).

A yield record denotes a material-by-site observation in NAM/MAGIC or a material-by-season replicate mean in IRRI; figure keys and archived code call these records labels. Source refers to the non-target records available for model training and development: other NAM families in NAM/MAGIC tasks and historical trials in IRRI. Calibration includes a limited set of target-population yield records in fitting; pooling combines them with source records. In NAM/MAGIC, target records replace source records within a total of 2000 yield records, including those used for tuning, so the main strategies differ in which lines supply records, not in how many. IRRI instead retains 626 historical observations and adds 25 current-season records. Each NAM comparison concerns one target family at a time; allocation of a shared field budget across several new families was not optimized. Detailed designs and phenotype definitions are provided in Additional file 1; full-precision results are provided as a workbook in Additional file 2 and as canonical comma-separated values (CSV) tables in Additional file 3.

NAM and MAGIC deployments

At SH, a deterministic ID-hash rule assigned floor(N/2) of the N eligible materials in each NAM family to a permanent scoring set S, totaling 7793. The remaining materials formed adaptation pool A. Five paired configurations shared S and used nested C25/C100 calibration sets (C); source/core samples and the pair of stand-in target families also varied across configurations. The suffix denotes the number of calibration yield records. In each configuration, the lines of A were ordered by a hash of target family, draw and line identifier fixed in advance, and C25 and C100 were the first 25 and 100 lines of that order. At HZ, C was taken in the same way from HZ-eligible lines outside the SH scoring sets, so each calibrated strategy had its full 25 or 100 target-family records at both sites. HZ used the SH scoring IDs intersected with HZ eligibility, totaling 7465. A restricted SH view reranked the same SH predictions on that intersection, without refitting; it is a paired-domain sensitivity rather than a third site.

Each outer NAM task excluded its target-family yield records from source fitting and tuning. For tuning, two source families acted as stand-in target families, called pseudo-targets, each supplying 100 pseudo-calibration and 250 pseudo-validation yield records; the other 12 supplied a 1200-record core. The 700

development yield records are included in the common 1900-record bank. Final source allocations of 2000, 1975 or 1900 accompany C0, C25 or C100, respectively. Each strategy selected its own hyperparameters by equal mean pseudo-family top-10% U. Actual target S yield records did not enter this choice. A family's yield records could serve as source records in a different outer task.

Original NAM strategies comprised source ridge, source Gaussian, target-mass-weighted ridge with C25 or C100, and pooled ridge with C100, together with the exploratory target-only100 comparator described below. The original primary contrast was weighted ridge100 minus source Gaussian0; pool100 minus ridge0 was a checking contrast. A later, post-hoc completion added calibrated Gaussian100 under the same 2000-record budget and source-family pseudo-target tuning, yielding four cells: source ridge0, source Gaussian0, pooled ridge100 and calibrated Gaussian100. Like pooled ridge100, calibrated Gaussian100 weighted all 2000 training records equally; the corresponding Gaussian strategies with target records are called pooled Gaussian100 in MAGIC and pooled Gaussian25 in IRRI. This completion reused the original predictions after technical verification and added the missing Gaussian fits (Additional file 1).

The MAGIC population derives from (Huanghuazhan × Kasalath) × (Nipponbare × IAC25), followed by selfing (Wei et al. 2024). MAGIC used permanent adaptation and scoring halves of 1097 jointly eligible materials, with q = 110 selected lines (10%). Five materials in a high-concordance genotype cluster were kept together in A under the predefined group rule. Five overlapping C100 draws, with nested C25 prefixes, shared the same S across both sites. With NAM families labeled N1–N15, pseudo-target families N12 and N10 and a 1200-record core from the other 13 families supplied source development. Seven original arms included pooled Gaussian100 as well as the linear and weighted strategies. Its primary remained weighted ridge100 minus source Gaussian0. In NAM and MAGIC, target-only100 was a separately labeled exploratory comparator using 700 development yield records plus C100, an 800-record budget; it was not an equal-budget test of removing source phenotypes.

Prediction models and regularization

NAM/MAGIC markers use source-core mean imputation, population-standard-deviation (SD) scaling and removal of degenerate features. With standardized marker matrix Z and p retained markers, the linear kernel is ZZ'/p. Here Z_i denotes row i of Z. Gaussian distances and the kernel are:

$$d^2(i,j) = \frac{\|\varvec{Z}_i - \varvec{Z}_j\|^2}{p}$$

$$K(i,j) = \exp\left[-\frac{d^2(i,j)}{2b^2}\right]$$

Squared bandwidth b² is 0.5, 1 or 2 times the median non-diagonal source-core distance. An unpenalized intercept and normalized weighted loss are common:

$$\sum_i w_i \left[y_i - a - \mathrm{f}\left(\backslash \mathrm{varvec} Z_i \right) \right]^2 + \lambda \left\| f \right\|_K^2$$

Weights sum to one. Uniform pooling uses equal row weights; target-mass weighting gives each C row η/k and each source row (1 − η)/n_source, with η = 0.05/0.20/0.50. The grid for the regularization penalty λ is 0.0001/0.001/0.01/0.1. At source1900 + C100, η = 0.05 equals uniform final pooling, although different pseudo-target fit compositions can select different λ values.

The post-hoc NAM fixed-grid comparison evaluated source ridge0 and pool100 at all four λ values, retaining their original source banks, transformation, scoring sets and selection rule. Source ridge used the common 1900 plus 100 extra source yield records; pooling replaced those 100 with C100. Both had 2000 equal-weight yield records. Every setting was retained; S did not select a best λ. Under a specific equal-residual-variance relationship-kernel model, nλ corresponds to a variance-ratio convention, giving 0.2/2/20/200 for n = 2000. These are not heritability estimates (Additional file 1).

IRRI seasonal calibration and tuning policies

Historical 2009/2010 dry-season (DS) and wet-season (WS) trials supplied 626 observations from 201 IDs. The target cohort first appeared in 2011 DS and was absent from qualified historical source records. The original deterministic C25 set supplied current-season yield records in 2011 DS, 2011 WS, 2012 DS and 2012 WS. Seasonal S sizes were 96/96/94/96, with q = 10 selected lines at the original selection fraction. Each calibrated fit added 25 current-season observations to all 626 source observations; earlier target seasons were not accumulated. This evaluates the specified seasonal update rather than a cumulative breeding program.

The five IRRI arms were source ridge0, source Gaussian0, pooled ridge25, target-similarity-weighted source ridge25 and pooled Gaussian25. Source-unique genotypes define transforms and kernel scales. Full season-indicator columns are jointly unpenalized, without an additional global intercept. Source-only unseen-season indicators are zero; a calibrated target-season coefficient is supported by C only. The Gaussian kernel is:

$$K = \exp\left(-\frac{\gamma\, d^2}{\mathrm{median}_{\mathrm{source}\backslash:\mathrm{pairs}}\left(d^2 \right)} \right)$$

Target-similarity weighting uses source-genotype distance to the C centroid, with source weights exp(−d²/scale), clipped to [0.5, 2] and normalized to source-row mean one; C weights are one. This differs from NAM/MAGIC target-mass weighting. The loss is weighted squared residual divided by total weight plus the kernel penalty.

The original IRRI analysis and later fixed policy inherited source-only parameters selected by equal mean Pearson r in two historical wet-season pseudo-target batches. After reserving pseudo-C25, their

validation sets contained 24 and 15 IDs. The linear grid was λ = 0.01/0.1/1; Gaussian candidates crossed these values with γ = 0.5/1/2. Calibrated arms inherited the selected source ridge (λ = 0.01) or source Gaussian (λ = 1, γ = 2) setting. Pseudo-C yield records were reserved rather than fitted during this original development stage.

The post-hoc partition sensitivity used 20 deterministic C25 sets drawn from the retrospective intersection of 119 four-season-complete target IDs. Seasonal S retained all eligible non-C IDs, including the two IDs unavailable in 2012 DS in their other eligible seasons. Fixed policy required 242 final fits and no new tuning. A separately specified adaptive policy retuned each of the five arms under its own calibrated or source-only deployment in the two historical batches, maximizing equal mean pseudo-target U. It added 54 inner and 242 final fits. One setting per arm was locked for all 20 configurations and seasons before their new outer scores were revealed; no target-S tuning occurred. This policy changes the selection criterion and calibrated-arm reoptimization together. Complete grids and tie rules are in Additional file 1.

Endpoints and summaries

For a scoring set S of N_S lines and selection fraction α, the $q_\alpha = \text{ceiling}(\alpha \times N_S)$ highest predictions were selected. The original endpoint uses α = 0.10. Exact prediction ties use the frozen ID-hash rule. U is the selected mean recorded yield minus the mean of S; r is Pearson correlation between predictions and recorded yield within the same S. For NAM this is within-family, within-site predictive ability with recorded phenotypes; across designs it is not heritability-corrected accuracy against unobserved breeding values. NAM averages five paired draw contrasts within family and then weights 15 family means equally. MAGIC draw means and ranges describe five overlapping allocations on common S. IRRI summaries describe its finite overlapping partition grid, separately by season and policy. Supplementary figures for the NAM, MAGIC and IRRI comparisons are in Additional file 1: Figures S1–S11.

Frozen-prediction sensitivity analyses

Frozen predictions were evaluated at selection fractions of 5%, 10% and 20%, retaining $q = \text{ceiling}(\alpha \times N_S)$, the original ID-hash tie rule, S membership and parameters. NAM and adaptive IRRI retained settings selected using pseudo-target top-10%-U; fixed IRRI retained its r-oriented settings. No fraction-specific retuning was performed. Spearman correlation used average tied ranks over all S; constant inputs would remain undefined, with definedness counts retained. Pearson and Spearman therefore repeat the same whole-S result across fractions and are not three independent checks.

Alongside the four-strategy analysis, the original weighted ridge100 − source Gaussian0 comparison was rescored at the same fractions and with Spearman correlation; its 10% U estimate and original intervals remain the primary result. Between-family SDs and change-versus-midpoint correlations were also calculated from frozen five-draw family means. These are post-hoc dispersion descriptors (Additional file 1: Section S7).

