

Additional file 1 (Supplementary Materials)

Supplementary derivations, proofs, algorithm specification and results

A Finite-Sample Inferential Framework for the Weighted Kappa Coefficient (FWKC) in Binary Diagnostic Agreement Studies

Ismail Radi Ismail Fkhadah^{1,2,*}

¹ Department of Biostatistics, School of Medicine, University of Granada, Granada, Spain

² Universal Institute of Applied and Health Research, Nablus, State of Palestine

* Correspondence: ismailfk2021@correo.ugr.es

About this file

This file supports the revised manuscript and is organised to answer the methodological concerns raised by both reviewers. FWKC denotes the finite-sample inferential framework; $\kappa(c)$ remains the conventional binary weighted kappa estimand and is not redefined. The file gives (i) derivations or proofs for the formulas used in the main text, (ii) a complete specification of the FWKC algorithm and its safeguards, (iii) the simulation design and data-generating mechanisms, and (iv) the complete numerical results behind the tables and figures of the main text: 243 primary scenarios with 120 independent datasets each, 162 alternative-mechanism scenarios, 34 sensitivity variants, the $n = 620$ bridge, the Weiner et al. analysis and 126 AuditC dataset-by- c fits. Equations numbered (1)–(22) are those of the main text; equations numbered (S1), (S2), ... belong to this file. Numbers in square brackets refer to the reference list of the main text, and references prefixed with A are listed at the end of this file. Every numerical value was computed from the analysis output files released with the article, and every table was checked cell by cell against the summary workbook of the final analysis. The supplementary file of the original submission

was based on an analysis that has been withdrawn; Section S6 lists each of its items and states what was retained, what was corrected and what was withdrawn, and why.

Table S1 maps every numbered comment of Reviewer 1 and Reviewer 2 to the derivation, table or figure in this file and to the corresponding section of the revised manuscript. Table S2 fixes the notation used for the implemented analysis.

Table S1. Where each reviewer comment is addressed in this file

Comment	Topic	This file	Main text
R1-1	Bias introduced by the pipeline	S3.2 (Proposition 12), S3.6, Tables S9, S23	3.3, Table 10
R1-2	Choice of Beta distributions	S3.1 (Propositions 9, 10), Tables S8, S21	2.6.1, 3.5
R1-3	Fusion rule	S3.6 (Propositions 15, 16), Tables S11, S21	2.6.5, 3.5
R1-4	Random-forest inputs and overfitting	S3.5, S3.10 (Step 4), Table S21	2.6.4, Table 2
R1-5	Transformation of kappa	S2.3 (Proposition 6), S3.3 (Proposition 13)	2.5, 2.6.3, 3.5
R1-6	Covariance across pseudo-raters	S3.2 (Proposition 11), S3.9 (Proposition 19)	2.6.6, 3.3
R1-7	Threshold n_0	S3.8 (Propositions 17, 18), Tables S12, S21	2.6.6, 3.5
R1-8	Single anchor dataset	S4.2, S5.4–S5.7, Tables S19, S20, S16, S22	2.9, 3.4, 3.6
R1-9	Misspecification of the DGM	S4.2 (Propositions 21, 22), Table S19	2.9, 3.4
R1-10	Choice of comparison test	S4.5 (Proposition 24), Table S15	2.8
R1-11	Theoretical support for coverage	S2.3, S3.8 (Proposition 18), Table S12	4.2
R1-12	Computational cost	S3.11 (Proposition 20), Table S13	3.7
R1-13	Small real-data difference	S5.1, S5.5, Tables S9, S20	3.1, 3.4
R1-14	Ordinal and multi-class settings	S1.3 (Remark 1)	4.6, Future work
R1-15	Prevalence-adjusted measures	S2.4 (Proposition 8), Tables S5, S22	3.1, 4.5
R2-1	Is $\kappa(c)$ a weighted kappa?	S1.2–S1.5 (Propositions 1–4), Tables S3, S4	2.4, 3.1
R2-2	Beta parameterisation	S3.1 (Propositions 9, 10), Table S8	2.6.1
R2-3	Coverage from one dataset	S4.3 (Proposition 23), Tables S13, S23	2.7, 3.2
R2-4	Random-forest specification	S3.5, S3.10 (Step 4)	Table 2
R2-5	Circular grid, variance claim	S4.2, S4.6, S5.2–S5.4, Figures S5–S8	3.3, 3.4, Table 11
R2-6	Figure, single dataset, validation	S5.7, Tables S16, S22, Figure S11; S6	3.6, Declarations

Comment	Topic	This file	Main text
R2-7	Overall assessment: novelty, competitors, design	Synthesis of S1–S5 (estimand, Beta, coverage, RF, grid)	2.5, 4.6, Future Work

R1-k and R2-k denote comment k of Reviewer 1 and Reviewer 2.

Table S2. Notation used in the main text and in this file

Symbol	Meaning	Defined in
D, T	Disease status by the gold standard; binary test result	Section 2.3
n_{dt}, π_{dt}	Count and probability of cell ($D = d, T = t$); $n_{11} = \text{TP}$, $n_{10} = \text{FN}$, $n_{01} = \text{FP}$, $n_{00} = \text{TN}$	Table 1
n_{d+}, n_{+t}, n	Row margins, column margins and total	Table 1
p, Se, Sp	Prevalence $P(D = 1)$, sensitivity, specificity	Eq. (1)
Q	Probability of a positive test, $pSe + (1 - p)(1 - Sp)$	Eq. (2)
J	Youden index, $Se + Sp - 1$	Section S1.1
c	Weighting index, $L_{FN}/(L_{FN} + L_{FP})$	Section 2.4
w_{dt}, A_o, A_e	Agreement weights; observed and chance-expected weighted agreement	Eq. (3)
$\kappa(c), \hat{\kappa}_D(c)$	Estimand; direct plug-in estimator	Eqs. (6), (7)
η	$\log\{(1 + \kappa)/(1 - \kappa)\} = 2 \operatorname{artanh}\kappa$	Eqs. (9), (12)
$j = 1, \dots, n_{mc}$	Posterior draw of (Se_j, Sp_j)	Eq. (10)
$r = 1, \dots, R$	Pseudo-rater ($R = 3$)	Eq. (11)
κ_{rj}^*, η_{rj}	Pseudo-rater kappa and its transform	Eqs. (11), (12)
$\mu, u_r, v_j, \varepsilon_{rj}$	Grand mean, rater effect, draw effect, residual	Eq. (14)
$\tau_r^2, \tau_{mc}^2, \sigma^2$	Variances of $u_r, v_j, \varepsilon_{rj}$	Eq. (14)
$\hat{\kappa}_B, SE_B$	Bayesian component and its posterior standard deviation	Section 2.6.3
$\hat{\kappa}_{RF}, SE_{RF}$	Random-forest component and the standard deviation of its predictions	Table 2
w_B	Fusion weight of the Bayesian component	Eq. (15)
$\hat{\kappa}_{FWKC}(c), \hat{\eta}_{FWKC}$	Fused estimate and its transform	Eqs. (15), (17)
$b = 1, \dots, B$	Bootstrap resample $D^{*(b)}$	Eq. (16)
$SE_{\eta}^*, \lambda_n, n_0$	Bootstrap SE of $\hat{\eta}_{FWKC}$; expansion factor; threshold	Eq. (17)

Symbol	Meaning	Defined in
$SE_{cov}, \hat{\Sigma}_K, \epsilon$	Covariance diagnostic, pseudo-rater covariance, ridge constant (10^{-10})	Eq. (18)
ϵ_t	Truncation constant for pseudo-rater values (10^{-8})	Section S3.3
$m = 1, \dots, M$	Independent simulated dataset	Eqs. (19)–(21)
$\Delta_m, \delta_m, \hat{\Delta}$	Paired differences between FWKC and direct estimates	Eqs. (21), (22)
$z_{0.975}, \Phi$	Standard normal quantile and distribution function	—

The indices r (pseudo-rater) and j (posterior draw), the draw effect v_j and the variance parameters τ_r^2 , τ_{mc}^2 and σ^2 are those of the implementation; Eqs. (11)–(14) of the revised main text use the same notation. The index m is reserved for simulated datasets.

S1. The estimand: $\kappa(c)$ is Weighted Kappa Coefficient

S1.1 Setting

Let D denote disease status by the gold standard and T the binary test result, and let $\pi_{dt} = P(D = d, T = t)$ for $d, t \in \{0,1\}$, with margins π_{d+} and π_{+t} . With $p = \pi_{1+}$, $Se = \pi_{11}/\pi_{1+}$ and $Sp = \pi_{00}/\pi_{0+}$, the four cell probabilities are

$$\pi_{11} = p Se, \quad \pi_{10} = p (1 - Se), \quad \pi_{01} = (1 - p)(1 - Sp), \quad \pi_{00} = (1 - p) Sp, \quad (S1)$$

and the column margins are $\pi_{+1} = Q = p Se + (1 - p)(1 - Sp)$ (Eq. 2) and $\pi_{+0} = 1 - Q$. The Youden index is written $J = Se + Sp - 1$.

S1.2 Derivation from the Cohen weight matrix

Let $L_{FN} > 0$ and $L_{FP} > 0$ be the losses attached to a false negative and a false positive, and let correct classifications carry no loss. Dividing by $L_{FN} + L_{FP}$ and setting $w_{dt} = 1 - (\text{normalised loss})_{dt}$ gives the agreement weights

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{10} \\ w_{01} & w_{00} \end{pmatrix} = \begin{pmatrix} 1 & 1 - c \\ c & 1 \end{pmatrix}, \quad c = \frac{L_{FN}}{L_{FN} + L_{FP}}, \quad (S2)$$

with rows indexed by $D = 1,0$ and columns by $T = 1,0$. The weighting index enters only the two off-diagonal (disagreement) cells; both agreement cells have weight one. Cohen's weighted kappa [2] with these weights is $\kappa_w = (A_o - A_e)/(1 - A_e)$, with $A_o = \sum_{d,t} w_{dt} \pi_{dt}$ and $A_e = \sum_{d,t} w_{dt} \pi_{d+} \pi_{+t}$ (Eq. 3).

Proposition 1 ($\kappa(c)$ is a weighted kappa). For every (p, Se, Sp) with $0 < p < 1$ and $0 < Q < 1$, Cohen's weighted kappa with the weights (S2) equals $\kappa(c)$ of Eq. (6):

$$\kappa_w = \kappa(c) = \frac{p(1-p)J}{c p (1-Q) + (1-c)(1-p)Q}. \quad (S3)$$

Proof. Because $\sum_{d,t} \pi_{dt} = 1$ and $\sum_{d,t} \pi_{d+} \pi_{+t} = 1$, the weighted disagreement observed and that expected by chance are

$$1 - A_o = \sum_{d,t} (1 - w_{dt}) \pi_{dt} = c \pi_{10} + (1 - c) \pi_{01}, \quad 1 - A_e = c \pi_{1+} \pi_{+0} + (1 - c) \pi_{0+} \pi_{+1}. \quad (S4)$$

Substituting (S1) gives $1 - A_o = c p (1 - Se) + (1 - c)(1 - p)(1 - Sp)$ and $1 - A_e = c p (1 - Q) + (1 - c)(1 - p)Q$, which is Eq. (4). Since $\kappa_w = \{(1 - A_e) - (1 - A_o)\} / (1 - A_e)$ (Eq. 5), the numerator is

$$(1 - A_e) - (1 - A_o) = c p (Se - Q) + (1 - c)(1 - p)\{Q - (1 - Sp)\}. \quad (S5)$$

Two identities follow directly from Eq. (2):

$$Se - Q = (1 - p)J, \quad Q - (1 - Sp) = pJ. \quad (S6)$$

Hence the numerator equals $c p (1 - p)J + (1 - c)(1 - p)pJ = p(1 - p)J$, and (S3) follows. $\square$

Proposition 2 (normalising the losses loses no generality). Let $w_{11} = w_{00} = 1$, $w_{10} = 1 - a$ and $w_{01} = 1 - b$ with $a, b > 0$. Then $\kappa_w = \kappa\{a/(a + b)\}$.

Proof. By (S4), $1 - A_o = a \pi_{10} + b \pi_{01} = (a + b)\{c \pi_{10} + (1 - c)\pi_{01}\}$ with $c = a/(a + b)$, and likewise $1 - A_e = (a + b)\{c \pi_{1+} \pi_{+0} + (1 - c)\pi_{0+} \pi_{+1}\}$. The common factor $(a + b)$ cancels in $\kappa_w = 1 - (1 - A_o)/(1 - A_e)$. $\square$

Propositions 1 and 2 answer the question of whether $\kappa(c)$ is a weighted kappa: it is Cohen's weighted kappa for the only two-category weight matrix with unit agreement weights, up to the scale of the losses. The loss-ratio form of Eq. (5) is the same statement, $1 - A_o$ being the expected normalised loss and $1 - A_e$ the loss expected under independence of D and T .

S1.3 Equivalent representations and properties

Proposition 3 (count form). Replacing the cell probabilities by the observed proportions n_{dt}/n in (S3) gives the direct estimator of Eq. (7),

$$\hat{\kappa}_D(c) = \frac{n_{11}n_{00} - n_{10}n_{01}}{c n_{1+}n_{0+} + (1-c) n_{0+}n_{1+}}. \quad (S7)$$

Proof. In terms of cell probabilities, $p(1-p)J = \pi_{1+}\pi_{0+}(\pi_{11}/\pi_{1+} + \pi_{00}/\pi_{0+} - 1) = \pi_{11}\pi_{0+} + \pi_{00}\pi_{1+} - \pi_{1+}\pi_{0+}$. Expanding the margins,

$$p(1-p)J = \pi_{11}(\pi_{01} + \pi_{00}) + \pi_{00}(\pi_{11} + \pi_{10}) - (\pi_{11} + \pi_{10})(\pi_{01} + \pi_{00}) = \pi_{11}\pi_{00} - \pi_{10}\pi_{01}, \quad (S8)$$

and the denominator of (S3) is $c \pi_{1+}\pi_{0+} + (1-c)\pi_{0+}\pi_{1+}$. Numerator and denominator are homogeneous of degree two in the cell probabilities, so substituting n_{dt}/n and multiplying both by n^2 gives (S7). $\square$

Proposition 4 (properties of $\kappa(c)$). Assume $0 < p < 1$ and $0 < Q < 1$, and write $\kappa(0) = pJ/Q = \{Sp - (1-Q)\}/Q$ and $\kappa(1) = (1-p)J/(1-Q) = (Se - Q)/(1-Q)$.

- (a) *Cohen's kappa.* $\kappa(1/2)$ equals Cohen's unweighted kappa [1].
- (b) *Sign and extremes.* $\kappa(c)$ has the sign of J ; $\kappa(c) = 0$ if and only if $Se + Sp = 1$; for $0 < c < 1$, $\kappa(c) = 1$ if and only if $Se = Sp = 1$ (at the end points, $\kappa(1) = 1$ if and only if $Se = 1$, and $\kappa(0) = 1$ if and only if $Sp = 1$).
- (c) *Weighted average.*

$$\kappa(c) = \frac{c p (1-Q) \kappa(1) + (1-c)(1-p) Q \kappa(0)}{c p (1-Q) + (1-c)(1-p) Q}, \quad (S9)$$

so $\kappa(c)$ lies between $\kappa(0)$ and $\kappa(1)$.

- (d) *Monotonicity in c .*

$$\frac{\partial \kappa(c)}{\partial c} = - \frac{p(1-p)J(p-Q)}{\{c p (1-Q) + (1-c)(1-p) Q\}^2}, \quad (S10)$$

so for $J > 0$ $\kappa(c)$ decreases in c when $Q < p$, increases when $Q > p$, and is constant and equal to J when $Q = p$.

- (e) *Range.* If $J \geq 0$ then $0 \leq \kappa(c) \leq 1$. If $J < 0$, $\kappa(c)$ is negative and, unlike Cohen's kappa, is not bounded below by -1 when $c \neq 1/2$.

Proof. (a) For a 2×2 table, $P_o - P_e = (\pi_{11} + \pi_{00}) - (\pi_{1+}\pi_{+1} + \pi_{0+}\pi_{+0}) = 2(\pi_{11}\pi_{00} - \pi_{10}\pi_{01})$ and $1 - P_e = \pi_{1+}\pi_{+0} + \pi_{0+}\pi_{+1}$; by (S8) Cohen's kappa is therefore (S7) with $c = 1/2$ in population form. (b) The denominator is a convex combination of the positive numbers $p(1-Q)$ and $(1-p)Q$, so the sign is

that of J . If $Se = Sp = 1$ then $Q = p$, the denominator is $p(1 - p)$ and $\kappa(c) = 1$; conversely, for $0 < c < 1$ both weights in (c) are positive and both $\kappa(0)$ and $\kappa(1)$ are at most one by (e), so $\kappa(c) = 1$ requires $\kappa(0) = \kappa(1) = 1$, that is $Sp = 1$ and $Se = 1$; the end-point statements follow from $\kappa(1) = (Se - Q)/(1 - Q)$ and $\kappa(0) = \{Sp - (1 - Q)\}/Q$. (c) $c p(1 - Q)\kappa(1) = c p(1 - p)J$ and $(1 - c)(1 - p)Q \kappa(0) = (1 - c)p(1 - p)J$; their sum is the numerator of (S3). (d) Differentiate the denominator: $\partial/\partial c\{c p(1 - Q) + (1 - c)(1 - p)Q\} = p(1 - Q) - (1 - p)Q = p - Q$. (e) If $J \geq 0$, $\kappa(1) = (Se - Q)/(1 - Q) \leq 1$ because $Se \leq 1$, and $\kappa(0) = \{Sp - (1 - Q)\}/Q \leq 1$ because $Sp \leq 1$; (c) gives the bound. For $J < 0$ take $Se = Sp = 0$, so that $J = -1$, $Q = 1 - p$ and $\kappa(c) = -p(1 - p)/\{c p^2 + (1 - c)(1 - p)^2\}$. Then $\kappa(c) < -1$ if and only if $f(p) = p(1 - p) - c p^2 - (1 - c)(1 - p)^2 > 0$. The quadratic f has leading coefficient -2 and roots $p = \frac{1}{2}$ and $p = 1 - c$, so $\kappa(c) < -1$ for every p strictly between $\frac{1}{2}$ and $1 - c$, an interval that is non-empty whenever $c \neq \frac{1}{2}$.

For the Weiner et al. data ($\hat{Q} = 0.798 < \hat{p} = 0.894, \hat{f} = 0.520$), (d) predicts that $\hat{\kappa}_D(c)$ decreases in c ; it falls from 0.524 at $c = 0.1$ to 0.290 at $c = 0.9$ (Table 4 of the main text). Property (e) is the reason pseudo-rater values are truncated to $[-1 + \epsilon_t, 1 - \epsilon_t]$ before transformation (Section S3.3).

Remark 1 (more than two categories). Propositions 1 and 2 are the two-category case of Cohen's weighted kappa for a $K \times K$ weight matrix. An FWKC-type pipeline for K ordered or unordered categories would replace the two Beta posteriors of Eq. (10) by one Dirichlet posterior per disease category, $(\pi_{t|d})_{t=1}^K \sim \text{Dirichlet}(n_{d1} + 1/2, \dots, n_{dK} + 1/2)$, and generate pseudo-rater tables row by row from multinomial distributions; the pooling, forest, fusion and bootstrap layers would be unchanged. This extension has not been implemented or evaluated, and the main text makes no claim about it.

S1.4 Numerical verification

The population form (S3), the weight-matrix form κ_w , the loss form $1 - (1 - A_o)/(1 - A_e)$ and the count form (S7) were evaluated on 250 randomly generated 2×2 tables with c drawn uniformly from (0,1). The largest absolute difference between any two forms over the 250 tables was 9.2×10^{-1} , that is, floating-point precision, and $\hat{\kappa}_D(1/2)$ equalled Cohen's kappa to the same precision. Table S3 shows the first twelve tables.

Table S3. Numerical verification of the four representations of $\kappa(c)$: first 12 of 250 random tables

$n_{11}/n_{10}/n_{01}/n_{00}$	c	Count (S7)	Loss (5)	Weight matrix (3)	Population (6)	$\hat{\kappa}_D\left(\frac{1}{2}\right)$	Cohen	Max. diff.
8/6/46/86	0.176	0.067558	0.067558	0.067558	0.067558	0.097909	0.097909	4.2e-16
48/54/12/34	0.801	0.127115	0.127115	0.127115	0.127115	0.167689	0.167689	1.7e-16
22/30/27/60	0.082	0.118686	0.118686	0.118686	0.118686	0.114056	0.114056	5.7e-16
16/58/45/59	0.634	-0.213285	-0.213285	-0.213285	-0.213285	-0.222104	-0.222104	8.3e-17
12/21/15/77	0.142	0.235058	0.235058	0.235058	0.235058	0.213012	0.213012	1.4e-16
20/5/20/50	0.502	0.431616	0.431616	0.431616	0.431616	0.431138	0.431138	3.3e-16
12/38/49/15	0.937	-0.616217	-0.616217	-0.616217	-0.616217	-0.513274	-0.513274	1.1e-16
30/51/7/81	0.663	0.253130	0.253130	0.253130	0.253130	0.297247	0.297247	1.1e-16
80/1/59/67	0.343	0.395661	0.395661	0.395661	0.395661	0.460516	0.460516	2.2e-16
75/53/21/96	0.407	0.422132	0.422132	0.422132	0.422132	0.401729	0.401729	1.7e-16
15/6/29/38	0.080	0.142147	0.142147	0.142147	0.142147	0.204545	0.204545	2.8e-17
20/6/15/45	0.544	0.482429	0.482429	0.482429	0.482429	0.472855	0.472855	3.3e-16

Max. diff. is the largest absolute difference among the four representations and between $\hat{\kappa}_D\left(\frac{1}{2}\right)$ and Cohen's kappa in that table.

Over all 250 tables the largest difference was 9.2×10^{-16} .

S1.5 Reproduction of the published analysis

The point values printed in Table 3 of [3] for the Weiner et al. data are not reproduced by the formula and counts printed in the same paper (Table S4). The discrepancy is systematic, grows with c , and is 0.003 at $c = 0.1$ and 0.041 at $c = 0.9$. The confidence limits printed in [3] are, in contrast, reproduced to the third decimal at eight of the nine values of c by the logit interval recomputed from our estimates (Table 5 of the main text and Table S17), which indicates that the printed point values contain an inconsistency and that our implementation of the estimator is correct.

Table S4. Point values printed in [3] and values computed from the formula and counts printed there

c	Printed in [3, Table 3]	Computed from printed counts	Absolute difference
0.1	0.527	0.524	0.003
0.2	0.462	0.476	0.014
0.3	0.412	0.436	0.024

c	Printed in [3, Table 3]	Computed from printed counts	Absolute difference
0.4	0.371	0.402	0.031
0.5	0.338	0.373	0.035
0.6	0.310	0.348	0.038
0.7	0.287	0.327	0.040
0.8	0.267	0.307	0.040
0.9	0.249	0.290	0.041

Weiner et al. data [11], $n = 620$. The computed values are those of Table 4 of the main text.

S2. Direct estimator and closed-form intervals

S2.1 The Weiner et al. data

Table S5 gives the observed table of the motivating example [11] and the quantities derived from it that are used in the main text and in this file.

Table S5. Weiner et al. data: observed table and derived quantities

Quantity	Value
True positives, n_{11}	473
False negatives, n_{10}	81
False positives, n_{01}	22
True negatives, n_{00}	44
Total, n	620
Prevalence, $\hat{p} = n_{1+}/n$	0.8935
Sensitivity, $\hat{Se}$	0.8538
Specificity, $\hat{Sp}$	0.6667
Positive-test probability, $\hat{Q} = n_{+1}/n$	0.7984
Youden index, $\hat{J}$	0.5205
Positive predictive value	0.9556

Quantity	Value
Negative predictive value	0.3520
Observed agreement, P_o	0.8339
$\hat{\kappa}(0) = \{\hat{S}p - (1 - \hat{Q})\}/\hat{Q}$	0.5825
$\hat{\kappa}(1) = (\hat{S}e - \hat{Q})/(1 - \hat{Q})$	0.2748
Cohen's kappa, $\hat{\kappa}_D \left(\frac{1}{2}\right)$	0.3734
PABAK, (S16)	0.6677
Gwet's AC1, (S16)	0.7753

Coronary artery disease (gold standard: angiography) and exercise stress test [11]. Because $\hat{Q} < \hat{p}$, $\hat{\kappa}_D(c)$ decreases from $\hat{\kappa}(0)$ towards $\hat{\kappa}(1)$ as c increases (Proposition 4(d)).

S2.2 Delta-method variance in closed form

Write $\boldsymbol{\pi} = (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})^T$ and $\kappa(c) = g(\boldsymbol{\pi}) = N(\boldsymbol{\pi})/D_c(\boldsymbol{\pi})$ with

$$N = \pi_{11}\pi_{00} - \pi_{10}\pi_{01}, \quad D_c = c\pi_{1+}\pi_{+0} + (1-c)\pi_{0+}\pi_{+1}. \quad (\text{S11})$$

The main text evaluates the gradient in Eq. (8) numerically. It has a closed form.

Proposition 5 (gradient and variance). The partial derivatives of g are $g_{dt} = (N_{dt}D_c - N D_{c,dt})/D_c^2$, where

$$(N_{11}, N_{10}, N_{01}, N_{00}) = (\pi_{00}, -\pi_{01}, -\pi_{10}, \pi_{11}), \quad (\text{S12})$$

$$D_{c,11} = c\pi_{+0} + (1-c)\pi_{0+}, \quad D_{c,10} = c(\pi_{1+} + \pi_{+0}), \quad (\text{S13})$$

$$D_{c,01} = (1-c)(\pi_{0+} + \pi_{+1}), \quad D_{c,00} = c\pi_{1+} + (1-c)\pi_{+1}.$$

Moreover $\sum_{d,t} \pi_{dt} g_{dt} = 0$, so that Eq. (8) reduces to

$$\widehat{\text{Var}}\{\hat{\kappa}_D(c)\} = \frac{1}{n} \sum_{d,t} \hat{\pi}_{dt} \hat{g}_{dt}^2. \quad (\text{S14})$$

Proof. (S12) is immediate. For (S13), π_{11} enters π_{1+} and π_{+1} , π_{10} enters π_{1+} and π_{+0} , π_{01} enters π_{0+} and π_{+1} , and π_{00} enters π_{0+} and π_{+0} ; differentiate each product. N and D_c are homogeneous of degree two, so g is homogeneous of degree zero, and Euler's theorem gives $\sum_{d,t} \pi_{dt} \partial g / \partial \pi_{dt} = 0 \cdot g = 0$. Finally $\nabla g^T \{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\} \nabla g = \sum \pi_{dt} g_{dt}^2 - (\sum \pi_{dt} g_{dt})^2$.

For the Weiner et al. data the closed form (S14) and the numerical gradient used in the analysis give the same standard error to six decimals at every c (Table S6), and the Euler identity holds to machine precision. The closed form is the variance of Bloch and Kraemer [22] and Roldán-Nofuentes et al. [9] written in cell probabilities.

Table S6. Closed-form delta-method standard error and the transformed interval for the Weiner et al. data

c	$\hat{\kappa}_D(c)$	SE, closed form (S14)	SE, numerical gradient	Eq. (9) interval	Fisher-z interval (S15)
0.1	0.5238	0.060716	0.060716	(0.3949, 0.6325)	(0.3949, 0.6325)
0.2	0.4759	0.055384	0.055384	(0.3604, 0.5770)	(0.3604, 0.5770)
0.3	0.4360	0.052042	0.052042	(0.3287, 0.5322)	(0.3287, 0.5322)
0.4	0.4023	0.049719	0.049719	(0.3006, 0.4950)	(0.3006, 0.4950)
0.5	0.3734	0.047924	0.047924	(0.2759, 0.4633)	(0.2759, 0.4633)
0.6	0.3484	0.046408	0.046408	(0.2544, 0.4359)	(0.2544, 0.4359)
0.7	0.3265	0.045050	0.045050	(0.2356, 0.4118)	(0.2356, 0.4118)
0.8	0.3073	0.043788	0.043788	(0.2191, 0.3904)	(0.2191, 0.3904)
0.9	0.2901	0.042594	0.042594	(0.2046, 0.3712)	(0.2046, 0.3712)

The numerical standard error is the one used in the analysis. The two interval columns agree to every printed digit, as Proposition 6 requires, and equal the transformed intervals of Table 5 of the main text.

S2.3 The transformed interval of Eq. (9) and the logit interval of [3, 9]

Proposition 6 (the transformed interval is the Fisher-z interval). Let $\eta = \log\{(1 + \kappa)/(1 - \kappa)\}$. Then $\eta = 2\text{artanh}\kappa$, $d\eta/d\kappa = 2/(1 - \kappa^2)$ and $\widehat{SE}(\hat{\eta}) = 2\widehat{SE}\{\hat{\kappa}_D(c)\}/\{1 - \hat{\kappa}_D(c)^2\}$. The interval of Eq. (9) is identical to the Fisher-z interval,

$$\tanh\left\{\frac{\hat{\eta} \pm z_{0.975}\widehat{SE}(\hat{\eta})}{2}\right\} = \tanh\left\{\text{artanh}\hat{\kappa}_D(c) \pm z_{0.975}\frac{\widehat{SE}\{\hat{\kappa}_D(c)\}}{1 - \hat{\kappa}_D(c)^2}\right\}. \quad (\text{S15})$$

It (i) always lies inside $(-1, 1)$; (ii) is defined whenever $|\hat{\kappa}_D(c)| < 1$, including $\hat{\kappa}_D(c) \leq 0$; (iii) is asymmetric about $\hat{\kappa}_D(c)$, with the shorter arm on the side of the nearer boundary; and (iv) has asymptotic coverage $1 - \alpha$ whenever $\sqrt{n}\{\hat{\kappa}_D(c) - \kappa(c)\}$ is asymptotically normal with a consistently estimated variance and $|\kappa(c)| < 1$.

Proof. $\log\{(1 + \kappa)/(1 - \kappa)\} = 2\text{artanh}\kappa$ is the definition of the inverse hyperbolic tangent; differentiate to obtain the derivative. Substituting $\hat{\eta} = 2\text{artanh}\hat{\kappa}_D$ and $\widehat{SE}(\hat{\eta})$ into the left side of (S15) and dividing by two gives the right side. (i) holds because $\tanh$ maps $\mathbb{R}$ into $(-1,1)$; (ii) because artanh is defined on $(-1,1)$; (iii) with $a = \text{artanh}\hat{\kappa}_D > 0$ and half-width $h > 0$ on the artanh scale,

$$\tanh(a + h) + \tanh(a - h) = \frac{\sinh 2a}{\cosh(a + h)\cosh(a - h)} = \frac{\sinh 2a}{\cosh^2 a + \sinh^2 h} \leq \frac{\sinh 2a}{\cosh^2 a} = 2\tanh a,$$

so the upper arm $\tanh(a + h) - \tanh a$ is no longer than the lower arm $\tanh a - \tanh(a - h)$; the case $a < 0$ is symmetric. (iv) follows from the delta method and the continuous-mapping theorem, since artanh is continuously differentiable on $(-1,1)$ and its inverse is continuous.

Proposition 6 is the basis for the attribution in Section 2.5 of the main text: Eq. (9) is the Fisher-z interval [A1] recommended for weighted kappa coefficients in general [23], applied here to the $\kappa(c)$ of a binary diagnostic test. The identity is exact in the implementation (Table S6).

Proposition 7 (the logit interval). The interval of Roldán-Nofuentes et al. [3, 9] is based on $l = \log\{\kappa/(1 - \kappa)\}$ with $\widehat{SE}(\hat{l}) = \widehat{SE}\{\hat{\kappa}_D(c)\}/[\hat{\kappa}_D(c)\{1 - \hat{\kappa}_D(c)\}]$. It is defined if and only if $0 < \hat{\kappa}_D(c) < 1$, and it always lies in $(0,1)$; $\hat{\kappa}_D(c) = 1$ occurs when $n_{10} = n_{01} = 0$. The event $\hat{\kappa}_D(c) \leq 0$ does not depend on c , because the sign of (S7) is that of $n_{11}n_{00} - n_{10}n_{01}$.

Proof. The logarithm requires $\kappa/(1 - \kappa) > 0$; the inverse logit maps into $(0,1)$; the denominator of (S7) is positive whenever it is defined.

The probability that the logit interval is unavailable can therefore be computed exactly, by summing over $n_{1+} \sim \text{Bin}(n, p)$, $n_{11} \mid n_{1+} \sim \text{Bin}(n_{1+}, Se)$ and $n_{00} \mid n_{0+} \sim \text{Bin}(n_{0+}, Sp)$ the indicator that $n_{11}n_{00} \leq n_{10}n_{01}$, that $n_{10} = n_{01} = 0$, or that a disease margin is zero. Table S7 gives it under the data-generating mechanism of the primary grid. At $n = 30$ with $p = 0.9$ the interval is unavailable in 13.4% of samples (8.7% with $\hat{\kappa}_D \leq 0$, 0.4% with $\hat{\kappa}_D = 1$ and 4.2% with no non-diseased subject), and at $p = 0.1$ in 8.3%; this is why the logit interval was verified against the published analysis but not carried into the simulation (Section 4.6 of the main text).

Table S7. Probability that the logit interval of [3, 9] is unavailable under the primary data-generating mechanism

p	$n = 30$	$n = 50$	$n = 100$
0.1	0.040 / 0.000 / 0.042 / 0.083	0.014 / 0.000 / 0.005 / 0.020	0.001 / 0.000 / 0.000 / 0.001
0.2	0.011 / 0.000 / 0.001 / 0.012	0.001 / 0.000 / 0.000 / 0.001	0.000 / 0.000 / 0.000 / 0.000
0.3	0.003 / 0.000 / 0.000 / 0.003	0.000 / 0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000 / 0.000
0.4	0.001 / 0.000 / 0.000 / 0.001	0.000 / 0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000 / 0.000
0.5	0.001 / 0.000 / 0.000 / 0.001	0.000 / 0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000 / 0.000
0.6	0.002 / 0.001 / 0.000 / 0.002	0.000 / 0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000 / 0.000
0.7	0.005 / 0.001 / 0.000 / 0.006	0.000 / 0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000 / 0.000
0.8	0.019 / 0.002 / 0.001 / 0.022	0.002 / 0.000 / 0.000 / 0.002	0.000 / 0.000 / 0.000 / 0.000
0.9	0.087 / 0.004 / 0.042 / 0.134	0.030 / 0.000 / 0.005 / 0.035	0.002 / 0.000 / 0.000 / 0.002

Each cell gives, in order, the probability that $\hat{\kappa}_D(c) \leq 0$; that $\hat{\kappa}_D(c) = 1$ ($n_{10} = n_{01} = 0$); that a disease margin is empty, so that $\hat{\kappa}_D(c)$ is undefined; and the total. Exact probabilities by enumeration, with $Se = 0.854$ and $Sp = 0.667$; none of the events depends on c (Proposition 7). The transformed interval of Eq. (9) is undefined in the third case and degenerate in the second.

S2.4 Prevalence-adjusted descriptive measures

The prevalence-adjusted bias-adjusted kappa [7] and Gwet's first-order agreement coefficient [8] are

$$\text{PABAK} = 2P_o - 1, \quad \text{AC1} = \frac{P_o - P_e^Y}{1 - P_e^Y}, \quad (\text{S16})$$

$$P_o = \pi_{11} + \pi_{00}, \quad P_e^Y = 2\bar{\pi}(1 - \bar{\pi}), \quad \bar{\pi} = \frac{Q + p}{2}.$$

Proposition 8 (PABAK and AC1 are not substitutes for $\kappa(c)$). (a) Neither depends on c . (b) In terms of the diagnostic parameters, $\text{PABAK} = 2\{p Se + (1 - p)Sp\} - 1$. (c) Neither is zero, in general, for a test with no diagnostic value ($J = 0$), for which $\kappa(c) = 0$ for every c .

Proof. (a) Neither formula contains c . (b) Substitute (S1) in P_o . (c) Take $Se = 0.1$, $Sp = 0.9$ ($J = 0$) and $p = 0.1$. Then $P_o = 0.01 + 0.81 = 0.82$, so $\text{PABAK} = 0.64$; $Q = 0.1$, $\bar{\pi} = 0.1$, $P_e^Y = 0.18$ and $\text{AC1} = 0.78$.

Both measures answer a different question from $\kappa(c)$: they describe raw agreement adjusted for skewed margins, not the chance-corrected, loss-weighted accuracy of a test against a reference standard. They are

therefore reported in the main text as complementary descriptive measures (Weiner et al. data: 0.668 and 0.775 against $\hat{\kappa}_D(0.5) = 0.373$; AuditC data: Table S22), and not as competitors in the simulation.

S3. The FWKC framework: derivations and complete specification

S3.1 Jeffreys posteriors for sensitivity and specificity (Eq. 10)

Proposition 9 (Jeffreys posterior). Under a cross-sectional design, $n_{11} \mid n_{1+} \sim \text{Bin}(n_{1+}, Se)$. The Jeffreys prior for Se is $\text{Beta}(1/2, 1/2)$, and the posterior is

$$Se \mid \text{data} \sim \text{Beta}(n_{11} + 1/2, n_{10} + 1/2), \quad (\text{S17})$$

$$E(Se \mid \text{data}) = \frac{n_{11} + \frac{1}{2}}{n_{1+} + 1}, \quad \text{Var}(Se \mid \text{data}) = \frac{(n_{11} + \frac{1}{2})(n_{10} + \frac{1}{2})}{(n_{1+} + 1)^2(n_{1+} + 2)}.$$

The posterior is proper for every $n_{11} \in \{0, \dots, n_{1+}\}$, including tables with a zero cell. The same holds for Sp with (n_{00}, n_{01}, n_{0+}) .