The archived two-arm fixed-λ predictions were additionally rescored for r at all four penalties in all three NAM scoring views. The 150 original tasks supplied 1200 prediction vectors and 1800 arm-by-view scores; all 5400 numerical checks against the original 900 paired U rows reproduced exactly before new results were accepted. Five draw-specific r differences were averaged within family, followed by equal averaging over all 15 families. All 12 grid summaries and their positive/exact-zero/negative family counts are reported descriptively, without new intervals or p-values (table Grid_r_summary in Additional files 2 and 3).

Statistical analysis and software

The study is retrospective and was not publicly preregistered. The original internally frozen NAM/MAGIC primary comparisons used U for weighted ridge100 minus source Gaussian0; original IRRI used 2011 DS pooled ridge25 minus source Gaussian0. These comparisons followed earlier trait and model exploration. Later fixed-λ, MAGIC matched-λ/domain, interval-method, four-arm and partition/policy analyses are post-hoc extensions; their protocols preceded their own new computations but followed knowledge of earlier results. Companion r summaries, interval sensitivities, the direct strategy contrast and the frozen-prediction sensitivities above were added after inspection of earlier findings and are reported as post-hoc analyses alongside the original U primary endpoint.

The five four-arm contrasts are ΔL = pool100 − ridge0, ΔG = RKHS100 − RKHS0, K0 = RKHS0 − ridge0, K100 = RKHS100 − pool100, and I = ΔG − ΔL. K denotes a kernel contrast. For these NAM contrasts, ridge0 denotes source ridge, pool100 denotes pooled ridge100, and the archive codes RKHS0/RKHS100 denote source/calibrated Gaussian, respectively. The diagonal strategy contrast is D = pool100 − RKHS0 = ΔL − K0 = ΔG − K100 = (ΔL + ΔG − K0 − K100)/2. Thus D measures the difference between the average target-record and kernel increments in this four-strategy design, including each arm's development and reoptimization. The original 10000 archived bootstrap index rows in family order N1–N15 were reused, with the same rows applied to paired family means in U or r. Linear-interpolated 2.5th/97.5th percentiles give conditional pointwise intervals. Sensitivities are mean ± t(0.975,14) × sample SD/√15, and the range of 14-family means after deleting each family once. Correlations are averaged arithmetically, without Fisher transformation or pooled candidate predictions. With 15 families, percentile-bootstrap intervals are expected to be narrower than t intervals and to under-cover (Hesterberg 2015); here they were 10–13% narrower pointwise and 17–19% narrower after the Bonferroni adjustment described below. Where the two methods disagree, the wider t interval is the more conservative guide.

A further post-hoc sensitivity considered the six r comparisons ΔL, ΔG and D at SH/HZ together, using Bonferroni tail probabilities 0.05/(2 × 6) for the archived-bootstrap percentiles and corresponding t_{14} intervals (Additional file 1: Section S7). This selected-six sensitivity does not cover all study comparisons. All intervals condition on the frozen source banks, predictions, tuning and five draws. Shared parentage, overlapping source records and field structure limit independent-family sampling interpretations, and the intervals omit fitting and tuning uncertainty. IRRI distributions and literal sign

counts describe overlapping partitions and repeated seasons of one cohort without independence-based tests.

Models were fitted and scored with custom Python/NumPy implementations; the archived reproduction package specifies Python 3.12 and NumPy 2.3.5. Checks included separately implemented scoring, preselected technical replays and selected model-to-prediction reconstructions; their scope is detailed in Additional file 1: Section S11. Selection-fraction and rank-correlation analyses used stored predictions without refitting or retuning. The original 10% metrics and selected-ID ordering were replayed. The generative artificial intelligence tools OpenAI Codex and Anthropic Claude were used to assist with code debugging, optimizing, and checking calculations against archived results. Further details of AI assistance are provided under Declarations.

Results

Target-family records improve mean predictive ability for both NAM models

The four-strategy comparison shows the performance achieved by reallocating 100 of the 2000 yield records to the target family (Fig. 2; Table 2). In source ridge, source Gaussian, pooled ridge100 and calibrated Gaussian100 order, mean r was 0.187/0.199/0.237/0.257 at SH and 0.163/0.168/0.190/0.213 at HZ. Mean U followed the same ordering: 5.437/5.541/6.256/6.391 g/plant at SH and 2.948/3.149/3.273/3.578 at HZ. These are means of 15 equally weighted five-draw family means. For reference, the equally weighted family means of recorded yield were 32.6 g/plant at SH and 20.9 g/plant at HZ, with individual family means of 23.9–49.0 and 12.5–30.0 g/plant, respectively (table AF38 in Additional files 2 and 3).

Table 2
Post-hoc NAM four-strategy contrasts

Domain	Contrast	ΔU [conditional 95%]	Δr [conditional 95%]	U +/0/−	r +/0/−
SH	Target: ridge ΔL	+ 0.819 [0.296, 1.332]	+ 0.049 [0.029, 0.069]	12/0/3	12/0/3
SH	Target: Gaussian ΔG	+ 0.850 [0.191, 1.603]	+ 0.059 [0.032, 0.091]	10/0/5	13/0/2
SH	Gaussian: source K0	+ 0.104 [− 0.273, 0.478]	+ 0.011 [− 0.007, 0.027]	7/0/8	10/0/5
SH	Gaussian: calibrated K100	+ 0.134 [− 0.298, 0.564]	+ 0.020 [0.008, 0.033]	8/0/7	10/0/5
SH	Interaction I	+ 0.031 [− 0.468, 0.580]	+ 0.009 [− 0.010, 0.033]	5/0/10	10/0/5
HZ	Target: ridge ΔL	+ 0.325 [− 0.094, 0.797]	+ 0.027 [0.009, 0.049]	8/0/7	11/0/4
HZ	Target: Gaussian ΔG	+ 0.428 [− 0.000220, 0.915]	+ 0.045 [0.024, 0.071]	8/0/7	14/0/1
HZ	Gaussian: source K0	+ 0.201 [− 0.026, 0.466]	+ 0.005 [− 0.009, 0.018]	8/0/7	9/0/6
HZ	Gaussian: calibrated K100	+ 0.304 [− 0.015, 0.681]	+ 0.022 [0.005, 0.041]	10/0/5	11/0/4
HZ	Interaction I	+ 0.103 [− 0.242, 0.498]	+ 0.018 [− 0.006, 0.042]	9/0/6	13/0/2
Restricted SH	Target: ridge ΔL	+ 0.813 [0.324, 1.308]	+ 0.050 [0.029, 0.070]	12/0/3	12/0/3
Restricted SH	Target: Gaussian ΔG	+ 0.977 [0.269, 1.775]	+ 0.058 [0.032, 0.089]	11/0/4	13/0/2
Restricted SH	Gaussian: source K0	+ 0.015 [− 0.350, 0.386]	+ 0.012 [− 0.004, 0.028]	7/0/8	11/0/4
Restricted SH	Gaussian: calibrated K100	+ 0.179 [− 0.237, 0.571]	+ 0.021 [0.008, 0.033]	9/0/6	10/0/5
Restricted SH	Interaction I	+ 0.164 [− 0.355, 0.732]	+ 0.008 [− 0.010, 0.031]	5/0/10	10/0/5

U is g/plant; r is dimensionless. Counts refer to 15 five-draw family means. The five four-strategy contrasts and the later diagonal contrast D retain conditional pointwise intervals; ΔL and K0 in U reproduce original checking contrasts (Table 3), and all other entries were added post hoc; complete t_{14} and delete-one-family summaries are in tables AF33 and N05 of Additional files 2 and 3. K0/K100 denote kernel contrasts, not model arms. Restricted SH is the same SH predictions reranked on HZ-valid S and is a paired-domain sensitivity. D is not the original pool100 − ridge0 checking contrast.

Domain	Contrast	ΔU [conditional 95%]	Δr [conditional 95%]	U +/0/−	r +/0/−
SH	Diagonal strategy D	+ 0.716 [0.099, 1.362]	+ 0.038 [0.014, 0.065]	10/0/5	9/0/6
HZ	Diagonal strategy D	+ 0.124 [− 0.342, 0.577]	+ 0.023 [0.005, 0.041]	8/0/7	10/0/5
Restricted SH	Diagonal strategy D	+ 0.798 [0.202, 1.417]	+ 0.037 [0.013, 0.064]	10/0/5	9/0/6

U is g/plant; r is dimensionless. Counts refer to 15 five-draw family means. The five four-strategy contrasts and the later diagonal contrast D retain conditional pointwise intervals; ΔL and K0 in U reproduce original checking contrasts (Table 3), and all other entries were added post hoc; complete t_{14} and delete-one-family summaries are in tables AF33 and N05 of Additional files 2 and 3. K0/K100 denote kernel contrasts, not model arms. Restricted SH is the same SH predictions reranked on HZ-valid S and is a paired-domain sensitivity. D is not the original pool100 − ridge0 checking contrast.

Target-record replacement improved mean r for both models. The paired ΔL/ΔG gains were + 0.049 [0.029, 0.069]/+0.059 [0.032, 0.091] at SH and + 0.027 [0.009, 0.049]/+0.045 [0.024, 0.071] at HZ. Family means were positive in 12/15 and 13/15 families at SH and in 11/15 and 14/15 at HZ for ridge and Gaussian prediction, respectively (Table 2). Both pointwise interval methods excluded zero for these four contrasts; all 15 family values are shown with the paired intervals in Fig. 2.

For the selected-line yield advantage (U), SH ΔL/ΔG were + 0.819 [0.296, 1.332]/+0.850 [0.191, 1.603] g/plant, with 12/15 and 10/15 positive family means. At HZ they were + 0.325 [− 0.094, 0.797]/+0.428 [− 0.000220, 0.915], with 8/15 positive families for each. The t intervals showed the same pattern (Additional file 1: Figure S11).