Proof. The Fisher information of $\text{Bin}(n_{1+}, Se)$ is $I(Se) = n_{1+}/\{Se(1 - Se)\}$, so Jeffreys' prior [A2] is proportional to $I(Se)^{1/2} \propto Se^{-1/2}(1 - Se)^{-1/2}$, the $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ kernel. Multiplying by the likelihood $Se^{n_{11}}(1 - Se)^{n_{10}}$ gives the $\text{Beta}(n_{11} + \frac{1}{2}, n_{10} + \frac{1}{2})$ kernel; both parameters are at least $\frac{1}{2} > 0$, so the posterior is proper. The moments are those of the Beta distribution.

The equal-tailed intervals of this posterior are the Jeffreys intervals for a binomial proportion, whose frequentist coverage is close to nominal even for small counts [24]. Using the posterior rather than a plug-in distribution is what makes the pseudo-rater layer a propagation of estimation uncertainty.

Proposition 10 (the naive parameterisation). A naive alternative draws Se from $\text{Beta}(n\widehat{Se}, n(1 - \widehat{Se}))$, with n the total sample size. This distribution (a) is centred at $\widehat{Se}$ with variance $\widehat{Se}(1 - \widehat{Se})/(n + 1)$; (b) is undefined when $\widehat{Se} \in \{0, 1\}$ (the naive implementation replaced a zero parameter by 10^{-6}); and (c) has a standard deviation that is, to first order, a fraction

$$\psi_{Se} = \sqrt{\frac{n_{1+} + 2}{n + 1}} \quad (\text{S18})$$

of the posterior standard deviation (S17), because it is governed by n rather than by n_{1+} , the only part of the sample that is informative about Se . The analogous fraction for Sp is $\psi_{Sp} = \{(n_{0+} + 2)/(n + 1)\}^{1/2}$.

Proof. (a) and (b) are properties of the Beta distribution. For (c), $(n_{11} + \frac{1}{2})(n_{10} + \frac{1}{2})/(n_{1+} + 1)^2 = \widehat{Se}(1 - \widehat{Se}) + O(n_{1+}^{-1})$, so the posterior variance is $\widehat{Se}(1 - \widehat{Se})/(n_{1+} + 2)$ to first order, and the ratio of the two standard deviations is $\{(n_{1+} + 2)/(n + 1)\}^{1/2}$.

Table S8 evaluates ψ_{Se} and ψ_{Sp} at the expected margins. The naive distribution is not a posterior, it understates the uncertainty of Se most when disease was rare (by a factor of 0.40 at $n = 30, p = 0.1$) and that of Sp when disease was common, and it failed for zero cells. Eq. (10) removes all three defects. The sensitivity analysis retained the naive parameterisation as a comparator (Table S21, variant 3).

Table S8. Standard deviation of the naive pseudo-rater distribution relative to the Jeffreys posterior

n	p	n_{1+} (expected)	n_{0+} (expected)	ψ_{Se} (S18)	ψ_{Sp}
30	0.1	3	27	0.40	0.97
30	0.5	15	15	0.74	0.74
30	0.9	27	3	0.97	0.40
50	0.1	5	45	0.37	0.96
50	0.5	25	25	0.73	0.73
50	0.9	45	5	0.96	0.37
100	0.1	10	90	0.34	0.95
100	0.5	50	50	0.72	0.72
100	0.9	90	10	0.95	0.34

Values below one mean that the naive distribution $\text{Beta}(n\widehat{Se}, n(1 - \widehat{Se}))$ was more concentrated than the posterior, understating the uncertainty in Se (or Sp).

S3.2 Posterior-predictive pseudo-raters (Eq. 11)

For each draw j and pseudo-rater r , $n_{11,rj}^* \sim \text{Bin}(n_{1+}, Se_j)$ and $n_{01,rj}^* \sim \text{Bin}(n_{0+}, 1 - Sp_j)$, with $n_{10,rj}^* = n_{1+} - n_{11,rj}^*$ and $n_{00,rj}^* = n_{0+} - n_{01,rj}^*$; the disease margins are held at their observed values. Integrating over Se_j gives a beta-binomial distribution.

Proposition 11 (dispersion of the pseudo-rater tables). Marginally over the posterior, $n_{11,rj}^* \sim \text{BetaBin}(n_{1+}, n_{11} + 1/2, n_{10} + 1/2)$, with

$$\text{Var}(n_{11,rj}^*) = n_{1+} \frac{\left(\frac{n_{11}+1}{2}\right)\left(\frac{n_{10}+1}{2}\right)}{(n_{1+}+1)^2} \cdot \frac{2n_{1+}+1}{n_{1+}+2} \approx 2 n_{1+} \widehat{Se}(1 - \widehat{Se}). \quad (\text{S19})$$

Two pseudo-raters that share draw j are conditionally independent given (Se_j, Sp_j) and positively correlated marginally, with $\text{Cov}(n_{11,rj}^*, n_{11,r'j}^*) = n_{1+}^2 \text{Var}(Se_j \mid \text{data})$ for $r \neq r'$.

Proof. The variance of $\text{BetaBin}(m, \alpha, \beta)$ is $m\alpha\beta(\alpha + \beta + m)/\{(\alpha + \beta)^2(\alpha + \beta + 1)\}$; set $m = n_{1+}$, $\alpha + \beta = n_{1+} + 1$. For the covariance, by the law of total covariance, $\text{Cov} = E\{\text{Cov}(\cdot, \cdot \mid Se_j)\} + \text{Cov}\{E(\cdot \mid Se_j), E(\cdot \mid Se_j)\} = 0 + \text{Var}(n_{1+} Se_j)$.

Two consequences are used in the main text. First, each pseudo-rater table is about twice as dispersed as a new sample with the same margins, because it combines estimation uncertainty (the posterior of Se) with sampling uncertainty (the binomial draw). Pseudo-rater dispersion is therefore not an estimate of the sampling variance of any estimator, which is why it does not enter the primary interval (Section 2.6.6). Second, the draw index j induces a common effect shared by the R pseudo-raters, which is what the crossed term v_j of Eq. (14) represents; the pseudo-raters themselves are exchangeable, which the rater effect u_r represents.

Proposition 12 (large-sample behaviour of the pipeline). Fix (p, Se, Sp, c) with $0 < p < 1$ and $0 < Q < 1$, and fix R and n_{mC} . As $n \rightarrow \infty$:

- (a) every pseudo-rater value satisfies $\kappa_{rj}^* \rightarrow \kappa(c)$ in probability, and hence $\eta_{rj} \rightarrow \eta(c)$;
- (b) $\hat{\kappa}_{RF} \rightarrow \kappa(c)$ and $SE_{RF} \rightarrow 0$ in probability;
- (c) SE_B does not tend to zero, because with R fixed the posterior of τ_r^2 does not degenerate;
- (d) consequently $w_B \rightarrow 0$, $\hat{\kappa}_{FWKC}(c) - \hat{\kappa}_{RF} \rightarrow 0$, and $\hat{\kappa}_{FWKC}(c) \rightarrow \kappa(c)$ in probability. In finite samples

$$\hat{\kappa}_{FWKC}(c) - \hat{\kappa}_D(c) = w_B(\hat{\kappa}_B - \hat{\kappa}_{RF}) + (\hat{\kappa}_{RF} - \hat{\kappa}_D(c)). \quad (\text{S20})$$

Proof. (a) By (S17), $E(Se_j \mid \text{data}) \rightarrow Se$ and $\text{Var}(Se_j \mid \text{data}) = O(n^{-1})$, so $Se_j \rightarrow Se$ in probability, and likewise $Sp_j \rightarrow Sp$. Given the draw, $n_{11,rj}^*/n_{1+} - Se_j \rightarrow 0$ and $n_{01,rj}^*/n_{0+} - (1 - Sp_j) \rightarrow 0$ by the law of large numbers, while $n_{1+}/n \rightarrow p$. The count form (S7) is a continuous function of the cell proportions wherever its denominator is positive, so $\kappa_{rj}^* \rightarrow \kappa(c)$; η is continuous on $(-1,1)$. (b) A regression-forest prediction is a weighted average of training responses with non-negative weights summing to one [17], so every out-of-bag prediction lies between the smallest and the largest η_{rj} . All $R n_{mc}$ responses converge to $\eta(c)$ by (a), so all predictions converge to $\eta(c)$; $\hat{\kappa}_{RF}$, their back-transformed mean, converges to $\kappa(c)$, and SE_{RF} , their standard deviation, to zero. (c) By Proposition 14(a) below, $\text{Var}(\mu \mid \sigma^2, \tau_r^2, \tau_{mc}^2, \eta) = \left(\frac{1}{4} + N/\lambda_\mu\right)^{-1}$ with $\lambda_\mu \geq n_{mc}\tau_r^2$ and $N = Rn_{mc}$, so $\text{Var}(\mu \mid \cdot) \geq \left(\frac{1}{4} + R/\tau_r^2\right)^{-1}$. The full conditional of τ_r^2 is $\text{IG}(2 + R/2, 1 + \sum_r u_r^2/2)$, which is stochastically larger than $\text{IG}(2 + R/2, 1)$ whatever the data; hence the posterior variance of μ , and of $\tanh(\mu/2)$ when $|\kappa(c)| < 1$, is bounded away from zero. (d) $w_B = SE_{RF}/(SE_B + SE_{RF}) \leq SE_{RF}/SE_B \rightarrow 0$ by (b) and (c), and $\hat{\kappa}_B \in (-1,1)$ is bounded, so $w_B(\hat{\kappa}_B - \hat{\kappa}_{RF}) \rightarrow 0$. (S20) is Eq. (15) rearranged.

Proposition 12 locates the bias that Reviewer 1 asked about. With only three rater levels the posterior of the rater variance is dominated by its $\text{IG}(2,1)$ prior, so the Bayesian component is shrunk towards zero on the η scale and carries a posterior standard deviation of about 0.15–0.18 even at $n = 620$. Integrating over $\tau_r^2 \sim \text{IG}(3.5,1)$ in the bound of part (c) gives 0.133 to 0.160 against the observed 0.147 to 0.178 in the Weiner et al. analysis (Table S9); the approximation neglects the σ^2 and τ_{mc}^2 terms and understates SE_B by up to about 14%. The fused estimator inherits a fraction w_B of this shrinkage. Table S9 applies (S20) to the Weiner et al. data: at every c the Bayesian term $w_B(\hat{\kappa}_B - \hat{\kappa}_{RF})$ is negative (−0.012 to −0.002), while the forest term $\hat{\kappa}_{RF} - \hat{\kappa}_D$ lies between −0.004 and +0.002; summed over the nine values of c the Bayesian term contributes −0.047 of the total −0.050. The systematic difference between the FWKC and direct estimates reported in Section 3.1 of the main text therefore comes mainly from the Bayesian layer. The same mechanism accounts for the negative bias of the Bayesian component in the simulation at $n = 50$ and $n = 100$ (−0.023 and −0.026; Table 10 of the main text); at $n = 30$ other small-sample effects dominate and its mean bias is +0.004. Because $SE_{RF} = O_p(n^{-1/2})$, the Bayesian term is of the same order as the standard error: the fused estimator is consistent, but not unbiased to order $n^{-1/2}$.

Table S9. Decomposition of the difference between the FWKC and direct estimates, Weiner et al. data

c	$\hat{\kappa}_D$	$\hat{\kappa}_{RF}$	$\hat{\kappa}_B$	w_B	$w_B(\hat{\kappa}_B - \hat{\kappa}_{RF})$	$\hat{\kappa}_{RF} - \hat{\kappa}_D$	$\hat{\kappa}_{FWKC} - \hat{\kappa}_D$	SE_B observed	SE_B approx.
0.1	0.5238	0.5244	0.4855	0.301	-0.0117	+0.0006	-0.0112	0.152	0.133
0.2	0.4759	0.4747	0.4559	0.269	-0.0051	-0.0012	-0.0063	0.147	0.140
0.3	0.4360	0.4366	0.4137	0.235	-0.0054	+0.0006	-0.0048	0.154	0.145
0.4	0.4023	0.4013	0.3753	0.200	-0.0052	-0.0010	-0.0062	0.166	0.149
0.5	0.3734	0.3739	0.3513	0.200	-0.0045	+0.0005	-0.0040	0.170	0.152
0.6	0.3484	0.3494	0.3243	0.172	-0.0043	+0.0010	-0.0033	0.178	0.154
0.7	0.3265	0.3225	0.2964	0.196	-0.0051	-0.0041	-0.0092	0.171	0.157
0.8	0.3073	0.3097	0.2884	0.185	-0.0039	+0.0024	-0.0015	0.168	0.158
0.9	0.2901	0.2891	0.2758	0.157	-0.0021	-0.0010	-0.0031	0.175	0.160

The last three difference columns satisfy (S20) exactly. “ SE_B approx.” integrates the bound of Proposition 12(c) over $\tau_r^2 \sim \text{IG}(3.5, 1)$, the full conditional of τ_r^2 when the rater effects are near zero, neglecting σ^2 and τ_{mc}^2 .

The Monte Carlo summaries over j are averages of bounded quantities, so their Monte Carlo error is $O(n_{mc}^{-1/2})$; with $n_{mc} = 150$ it is at most $1/\sqrt{150} = 0.082$ on the κ scale and in practice far smaller. The sensitivity analysis varied n_{mc} from 150 to 2000 (Table S21, variants 28–31).

S3.3 Transformation and truncation (Eqs. 12–13)

Before transformation each pseudo-rater value is truncated to $[-1 + \epsilon_t, 1 - \epsilon_t]$ with $\epsilon_t = 10^{-8}$, so that

$$|\eta_{rj}| \leq \log \frac{2 - \epsilon_t}{\epsilon_t} \approx 19.1. \quad (\text{S21})$$

Truncation changes a value in $[-1, 1]$ by at most ϵ_t . By Proposition 4(e) a pseudo-rater value can fall below -1 when its table has $J^* < 0$ and $c \neq \frac{1}{2}$; such values are set to $-1 + \epsilon_t$.

Proposition 13 (the Fisher-z variant changes only the prior scale). If the crossed model of Eq. (14) is fitted to $z_{rj} = \eta_{rj}/2 = \text{artanh} \kappa_{rj}^*$ with priors $\mu_z \sim N(0, 2^2)$ and variance components $\sim \text{IG}(2, 1)$, the implied model for η_{rj} is Eq. (14) with priors $\mu \sim N(0, 4^2)$ and variance components $\sim \text{IG}(2, 4)$.

Proof. $\eta = 2z$, so $\mu = 2\mu_z \sim N(0, 4^2)$ and each variance is multiplied by 4; if $X \sim \text{IG}(a, b)$ then $4X \sim \text{IG}(a, 4b)$.

Because of Proposition 13, the change in the coverage of the primary interval under the Fisher-z variant (0.867 against 0.767, that is three datasets out of 30 and within Monte Carlo error; Table S21, variant 14) reflects a change in the effective prior, not in the transformation as such; the random forest and the SE-expanded interval are invariant to the rescaling $\eta \rightarrow \eta/2$. It is consistent with Proposition 12(c): the prior on the rater variance governs SE_B , and therefore w_B .

S3.4 The crossed Bayesian model and its exact Gibbs sampler (Eq. 14)

Arrange the transformed values η_{rj} in an $n_{mc} \times R$ matrix with rows indexed by j and columns by r , and let $N = Rn_{mc}$. The model is

$$\eta_{rj} = \mu + u_r + v_j + \varepsilon_{rj}, \quad u_r \sim N(0, \tau_r^2), \quad v_j \sim N(0, \tau_{mc}^2), \quad \varepsilon_{rj} \sim N(0, \sigma^2), \quad (\text{S22})$$

$$\mu \sim N(0, 2^2), \quad \sigma^2, \tau_r^2, \tau_{mc}^2 \sim \text{IG}(2, 1), \quad (\text{S23})$$

$$\hat{\kappa}_B = E\{\tanh(\mu/2) \mid \eta\}, \quad SE_B = \text{SD}\{\tanh(\mu/2) \mid \eta\}.$$

Proposition 14 (full conditional distributions). Let $\mathbf{s} = (s_1, \dots, s_R)^T$ with $s_r = \sum_j \eta_{rj} - n_{mc}\mu$, and let $\mathbf{1}$ be a vector of ones.

- (a) With u and v integrated out, $\mu \mid \sigma^2, \tau_r^2, \tau_{mc}^2, \eta \sim N(m_\mu, s_\mu^2)$, where

$$\lambda_\mu = \sigma^2 + n_{mc}\tau_r^2 + R\tau_{mc}^2, \quad s_\mu^2 = \left(\frac{1}{4} + \frac{N}{\lambda_\mu}\right)^{-1}, \quad m_\mu = s_\mu^2 \frac{\sum_{jr} \eta_{rj}}{\lambda_\mu}. \quad (\text{S24})$$

- (b) With v integrated out, $\mathbf{u} \mid \mu, \sigma^2, \tau_r^2, \tau_{mc}^2, \eta$ is normal with precision matrix $\mathbf{P} = \tau_r^{-2}\mathbf{I}_R + n_{mc}(\sigma^2\mathbf{I}_R + \tau_{mc}^2\mathbf{1}\mathbf{1}^T)^{-1}$ and mean $\mathbf{P}^{-1}(\sigma^2\mathbf{I}_R + \tau_{mc}^2\mathbf{1}\mathbf{1}^T)^{-1}\mathbf{s}$. $\mathbf{P}$ has eigenvalue

$$p_\parallel = \frac{1}{\tau_r^2} + \frac{n_{mc}}{\sigma^2 + R\tau_{mc}^2} \quad \text{on } \mathbf{1}, \quad p_\perp = \frac{1}{\tau_r^2} + \frac{n_{mc}}{\sigma^2} \quad \text{on } \mathbf{1}^\perp. \quad (\text{S25})$$

- (c) The v_j are conditionally independent with

$$v_j \mid \mu, \mathbf{u}, \sigma^2, \tau_{mc}^2, \eta \sim N\left(s_v^2 \frac{\sum_r (\eta_{rj} - \mu - u_r)}{\sigma^2}, s_v^2\right), \quad s_v^2 = \left(\frac{R}{\sigma^2} + \frac{1}{\tau_{mc}^2}\right)^{-1}. \quad (\text{S26})$$

- (d) The variance components have conjugate full conditionals

$$\sigma^2 \mid \cdot \sim \text{IG}(2 + N/2, 1 + \text{SSE}/2), \quad (\text{S27})$$

$$\tau_r^2 \mid \cdot \sim \text{IG}\left(2 + R/2, 1 + 1/2 \sum_r u_r^2\right), \quad \tau_{mc}^2 \mid \cdot \sim \text{IG}\left(2 + n_{mc}/2, 1 + 1/2 \sum_j v_j^2\right),$$

where $\text{SSE} = \sum_{r,j} (\eta_{rj} - \mu - u_r - v_j)^2$.

Proof. (a) Stacking the η_{rj} , $\boldsymbol{\eta} = \mu \mathbf{1}_N + \boldsymbol{\xi}$ with $\text{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_N + \tau_r^2 (\mathbf{J}_{n_{mc}} \otimes \mathbf{I}_R) + \tau_{mc}^2 (\mathbf{I}_{n_{mc}} \otimes \mathbf{J}_R)$, where $\mathbf{J}$ is a matrix of ones. Since $(\mathbf{J}_{n_{mc}} \otimes \mathbf{I}_R) \mathbf{1}_N = n_{mc} \mathbf{1}_N$ and $(\mathbf{I}_{n_{mc}} \otimes \mathbf{J}_R) \mathbf{1}_N = R \mathbf{1}_N$, $\mathbf{1}_N$ is an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue λ_μ , so $\mathbf{1}_N^T \boldsymbol{\Sigma}^{-1} = \lambda_\mu^{-1} \mathbf{1}_N^T$. The normal prior with variance 4 and the normal likelihood give posterior precision $\frac{1}{4} + \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1} = \frac{1}{4} + N/\lambda_\mu$ and mean $s_\mu^2 \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}$. (b) Given μ and $\mathbf{u}$, row j satisfies $\boldsymbol{\eta}_{j\cdot} - \mu \mathbf{1}_R - \mathbf{u} = v_j \mathbf{1}_R + \boldsymbol{\varepsilon}_{j\cdot} \sim N(\mathbf{0}, \mathbf{V})$ with $\mathbf{V} = \sigma^2 \mathbf{I}_R + \tau_{mc}^2 \mathbf{J}_R$, independently over j . Summing over j , $\mathbf{s} = n_{mc} \mathbf{u} + \mathbf{e}$ with $\mathbf{e} \sim N(\mathbf{0}, n_{mc} \mathbf{V})$. Combining with the prior $N(\mathbf{0}, \tau_r^2 \mathbf{I}_R)$ gives precision $\tau_r^{-2} \mathbf{I} + n_{mc} \mathbf{V}^{-1}$ and mean $\mathbf{P}^{-1} \mathbf{V}^{-1} \mathbf{s}$. $\mathbf{V}$ has eigenvalue $\sigma^2 + R\tau_{mc}^2$ on $\mathbf{1}$ and σ^2 on $\mathbf{1}^\perp$, which gives (S25). (c) and (d) are the standard normal–normal and normal–inverse-gamma conjugate updates. $\square$

The sampler draws $(\mu, \mathbf{u}, \mathbf{v})$ as one block through the factorisation $p(\mu \mid \cdot) p(\mathbf{u} \mid \mu, \cdot) p(\mathbf{v} \mid \mathbf{u}, \mu, \cdot)$ given by (a)–(c), which is an exact draw from their joint full conditional, and then updates the three variances by (d). This two-block scheme targets the same posterior as a single-site sampler but avoids the strong autocorrelation that a single-site sampler shows for the location parameters when there are only three rater levels [A3]. Four chains are run; a fit is accepted when $\max(\hat{R}_{\text{classical}}, \hat{R}_{\text{rank}}) \leq 1.05$, where $\hat{R}_{\text{rank}}$ is the rank-normalised statistic of [26]. The primary grid had a Bayesian convergence rate of 1.00 (main text, Section 3.2).

Cross-check. For the Weiner et al. data at $c = 0.2, 0.5$ and 0.8 , the same model with the same priors was coded independently in Stan, with the random effects integrated out, and run in CmdStan through cmdstanr [14, 27] (four chains, 600 iterations of which 300 warm-up, adapt_delta = 0.9999, maximum tree depth 18). Table S10 compares the two. The posterior means differ by 0.003 to 0.021, that is, by at most 0.13 posterior standard deviations; the HMC runs had no divergent transitions and no maximum-tree-depth hits.

Table S10. Cross-check of the Gibbs sampler against CmdStan HMC for the same model and priors (Weiner et al. data)

c	$\hat{\kappa}_B$, Gibbs / HMC	Abs. diff.	Diff./SD	SE_B , Gibbs / HMC	$\hat{R}$, Gibbs / HMC	HMC divergences	HMC min. E- BFMI	HMC ESS bulk / tail
0.2	0.4585 / 0.4389	0.0196	0.12	0.1543 / 0.1599	0.999 / 1.010	0	0.86	268 / 215
0.5	0.3515 / 0.3310	0.0205	0.13	0.1612 / 0.1636	1.004 / 1.019	0	0.91	208 / 363
0.8	0.2908 / 0.2879	0.0029	0.02	0.1740 / 0.1616	1.003 / 1.011	0	0.95	235 / 325

Both samplers were run on identical pseudo-rater arrays ($n_{mc} = 500$, $R = 3$). Diff./SD is the absolute difference divided by the larger posterior standard deviation. HMC: 4 chains $\times$ 600 iterations (300 warm-up), 1200 post-warm-up transitions, no maximum-tree-depth hits.

S3.5 Random-forest layer (Section 2.6.4 and Table 2 of the main text)

For each target pseudo-rater r , a regression forest is trained on the n_{mc} rows indexed by j , not on patients. The response is η_{rj} and the predictors are Se_j , Sp_j and $\eta_{r',j}$ for the other $R - 1$ pseudo-raters, so the target never predicts itself. Only out-of-bag predictions $\hat{\eta}_{rj}^{OOB}$ are used, and

$$\hat{\kappa}_{RF} = \frac{1}{R n_{mc}} \sum_{r=1}^R \sum_{j=1}^{n_{mc}} \tanh\left(\frac{\hat{\eta}_{rj}^{OOB}}{2}\right), \quad SE_{RF} = SD_{r,j} \left\{ \tanh\left(\frac{\hat{\eta}_{rj}^{OOB}}{2}\right) \right\}. \quad (\text{S28})$$

Overfitting is controlled in three ways: out-of-bag prediction, so that no row is predicted by trees trained on it; a minimum node size of 10 and at most $\min(100, \lceil \text{rows}/10 \rceil)$ terminal nodes; and a number of training rows that is fixed by n_{mc} and does not fall with n . A target is skipped when its response has five or fewer distinct values (as happens when n is so small that few distinct tables can occur); if every target is skipped, the Bayesian component is used alone. Cross-fitted natural cubic splines (3 degrees of freedom, 5 folds) and the absence of any machine-learning layer were examined as comparators (Table S21, variants 16–18).

The forest is the most accurate and least biased component of the pipeline. In the primary grid its mean bias was -0.001 , -0.009 and -0.004 at $n = 30$, 50 and 100, against -0.009 , -0.005 and -0.001 for the direct estimator and $+0.004$, -0.023 and -0.026 for the Bayesian component; its RMSE was 0.178, 0.135 and 0.094, against 0.177, 0.136 and 0.095 for the direct estimator and 0.198, 0.139 and 0.096 for the

Bayesian component, and it had the lower RMSE of the two in 26, 44 and 44 of the 81 scenarios. In the sensitivity analyses the random-forest-only fusion had the lowest RMSE of the five fusion rules (0.226, against 0.228 for the primary rule and 0.230 for the Bayesian-only estimator, which is the estimator obtained when the machine-learning layer is omitted). In the Weiner et al. data it matched $\hat{\kappa}_D(c)$ to within 0.005 at every c (Section 3.1 of the main text).

S3.6 Fusion rule (Eq. 15)

Proposition 15 (combining two correlated estimators). Let $\hat{\kappa}_B$ and $\hat{\kappa}_{RF}$ have common mean, variances σ_B^2 and σ_{RF}^2 and correlation ρ . For $\hat{\kappa}(w) = w\hat{\kappa}_B + (1 - w)\hat{\kappa}_{RF}$,

$$\text{Var}\{\hat{\kappa}(w)\} = w^2\sigma_B^2 + (1 - w)^2\sigma_{RF}^2 + 2w(1 - w)\rho\sigma_B\sigma_{RF}. \quad (\text{S29})$$

- (a) The variance-minimising weight is $w^* = (\sigma_{RF}^2 - \rho\sigma_B\sigma_{RF})/(\sigma_B^2 + \sigma_{RF}^2 - 2\rho\sigma_B\sigma_{RF})$; for $\rho = 0$ this is the inverse-variance weight $\sigma_B^{-2}/(\sigma_B^{-2} + \sigma_{RF}^{-2})$, and then $1/\text{Var}\{\hat{\kappa}(w^*)\} = \sigma_B^{-2} + \sigma_{RF}^{-2}$.
- (b) For every $w \in [0, 1]$,

$$\begin{aligned} \text{Var}\{\hat{\kappa}(w)\} &= \{w\sigma_B + (1 - w)\sigma_{RF}\}^2 - 2w(1 - w)(1 - \rho)\sigma_B\sigma_{RF} \\ &\geq \min(\sigma_B, \sigma_{RF})^2 - 1/2(1 - \rho)\sigma_B\sigma_{RF}, \end{aligned} \quad (\text{S30})$$

so when ρ is close to one no convex weighting can reduce the variance materially below that of the better component.

Proof. (S29) is the variance of a linear combination. (a) Setting the derivative $2w\sigma_B^2 - 2(1 - w)\sigma_{RF}^2 + 2(1 - 2w)\rho\sigma_B\sigma_{RF}$ to zero gives w^* ; for $\rho = 0$, substituting w^* gives $\sigma_B^2\sigma_{RF}^2/(\sigma_B^2 + \sigma_{RF}^2)$. (b) Write $2w(1 - w)\rho\sigma_B\sigma_{RF} = 2w(1 - w)\sigma_B\sigma_{RF} - 2w(1 - w)(1 - \rho)\sigma_B\sigma_{RF}$ and complete the square, which gives the identity; then use $w\sigma_B + (1 - w)\sigma_{RF} \geq \min(\sigma_B, \sigma_{RF})$ and $w(1 - w) \leq \frac{1}{4}$. $\square$

Proposition 15(a) is the derivation of inverse-variance weighting given for the general case; it requires $\rho = 0$. Across bootstrap resamples the two components are almost perfectly correlated (mean 0.989, 0.996 and 0.998 at $n = 30, 50$ and 100 in the simulation, and 0.993 to 0.998 in the Weiner et al. data), so part (b) applies. This is the formal reason why the fused estimator improved on the random-forest component

by only 0.0003 in RMSE at $n = 100$ (Section 4.4 of the main text), and why the inverse-standard-error weights of Eq. (15) are presented there as a pragmatic rule rather than as an optimal one.

Proposition 16 (sensitivity of the estimate to the fusion rule). For any two weights $w, w' \in [0,1]$, $|\hat{\kappa}(w) - \hat{\kappa}(w')| = |w - w'| |\hat{\kappa}_B - \hat{\kappa}_{RF}| \leq |\hat{\kappa}_B - \hat{\kappa}_{RF}|$.

Proof. $\hat{\kappa}(w) - \hat{\kappa}(w') = (w - w')(\hat{\kappa}_B - \hat{\kappa}_{RF})$. $\square$

For the Weiner et al. data $|\hat{\kappa}_B - \hat{\kappa}_{RF}| \leq 0.039$, so no fusion rule can move the estimate by more than 0.039 at any c ; Table S11 evaluates the inverse-SE, inverse-variance and equal weights. In the stress scenarios the fusion variants changed RMSE by at most 0.002 (Table S21, variants 8–12).

Table S11. Effect of the fusion rule on the FWKC estimate, Weiner et al. data

c	$\hat{\kappa}_B$	$\hat{\kappa}_{RF}$	w_B , inverse SE	w_B , inverse variance	Estimate, inverse SE (Eq. 15)	Estimate, inverse variance	Estimate, equal weights	Bound
0.1	0.4855	0.5244	0.301	0.157	0.5127	0.5183	0.5049	0.0389
0.2	0.4559	0.4747	0.269	0.120	0.4696	0.4724	0.4653	0.0188
0.3	0.4137	0.4366	0.235	0.086	0.4312	0.4346	0.4251	0.0229
0.4	0.3753	0.4013	0.200	0.059	0.3961	0.3998	0.3883	0.0260
0.5	0.3513	0.3739	0.200	0.059	0.3694	0.3726	0.3626	0.0227
0.6	0.3243	0.3494	0.172	0.041	0.3451	0.3484	0.3369	0.0251
0.7	0.2964	0.3225	0.196	0.056	0.3173	0.3210	0.3094	0.0260
0.8	0.2884	0.3097	0.185	0.049	0.3058	0.3087	0.2990	0.0213
0.9	0.2758	0.2891	0.157	0.034	0.2870	0.2887	0.2825	0.0133

Inverse-variance weights are computed from the same SE_B and SE_{RF} as the inverse-SE weights. Bound: the absolute difference between $\hat{\kappa}_B$ and $\hat{\kappa}_{RF}$; by Proposition 16 no rule with weights in $[0,1]$ can differ from another by more than this. Correlation between the components across bootstrap resamples: 0.993–0.998.

S3.7 Primary interval (Eq. 16)

The primary interval is the percentile interval of B full-pipeline bootstrap replicates: each resample $D^{*(b)}$ of the n subjects is passed through Steps 1–5 of Section S3.10. Because order statistics are equivariant under monotone maps, the interval obtained on the η scale and back-transformed equals the interval

obtained on the κ scale (exactly for order-statistic percentiles, and up to the linear interpolation of the type-7 quantiles used with $B = 29$), so Eq. (16) does not depend on the choice of transformation once the replicates are given; it also always lies inside $[-1, 1]$. Every source of variability in the pipeline — the posterior draws, the pseudo-rater tables, the Gibbs sampler and the forest — is re-generated within each replicate, so it is propagated into the interval without any variance formula.

S3.8 SE-expanded interval (Eq. 17)

$$\lambda_n = \sqrt{\min(1, n/n_0)} = \begin{cases} \sqrt{n/n_0}, & n < n_0, \\ 1, & n \geq n_0, \end{cases} \quad CI_{cal} = \tanh\left\{\frac{\hat{\eta}_{FWKC} \pm z_{0.975} SE_{\eta}^*/\lambda_n}{2}\right\}. \quad (S31)$$

Proposition 17 (properties of the expansion factor). (a) λ_n is continuous and non-decreasing in n and equal to one for $n \geq n_0$, so the secondary interval coincides with an unexpanded normal interval on the η scale in large samples. (b) For $n < n_0$ the expansion factor $1/\lambda_n = \sqrt{n_0/n}$ satisfies

$$\frac{\partial(1/\lambda_n)}{\partial n_0} = \frac{1}{2\sqrt{n} n_0}, \quad \frac{1/\lambda_n(n_0')}{1/\lambda_n(n_0)} = \sqrt{\frac{n_0'}{n_0}}, \quad (S32)$$

so replacing $n_0 = 100$ by 75, 125 or 150 changes the half-width on the η scale by -13.4% , $+11.8\%$ or $+22.5\%$ whenever $n < \min(n_0, n_0')$, and not at all when $n \geq \max(n_0, n_0')$.

Proof. (a) $\sqrt{n/n_0}$ is increasing and equals one at $n = n_0$, where it joins the constant branch. (b)

Differentiate $\sqrt{n_0/n}$ with respect to n_0 ; $\sqrt{0.75} - 1 = -0.134$, $\sqrt{1.25} - 1 = 0.118$, $\sqrt{1.5} - 1 = 0.225$. $\square$

The observed change in the coverage of the SE-expanded interval when n_0 was varied was at most 0.033, one dataset in 30 (Table S21, variants 24–27), consistent with (S32).

Proposition 18 (why expansion protects coverage and contraction destroys it). Suppose that $\hat{\eta}_{FWKC}$ is approximately $N(\eta, s^2)$ and that the bootstrap standard error underestimates s by a factor $\gamma \geq 1$, $SE_{\eta}^* = s/\gamma$. Then the coverage of an interval $\hat{\eta}_{FWKC} \pm z_{0.975} a SE_{\eta}^*$ with multiplier $a > 0$, mapped back by $\tanh(\cdot/2)$, is

$$\mathcal{C}(a, \gamma) = 2 \Phi\left(\frac{a z_{0.975}}{\gamma}\right) - 1. \quad (S33)$$

- (a) For the SE-expanded interval, $a = 1/\lambda_n$, and $\mathcal{C} \geq 0.95$ if and only if $\gamma \leq 1/\lambda_n = \sqrt{n_0/n}$: the interval keeps nominal coverage as long as the bootstrap standard error underestimates the true one by a factor of at most 1.83 at $n = 30$ and 1.41 at $n = 50$.
- (b) For a contraction rule, $a < 1$, and $\mathcal{C} < 0.95$ even when $\gamma = 1$; for the multiplier $a = \min(1, n/n_0)$, implemented on the η scale of the present pipeline, $\mathcal{C} = 0.443$ at $n = 30$ and 0.673 at $n = 50$. (This multiplier was also examined applied to a κ -scale Wald half-width built on the covariance diagnostic of Section S3.9; the conclusion $\mathcal{C} < 0.95$ for every $a < 1$ does not depend on that choice.)

Proof. $\tanh(\cdot/2)$ is strictly increasing, so coverage on the κ scale equals coverage on the η scale, which is $P(|\hat{\eta} - \eta| \leq az_{0.975}S/\gamma) = 2\Phi(az_{0.975}/\gamma) - 1$. (a) $2\Phi(z_{0.975}/(\lambda_n\gamma)) - 1 \geq 0.95$ if and only if $1/(\lambda_n\gamma) \geq 1$. (b) Evaluate (S33) at $\gamma = 1$.

Proposition 18 is the theoretical support requested in comment R1-11, and it is also its limit. It explains why the expansion rule gives worst-case protection in small samples — it absorbs an underestimated standard error up to a known factor — and why a contraction rule could not, but it rests on approximate normality on the η scale and does not guarantee coverage in any particular scenario. Table S12 compares the levels implied by (S33) at $\gamma = 1$ with the coverage observed in the primary grid for the three rules that were run. The ordering predicted by the proposition is reproduced exactly. At $n = 30$ and $n = 50$ the observed values lie below the implied ones for the expansion rules and above them for the contraction rule; this is the pattern produced by a sampling distribution of $\hat{\eta}_{FWKC}$ with heavier tails than the normal (more mass both near the centre and far from it), to which the error of a bootstrap standard error based on $B = 29$ replicates is added. The worst-case result of the main text — no scenario below 0.85 for the SE-expanded interval — is empirical and is reported as such.

Table S12. Coverage implied by Proposition 18 and coverage observed in the primary grid for three interval rules

Rule	Multiplier a	a at $n = 30 / 50 / 100$	Implied coverage	Observed coverage
SE-expanded, Eq. (17)	$\sqrt{n_0/n}$	1.826 / 1.414 / 1.000	0.9997 / 0.9944 / 0.9500	0.977 / 0.977 / 0.934

Rule	Multiplier a	a at $n = 30 / 50 / 100$	Implied coverage	Observed coverage
Linear SE expansion	n_0/n	3.333 / 2.000 / 1.000	1.0000 / 0.9999 / 0.9500	0.987 / 0.992 / 0.934
Original contraction	$\min(1, n/n_0)$	0.300 / 0.500 / 1.000	0.4435 / 0.6729 / 0.9500	0.518 / 0.708 / 0.934

Implied coverage is $2\Phi(az_{0.975}) - 1$, that is (S33) with $\gamma = 1$. Observed coverage is the mean over the 81 scenarios at each sample size ($M = 120$ datasets per scenario); the three rules were computed on the same datasets and bootstrap replicates.