The source-only Gaussian increment K0 averaged + 0.011/+0.005 in r and + 0.104/+0.201 g/plant in U at SH/HZ, with pointwise intervals spanning zero. With target yield records, K100 added + 0.020/+0.022 in r, with both interval methods excluding zero. Its mean U increments were + 0.134/+0.304 g/plant, with intervals spanning zero and negative family means in 7/15 SH and 5/15 HZ families. The interaction I averaged + 0.009/+0.018 in r and + 0.031/+0.103 g/plant in U; its intervals spanned zero. Thus the Gaussian gain in mean r was clear with target-family records (K100) but not without them (K0), and their difference, the interaction I, was not clear. Family-mean and draw-level U values for these five contrasts are shown in Additional file 1: Figures S1 and S2; Figure S3 compares their paired scoring views.

The direct diagonal comparison D favored pooled ridge100 over source Gaussian0 in mean r: +0.038 [0.014, 0.065] at SH and + 0.023 [0.005, 0.041] at HZ, with 9/15 and 10/15 positive family means. Both pointwise interval methods excluded zero. The corresponding U differences were + 0.716 [0.099, 1.362] and + 0.124 [− 0.342, 0.577] g/plant; SH's t_{14} interval was [0.003, 1.429]. These results directly compare linear prediction with target-family records and Gaussian prediction without them at equal numbers of yield records. In the selected-six r sensitivity, bootstrap intervals excluded zero for five comparisons, except HZ D; t_{14} intervals excluded zero for SH ΔL/ΔG and HZ ΔG, while HZ ΔL and both D intervals spanned zero (Additional file 1: Section S7).

The original weighted ridge100 − source Gaussian0 U primary contrast was + 0.732 g/plant at SH and + 0.278 at HZ (Table 3; Additional file 1: Figure S8). Its conditional family-bootstrap intervals were [0.014, 1.442]/[− 0.193, 0.827], and t_{14} intervals were [− 0.070, 1.534]/[− 0.302, 0.858]. Thus the SH interval's exclusion of zero depended on the interval method. The original pooling check corresponds to the ΔL U comparison above.

Table 3
Original primary and checking comparisons

Context and role	Comparison	Original U outcome	Primary t sensitivity
NAM, frozen primary	Weighted ridge100 − source Gaussian0	SH + 0.732 [0.014, 1.442]; HZ + 0.278 [− 0.193, 0.827] g/plant	SH [− 0.070, 1.534]; HZ [− 0.302, 0.858] g/plant
NAM, frozen check	Pooled ridge100 − source ridge0	SH + 0.819 [0.296, 1.332]; HZ + 0.325 [− 0.094, 0.797] g/plant; 12/15 and 8/15 family means positive	—
NAM, frozen checks	Weighted ridge100 − pooled ridge100; source Gaussian0 − source ridge0	Weighting: SH + 0.016 [− 0.339, 0.430]; HZ + 0.154 [− 0.168, 0.544]. Gaussian: SH + 0.104 [− 0.273, 0.478]; HZ + 0.201 [− 0.026, 0.466] g/plant	—
MAGIC, frozen primary	Weighted ridge100 − source Gaussian0	SH − 1.53; HZ − 0.01 g/plant	—
MAGIC, frozen check	Pooled ridge100 − source ridge0	SH − 0.97; HZ + 0.75 g/plant	—
MAGIC, frozen checks	Pooled Gaussian100 − source Gaussian0; allocation interaction	SH − 0.97 / −0.004; HZ − 0.87 / −1.62 g/plant	—
IRRI, frozen primary	Pooled ridge25 − source Gaussian0	2011 DS + 231.5 kg/ha	—
IRRI, frozen check	Pooled ridge25 − source ridge0	2011 DS + 255.1; 2012 WS − 151.8 kg/ha	—
IRRI, frozen checks	SW-source ridge25 − pooled ridge25; pooled Gaussian25 − source Gaussian0	Primary − 36.1; 0.0 kg/ha	—

NAM bracketed original intervals are conditional 95% family-bootstrap descriptions; the t intervals are a later method sensitivity. MAGIC and IRRI values describe their observed allocations. NAM/MAGIC weighted ridge uses target-mass weighting; IRRI SW denotes target-similarity weighting. The SH MAGIC allocation interaction is shown to three decimal places.

Regularization changes predictive ability and selected-yield benefits

Across $\lambda = 10^{-4}$, 10^{-3}, 10^{-2} and 10^{-1}, the fixed-grid mean U advantage of pooled ridge100 over source ridge0 declined monotonically in all three NAM scoring views (Fig. 3; Table 4). SH values were +

1.743/+1.309/+0.482/+0.015 g/plant; HZ values were + 0.755/+0.624/+0.317/−0.092; restricted SH values were + 1.762/+1.214/+0.478/−0.018. The three weaker settings had 11–14 positive family means in each view, whereas λ = 0.1 gave means near zero or negative. Across the four observed grid values, pooled ridge also attained a higher maximum U than source ridge: 6.766 versus 5.810 g/plant at SH and 3.502 versus 3.186 at HZ. Because these maxima were picked on the evaluation lines, they are descriptive only; deployed settings were chosen on other families.

Table 4

NAM selected-line yield advantage and predictive ability across the fixed penalty grid A. Selected-line yield advantage

Scoring view	λ	nλ	Ridge0 U	Pool100 U	ΔU [conditional 95%]	Positive families
SH original	0.0001	0.2	3.767	5.510	+ 1.743 [1.111, 2.354]	13/15
SH original	0.001	2	5.457	6.766	+ 1.309 [0.581, 2.152]	13/15
SH original	0.01	20	5.810	6.292	+ 0.482 [− 0.272, 1.131]	12/15
SH original	0.1	200	4.750	4.765	+ 0.015 [− 0.499, 0.408]	8/15
HZ paired	0.0001	0.2	1.836	2.591	+ 0.755 [0.263, 1.315]	11/15
HZ paired	0.001	2	2.755	3.378	+ 0.624 [0.335, 0.896]	13/15
HZ paired	0.01	20	3.186	3.502	+ 0.317 [0.077, 0.568]	12/15
HZ paired	0.1	200	3.089	2.997	−0.092 [− 0.235, 0.057]	4/15
Restricted SH	0.0001	0.2	3.874	5.636	+ 1.762 [1.085, 2.412]	14/15
Restricted SH	0.001	2	5.616	6.830	+ 1.214 [0.531, 2.011]	13/15
Restricted SH	0.01	20	5.858	6.336	+ 0.478 [− 0.280, 1.137]	12/15
Restricted SH	0.1	200	4.866	4.848	−0.018 [− 0.554, 0.383]	9/15

U and ΔU are g/plant. Brackets retain the existing conditional pointwise 95% family-bootstrap intervals; 15 family means average five draws each. nλ is model-implied under the stated equal-weight assumptions, not estimated heritability.

B. Predictive ability at matched λ

Scoring view	λ	Source ridge r	Pooled ridge r	Δr	Family +/0/−
SH original	0.0001	0.158	0.229	+0.071	14/0/1
SH original	0.001	0.210	0.258	+0.048	14/0/1
SH original	0.01	0.193	0.218	+0.025	14/0/1
SH original	0.1	0.149	0.160	+0.011	13/0/2
HZ paired	0.0001	0.122	0.181	+0.059	13/0/2
HZ paired	0.001	0.169	0.212	+0.043	13/0/2
HZ paired	0.01	0.171	0.197	+0.026	14/0/1
HZ paired	0.1	0.146	0.154	+0.009	13/0/2
Restricted SH	0.0001	0.162	0.231	+0.069	14/0/1
Restricted SH	0.001	0.213	0.261	+0.048	14/0/1
Restricted SH	0.01	0.196	0.221	+0.025	14/0/1
Restricted SH	0.1	0.152	0.162	+0.010	13/0/2

r is Pearson predictive ability. Entries are equal means of 15 five-draw family means, using the same archived predictions as panel A. All 12 rows are post-hoc descriptive summaries without new intervals or tests.

At the same numerical λ in the two arms, pooling also improved mean r at every evaluated setting (Table 4B). In ascending λ order, Δr was + 0.071/+0.048/+0.025/+0.011 at SH, + 0.059/+0.043/+0.026/+0.009 at HZ, and + 0.069/+0.048/+0.025/+0.010 in restricted SH; 13–14 of 15 family means were positive in each of the 12 combinations. Thus the r difference persisted throughout this finite grid even when strong regularization reduced the selected-yield difference to near zero.

NAM-to-MAGIC transfer varies with site and regularization

MAGIC comparisons are reported for U only. The original MAGIC primary contrast was − 1.53 g/plant at SH and − 0.01 at HZ (Table 3; Fig. 4). All five SH draw contrasts were negative, ranging from − 2.15 to − 0.59. Source Gaussian0 had the highest SH U (6.16 g/plant). Pooled ridge100 − source ridge0 was − 0.97 at SH but + 0.75 at HZ; pooled Gaussian100 − source Gaussian0 was − 0.97/−0.87. Thus the favorable NAM pattern did not carry over consistently to this multiparent target population.

At matched λ, SH pooled ridge100 − source ridge0 was + 0.43/+0.62/+0.35/−0.33 g/plant across the four grid values; HZ was − 0.16/−0.57/+0.75/−0.04. At HZ, η = 0.05 made weighted and uniform pooling predictions identical at matched λ. Removing from S the 25 lines whose common-allele proxy score was at least 3 SD above the scoring-set mean changed N/q from 1097/110 to 1072/108, attenuated the SH primary contrast from − 1.53 to − 0.74 and changed HZ pooling from + 0.75 to + 0.69, with the same

fitted models and predictions (Additional file 1: Figure S10). This proxy score is a marker-similarity measure, not verified ancestry.

Current-season records improve IRRI predictive ability, while selected-line yield varies

For the original IRRI allocation, pooled ridge25 − source Gaussian0 was + 231.5 kg/ha in primary 2011 DS, followed by + 89.8/+373.7/+65.0 in the three correlated seasons (Table 3; Additional file 1: Figure S9). Pooled ridge25 − source ridge0 was + 255.1/+86.0/+172.3/−151.8. In primary 2011 DS, pooling and source Gaussian shared six selected IDs and had four unique IDs each. Calibrated and source Gaussian had identical selected top-10 sets despite different predictions and r.

Across the 20 fixed-policy configurations, the original pooled ridge25 − source Gaussian0 comparison was positive in 16/20 primary-season partitions and 17/20 in 2012 DS (Fig. 5; Table 5). Mean U contrasts in chronological season order were + 172.2/−94.9/+205.0/+1.2 kg/ha, with 16/7/17/12 positive partitions. Against source ridge, pooling means were + 66.7/−42.7/−58.6/−36.0 kg/ha, with 13/10/8/9 positive partitions. The historical allocation is displayed separately from these 20-configuration summaries. Partition-level U contrasts under both policies are shown against source Gaussian and source ridge in Additional file 1: Figures S4 and S5, respectively.

Table 5
IRRI seasonal contrasts across the 20 configurations A. Pooled ridge25 minus source Gaussian0

Policy / season	Mean ΔU	Median [min, max] ΔU	U +/0/−	Mean Δr	r +/0/−
fixed / 2011DS	+ 172.2	206.8 [− 166.8, 453.1]	16/0/4	+ 0.117	19/0/1
fixed / 2011WS	−94.9	−120.4 [− 353.7, 201.8]	7/0/13	+ 0.042	15/0/5
fixed / 2012DS	+ 205.0	227.5 [− 220.6, 480.4]	17/0/3	+ 0.074	18/0/2
fixed / 2012WS	+ 1.2	10.7 [− 212.2, 123.1]	12/0/8	+ 0.128	19/0/1
adaptive / 2011DS	−20.6	6.5 [− 403.6, 242.8]	11/2/7	+ 0.124	19/0/1
adaptive / 2011WS	−1.4	0.5 [− 244.3, 356.5]	10/0/10	−0.009	11/0/9
adaptive / 2012DS	−39.6	−66.2 [− 185.9, 188.4]	6/0/14	+ 0.049	15/0/5
adaptive / 2012WS	−7.1	−13.0 [− 172.1, 151.3]	9/0/11	+ 0.069	16/0/4

B. Pooled ridge25 minus source ridge0

Policy / season	Mean ΔU	Median [min, max] ΔU	U +/0/−	Mean Δr	r +/0/−
fixed / 2011DS	+66.7	52.4 [−275.9, 373.3]	13/0/7	+0.147	20/0/0
fixed / 2011WS	−42.7	−27.6 [−203.9, 118.6]	10/0/10	+0.010	11/0/9
fixed / 2012DS	−58.6	−86.8 [−304.7, 199.6]	8/0/12	+0.048	18/0/2
fixed / 2012WS	−36.0	−49.8 [−228.9, 192.0]	9/0/11	+0.045	16/0/4
adaptive / 2011DS	+26.6	79.1 [−244.9, 419.9]	12/0/8	+0.198	20/0/0
adaptive / 2011WS	+18.1	16.2 [−231.0, 234.1]	11/0/9	+0.019	15/0/5
adaptive / 2012DS	−97.8	−80.3 [−441.1, 239.0]	7/0/13	+0.077	17/0/3
adaptive / 2012WS	−30.0	22.0 [−327.4, 158.9]	12/0/8	+0.068	16/0/4

Differences in U are kg/ha; Δr is dimensionless. Counts describe 20 overlapping configurations per season, with literal exact zeros and the original allocation excluded. Same-model Gaussian contrasts, pooled Gaussian25 minus source Gaussian0, had positive mean Δr in all eight policy–season summaries (+0.023 to +0.117; AF34). Complete six-contrast summaries, endpoint sign agreement and same-S policy differences are retained in Additional files 2 and 3.

Under fixed tuning, pooling raised mean r over source Gaussian (+ 0.117/+0.042/+0.074/+0.128, with 19/15/18/19 positive partitions) and over source ridge (+ 0.147/+0.010/+0.048/+0.045, with 20/11/18/16 positive partitions) in all four seasons. Same-model comparisons were positive in all four seasons under both tuning policies: current-season records raised mean r by + 0.010 to + 0.198 for ridge regression (Table 5B) and by + 0.023 to + 0.117 for the Gaussian model (table AF34 in Additional files 2 and 3). For pooled ridge25 − source ridge0, U/r contrast signs agreed in 13/7/10/11 of 20 cells per season, totaling 41/80; the adaptive-policy total was 38/80. These paired results demonstrate that an improvement in candidate-wide correlation can accompany a reduction in the yield of selected lines.

Under the adaptive policy, mean pooled ridge25 − source Gaussian0 U was − 20.6/−1.4/−39.6/−7.1 kg/ha, with positive/exact-zero/negative counts of 11/2/7, 10/0/10, 6/0/14 and 9/0/11. The first two medians were positive (+ 6.5/+0.5). For source-only models, the historical training/validation partitions, genotype transformations, candidate grids and final source yield records were retained. The U-based rule selected ridge λ = 0.1 instead of 0.01; this parameter change lowered source-ridge r in all 80 overlapping cells, with mean seasonal changes of − 0.057 to − 0.033. For source Gaussian, λ remained 1 while γ changed from 2 to 1 because six settings tied exactly on U and the pre-declared tie rule chose among them; r increased in 59/80 cells. For strategies with target-season records, the adaptive policy also re-tuned them with stand-in target records included; calibrated-Gaussian r increased in 14–19 partitions per season. Same-S adaptive-minus-fixed policy differences in the pooled ridge25 U contrasts against source Gaussian0 and source ridge0 are shown in Additional file 1: Figures S6 and S7, respectively.

Decision-threshold and rank-correlation sensitivities

Mean NAM target-record U advantages remained positive across 5%/10%/20% selection. Pooled ridge100 − source ridge0 differences were + 0.910/+0.819/+0.984 g/plant at SH and + 0.398/+0.325/+0.237 at HZ; calibrated Gaussian differences were + 1.370/+0.850/+1.005 and + 0.640/+0.428/+0.539. D means were + 0.975/+0.716/+0.661 and + 0.207/+0.124/+0.226. Across these three contrasts, fractions and scoring views, 2–7 of 15 family means were negative in each combination. Original-primary U means were also positive across the three fractions: SH + 1.064/+0.732/+0.582 and HZ + 0.327/+0.278/+0.415 g/plant.

Spearman corroborated the target-record pattern: ridge/Gaussian increments were + 0.053/+0.063 at SH, with 13/15 and 14/15 positive families, and + 0.029/+0.046 at HZ, with 11/15 and 14/15 positive families. For the original weighted ridge100 − source Gaussian0 comparison, mean Pearson/Spearman increments were + 0.042/+0.049 at SH (11/15 and 12/15 positive families) and + 0.042/+0.051 at HZ (14/15 for both). At HZ, switching models alone left Spearman correlation essentially unchanged (+ 0.00002; 8/15 positive families).

At top 5%, pooled ridge improved mean U relative to source ridge in every IRRI season under both policies. Fixed-policy differences were + 245.9/+227.4/+231.6/+212.6 kg/ha, with positive medians and 16/15/15/16 positive partitions (62/80 cells); adaptive means were + 167.0/+145.9/+93.6/+56.0, also with positive medians (Additional file 1: Section S7; table IRRI_top5 in Additional files 2 and 3). Relative to source Gaussian, however, fixed means were + 57.9/−347.4/+101.6/−290.4 kg/ha (12/4/14/1 positive partitions), and adaptive means were + 65.5/−39.4/−115.8/−144.1 (Additional file 1: Section S7). The fixed/adaptive pooled ridge25 − source ridge0 means at the original top 10% were + 66.7/−42.7/−58.6/−36.0 and + 26.6/+18.1/−97.8/−30.0; at top 20% they were + 128.9/−3.3/+12.5/−31.1 and + 31.0/−91.6/−81.3/−46.7. For that ridge comparison across both policies, positive/exact-zero/negative cell counts were 107/6/47, 82/0/78 and 60/3/97 at 5%/10%/20%.

Pearson and Spearman also distinguished different aspects of prediction: fixed 2011 DS pooling relative to source Gaussian gave mean increments + 0.117/−0.040, and 2012 DS gave + 0.074/−0.025. At top 10%, the two ridge strategies shared an average 6.10–7.20 of ten selected IDs across policy-season summaries. Their U difference equals the recorded-yield contributions of entered minus exited IDs, divided by q; common IDs cancel exactly.

Discussion

The main finding is that, within these 15 related families, measuring yield in 100 lines of the family being selected raised average within-family predictive ability for both models (Fig. 2; Table 2). These 100 lines supplied 5% of the 2000 records and made up 4.7–15.2% of each family's lines at Shanghai. The mean correlation between predicted and recorded yield rose by 0.027–0.059, from 0.163–0.199 to 0.190–0.257 across the two sites and both models, with gains in 11–14 of the 15 families and a similar pattern in Spearman rank correlation. Because both options were measured on the same evaluation lines with the same number of yield records, their effects can be compared directly: switching to the Gaussian

model alone added only 0.005–0.011. The gain held up under two further checks. After a Bonferroni correction across six selected correlation comparisons, namely these four gains and the two direct comparisons of target-family records with model switching (D), three of the four gains stayed above zero by the more conservative t interval and all four by the bootstrap interval (Additional file 1: Section S7). For ridge regression, the gain also appeared at each of four fixed penalty settings, in 13–14 of the 15 families, so it did not depend on the two strategies choosing different penalties (Table 4B).

Relative to the mean without target-family records, the gain is 17–30% as a ratio of means (Additional file 1: Section S7), a substantial relative gain on the modest scale typical of within-family correlations. The Gaussian model with target-family records gave the highest correlation, and ridge regression with target-family records exceeded the Gaussian model without them by 0.038 at Shanghai and 0.023 at Hangzhou. This pattern agrees with studies in which added model complexity brought little consistent benefit (Azodi et al. 2019; Nguyen et al. 2023) and with ridge regression being among the leading models for grain yield in elite IRRI lines (Spindel et al. 2015), whereas the relatedness or composition of the training data had clearer effects (Zhu et al. 2021; Fernández-González et al. 2023). In rice, sparse phenotyping of multi-environment training sets maintained accuracy (He et al. 2023). The present results add that where a few records are placed matters: records from the family being selected raised mean predictive ability more than switching to the Gaussian model did. The value of full-sib records agrees with soybean families in which substantially more half-sib than full-sib lines were needed for comparable accuracy (Zhu et al. 2021).