S3.9 Covariance diagnostic (Eq. 18)

Let $\mathbf{K}$ be the $n_{mc} \times R$ matrix of pseudo-rater values, $\hat{\Sigma}_K$ its sample covariance matrix and $\mathbf{w} = R^{-1}\mathbf{1}_R$.

Proposition 19. (a) $\mathbf{w}^T \hat{\Sigma}_K \mathbf{w} = \sum_{r,r'} w_r w_{r'} \hat{\Sigma}_{K,rr'}$ estimates $\text{Var}(\mathbf{w}^T \mathbf{k}_j)$, the variance over posterior draws of the equally weighted rater average of a single draw j , and includes every cross-rater covariance; the variance of the average over all n_{mc} draws would be smaller by the factor n_{mc} . (b) $\text{rank}(\hat{\Sigma}_K) \leq \min(n_{mc} - 1, R)$. (c) For $\epsilon > 0$, $\hat{\Sigma}_K + \epsilon \mathbf{I}$ is positive definite with eigenvalues $\ell_i + \epsilon \geq \epsilon$, determinant at least ϵ^R and condition number at most $(\ell_{\max} + \epsilon)/\epsilon$, so SE_{cov} of Eq. (18) is always defined and strictly positive.

Proof. (a) Variance of a linear combination. (b) $\hat{\Sigma}_K = (n_{mc} - 1)^{-1} \mathbf{K}_c^T \mathbf{K}_c$ with $\mathbf{K}_c$ the column-centred matrix, whose rows sum to zero, so its rank is at most $\min(n_{mc} - 1, R)$. (c) $\hat{\Sigma}_K$ is positive semi-definite, so its eigenvalues ℓ_i are non-negative and those of the sum are $\ell_i + \epsilon$.

The implementation uses $\epsilon = 10^{-10}$. With $R = 3$ and $n_{mc} \geq 150$, rank deficiency occurs only when pseudo-rater columns are collinear, and the ridge term then prevents a zero or undefined value. SE_{cov} agreed with a bootstrap-over-draws estimate (mean over the 81 scenarios 0.170 against 0.169 at $n = 30$, 0.137 against 0.136 at $n = 50$ and 0.099 against 0.099 at $n = 100$). By Proposition 11 it measures posterior-predictive dispersion, not sampling variance, and it is therefore reported as a diagnostic and not used for inference.

S3.10 Complete algorithm

Settings that differ between analyses are listed in Table S13. The random seed was 20260912.

Input. The observed table $(n_{11}, n_{10}, n_{01}, n_{00})$; the weighting index c ; nominal level $1 - \alpha = 0.95$.

Step 0 (direct estimator and established intervals). 0.1 $\hat{\kappa}_D(c)$ from (S7). 0.2 Wald interval $\hat{\kappa}_D(c) \pm z_{0.975}\widehat{SE}$ with $\widehat{SE}$ from (S14). 0.3 Transformed interval (S15). 0.4 Percentile and basic bootstrap intervals from the same B resamples used in Step 6, so that direct and FWKC intervals are paired.

Step 1 (posterior draws, Eq. 10). For $j = 1, \dots, n_{mc}$ draw $Se_j \sim \text{Beta}\left(n_{11} + \frac{1}{2}, n_{10} + \frac{1}{2}\right)$ and $Sp_j \sim \text{Beta}\left(n_{00} + \frac{1}{2}, n_{01} + \frac{1}{2}\right)$, each restricted numerically to $[10^{-6}, 1 - 10^{-6}]$.

Step 2 (pseudo-raters, Eq. 11). For each j and $r = 1, \dots, R$ ($R = 3$): $n_{11,rj}^* \sim \text{Bin}(n_{1+}, Se_j)$, $n_{01,rj}^* \sim \text{Bin}(n_{0+}, 1 - Sp_j)$, $n_{10,rj}^* = n_{1+} - n_{11,rj}^*$, $n_{00,rj}^* = n_{0+} - n_{01,rj}^*$, and κ_{rj}^* by (S7).

Step 3 (transformation and Bayesian pooling, Eqs. 12–14). 3.1 Truncate to $[-1 + 10^{-8}, 1 - 10^{-8}]$ and set $\eta_{rj} = \log\{(1 + \kappa_{rj}^*)/(1 - \kappa_{rj}^*)\}$. 3.2 Fit (S22)–(S23) with the blocked Gibbs sampler of Proposition 14: four chains, iterations and burn-in as in Table S13. 3.3 Accept the fit if $\max(\hat{R}_{\text{classical}}, \hat{R}_{\text{rank}}) \leq 1.05$; otherwise record an FWKC failure for the dataset. 3.4 $\hat{\kappa}_B$ and SE_B : posterior mean and standard deviation of $\tanh(\mu/2)$.

Step 4 (random forest). 4.1 For each target r : rows $j = 1, \dots, n_{mc}$; response η_{rj} ; predictors $Se_j, Sp_j, \eta_{r'j}$ ($r' \neq r$). 4.2 300 trees; mtry equal to $\min(3, \text{number of predictors})$; minimum node size 10; at most $\min(100, \lfloor \text{rows}/10 \rfloor)$ terminal nodes, using the randomForest package [28] of the main text. 4.3 Out-of-bag predictions only; $\hat{\kappa}_{RF}$ and SE_{RF} by (S28). 4.4 Skip a target whose response has at most five distinct values; if all are skipped, use the Bayesian component alone.

Step 5 (fusion, Eq. 15). $w_B = SE_B^{-1}/(SE_B^{-1} + SE_{RF}^{-1})$; $\hat{\kappa}_{FWKC}(c) = w_B \hat{\kappa}_B + (1 - w_B) \hat{\kappa}_{RF}$.

Step 6 (intervals, Eqs. 16–17). 6.1 Draw B nonparametric bootstrap resamples of the n subjects and repeat Steps 1–5 on each. 6.2 Primary interval: the 2.5th and 97.5th percentiles of $\hat{\kappa}_{FWKC}^{*(b)}$. 6.3 Secondary interval: (S31) with $n_0 = 100$ and SE_η^* the standard deviation of the transformed replicates; for $n \geq n_0$ it is an unexpanded normal interval on the η scale, not the percentile interval of 6.2. 6.4 At least $\max(20, \lfloor 0.8B \rfloor)$ valid replicates are required (24 of 29 in the simulations, 400 of 499 for the Weiner et al. data); otherwise the interval is recorded as a failure.

Step 7 (diagnostic only). SE_{cov} of Eq. (18) with $\epsilon = 10^{-10}$, compared with a bootstrap-over-draws estimate; not used for inference.

S3.11 Computational cost

Proposition 20 (cost per interval). With C chains of I iterations and forests of n_{tree} trees, the cost of one FWKC interval is

$$O(B [n + R n_{mc} C I + R n_{tree} n_{mc} \log n_{mc}]), \quad (S34)$$

which, apart from the $O(Bn)$ cost of resampling and tabulating, does not depend on n . The covariance diagnostic adds a bootstrap over draws of the same order, $O(B B_{cov} R n_{mc})$, which is also independent of n .

Proof. Each bootstrap replicate costs $O(n)$ to resample and tabulate; Steps 1–2 cost $O(R n_{mc})$ because only four counts are drawn per pseudo-rater table; each Gibbs sweep costs $O(R n_{mc})$ through the sufficient statistics of Proposition 14; growing n_{tree} trees on n_{mc} rows costs $O(n_{tree} n_{mc} \log n_{mc})$ per target. $\square$

In the primary grid ($B = 29$, $n_{mc} = 150$, $C \times I = 4 \times 400$) one interval took 50.5 seconds on average, against 0.012 seconds for the direct percentile bootstrap. For the Weiner et al. data ($B = 499$, $n_{mc} = 500$, 4×1000) one interval took 49–72 minutes (Table S18). The ratio of the two settings predicted by (S34) is between $(499 \times 500)/(29 \times 150) = 57$ and $57 \times 2.5 = 143$, depending on the share of the Gibbs sampler; the observed ratio, 58 to 86, lies in that range.

S4. Simulation design and evaluation

S4.1 Structure and computational settings

The simulation follows the ADEMP structure [10] described in Section 2.7 of the main text. Table S13 lists the settings of every analysis. The primary grid crosses $n \in \{30, 50, 100\}$, $p \in \{0.1, \dots, 0.9\}$ and $c \in \{0.1, \dots, 0.9\}$ (243 scenarios) with $M = 120$ independent datasets per scenario, 29,160 datasets in all.

Table S13. Computational settings of each analysis

Analysis	n	M	B	n_{mc}	Gibbs chains $\times$ iterations (burn-in)
Primary simulation (243 scenarios)	30, 50, 100	120	29	150	4×400 (100)
Robustness (162 scenarios)	30, 50, 100	40	49	150	4×400 (100)
Sensitivity (34 variants $\times$ 3 scenarios)	30, 50	10	29*	150*	4×400 (100)
Anchor bridge (9 scenarios)	620	15	29	150	4×400 (100)
Weiner et al. data	620	—	499	500	4×1000 (300), thinning 2
AuditC data (14 studies $\times$ 9 values of c)	112–6,955	—	49	150	4×400 (100)

*Baseline values; the sensitivity variants used $B \in \{100, 200, 500\}$ and $n_{mc} \in \{250, 500, 1000, 2000\}$. All analyses used $R = 3$ pseudo-raters, forests of 300 trees and the random seed 20260912. The primary grid required 108 hours of wall-clock time on four parallel workers.

S4.2 Data-generating mechanisms

In every mechanism the true value is $\kappa(c)$ of (S3) at the marginal (p, Se, Sp) and c .

- (a) *Primary grid.* $D_i \sim \text{Bernoulli}(p)$, and $T_i \mid D_i = 1 \sim \text{Bernoulli}(Se)$, $T_i \mid D_i = 0 \sim \text{Bernoulli}(1 - Sp)$, independently for $i = 1, \dots, n$, with $Se = 0.854$ and $Sp = 0.667$.
- (b) *Four independent mechanisms.* As (a) with $(Se, Sp) = (0.75, 0.75)$, $(0.90, 0.90)$, $(0.95, 0.70)$ and $(0.70, 0.95)$.
- (c) *Clustered mechanism.* Subjects form clusters of $m = 10$. Cluster k draws $Se_k \sim \text{Beta}(Se \phi, (1 - Se)\phi)$ and $Sp_k \sim \text{Beta}(Sp \phi, (1 - Sp)\phi)$ with $Se = Sp = 0.75$ and $\phi = 25$, shared by its members; disease status is drawn independently for each subject.
- (d) *Mixture mechanism.* Each subject is assigned independently, with probability $\frac{1}{2}$, to a class with $(Se, Sp) = (0.90, 0.60)$ or $(0.60, 0.90)$.
- (e) *Anchor bridge.* Mechanism (a) at $n = 620$ and $p = 0.8935$, the observed prevalence of the Weiner et al. data.

Mechanisms (b)–(d) were evaluated across the full $3 \times 3 \times 3$ grid of $n \in \{30, 50, 100\}$, $p \in \{0.1, 0.5, 0.9\}$ and $c \in \{0.1, 0.5, 0.9\}$ (162 scenarios in total, 27 per mechanism, $M = 40$).

Proposition 21 (the clustered mechanism). (a) The marginal sensitivity and specificity are Se and Sp , so $\kappa(c)$ is unchanged. (b) Two members of the same cluster are dependent, with intra-cluster correlation $1/(\phi + 1) = 0.038$ for the test results of two diseased or two non-diseased members. (c) With clusters of equal size m , the variance of the false-positive count is

$$\text{Var}(n_{01}) = n \pi_{01} (1 - \pi_{01}) \left\{ 1 + \frac{(m-1)(1-p)^2 Sp(1-Sp)}{(\phi+1) \pi_{01} (1-\pi_{01})} \right\}, \quad (\text{S35})$$

and analogously for n_{11} with $p^2 Se(1 - Se)$ and π_{11} ; under independence the braces equal one.

Proof. (a) $E(Se_k) = Se$ and $E(Sp_k) = Sp$ for the Beta distributions used. (b) $\text{Var}(Se_k) = Se(1 - Se)/(\phi + 1)$, so for two diseased members $\text{Cov}(T_i, T_{i'}) = \text{Var}(Se_k)$ and the correlation is $1/(\phi + 1)$. (c) Let $X_i = (1 - D_i)T_i$, so $n_{01} = \sum_i X_i$ with $P(X_i = 1) = \pi_{01}$. For $i \neq i'$ in the same cluster, $\text{Cov}(X_i, X_{i'}) = (1 - p)^2 \text{Var}(1 - Sp_k) = (1 - p)^2 Sp(1 - Sp)/(\phi + 1)$, and members of different clusters are independent. Summing the n variances and the $n(m - 1)$ ordered within-cluster covariances gives (S35).

In the stress cell ($p = 0.1$, $Se = Sp = 0.75$, $m = 10$, $\phi = 25$), the design effect in (S35) is 1.30 for the false-positive count and 1.01 for the true-positive count: clustering inflates the variance of the count that carries most of the information at low prevalence by 30% (the design effect of cluster sampling [A4]; theoretical variance 6.81 against 5.23 under independence at $n = 30$). The clustered mechanism is thus the only one of the six that violates the independence assumption on which the delta-method intervals and the multinomial bootstrap rest.

Proposition 22 (the mixture mechanism is observationally equivalent to an independent mechanism). Under (d), the (D_i, T_i) are independent and identically distributed with $P(T_i = 1 | D_i = 1) = 0.75$ and $P(T_i = 0 | D_i = 0) = 0.75$. The distribution of the observed 2×2 table is therefore identical to that under mechanism (b) with $(Se, Sp) = (0.75, 0.75)$.

Proof. Class membership G_i is independent of D_i and of the other subjects, so $P(T_i = 1 | D_i = 1) = \frac{1}{2}(0.90) + \frac{1}{2}(0.60) = 0.75$ and $P(T_i = 1 | D_i = 0) = \frac{1}{2}(0.40) + \frac{1}{2}(0.10) = 0.25$. The pairs (D_i, T_i) are

independent across i , with the same four cell probabilities as under (b); the table is multinomial with those probabilities in both cases.

The mixture therefore checks that latent heterogeneity in accuracy that cannot be seen in the 2×2 table does not affect the methods; it does not test misspecification. It also provides a direct measure of Monte Carlo noise in the robustness tier: averaged over the 27-scenario grid used for each mechanism ($M=40$), the two equivalent mechanisms gave primary-interval coverage of 0.901 and 0.888, and direct-estimator RMSE of 0.152 and 0.145 (Table S19), showing that this noise is reduced by replication but not eliminated.

S4.3 Coverage from independent datasets

Proposition 23 (what the coverage estimate estimates). (a) With M independent datasets from mechanism F , $\widehat{Cov} = M^{-1} \sum_m I\{L_m \leq \kappa(c) \leq U_m\}$ (Eq. 19) is an unbiased estimator of the coverage probability $C(F) = P_F\{L \leq \kappa(c) \leq U\}$, with variance $C(1 - C)/M$. (b) An alternative quantity, computed from B bootstrap resamples $D^{*(b)}$ of a single simulated dataset D , $B^{-1} \sum_b I\{\kappa(c) \in CI(D^{*(b)})\}$, estimates $P_{\hat{F}_D}\{\kappa(c) \in CI(D^*)\}$: the probability that intervals computed from samples of the empirical distribution $\hat{F}_D$ contain the parameter of a different distribution, F . It depends on the particular D and is not an estimator of $C(F)$.

Proof. (a) The indicators are independent Bernoulli(C) variables. (b) Conditionally on D , the resamples are independent draws from $\hat{F}_D$, whose weighted kappa is $\hat{\kappa}_D(c)$, not $\kappa(c)$; the average of the indicators converges, as $B \rightarrow \infty$, to the stated conditional probability, which is a function of D . $\square$

All coverages in the main text and in this file use (a). With $M = 120$ the Monte Carlo standard error is 0.0199 at nominal coverage; across scenarios it is at most 0.038 for the primary FWKC interval and at most 0.045 for any of the six intervals; Wilson intervals [29] are given in the released output files. The robustness tier ($M = 20$) has a standard error of about 0.05 per scenario at nominal coverage and about 0.075 at the coverage actually observed (about 0.87); the sensitivity analyses ($M = 10$) have about 0.07 at nominal coverage and about 0.13 at the observed coverage (about 0.77). The values of 0.075 and 0.13 quoted in the main text refer to the observed coverage levels and to a single scenario. The coverages

reported for these two tiers are means over their three scenarios (60 and 30 datasets), for which the Monte Carlo standard errors are about 0.04 and about 0.08; the variants of the sensitivity tier are evaluated on the same simulated datasets, so differences between them are paired rather than independent.

S4.4 Failed replicates

A replicate whose interval could not be computed is excluded from the denominator in the main tables. Table S14 repeats the principal summaries with such replicates counted as non-coverage. Across the 29,160 primary datasets the FWKC fit failed in 6 and the direct estimator in 1. In all, 48 FWKC replicates (the 6 failed fits and 42 undefined intervals, in 28 scenarios, 26 of them at $n = 30$) and 5 direct percentile replicates (including the failed direct fit, in 4 scenarios) had no interval. No conclusion changes under the stricter convention.

Table S14. Coverage under two conventions for failed replicates

Interval	$n = 30$	$n = 50$	$n = 100$	Scenarios below 0.85 (of 243)	Minimum
FWKC, primary (Eq. 16)	0.875 / 0.871	0.882 / 0.882	0.885 / 0.885	29 / 31	0.775 / 0.775
FWKC, SE-expanded (Eq. 17)	0.977 / 0.973	0.977 / 0.977	0.934 / 0.934	0 / 0	0.850 / 0.850
Direct, percentile bootstrap	0.866 / 0.865	0.884 / 0.884	0.888 / 0.888	34 / 34	0.775 / 0.775

Each cell gives the value with failed replicates excluded from the denominator (main tables) / counted as non-coverage. Mean coverage over the 81 scenarios at each sample size; counts and minima over all 243 scenarios.

S4.5 Paired comparison of the FWKC and direct estimators (Eqs. 21–22)

Proposition 24 (variance of a paired difference). For two estimators computed on the same data, $\text{Var}(\hat{\kappa}_{FWKC} - \hat{\kappa}_D) = \sigma_F^2 + \sigma_D^2 - 2\sigma_{FD}$. If $\sigma_{FD} > 0$, a test that ignores the covariance uses a variance larger by the factor $(\sigma_F^2 + \sigma_D^2)/(\sigma_F^2 + \sigma_D^2 - 2\sigma_{FD})$ and is conservative.

Proof. Variance of a difference; the factor exceeds one when $\sigma_{FD} > 0$.