The advantage of target-family records over model switching was positive at both sites, with both pointwise interval methods excluding zero, but it is the less certain part of the comparison: after the Bonferroni correction only its Shanghai bootstrap interval excluded zero (Additional file 1: Section S7), and model rankings were evaluated under one tuning rule, a limitation discussed below. For selected-line yield, the evidence points the same way as for predictive ability but is weaker. The mean gain from target-family records was positive whether the top 5%, 10% or 20% were selected; at Shanghai, for example, the top 10% of lines chosen by ridge regression out-yielded the average evaluated line by 6.26 g per plant with target-family records and by 5.44 g without them, against an equally weighted family mean yield of 32.6 g per plant. At 10% selection, the 95% intervals for this gain stayed above zero at Shanghai but included zero at Hangzhou. The comparison specified in the original analysis, a weighted form of ridge regression with target-family records versus the Gaussian model without them, was positive at both sites but not conclusive (Table 3).

Target-family records improved the ranking of siblings, not merely the family's average: adding the same amount to every prediction in a family would leave the ranking and the within-family correlation unchanged. Genetic relationships and shared segregation are plausible contributors (Habier et al. 2007; Lorenz and Smith 2015): the measured and the evaluated lines come from the same cross, so the measured lines can inform marker effects that are poorly represented in other families. Target-family records also made predictive ability more even across families: the between-family SD of r fell by 11–28%, and families with zero or negative mean r fell from one to none at Shanghai and, at Hangzhou, from

two to one for ridge regression and to none for the Gaussian model. These descriptive values and Oldham's change-versus-midpoint descriptor (Oldham 1962; Chiolero et al. 2013) are given in Additional file 1: Section S7.

The penalty λ, which sets how strongly marker effects are shrunk, changed the size of both benefits, but only the selected-yield benefit disappeared: target-family records raised ridge-regression r at every tested penalty (Table 4B), whereas their selected-yield advantage shrank as the penalty increased and was near zero or slightly negative under the strongest (Table 4A). In MAGIC, the pooling effect also changed sign with the penalty: at matched penalties, pooling was positive at three of four values at Shanghai, where the deployed pooled ridge used λ = 0.1 against 0.01 for source ridge, but at only one of four at Hangzhou (Fig. 4C; Additional file 1: Section S4). Penalty choice should therefore be checked whenever records are transferred to a new population; more diverse haplotypes or a different relationship to the NAM training families may also change how many of its own records are needed.

Predictive ability and selected-line yield answer complementary breeding questions. Correlation evaluates all candidates; U evaluates the recorded yields of the selected lines. For a fixed set of evaluation lines and number selected, lines chosen by both models cancel, so a between-model U difference depends only on the few lines that enter or leave the selected group; a model can therefore track yield more closely across all candidates yet select a lower-yielding group, especially when few lines are selected (q = 10 in IRRI).

The comparison model and the selection intensity also shape the selected-yield result. In IRRI, pooled ridge regression raised mean and median U over ridge regression without target records in every season at the top 5% under both tuning policies, whereas its advantage over the Gaussian model varied with season and results at 10% and 20% were mixed; a strong model without target records can therefore remain competitive in an elite cohort. Because conclusions can depend on the threshold, breeding evaluations should report the intended selection fraction and comparison model alongside predictive ability.

How a model is tuned also matters. In IRRI, tuning ridge regression without target records for selected-line yield instead of correlation chose a ten-fold stronger penalty and lowered r in all 80 scored cases, whereas the same switch raised Gaussian-model r in 59 of 80. With only two or three selected lines in each historical validation set, several settings tied exactly and the pre-declared tie rule decided, so tuning on selected-line yield needs validation sets with enough selected lines to separate settings. In another rice dataset, genotype-by-environment models did not consistently outperform simpler ones (Nguyen et al. 2023); the added value of any more complex method is best measured where it will be used.

The intended application is partial phenotyping: yield measurements from some target-family lines are available when the remaining family members are ranked. Where such target records are available, pooled ridge provides a practical reference for assessing the added value of the tested Gaussian strategy. In NAM, ordinary equal-weight pooling did about as well as target-mass weighting (Table 3), which favors the simpler method. In the original analysis, weighted ridge with 25 target-family records

gave smaller mean selected-line yield advantages than with 100 (SH 5.53 versus 6.27 g/plant; HZ 3.28 versus 3.43; table AF02 in Additional files 2 and 3), in line with a dose response, although the two strategies were tuned separately.

Several limitations should be noted. This retrospective study examined one trait with fixed marker panels and custom ridge and Gaussian implementations, and all extensions, including the four-strategy comparison, were added after the original results had been seen. The four-strategy comparison tested one number of target-family records, 100 of 2000, so the best number to phenotype was not estimated. In NAM and MAGIC, model settings were tuned on other families to maximize selected-line yield rather than by restricted maximum likelihood (REML) variance-component estimation, and strategy rankings, including the advantage over model switching, were not evaluated under correlation-based tuning or REML-tuned GBLUP; the fixed-penalty check covered ridge regression only. NAM and MAGIC come from one field resource. The 15 NAM families share one parent, serve as training families for one another and were scored at both sites, so the two sites give paired rather than independent evidence. Hangzhou records may be missing non-randomly, marker quality control used all NAM genotypes, including later target families, and the sparse marker panel cannot fully confirm line identity. Because each family was grown in its own block, target-family records share the field conditions of the lines they help rank and, as the field year could not be determined, possibly their year and genotype-by-year effects, whereas other-family records might not; part of the gain may therefore be environmental and may not transfer unchanged to new seasons or trial designs. IRRI yields are replicate means rather than BLUEs, and the MAGIC and IRRI repeats reuse overlapping lines from single cohorts. The main NAM strategies used the same held-out lines and number of yield records, no evaluation record was used to train or tune the models that ranked it, and all original comparisons, including unfavorable ones, are reported (Table 3).

Programs adopting this approach should decide in advance which records will be used for training, what fraction of lines will be selected, and how models will be judged on both predictive ability and selected-line yield. Choosing the lines to phenotype from their genotypes (Rincent et al. 2012; Akdemir and Isidro-Sánchez 2019), comparing several numbers of target-family records and testing new populations prospectively are the next steps. The released predictions, evaluation-line identities and code (Li 2026) let others re-score the main results and test new models or phenotyping designs on exactly the same NAM, MAGIC and IRRI evaluation lines.

Conclusions

When a new breeding family has no yield records of its own, measuring yield in 100 of its lines, within a fixed total of 2000 training records, improved prediction of its held-out members in these related rice NAM families. Mean within-family predictive ability rose from 0.16–0.20 to 0.19–0.26 at two sites, with gains in 11–14 of the 15 families for both ridge regression and the Gaussian model, and the ridge gain persisted at every tested penalty setting. Switching to the Gaussian model alone changed predictive ability little (+ 0.005 to + 0.011), but combined with target-family records the Gaussian model gave the highest mean predictive ability; the gain from target-family records did not clearly differ between the two

models. Target-family records also increased mean selected-line yield at both sites, with 95% intervals above zero at Shanghai but not at Hangzhou. In an elite IRRI breeding cohort, adding 25 current-season records to 626 historical records raised the average predictive ability of both models in every season, whereas selected-line yield varied; in a multiparent population predicted from the NAM families, its own records lowered selected-line yield at one site.

For programs breeding related families like these, the results support three practices when part of a new family can be yield-tested before the rest are ranked: devote a small share of a fixed number of yield records, here 100 of 2000, to that family; use ridge regression trained with these records as the benchmark that more flexible models must beat; and compare strategies by both predictive ability and the yield of the lines selected at the intended fraction. Because target-family and evaluation records came from the same trials, at the same site in NAM and in the same season in IRRI, natural next steps are to test prediction in a later season and to determine how many lines to phenotype, at what cost and with what effect on genetic gain. The public predictions and code provide reference results for comparing new models and phenotyping designs.

Abbreviations

A
adaptation pool
BLUE
best linear unbiased estimate
C
calibration set
CSV
comma-separated values
DS
dry season
GBLUP
genomic best linear unbiased prediction
HZ
Hangzhou
ID
identifier
IRRI
International Rice Research Institute
MAGIC
multiparent advanced generation intercross
NAM
nested association mapping
q

number of selected lines
r
predictive ability,measured as Pearson correlation within a scoring set and,in NAM,within each family
REML
restricted maximum likelihood
RKHS
reproducing kernel Hilbert space
S
scoring set
SD
standard deviation
SH
Shanghai
U
selected-line yield advantage
WS
wet season.

Declarations

Ethics approval and consent to participate

Not applicable. This study reanalyzes rice genotype and phenotype records and does not involve human participants or animals.

Consent for publication

Not applicable. No individual person's data, images or videos are reported.

Competing interests

The author declares no competing interests.

Use of generative artificial intelligence

The author conceived and designed the study, made all analytical decisions, interpreted the results and finalized the manuscript. OpenAI Codex and Anthropic Claude Code were used as assistants for code debugging and optimization, along with text proofreading and language editing. The author reviewed and verified the AI-assisted outputs and takes full responsibility for the manuscript. The main results can be re-scored from the archived predictions using the public code (Li 2026).

Funding

No specific funding was received for this research.

Author Contribution

JL designed and conducted the reanalysis of public data, developed and checked the analysis and visualization code, interpreted the results, and wrote and revised the manuscript. JL read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

NAM/MAGIC source data are described by Wei et al. (2024; https://doi.org/10.1126/science.adm8762), with grain-yield records in Data S4 and genotypes in NAM_variations version 1 (Chen 2024; https://doi.org/10.6084/m9.figshare.19166475.v1). Family identities follow Rice Recombination Site Information version 1 (Chen 2026; https://doi.org/10.6084/m9.figshare.31881931.v1). IRRI source records are in Dryad version 1 (Spindel et al. 2016a; https://doi.org/10.5061/dryad.7369p). Additional files 1–3 contain the supplementary information, readable result workbook and canonical result tables.Frozen predictions, numerical model implementations and prediction-to-result scoring code are available in the GitHub release https://github.com/yuriljx/rice-genomic-selection-validation/releases/tag/v1.0.0, archived as software version 1.0.0 (Li 2026) at https://doi.org/10.5281/zenodo.23438608. The public package covers the analyses specified in its README and uses locally supplied source phenotype files; the reproduction entry does not refit or retune models. The reproduction scope and source-data requirements are described in Additional file 1: Section S11. The NAM plan file with the source, core and pseudo-target allocations needed for refitting is available from the author on reasonable request.Project name: Rice genomic selection: frozen-prediction reproduction. Project home page: https://github.com/yuriljx/rice-genomic-selection-validation. Archived version: 1.0.0, https://doi.org/10.5281/zenodo.23438608 (Li 2026). Operating system: platform independent; reference execution on macOS. Programming language: Python 3.12. Other requirements: NumPy 2.3.5; locally obtained Wei et al. (2024) Data S4 and RYT_plotdata_by_GHID_corrected_PUBLIC_ACCESS.zip from Dryad version 1. License: MIT for software and documentation; CC BY 4.0 for author-generated predictions and numerical results. Restrictions for non-academic use: no additional restrictions on the released software beyond the MIT license; source-data terms apply.

References

1. Akdemir D, Isidro-Sánchez J (2019) Design of training populations for selective phenotyping in genomic prediction. Sci Rep 9:1446. https://doi.org/10.1038/s41598-018-38081-6
2. Atanda SA, Olsen M, Burgueño J, Crossa J, Dzidzienyo D, Beyene Y, Gowda M, Dreher K, Zhang X, Prasanna BM, Tongoona P, Danquah EY, Olaoye G, Robbins KR (2021) Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. Theor Appl Genet 134:279–294. https://doi.org/10.1007/s00122-020-03696-9
3. Azodi CB, Bolger E, McCarren A, Roantree M, de los Campos G, Shiu S-H (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 (Bethesda) 9:3691–3702. https://doi.org/10.1534/g3.119.400498
4. Blondel M, Onogi A, Iwata H, Ueda N (2015) A ranking approach to genomic selection. PLoS ONE 10:e0128570. https://doi.org/10.1371/journal.pone.0128570
5. Brandariz SP, Bernardo R (2019) Small ad hoc versus large general training populations for genomewide selection in maize biparental crosses. Theor Appl Genet 132:347–353. https://doi.org/10.1007/s00122-018-3222-3
6. Brauner PC, Müller D, Molenaar WS, Melchinger AE (2020) Genomic prediction with multiple biparental families. Theor Appl Genet 133:133–147. https://doi.org/10.1007/s00122-019-03445-7
7. Chen (2024) NAM_variations. Version 1. figshare [dataset]. https://doi.org/10.6084/m9.figshare.19166475.v1
8. Chen M (2026) Rice Recombination Site Information. Version 1. figshare [dataset]. https://doi.org/10.6084/m9.figshare.31881931.v1
9. Chen S-P, Sung W-H, Liao C-T (2024) Constructing training sets for genomic selection to identify superior genotypes in candidate populations. Theor Appl Genet 137:270. https://doi.org/10.1007/s00122-024-04766-y
10. Chiolero A, Paradis G, Rich B, Hanley JA (2013) Assessing the relationship between the baseline value of a continuous variable and subsequent change over time. Front Public Health 1:29. https://doi.org/10.3389/fpubh.2013.00029
11. Cobb JN, Juma RU, Biswas PS, Arbelaez JD, Rutkoski J, Atlin G, Hagen T, Quinn M, Ng EH (2019) Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. Theor Appl Genet 132:627–645. https://doi.org/10.1007/s00122-019-03317-0
12. e los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385. https://doi.org/10.1534/genetics.109.101501
13. de Verdal H, Baertschi C, Frouin J, Quintero C, Ospina Y, Alvarez MF, Cao T-V, Bartholomé J, Grenier C (2023) Optimization of multi-generation multi-location genomic prediction models for recurrent genomic selection in an upland rice population. Rice 16:43. https://doi.org/10.1186/s12284-023-00661-0

14. Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255. https://doi.org/10.3835/plantgenome2011.08.0024
15. Fernández-González J, Akdemir D, Isidro y Sánchez J (2023) A comparison of methods for training population optimization in genomic selection. Theor Appl Genet 136:30. https://doi.org/10.1007/s00122-023-04265-6
16. Gianola D, van Kaam JBCHM (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303. https://doi.org/10.1534/genetics.107.084285
17. Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397. https://doi.org/10.1534/genetics.107.081190
18. He S, Liang S, Meng L, Cao L, Ye G (2023) Sparse phenotyping and haplotype-based models for genomic prediction in rice. Rice 16:27. https://doi.org/10.1186/s12284-023-00643-2
19. Hesterberg TC (2015) What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. Am Stat 69:371–386. https://doi.org/10.1080/00031305.2015.1089789
20. Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158. https://doi.org/10.1007/s00122-014-2418-4
21. Jarquin D, Howard R, Crossa J, Beyene Y, Gowda M, Martini JWR, Covarrubias Pazaran G, Burgueño J, Pacheco A, Grondona M, Wimmer V, Prasanna BM (2020) Genomic prediction enhanced sparse testing for multi-environment trials. G3 (Bethesda) 10:2725–2739. https://doi.org/10.1534/g3.120.401349
22. Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201:759–768. https://doi.org/10.1534/genetics.115.177907
23. Kitony JK, Reyes VP, Sunohara H, Tasaki M, Yamasaki M, Mori J-i, Shimazu A, Nishiuchi S, Michael TP, Doi K (2026) Efficient genomic prediction at reduced training size and moderate marker density in an expanded aus-NAM population of rice. bioRxiv, version 1, posted 1 May 2026 (preprint). https://doi.org/10.64898/2026.04.28.721500
24. Kusmec A, Negus KL, Yu J (2026) Critical evaluation of the theory and practice of feed-forward neural networks for genomic prediction. G3 (Bethesda) 16:jkaf314. https://doi.org/10.1093/g3journal/jkaf314
25. Lehermeier C, Krämer N, Bauer E et al (2014) Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198:3–16. https://doi.org/10.1534/genetics.114.161943
26. Li J (2026) Rice genomic selection: frozen-prediction reproduction. https://doi.org/10.5281/zenodo.23438608. Version 1.0.0. Zenodo [software]

27. Li J, Zhang D, Yang F, Zhang Q, Pan S, Zhao X, Zhang Q, Han Y, Yang J, Wang K, Zhao C (2024) TrG2P: A transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield. Plant Commun 5:100975. https://doi.org/10.1016/j.xplc.2024.100975
28. Li Y, Ren S, Li J, Lee J, Wan J, Gan X (2026) MeNet: A mixed-effect deep neural network for multi-environment genomic prediction of agronomic traits. Plant Commun 7:101620. https://doi.org/10.1016/j.xplc.2025.101620
29. Liu Z-J, Liao C-T (2026) Optimizing training sets for genomic selection to identify superior genotypes across multiple environments. G3 (Bethesda) 16:jkag031. https://doi.org/10.1093/g3journal/jkag031
30. Lorenz AJ, Smith KP (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Sci 55:2657–2667. https://doi.org/10.2135/cropsci2014.12.0827
31. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819
32. Montesinos-López OA, Barajas-Ramirez EA, Salinas-Ruiz J, Montesinos-López A, Gerard G, Vitale P, Dreisigacker S, Saint Pierre C, Crossa J (2025a) Optimizing genomic prediction with transfer learning under a ridge regression framework. Plant Genome 18:e70049. https://doi.org/10.1002/tpg2.70049
33. Montesinos-López OA, Solís-Covarrubias AE, Hernández-Suarez CM, Posadas LG, Luna I, Rathod S, Montesinos-López A, Kismiantini (2025b) A transfer learning method for enhanced genomic prediction. Discover Plants 2:278. https://doi.org/10.1007/s44372-025-00356-4
34. Montiel M, Moreno-Amores J, Punzalan J, Angira B, Cerioli T, Robbins K, McCouch S, Fritsche-Neto R, Famoso A (2025) The effect of haplotype size on genomic selection accuracy and epistasis: an empirical study in rice. Plant Genome 18:e70161. https://doi.org/10.1002/tpg2.70161
35. Nguyen VH, Morantte RIZ, Lopena V, Verdeprado H, Murori R, Ndayiragije A, Katiyar SK, Islam MR, Juma RU, Flandez-Galvez H, Glaszmann J-C, Cobb JN, Bartholomé J (2023) Multi-environment genomic selection in rice elite breeding lines. Rice 16:7. https://doi.org/10.1186/s12284-023-00623-6
36. Oldham PD (1962) A note on the analysis of repeated measurements of the same subjects. J Chronic Dis 15:969–977. https://doi.org/10.1016/0021-9681(62)90116-9
37. Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink J-L, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503. https://doi.org/10.1534/genetics.113.150227
38. Rincent R, Laloë D, Nicolas S et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L). Genetics 192:715–728. https://doi.org/10.1534/genetics.112.141473

39. Seck PA, Diagne A, Mohanty S, Wopereis MCS (2012) Crops that feed the world 7: Rice. Food Secur 4:7–24. https://doi.org/10.1007/s12571-012-0168-1
40. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink J-L, McCouch SR (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11:e1004982. https://doi.org/10.1371/journal.pgen.1004982
41. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink J-L, McCouch SR (2016a) Data from: Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. https://doi.org/10.5061/dryad.7369p. Version 1. Dryad [dataset]
42. Spindel JE, Begum H, Akdemir D, Collard B, Redoña E, Jannink J-L, McCouch S (2016b) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116:395–408. https://doi.org/10.1038/hdy.2015.113
43. VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980
44. Wang X, Yan J (2026) Selection-aware metrics provide complementary decision-level information in plant genomic prediction. Front Plant Sci 17:1951182. https://doi.org/10.3389/fpls.2026.1951182
45. Wei X, Chen M, Zhang Q et al (2024) Genomic investigation of 18,421 lines reveals the genetic architecture of rice. Science 385:eadm8762. https://doi.org/10.1126/science.adm8762
46. Werner CR, Gaynor RC, Gorjanc G, Hickey JM, Kox T, Abbadi A, Leckband G, Snowdon RJ, Stahl A (2020) How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. Front Plant Sci 11:592977. https://doi.org/10.3389/fpls.2020.592977
47. Zhu X, Leiser WL, Hahn V, Würschum T (2021) Training set design in genomic prediction with multiple biparental families. Plant Genome 14:e20124. https://doi.org/10.1002/tpg2.20124