The two estimators are computed from the same table and are very strongly correlated. In the Weiner et al. data the bootstrap standard error of the difference is 0.002 to 0.004 (Table S15), against 0.04 to 0.06 for either estimator, which implies a correlation above 0.99; a test ignoring the covariance would have used a

standard error more than twenty times too large. For this reason the revised analysis pairs by construction. Across datasets (Eq. 21) the Δ_m are independent; the sign test is exact under a zero median, and the Wilcoxon signed-rank test (computed with its normal approximation) is valid under symmetry about zero, neither requiring normality of the differences. Within a dataset (Eq. 22) the percentile interval of the B paired bootstrap differences and the centred bootstrap p-value

$$p = \frac{1 + \sum_{b=1}^B I\{|\Delta^{*(b)} - \hat{\Delta}| \geq |\hat{\Delta}|\}}{B+1} \quad (\text{S36})$$

are used; the normal-theory statistic $\hat{\Delta}/\text{SD}(\Delta^*)$ is reported only as a secondary summary. Table S15 gives all three for the Weiner et al. data. The two nonparametric summaries can disagree when the bootstrap distribution of the difference is shifted relative to $\hat{\Delta}$ (at $c = 0.7$ the point difference lies outside its own percentile interval); the centred p-value is not affected by such a shift.

Table S15. Paired comparison of the FWKC and direct estimates within the Weiner et al. data (Eq. 22)

c	$\hat{\Delta}$	SD of Δ^*	95% percentile interval	Centred p (S36)	z (secondary)	$P^*(\Delta^* > 0)$
0.1	-0.0112	0.0036	(-0.0149, -0.0016)	0.010	-3.14	0.008
0.2	-0.0063	0.0030	(-0.0136, -0.0020)	0.078	-2.08	0.002
0.3	-0.0048	0.0030	(-0.0133, -0.0013)	0.212	-1.60	0.010
0.4	-0.0062	0.0027	(-0.0113, -0.0006)	0.028	-2.26	0.012
0.5	-0.0040	0.0024	(-0.0104, -0.0009)	0.148	-1.66	0.010
0.6	-0.0033	0.0022	(-0.0094, -0.0005)	0.226	-1.48	0.014
0.7	-0.0092	0.0022	(-0.0080, +0.0001)	0.032	-4.19	0.030
0.8	-0.0015	0.0022	(-0.0079, +0.0007)	0.612	-0.68	0.054
0.9	-0.0031	0.0021	(-0.0072, +0.0012)	0.150	-1.50	0.084

$\hat{\Delta} = \hat{\kappa}_{FWKC} - \hat{\kappa}_D$; $B = 499$ paired bootstrap resamples. The negative sign at every c is explained by Table S9.

S4.6 Scenario-level comparative criteria (Figures S5–S8)

ADV1: primary FWKC interval closer to 0.95 than the direct percentile bootstrap. ADV2: primary FWKC interval closer to 0.95 than the transformed interval. ADV3: SE-expanded interval closer to 0.95 than the transformed interval. ADV4: SE-expanded interval within [0.91, 0.99] while the transformed interval is not. ADV5: primary FWKC interval narrower than the direct percentile bootstrap, with coverage lower by

at most 0.01. ADV6: lower RMSE than the direct estimator. ADV7: lower MAE than the direct estimator. ADV8: ADV6 together with a Wilcoxon signed-rank p-value below 0.05 for the paired absolute-error differences δ_m of Eq. (21). Each criterion is evaluated per scenario; the counts over the 243 scenarios are 98, 23, 77, 28, 82, 95, 98 and 13. These eight criteria are descriptive scenario-level comparisons chosen to expose different trade-offs in calibration, width and point accuracy; they are not combined into an inferential score, and no overall superiority claim is based on them.

S4.7 External data

The 14 AuditC diagnostic-accuracy tables of Kriston et al. [12] are listed in Table S16. Each was analysed at the nine values of c (126 fits) with the settings of Table S13.

Table S16. The 14 published AuditC 2×2 tables used for the external cross-domain analysis

Study	n	TP	FN	FP	TN	Prevalence	$\widehat{se}$	$\widehat{sp}$
AuditC 01	895	47	9	101	738	0.063	0.839	0.880
AuditC 02	1992	126	51	272	1543	0.089	0.712	0.850
AuditC 03	233	19	10	12	192	0.124	0.655	0.941
AuditC 04	393	36	3	78	276	0.099	0.923	0.780
AuditC 05	1319	130	19	211	959	0.113	0.872	0.820
AuditC 06	243	84	2	68	89	0.354	0.977	0.567
AuditC 07	603	68	0	112	423	0.113	1.000	0.791
AuditC 08	6955	752	0	3226	2977	0.108	1.000	0.480
AuditC 09	255	59	5	55	136	0.251	0.922	0.712
AuditC 10	3551	142	50	571	2788	0.054	0.740	0.830
AuditC 11	626	137	24	107	358	0.257	0.851	0.770
AuditC 12	600	57	3	103	437	0.100	0.950	0.809
AuditC 13	112	34	1	21	56	0.312	0.971	0.727
AuditC 14	555	152	51	88	264	0.366	0.749	0.750

Counts as published by Kriston et al. [12].

S5. Additional results

S5.1 The Weiner et al. data

Table S17 gives the direct estimate with all five direct intervals, and Table S18 the FWKC estimate, both FWKC intervals and the component diagnostics, at each value of c . Tables S9 and S11 (Section S3) decompose the difference between the two estimators, and Table S15 (Section S4.5) tests it. At $n = 620$ the expansion factor of Eq. (17) equals one, and both closed-form intervals are in their asymptotic regime; this dataset motivates the problem but does not represent the small-sample setting in which the framework is designed to help (Section 3.1 of the main text).

Table S17. Weiner et al. data: direct estimate and five 95% intervals by weighting index

c	$\hat{\kappa}_D(c)$	Wald (Eq. 8)	Logit of [3, 9]	Transformed (Eq. 9)	Percentile bootstrap	Basic bootstrap
0.1	0.524	(0.405, 0.643)	(0.406, 0.639)	(0.395, 0.633)	(0.397, 0.628)	(0.419, 0.651)
0.2	0.476	(0.367, 0.584)	(0.370, 0.584)	(0.360, 0.577)	(0.361, 0.596)	(0.356, 0.590)
0.3	0.436	(0.334, 0.538)	(0.338, 0.539)	(0.329, 0.532)	(0.332, 0.533)	(0.339, 0.540)
0.4	0.402	(0.305, 0.500)	(0.310, 0.502)	(0.301, 0.495)	(0.297, 0.499)	(0.305, 0.507)
0.5	0.373	(0.280, 0.467)	(0.285, 0.471)	(0.276, 0.463)	(0.278, 0.458)	(0.288, 0.469)
0.6	0.348	(0.257, 0.439)	(0.264, 0.444)	(0.254, 0.436)	(0.256, 0.441)	(0.256, 0.441)
0.7	0.327	(0.238, 0.415)	(0.245, 0.420)	(0.236, 0.412)	(0.240, 0.420)	(0.233, 0.413)
0.8	0.307	(0.221, 0.393)	(0.229, 0.399)	(0.219, 0.390)	(0.227, 0.394)	(0.220, 0.387)
0.9	0.290	(0.207, 0.374)	(0.214, 0.380)	(0.205, 0.371)	(0.211, 0.370)	(0.210, 0.369)

The logit interval is recomputed from our estimates and delta-method standard errors and reproduces the published intervals at eight of the nine values of c (Table 5 of the main text). Bootstrap intervals use $B = 499$ resamples.

Table S18. Weiner et al. data: FWKC estimate, both FWKC intervals and component diagnostics by weighting index

c	$\hat{\kappa}_{FWKC}(c)$	Primary (Eq. 16)	SE-expanded (Eq. 17)	$\hat{\kappa}_B$	$\hat{\kappa}_{RF}$	w_B	$\hat{R}$	Corr(B , RF)	Minutes
0.1	0.513	(0.389, 0.619)	(0.384, 0.622)	0.485	0.524	0.30	0.999	0.998	72
0.2	0.470	(0.355, 0.585)	(0.347, 0.577)	0.456	0.475	0.27	1.002	0.997	72
0.3	0.431	(0.326, 0.526)	(0.326, 0.526)	0.414	0.437	0.23	1.000	0.996	72

c	$\hat{\kappa}_{FWKC}(c)$	Primary (Eq. 16)	SE-expanded (Eq. 17)	$\hat{\kappa}_B$	$\hat{\kappa}_{RF}$	w_B	$\hat{R}$	Corr(B , RF)	Minutes
0.4	0.396	(0.289, 0.492)	(0.295, 0.489)	0.375	0.401	0.20	1.000	0.996	72
0.5	0.369	(0.276, 0.452)	(0.278, 0.454)	0.351	0.374	0.20	0.999	0.995	65
0.6	0.345	(0.253, 0.434)	(0.253, 0.431)	0.324	0.349	0.17	1.001	0.995	49
0.7	0.317	(0.238, 0.415)	(0.225, 0.404)	0.296	0.322	0.20	1.000	0.995	65
0.8	0.306	(0.225, 0.389)	(0.220, 0.387)	0.288	0.310	0.18	1.000	0.994	65
0.9	0.287	(0.210, 0.365)	(0.205, 0.365)	0.276	0.289	0.16	1.001	0.993	65

Primary: full-pipeline percentile interval ($B = 499$, $n_{mc} = 500$). SE-expanded: Eq. (17) with $\lambda_n = 1$ at $n = 620$. Corr(B , RF): correlation of the two components across bootstrap resamples. Minutes: wall-clock time for one FWKC interval. All 499 resamples were valid at every c .

S5.2 Primary grid: coverage by prevalence and by weighting index

Figures S1–S4 complement Figures 3 and 4 of the main text. Scenario-level results for all 243 scenarios are given in Table S23 (Section S5.8).

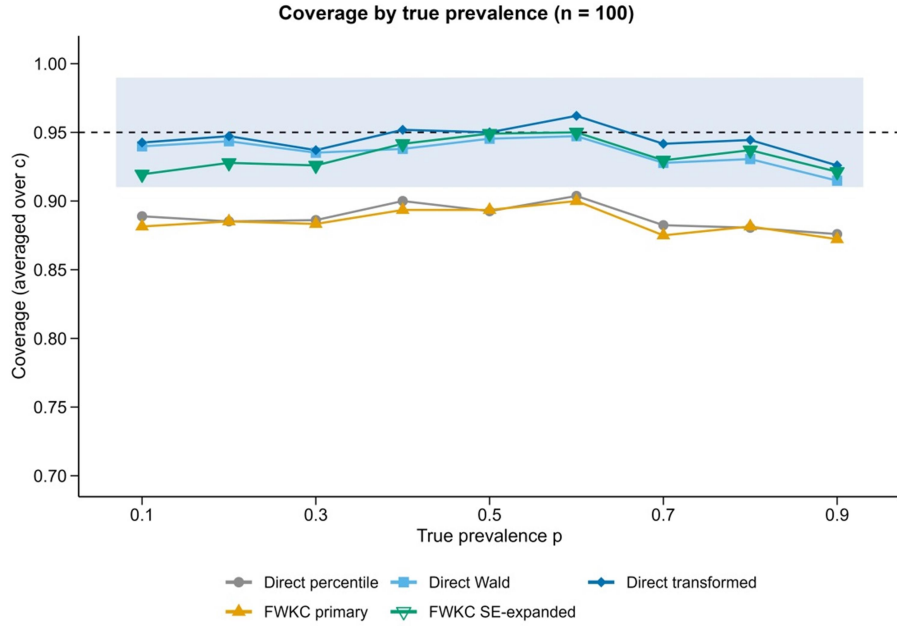


Figure S1. Coverage by true prevalence at $n = 100$

Averaged over the nine values of c . The shaded band marks 0.95 ± 2 Monte Carlo standard errors ($M = 120$). Complements Figure 4 of the main text, which shows $n = 30$ and $n = 50$.

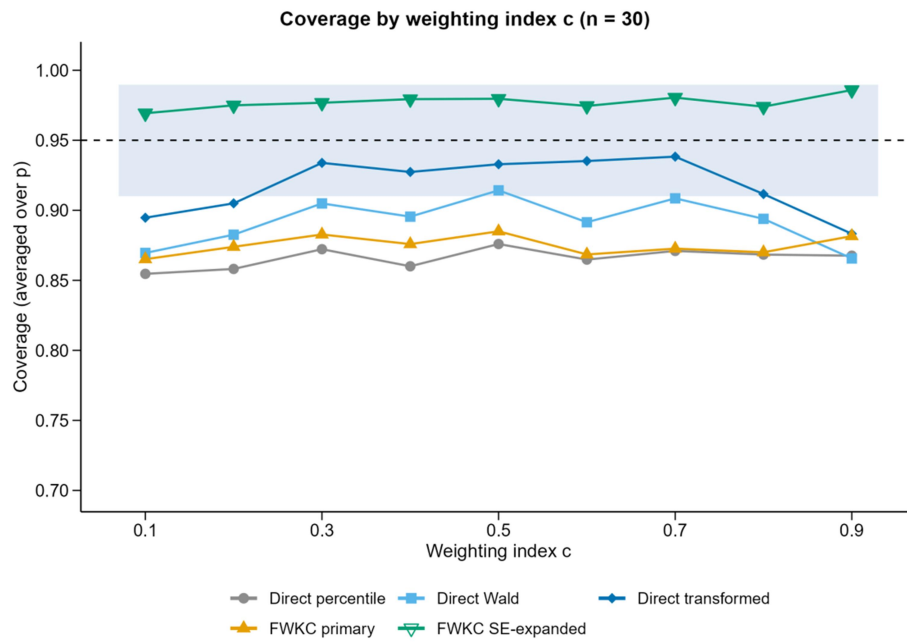


Figure S2. Coverage by weighting index at $n = 30$

Averaged over the nine prevalences; shading as in Figure S1.

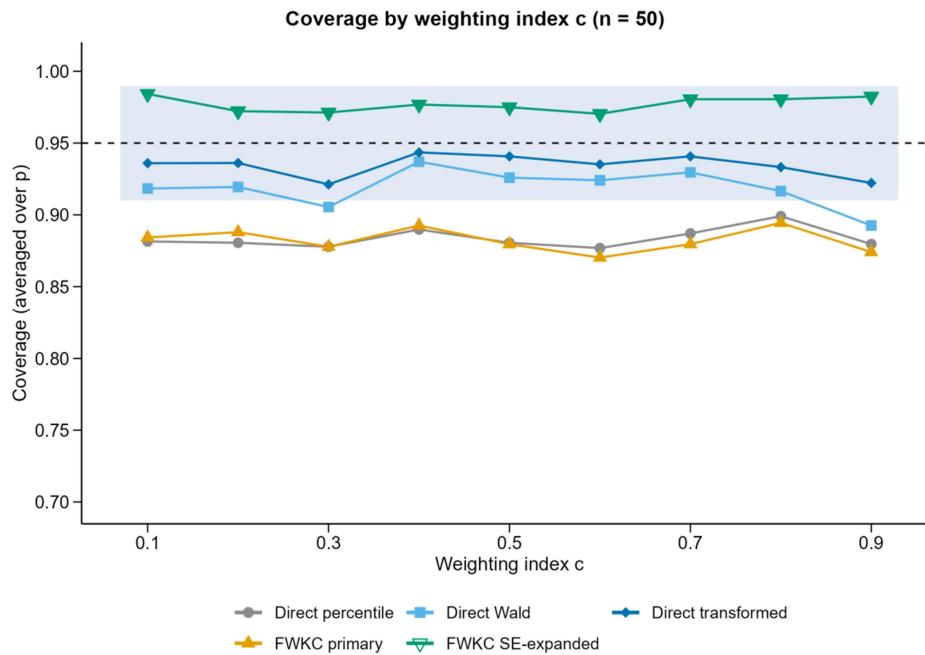


Figure S3. Coverage by weighting index at $n = 50$

Averaged over the nine prevalences; shading as in Figure S1.

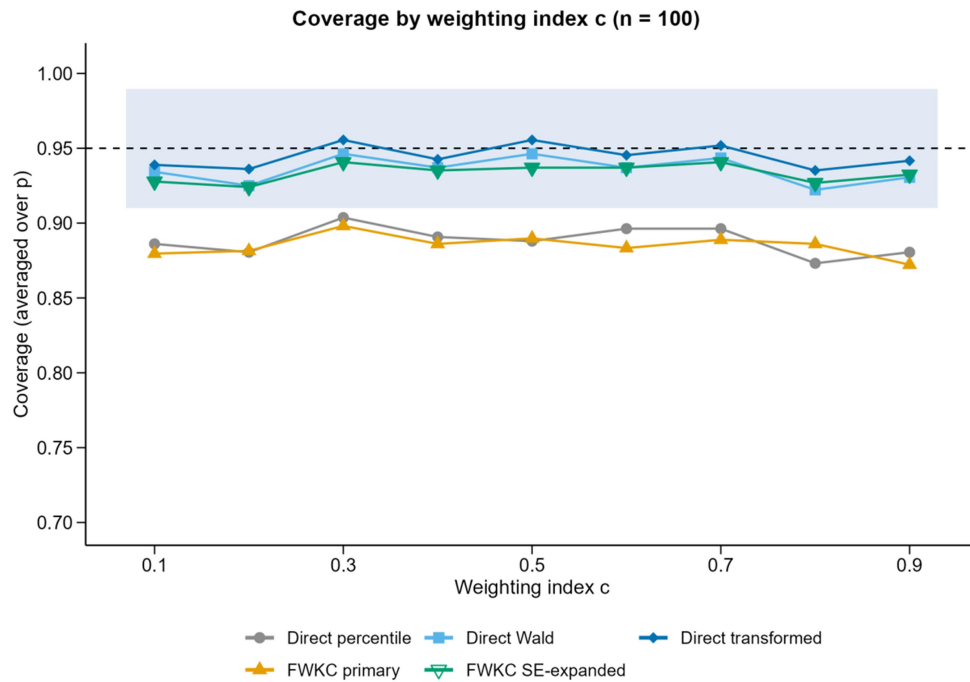


Figure S4. Coverage by weighting index at $n = 100$
Averaged over the nine prevalences; shading as in Figure S1.

S5.3 Scenario-level advantage criteria

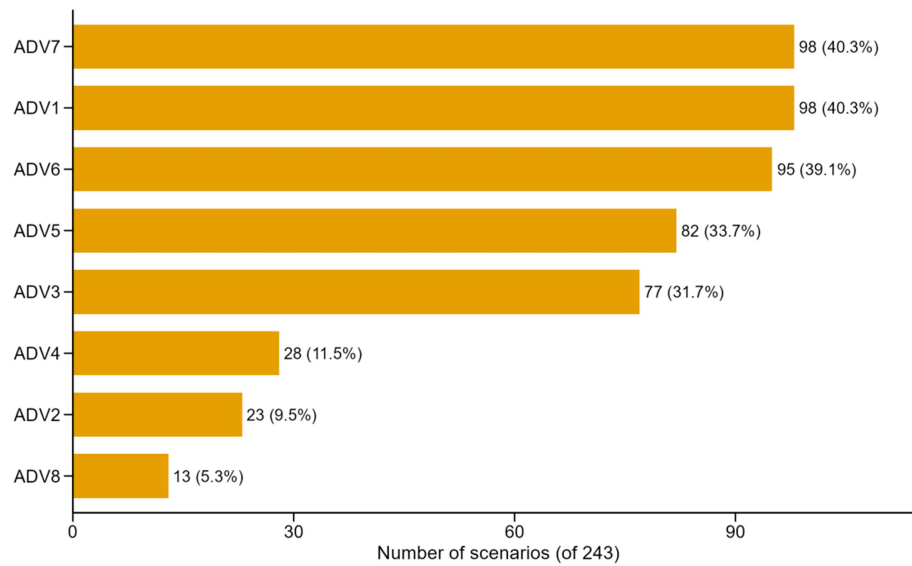


Figure S5. Number of primary scenarios satisfying each advantage criterion
Out of 243 scenarios; criteria ADV1–ADV8 are defined in Section S4.6. They are head-to-head and scenario-specific, and should be read together with the worst-case results of Tables 7 and 11 of the main text.

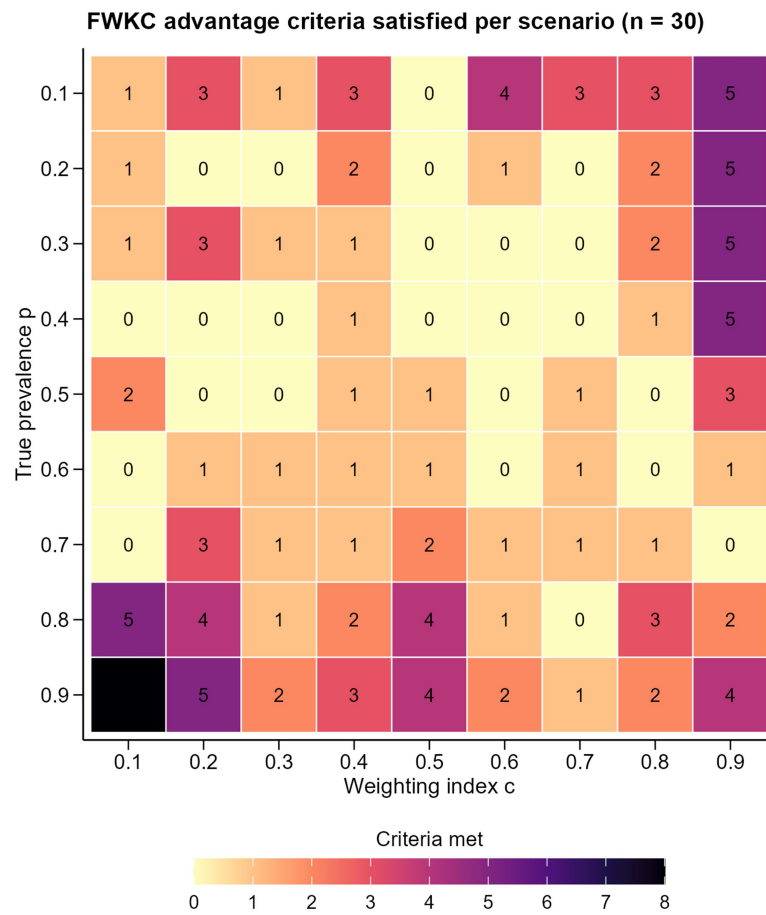


Figure S6. Advantage criteria satisfied per scenario at $n = 30$

Number of criteria ADV1–ADV8 satisfied, by true prevalence and weighting index.

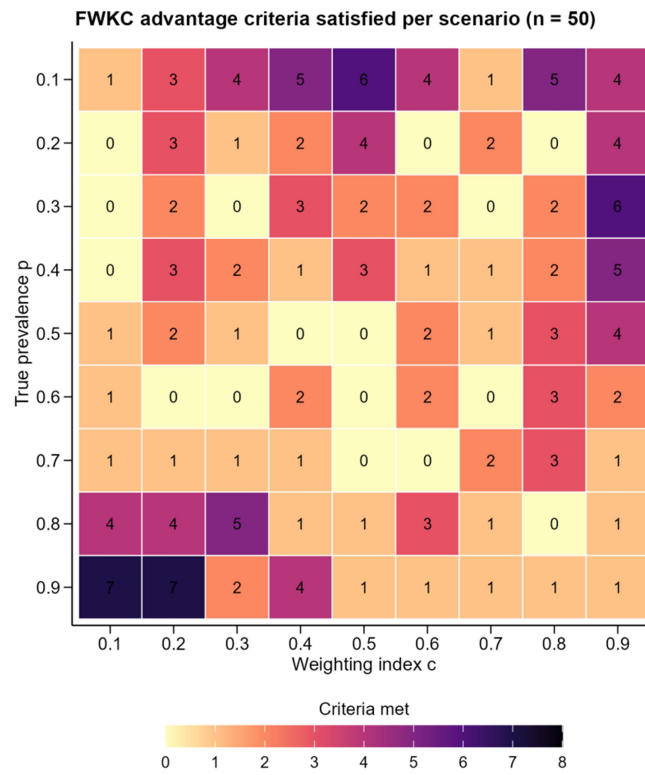


Figure S7. Advantage criteria satisfied per scenario at $n = 50$

As Figure S6.

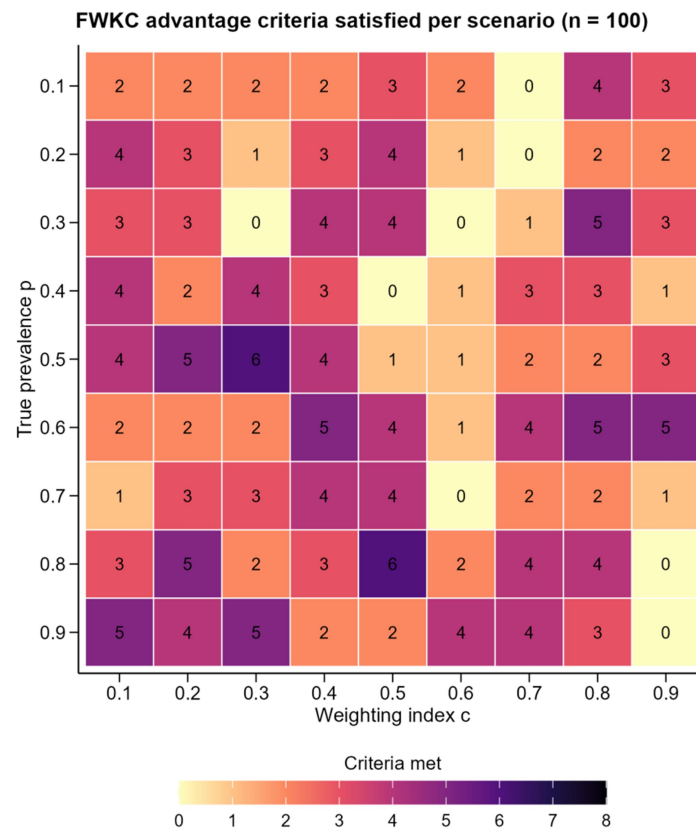


Figure S8. Advantage criteria satisfied per scenario at $n = 100$

As Figure S6.

S5.4 Robustness to the data-generating mechanism

Table S19 gives every one of the 162 robustness scenarios and Figure S9 summarises them by mechanism×n cell; Table 12 of the main text gives the means by mechanism group. Across the full grid the SE-expanded interval's coverage fell below 0.85 in only 1 of the 162 scenarios (0.825, under the two-component mixture mechanism at n = 30, p = 0.1, c = 0.5). Counting failed FWKC fits as non-coverage — which reaches 10% in the most demanding cells (large sensitivity–specificity imbalance at n = 30) — lowers the minimum SE-expanded coverage in this tier to 0.675, in 9 of the 162 scenarios; the worst-case claim of “never below 0.85” is stated in the main text only for the 243 primary scenarios. With M = 40, one replicate changes a scenario's coverage by 0.025.

Table S19. Coverage and RMSE across the 162 robustness scenarios (full 3×3×3 grid of n, p and c)

Mechanism (Se, Sp)	n	P	c	M	kc	FWKC primary	FWKC SE-exp.	Delta-method*	Percentile	Basic	RMSE direct	RMSE FWKC
Independent (0.75, 0.75)	30	0.1	0.1	40	0.180	0.900	0.950	0.900	0.875	0.875	0.138	0.173
Independent (0.75, 0.75)	30	0.1	0.5	40	0.265	0.850	0.950	0.850	0.850	0.775	0.204	0.203
Independent (0.75, 0.75)	30	0.1	0.9	40	0.500	0.800	0.900	0.675	0.825	0.500	0.339	0.285
Independent (0.75, 0.75)	30	0.5	0.1	40	0.500	0.900	0.975	0.925	0.900	0.875	0.187	0.202
Independent (0.75, 0.75)	30	0.5	0.5	40	0.500	0.950	1.000	0.950	0.950	0.900	0.164	0.166
Independent (0.75, 0.75)	30	0.5	0.9	40	0.500	0.950	1.000	0.925	0.950	0.825	0.179	0.200
Independent (0.75, 0.75)	30	0.9	0.1	40	0.500	0.825	0.900	0.769	0.800	0.675	0.315	0.251
Independent (0.75, 0.75)	30	0.9	0.5	40	0.265	0.875	0.900	0.850	0.850	0.725	0.187	0.177
Independent (0.75, 0.75)	30	0.9	0.9	40	0.180	0.825	0.875	0.816	0.800	0.750	0.162	0.212
Independent (0.75, 0.75)	50	0.1	0.1	40	0.180	0.975	1.000	0.975	0.975	0.950	0.098	0.100
Independent (0.75, 0.75)	50	0.1	0.5	40	0.265	0.950	0.975	0.950	0.950	0.900	0.120	0.117
Independent (0.75, 0.75)	50	0.1	0.9	40	0.500	0.750	0.925	0.750	0.750	0.650	0.257	0.235
Independent (0.75, 0.75)	50	0.5	0.1	40	0.500	0.950	1.000	0.950	0.950	0.875	0.135	0.134
Independent (0.75, 0.75)	50	0.5	0.5	40	0.500	0.850	1.000	0.925	0.875	0.875	0.127	0.127
Independent (0.75, 0.75)	50	0.5	0.9	40	0.500	0.975	1.000	0.950	0.975	0.850	0.136	0.135
Independent (0.75, 0.75)	50	0.9	0.1	40	0.500	0.850	0.925	0.850	0.825	0.775	0.239	0.213
Independent (0.75, 0.75)	50	0.9	0.5	40	0.265	0.825	0.950	0.900	0.875	0.800	0.128	0.123
Independent (0.75, 0.75)	50	0.9	0.9	40	0.180	0.900	0.975	0.925	0.900	0.875	0.121	0.127
Independent (0.75, 0.75)	100	0.1	0.1	40	0.180	0.950	0.975	1.000	0.950	0.950	0.063	0.062
Independent (0.75, 0.75)	100	0.1	0.5	40	0.265	0.900	0.950	0.925	0.900	0.875	0.095	0.094
Independent (0.75, 0.75)	100	0.1	0.9	40	0.500	0.800	0.900	0.825	0.825	0.775	0.177	0.166
Independent (0.75, 0.75)	100	0.5	0.1	40	0.500	0.975	0.975	1.000	0.975	0.975	0.083	0.085
Independent (0.75, 0.75)	100	0.5	0.5	40	0.500	1.000	0.975	1.000	1.000	0.950	0.076	0.075

Mechanism (Se, Sp)	n	P	c	M	kc	FWKC primary	FWKC SE-exp.	Delta-method*	Percentile	Basic	RMSE direct	RMSE FWKC
Independent (0.75, 0.75)	100	0.5	0.9	40	0.500	0.950	0.950	0.975	0.950	0.925	0.086	0.086
Independent (0.75, 0.75)	100	0.9	0.1	40	0.500	0.950	0.950	0.975	0.975	0.975	0.134	0.133
Independent (0.75, 0.75)	100	0.9	0.5	40	0.265	0.975	1.000	0.975	0.975	0.975	0.080	0.074
Independent (0.75, 0.75)	100	0.9	0.9	40	0.180	0.925	0.925	0.925	0.950	0.900	0.073	0.072
Independent (0.90, 0.90)	30	0.1	0.1	39	0.468	0.838	0.919	0.750	0.775	0.625	0.247	0.321
Independent (0.90, 0.90)	30	0.1	0.5	40	0.590	0.912	0.941	0.825	0.895	0.658	0.260	0.303
Independent (0.90, 0.90)	30	0.1	0.9	38	0.800	0.944	0.972	0.789	0.921	0.658	0.201	0.198
Independent (0.90, 0.90)	30	0.5	0.1	38	0.800	0.778	0.972	0.675	0.675	0.675	0.119	0.123
Independent (0.90, 0.90)	30	0.5	0.5	40	0.800	0.829	0.971	0.775	0.850	0.675	0.119	0.151
Independent (0.90, 0.90)	30	0.5	0.9	39	0.800	0.862	0.966	0.675	0.650	0.675	0.146	0.152
Independent (0.90, 0.90)	30	0.9	0.1	38	0.800	0.971	1.000	0.632	0.921	0.553	0.224	0.163
Independent (0.90, 0.90)	30	0.9	0.5	40	0.590	0.825	0.875	0.943	0.846	0.795	0.265	0.304
Independent (0.90, 0.90)	30	0.9	0.9	39	0.468	0.917	0.944	0.846	0.825	0.750	0.235	0.321
Independent (0.90, 0.90)	50	0.1	0.1	40	0.468	0.921	1.000	0.875	0.875	0.800	0.183	0.239
Independent (0.90, 0.90)	50	0.1	0.5	39	0.590	0.923	0.974	0.925	0.925	0.900	0.175	0.184
Independent (0.90, 0.90)	50	0.1	0.9	40	0.800	0.950	1.000	0.725	0.800	0.625	0.135	0.132
Independent (0.90, 0.90)	50	0.5	0.1	39	0.800	0.868	0.947	0.900	0.850	0.675	0.105	0.117
Independent (0.90, 0.90)	50	0.5	0.5	40	0.800	0.947	0.974	0.850	0.875	0.825	0.079	0.102
Independent (0.90, 0.90)	50	0.5	0.9	40	0.800	0.974	1.000	0.950	0.925	0.850	0.099	0.113
Independent (0.90, 0.90)	50	0.9	0.1	40	0.800	0.900	0.975	0.825	0.900	0.725	0.140	0.132
Independent (0.90, 0.90)	50	0.9	0.5	40	0.590	0.921	0.974	0.900	0.900	0.875	0.174	0.218
Independent (0.90, 0.90)	50	0.9	0.9	40	0.468	0.897	0.949	0.825	0.875	0.800	0.181	0.222
Independent (0.90, 0.90)	100	0.1	0.1	40	0.468	0.850	0.950	0.875	0.875	0.825	0.139	0.154
Independent (0.90, 0.90)	100	0.1	0.5	40	0.590	0.900	0.975	0.900	0.875	0.850	0.108	0.113
Independent (0.90, 0.90)	100	0.1	0.9	40	0.800	0.925	0.950	0.725	0.875	0.650	0.122	0.116
Independent (0.90, 0.90)	100	0.5	0.1	40	0.800	0.825	0.900	0.825	0.825	0.825	0.079	0.082
Independent (0.90, 0.90)	100	0.5	0.5	40	0.800	0.975	0.975	0.950	0.950	0.925	0.048	0.051
Independent (0.90, 0.90)	100	0.5	0.9	40	0.800	0.875	0.975	0.950	0.925	0.925	0.073	0.079
Independent (0.90, 0.90)	100	0.9	0.1	40	0.800	0.875	0.950	0.800	0.800	0.725	0.117	0.114
Independent (0.90, 0.90)	100	0.9	0.5	40	0.590	0.875	0.975	0.900	0.850	0.825	0.126	0.131
Independent (0.90, 0.90)	100	0.9	0.9	40	0.468	0.950	0.950	0.950	0.950	0.900	0.118	0.128
Independent (0.95, 0.70)	30	0.1	0.1	40	0.194	0.875	0.900	0.875	0.875	0.825	0.098	0.116
Independent (0.95, 0.70)	30	0.1	0.5	40	0.298	0.975	0.975	0.975	0.950	0.800	0.129	0.142
Independent (0.95, 0.70)	30	0.1	0.9	40	0.650	0.750	0.975	0.825	0.875	0.750	0.234	0.267
Independent (0.95, 0.70)	30	0.5	0.1	40	0.542	0.950	1.000	0.975	0.950	0.875	0.140	0.165
Independent (0.95, 0.70)	30	0.5	0.5	40	0.650	0.974	1.000	0.875	0.925	0.850	0.132	0.166
Independent (0.95, 0.70)	30	0.5	0.9	40	0.812	0.897	0.974	0.700	0.750	0.700	0.129	0.130
Independent (0.95, 0.70)	30	0.9	0.1	37	0.650	0.882	0.941	0.694	0.811	0.649	0.329	0.289
Independent (0.95, 0.70)	30	0.9	0.5	38	0.609	0.935	0.935	0.865	0.892	0.676	0.237	0.283
Independent (0.95, 0.70)	30	0.9	0.9	36	0.574	0.875	0.938	0.711	0.725	0.575	0.308	0.300
Independent (0.95, 0.70)	50	0.1	0.1	40	0.194	0.850	0.925	0.900	0.825	0.850	0.110	0.115
Independent (0.95, 0.70)	50	0.1	0.5	40	0.298	0.850	0.975	0.900	0.925	0.850	0.122	0.126
Independent (0.95, 0.70)	50	0.1	0.9	40	0.650	0.900	1.000	0.925	0.925	0.875	0.140	0.167