Figures

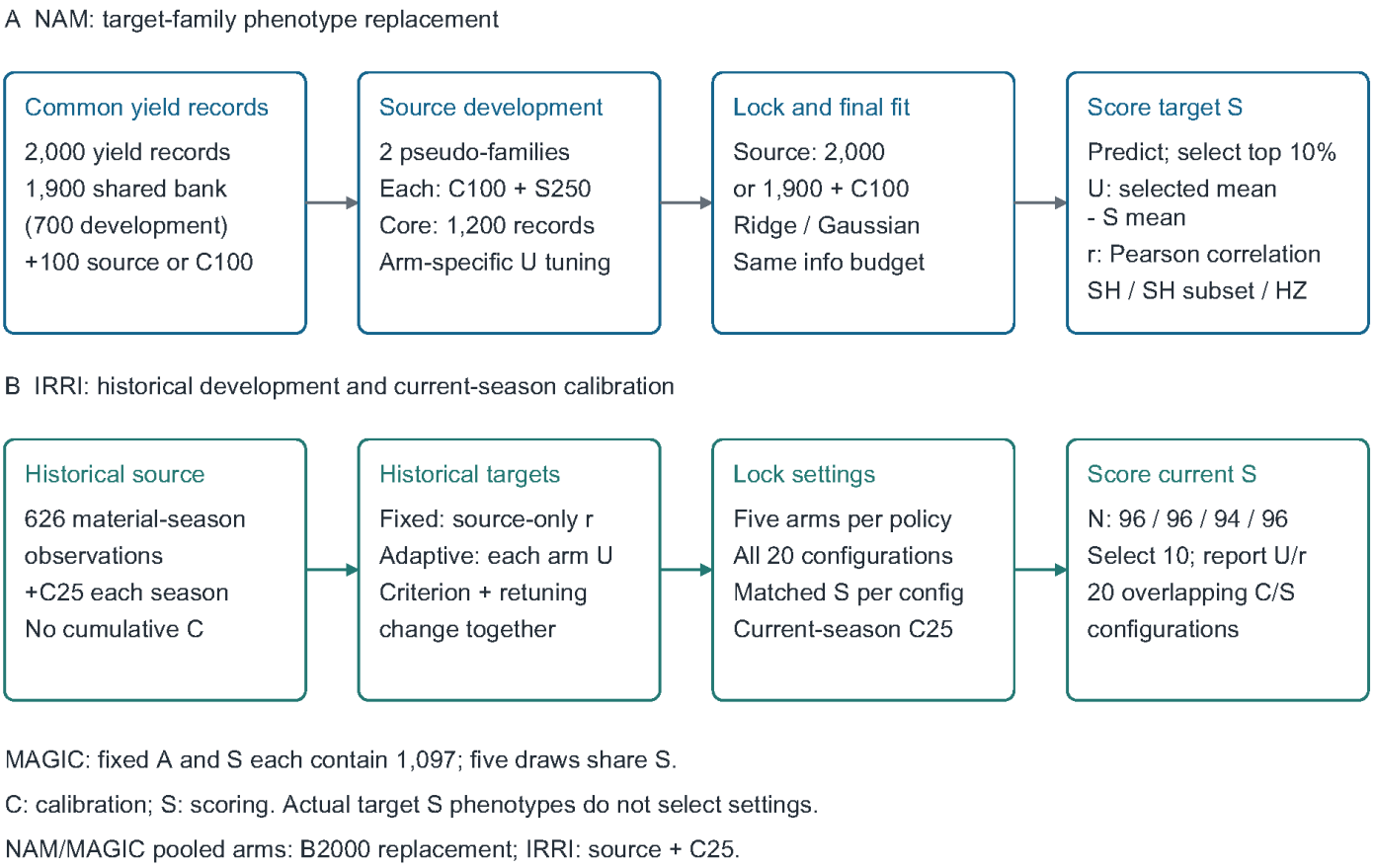


Figure 1

Deployment designs and phenotype budgets. For each NAM target family, the question is whether measuring a subset improves prediction of its held-out members at a fixed record budget. A, NAM design; B, IRRI design; the notes below the panels give MAGIC set sizes and abbreviations. Yield records flow from permitted source development to locked hyperparameters, final fitting and separate scoring. NAM/MAGIC substitute target yield records within 2000 permitted yield records, labeled B2000 in the diagram, including 700 source-development yield records; the target-only comparator uses 800. Development yield records may reenter final source banks. IRRI retains 626 source material-season observations and adds C25 for each season without accumulating earlier target seasons. The diagram is schematic. The original/fixed policy inherits source-only-r settings; the adaptive policy uses deployment-specific pseudo-target U. Within each task, target scoring records (S) are used only for evaluation, not for choosing parameters. IRRI configurations overlap; the analysis chronology is given in Additional file 1: Section S10.

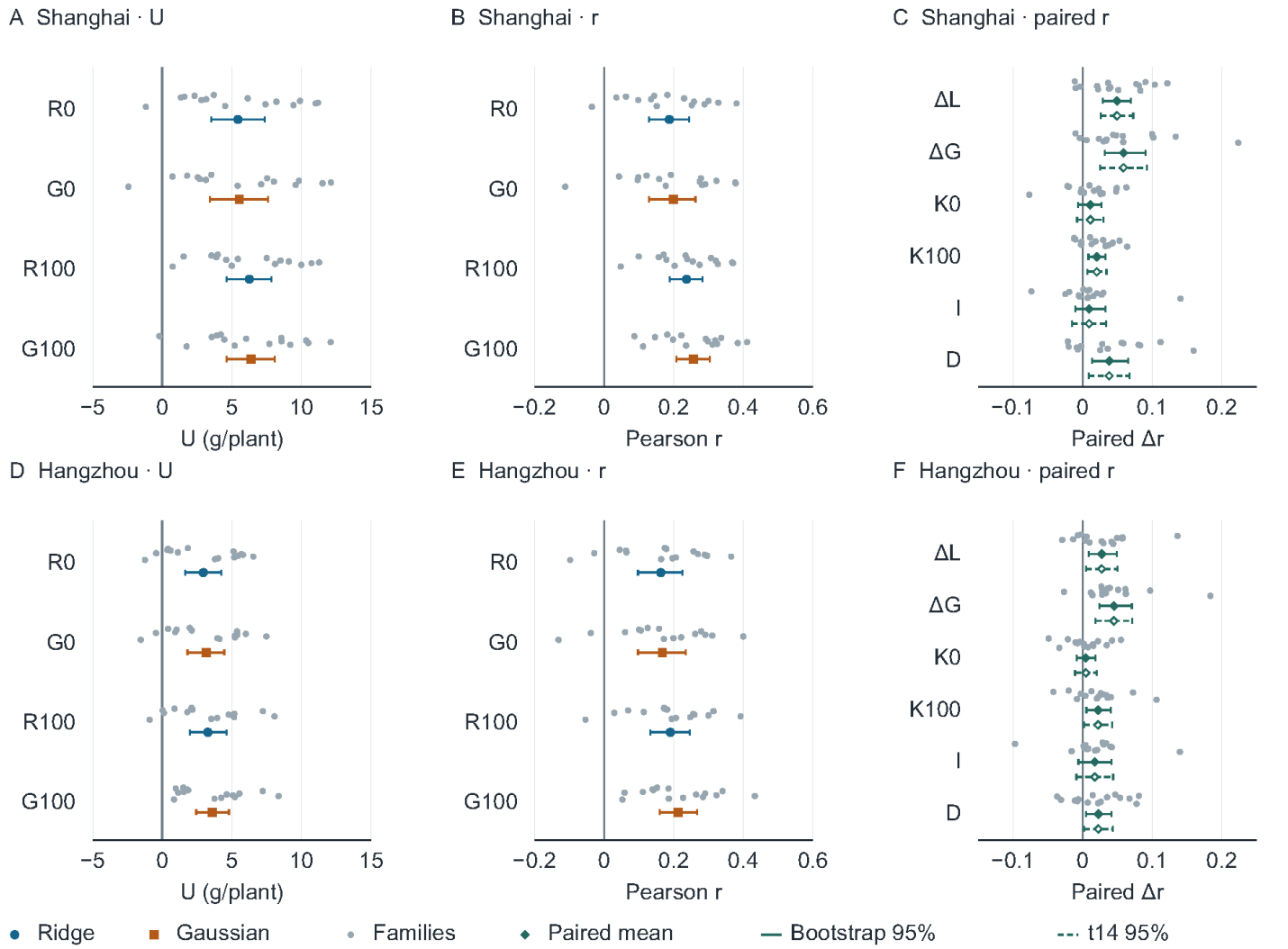


Figure 2

NAM absolute performance and paired predictive-ability contrasts. A/B/C show SH U/r/paired Δr; D/E/F show HZ U/r/paired Δr. Absolute panels show source and pooled ridge as blue circles and Gaussian models as orange squares; R0/G0 use no target yield records and R100/G100 use 100. U is g/plant. Absolute markers are means of 15 five-draw family means, with conditional 95% family-bootstrap intervals. Paired panels show all 15 family contrasts, their mean, solid bootstrap intervals and vertically offset dashed t_{14} intervals. ΔL and ΔG are the gains from target yield records for ridge and Gaussian prediction, K0 and K100 are the Gaussian increments without and with calibration, I is the interaction, and D is pooled ridge minus source Gaussian. Intervals are pointwise; the selected-six multiplicity sensitivity is in Additional file 1: Section S7. Paired U contrasts are in Additional file 1: Figure S11, and the original primary comparison is in Table 3 and Additional file 1: Figure S8.

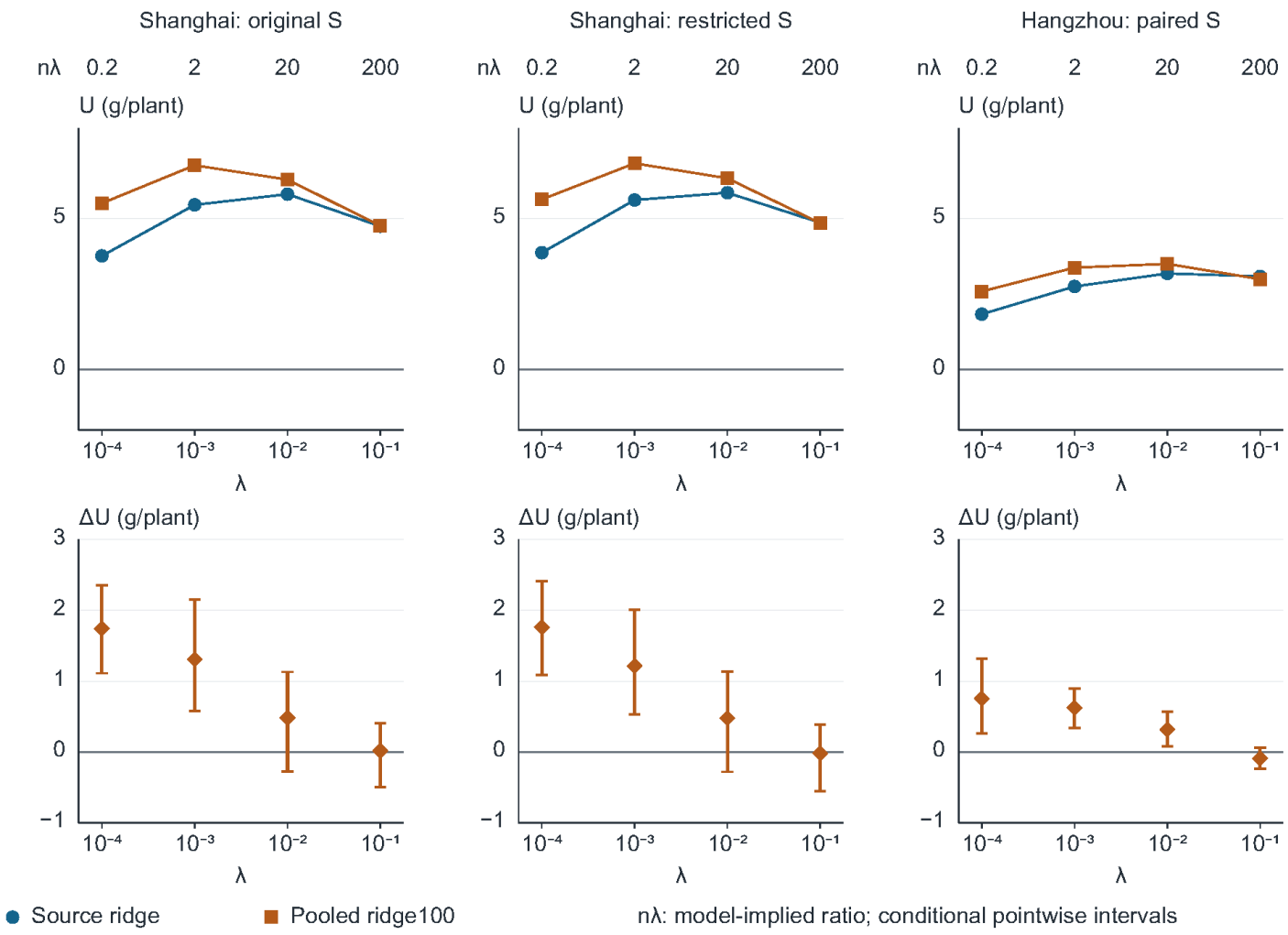
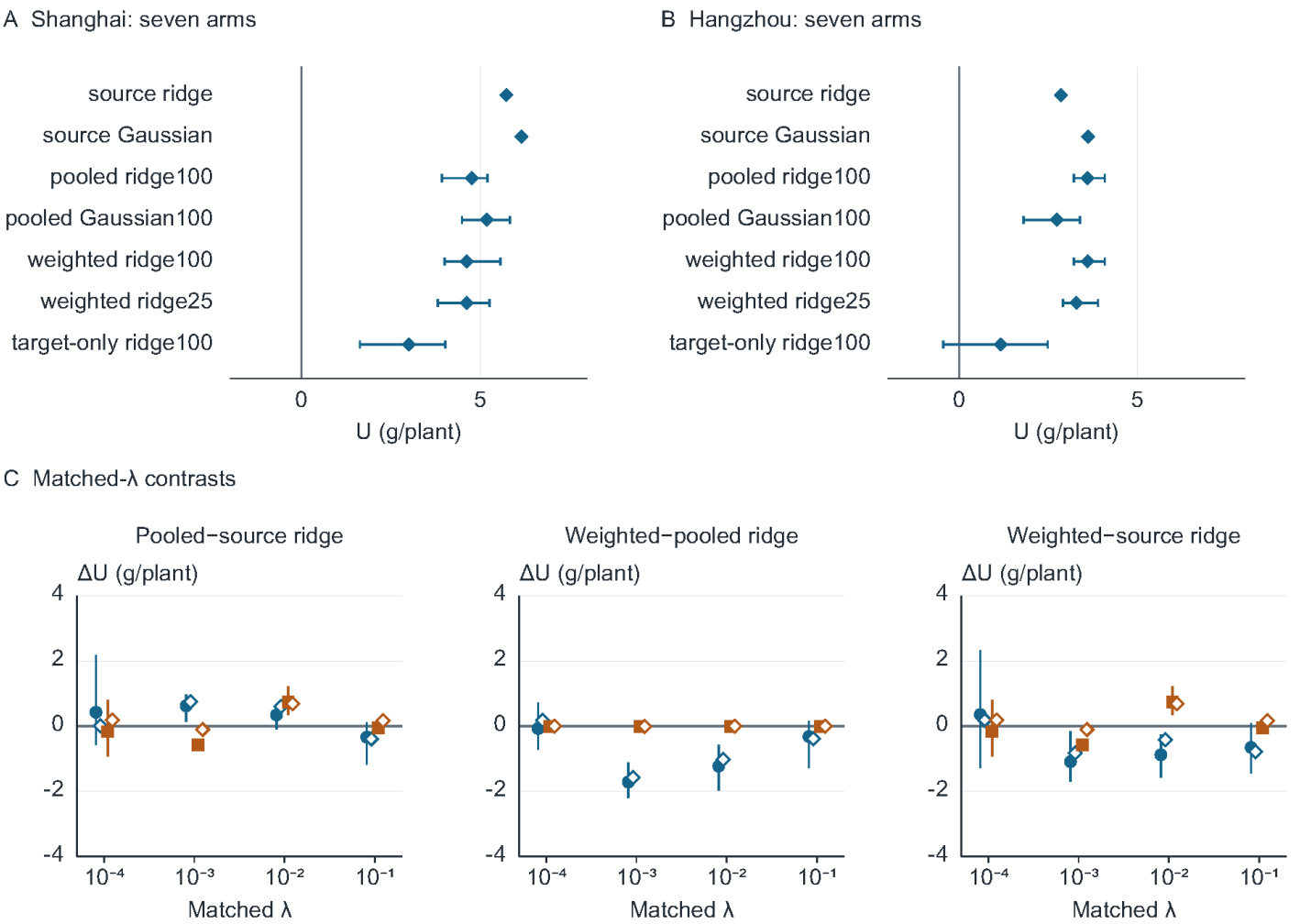


Figure 3

NAM pooling across the fixed regularization grid. Larger λ imposes stronger regularization, shrinking fitted effects more strongly. Columns show the SH original, restricted SH and HZ paired scoring views at all four λ values; the top row shows U and the bottom row the pool100 − ridge0 difference. U and pool100 − ridge0 differences are g/plant. Intervals are existing conditional pointwise 95% family-bootstrap summaries. Source-only fits use 2000 source yield records and pooling uses source1900+C100, with the same core transform. Five paired draws are averaged within family before equal aggregation. The nλ values 0.2/2/20/200 are model-implied variance-ratio conventions, not heritability estimates. This grid is a sensitivity analysis; deployment settings were selected by source-family pseudo-target U. Matched-λ predictive-ability contrasts are given in Table 4B.



SH: blue circles; HZ: orange squares. Open diamonds: no-high25.
A/B: observed five-draw ranges; C: matched-λ comparisons. Full proxy contrasts: Figure S10.

Figure 4

MAGIC transfer and sensitivity analyses. Panels A/B show seven-arm U means at SH/HZ; ranges describe five overlapping draws, not confidence intervals. Source-only predictions are fixed across draws. Target-only100 uses 800 permitted yield records. Panel C shows three matched-λ contrasts at four common settings: blue circles are SH, orange squares are HZ and offset open diamonds are the corresponding values after removing 25 high-proxy scoring IDs. Vertical bars in C show the observed range across five overlapping draws, not confidence intervals. The SH/HZ target masses are 0.50/0.05; HZ weighted and uniform pooling coincide at matched λ. Removing 25 high-proxy scoring IDs changes N/q from 1097/110 to 1072/108 without changing training, parameters or predictions. The proxy is not verified ancestry. The complete original-versus-reduced-domain contrast comparison is in Additional file 1: Figure S10.



Figure 5

IRRI correlation and selection contrasts across calibration partitions. Panels A and B show pool25 − source Gaussian0 and panels C and D show pool25 − source ridge0; A and C give ΔU (kg/ha), and B and D give dimensionless Δr. Both fixed and adaptive policies retain every C01–C20 configuration in each season; summary markers distinguish means and medians. Seasonal S sizes are 96/96/94/96 and q = 10. Historical original-allocation U markers are separate references, excluded from the 20-configuration summaries. C and S change together, configurations overlap, and seasons follow the same cohort. Thus ranges and sign counts describe this finite grid rather than independent validation populations. Policy differences jointly change the tuning criterion and calibrated-arm reoptimization. The complete five-arm metrics and policy differences are in Additional files 2 and 3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile2.xlsx
- Additionalfile3.zip