Mechanism (Se, Sp)	n	P	c	M	kc	FWKC primary	FWKC SE-exp.	Delta-method*	Percentile	Basic	RMSE direct	RMSE FWKC
Independent (0.95, 0.70)	50	0.5	0.1	40	0.542	0.850	0.975	0.900	0.850	0.900	0.119	0.133
Independent (0.95, 0.70)	50	0.5	0.5	40	0.650	0.950	0.975	0.950	0.925	0.900	0.094	0.099
Independent (0.95, 0.70)	50	0.5	0.9	40	0.812	0.875	1.000	0.700	0.700	0.700	0.094	0.086
Independent (0.95, 0.70)	50	0.9	0.1	40	0.650	0.892	0.946	0.800	0.850	0.650	0.213	0.208
Independent (0.95, 0.70)	50	0.9	0.5	40	0.609	0.889	0.972	0.846	0.850	0.675	0.223	0.263
Independent (0.95, 0.70)	50	0.9	0.9	37	0.574	0.917	0.972	0.769	0.821	0.692	0.260	0.273
Independent (0.95, 0.70)	100	0.1	0.1	40	0.194	0.875	0.925	0.900	0.875	0.900	0.067	0.068
Independent (0.95, 0.70)	100	0.1	0.5	40	0.298	0.900	0.900	0.900	0.925	0.900	0.070	0.072
Independent (0.95, 0.70)	100	0.1	0.9	40	0.650	0.900	0.900	0.925	0.900	0.850	0.098	0.112
Independent (0.95, 0.70)	100	0.5	0.1	40	0.542	0.925	0.950	0.875	0.900	0.875	0.094	0.094
Independent (0.95, 0.70)	100	0.5	0.5	40	0.650	0.925	0.950	0.950	0.900	0.875	0.077	0.076
Independent (0.95, 0.70)	100	0.5	0.9	40	0.812	0.950	1.000	0.925	0.975	0.900	0.053	0.055
Independent (0.95, 0.70)	100	0.9	0.1	40	0.650	0.949	1.000	0.875	0.875	0.750	0.157	0.160
Independent (0.95, 0.70)	100	0.9	0.5	40	0.609	0.875	0.975	0.950	0.900	0.875	0.134	0.152
Independent (0.95, 0.70)	100	0.9	0.9	40	0.574	0.900	0.925	0.925	0.875	0.900	0.133	0.146
Independent (0.70, 0.95)	30	0.1	0.1	36	0.574	0.912	0.941	0.737	0.789	0.711	0.308	0.288
Independent (0.70, 0.95)	30	0.1	0.5	39	0.609	0.970	0.970	0.775	0.795	0.667	0.299	0.269
Independent (0.70, 0.95)	30	0.1	0.9	37	0.650	0.900	0.900	0.538	0.694	0.556	0.330	0.320
Independent (0.70, 0.95)	30	0.5	0.1	40	0.812	0.900	1.000	0.500	0.525	0.500	0.133	0.129
Independent (0.70, 0.95)	30	0.5	0.5	40	0.650	0.872	0.949	0.850	0.850	0.825	0.173	0.196
Independent (0.70, 0.95)	30	0.5	0.9	40	0.542	0.842	0.974	0.850	0.850	0.825	0.173	0.216
Independent (0.70, 0.95)	30	0.9	0.1	40	0.650	0.875	0.975	0.923	0.950	0.700	0.194	0.219
Independent (0.70, 0.95)	30	0.9	0.5	40	0.298	0.925	0.975	0.974	0.950	0.825	0.144	0.159
Independent (0.70, 0.95)	30	0.9	0.9	40	0.194	0.925	0.975	0.923	0.950	0.825	0.113	0.137
Independent (0.70, 0.95)	50	0.1	0.1	39	0.574	0.865	0.973	0.775	0.875	0.700	0.212	0.242
Independent (0.70, 0.95)	50	0.1	0.5	40	0.609	0.892	0.946	0.800	0.850	0.700	0.183	0.227
Independent (0.70, 0.95)	50	0.1	0.9	39	0.650	0.919	0.973	0.750	0.842	0.763	0.230	0.230
Independent (0.70, 0.95)	50	0.5	0.1	40	0.812	0.825	1.000	0.800	0.800	0.700	0.108	0.108
Independent (0.70, 0.95)	50	0.5	0.5	40	0.650	0.875	1.000	0.925	0.900	0.875	0.102	0.111
Independent (0.70, 0.95)	50	0.5	0.9	40	0.542	0.950	0.975	0.950	0.950	0.775	0.123	0.136
Independent (0.70, 0.95)	50	0.9	0.1	40	0.650	0.825	0.950	0.925	0.900	0.775	0.138	0.175
Independent (0.70, 0.95)	50	0.9	0.5	40	0.298	0.950	1.000	0.975	0.950	0.925	0.096	0.095
Independent (0.70, 0.95)	50	0.9	0.9	40	0.194	0.850	0.900	0.800	0.850	0.750	0.094	0.102
Independent (0.70, 0.95)	100	0.1	0.1	40	0.574	0.900	1.000	0.950	0.900	0.950	0.133	0.142
Independent (0.70, 0.95)	100	0.1	0.5	40	0.609	0.975	1.000	1.000	0.975	1.000	0.109	0.117
Independent (0.70, 0.95)	100	0.1	0.9	40	0.650	0.925	0.975	0.950	0.925	0.900	0.134	0.139
Independent (0.70, 0.95)	100	0.5	0.1	40	0.812	0.875	0.875	0.925	0.900	0.775	0.069	0.067
Independent (0.70, 0.95)	100	0.5	0.5	40	0.650	0.925	0.950	0.925	0.900	0.875	0.073	0.077
Independent (0.70, 0.95)	100	0.5	0.9	40	0.542	0.900	0.950	0.950	0.900	0.900	0.083	0.083
Independent (0.70, 0.95)	100	0.9	0.1	40	0.650	0.850	0.900	1.000	0.900	0.925	0.111	0.129
Independent (0.70, 0.95)	100	0.9	0.5	40	0.298	0.925	0.925	0.950	0.925	0.925	0.073	0.077
Independent (0.70, 0.95)	100	0.9	0.9	40	0.194	0.950	0.950	0.950	0.950	0.925	0.049	0.052
Clustered (0.75, 0.75)	30	0.1	0.1	40	0.180	0.850	0.900	0.850	0.846	0.821	0.156	0.206

Mechanism (Se, Sp)	n	P	c	M	kc	FWKC primary	FWKC SE-exp.	Delta-method*	Percentile	Basic	RMSE direct	RMSE FWKC
Clustered (0.75, 0.75)	30	0.1	0.5	40	0.265	0.875	0.900	0.875	0.900	0.775	0.178	0.204
Clustered (0.75, 0.75)	30	0.1	0.9	40	0.500	0.825	0.925	0.800	0.850	0.625	0.320	0.290
Clustered (0.75, 0.75)	30	0.5	0.1	40	0.500	0.872	1.000	0.850	0.825	0.775	0.191	0.202
Clustered (0.75, 0.75)	30	0.5	0.5	40	0.500	0.900	1.000	0.900	0.875	0.875	0.168	0.184
Clustered (0.75, 0.75)	30	0.5	0.9	40	0.500	0.850	1.000	0.875	0.850	0.800	0.199	0.212
Clustered (0.75, 0.75)	30	0.9	0.1	40	0.500	0.923	0.974	0.718	0.900	0.625	0.279	0.256
Clustered (0.75, 0.75)	30	0.9	0.5	40	0.265	0.875	0.900	0.919	0.875	0.775	0.157	0.170
Clustered (0.75, 0.75)	30	0.9	0.9	40	0.180	0.850	0.950	0.850	0.800	0.800	0.174	0.221
Clustered (0.75, 0.75)	50	0.1	0.1	40	0.180	0.800	1.000	0.875	0.825	0.825	0.116	0.117
Clustered (0.75, 0.75)	50	0.1	0.5	40	0.265	0.900	1.000	0.975	0.925	0.900	0.120	0.112
Clustered (0.75, 0.75)	50	0.1	0.9	40	0.500	0.950	1.000	0.900	0.950	0.750	0.175	0.150
Clustered (0.75, 0.75)	50	0.5	0.1	40	0.500	0.950	0.975	0.875	0.950	0.875	0.162	0.166
Clustered (0.75, 0.75)	50	0.5	0.5	40	0.500	0.900	0.975	0.875	0.825	0.825	0.147	0.153
Clustered (0.75, 0.75)	50	0.5	0.9	40	0.500	0.875	0.975	0.875	0.875	0.875	0.150	0.149
Clustered (0.75, 0.75)	50	0.9	0.1	40	0.500	0.975	1.000	0.850	0.925	0.750	0.209	0.186
Clustered (0.75, 0.75)	50	0.9	0.5	40	0.265	0.900	0.925	0.900	0.900	0.850	0.140	0.129
Clustered (0.75, 0.75)	50	0.9	0.9	40	0.180	0.825	0.925	0.800	0.825	0.775	0.090	0.089
Clustered (0.75, 0.75)	100	0.1	0.1	40	0.180	0.900	0.875	0.875	0.875	0.825	0.070	0.070
Clustered (0.75, 0.75)	100	0.1	0.5	40	0.265	0.950	0.950	0.950	0.950	0.925	0.091	0.090
Clustered (0.75, 0.75)	100	0.1	0.9	40	0.500	0.925	0.975	0.950	0.975	0.950	0.130	0.127
Clustered (0.75, 0.75)	100	0.5	0.1	40	0.500	0.900	0.925	0.925	0.875	0.900	0.104	0.102
Clustered (0.75, 0.75)	100	0.5	0.5	40	0.500	0.850	0.925	0.875	0.850	0.875	0.098	0.094
Clustered (0.75, 0.75)	100	0.5	0.9	40	0.500	0.975	0.925	0.950	0.925	0.850	0.098	0.098
Clustered (0.75, 0.75)	100	0.9	0.1	40	0.500	0.900	0.925	0.925	0.925	0.875	0.158	0.153
Clustered (0.75, 0.75)	100	0.9	0.5	40	0.265	0.875	0.850	0.825	0.850	0.850	0.102	0.101
Clustered (0.75, 0.75)	100	0.9	0.9	40	0.180	0.950	0.925	0.925	0.925	0.875	0.062	0.063
Mixture (0.75, 0.75)	30	0.1	0.1	40	0.180	0.850	0.875	0.825	0.850	0.750	0.125	0.151
Mixture (0.75, 0.75)	30	0.1	0.5	40	0.265	0.800	0.825	0.725	0.775	0.725	0.213	0.193
Mixture (0.75, 0.75)	30	0.1	0.9	40	0.500	0.850	0.900	0.750	0.800	0.700	0.288	0.250
Mixture (0.75, 0.75)	30	0.5	0.1	40	0.500	0.875	1.000	0.900	0.850	0.800	0.175	0.200
Mixture (0.75, 0.75)	30	0.5	0.5	40	0.500	0.950	1.000	0.975	0.925	0.925	0.128	0.148
Mixture (0.75, 0.75)	30	0.5	0.9	40	0.500	0.925	1.000	0.950	0.925	0.950	0.170	0.186
Mixture (0.75, 0.75)	30	0.9	0.1	40	0.500	0.875	0.950	0.821	0.875	0.700	0.281	0.262
Mixture (0.75, 0.75)	30	0.9	0.5	40	0.265	0.875	0.950	0.897	0.875	0.775	0.181	0.197
Mixture (0.75, 0.75)	30	0.9	0.9	40	0.180	0.875	0.900	0.872	0.875	0.850	0.116	0.163
Mixture (0.75, 0.75)	50	0.1	0.1	40	0.180	0.975	1.000	0.975	0.975	0.950	0.093	0.094
Mixture (0.75, 0.75)	50	0.1	0.5	40	0.265	0.825	0.950	0.850	0.875	0.775	0.171	0.167
Mixture (0.75, 0.75)	50	0.1	0.9	40	0.500	0.875	0.950	0.900	0.900	0.850	0.227	0.205
Mixture (0.75, 0.75)	50	0.5	0.1	40	0.500	0.925	1.000	0.900	0.925	0.875	0.142	0.143
Mixture (0.75, 0.75)	50	0.5	0.5	40	0.500	0.800	0.975	0.900	0.800	0.800	0.139	0.137
Mixture (0.75, 0.75)	50	0.5	0.9	40	0.500	0.850	1.000	0.925	0.875	0.825	0.152	0.152
Mixture (0.75, 0.75)	50	0.9	0.1	40	0.500	0.900	0.925	0.825	0.850	0.750	0.222	0.197
Mixture (0.75, 0.75)	50	0.9	0.5	40	0.265	0.900	0.925	0.923	0.900	0.900	0.123	0.121

Mechanism (Se, Sp)	n	P	c	M	kc	FWKC primary	FWKC SE-exp.	Delta-method*	Percentile	Basic	RMSE direct	RMSE FWKC
Mixture (0.75, 0.75)	50	0.9	0.9	40	0.180	0.825	0.900	0.872	0.825	0.825	0.114	0.120
Mixture (0.75, 0.75)	100	0.1	0.1	40	0.180	0.850	0.875	0.900	0.850	0.900	0.072	0.072
Mixture (0.75, 0.75)	100	0.1	0.5	40	0.265	0.925	0.975	0.975	0.900	0.975	0.095	0.093
Mixture (0.75, 0.75)	100	0.1	0.9	40	0.500	0.825	0.850	0.850	0.800	0.750	0.172	0.159
Mixture (0.75, 0.75)	100	0.5	0.1	40	0.500	0.900	0.975	0.925	0.950	0.900	0.089	0.091
Mixture (0.75, 0.75)	100	0.5	0.5	40	0.500	0.950	0.975	0.950	0.950	0.925	0.081	0.080
Mixture (0.75, 0.75)	100	0.5	0.9	40	0.500	0.975	0.975	1.000	0.975	0.950	0.085	0.086
Mixture (0.75, 0.75)	100	0.9	0.1	40	0.500	0.975	0.975	0.950	0.950	0.900	0.113	0.109
Mixture (0.75, 0.75)	100	0.9	0.5	40	0.265	0.950	0.975	1.000	0.975	1.000	0.086	0.084
Mixture (0.75, 0.75)	100	0.9	0.9	40	0.180	0.875	0.900	0.875	0.850	0.900	0.075	0.075

M = 40 independent datasets were generated per scenario (M column: datasets with a valid FWKC fit; the raw fitting-success rate fell to 0.90 in the most demanding cells); coverage denominators are the datasets with a valid interval for each method. The Monte Carlo standard error of a coverage value is approximately 0.047 near a coverage of 0.90 and approximately 0.034 near 0.95. Only the clustered mechanism violates independence (Proposition 21); the mixture is observationally equivalent to the independent (0.75, 0.75) mechanism (Proposition 22). *The “Delta-method” column reports the single delta-based interval recorded in the released per-scenario output and does not distinguish between the Wald (Eq. 8) and transformed (Eq. 9) variants; that distinction is disaggregated only at the mechanism-group level, in Table 12 of the main text.

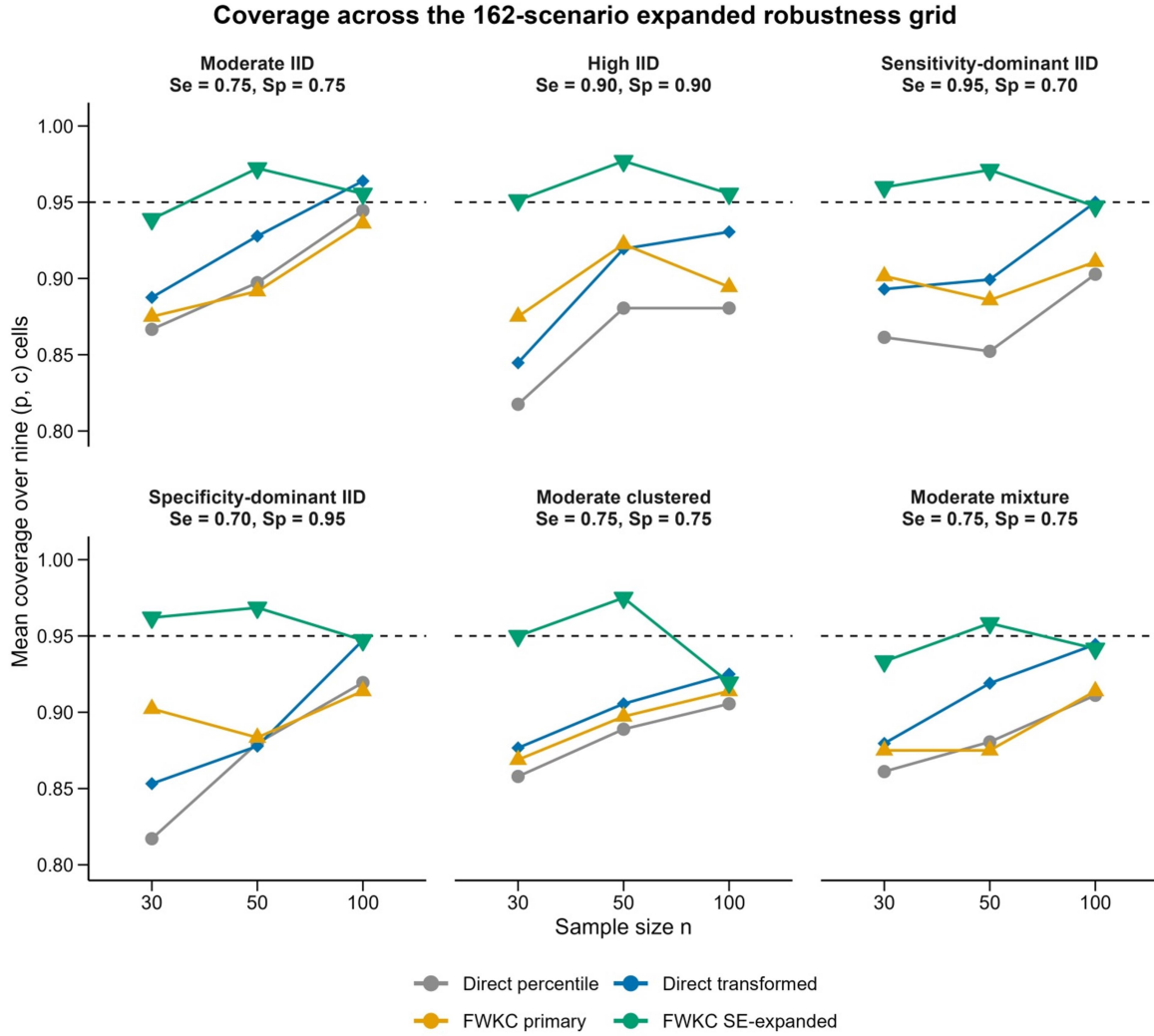


Figure S9. Coverage across the 162-scenario robustness grid

Six mechanisms, each evaluated across the full $3 \times 3 \times 3$ grid of $n \in \{30, 50, 100\}$, $p \in \{0.1, 0.5, 0.9\}$ and $c \in \{0.1, 0.5, 0.9\}$ ($M = 40$); each plotted point is the mean coverage over the 9 (p, c) combinations for that mechanism $\times$ n cell. Monte Carlo standard error is approximately 0.047 per scenario near a coverage of 0.90, smaller for these 9-cell means. Full per-scenario values are in Table S19.

S5.5 Anchor bridge at $n = 620$

Table S20 and Figure S10 give the bridge analysis of Section 3.4 of the main text. At this sample size the expansion factor of Eq. (17) is inactive and both closed-form intervals had higher coverage than both FWKC intervals.

Table S20. Anchor bridge at $n = 620$: coverage and RMSE by weighting index

c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	RMSE direct	RMSE FWKC
0.1	0.524	0.933	0.933	0.933	0.933	0.933	0.050	0.050
0.2	0.476	0.933	0.933	1.000	0.933	0.933	0.042	0.040
0.3	0.436	0.933	0.933	1.000	0.933	0.867	0.058	0.057
0.4	0.402	0.800	0.867	0.933	0.933	0.800	0.056	0.055
0.5	0.373	0.733	0.867	0.867	0.867	0.733	0.064	0.062
0.6	0.348	0.867	0.933	1.000	1.000	0.867	0.050	0.051
0.7	0.327	1.000	1.000	1.000	1.000	1.000	0.032	0.033
0.8	0.307	0.867	1.000	1.000	1.000	0.867	0.038	0.039
0.9	0.290	0.933	1.000	1.000	1.000	0.933	0.033	0.032
Mean		0.889	0.941	0.970	0.956	0.881	0.047	0.047

$p = 0.8935$, $Se = 0.854$, $Sp = 0.667$; $M = 15$ datasets per value of c . At this sample size $\lambda_n = 1$; the Wald interval was closest to nominal, the transformed interval reached 1.000 at six of the nine values of c , and the primary FWKC interval under-covered. With $M = 15$ the Monte Carlo standard error of a coverage value is about 0.056.

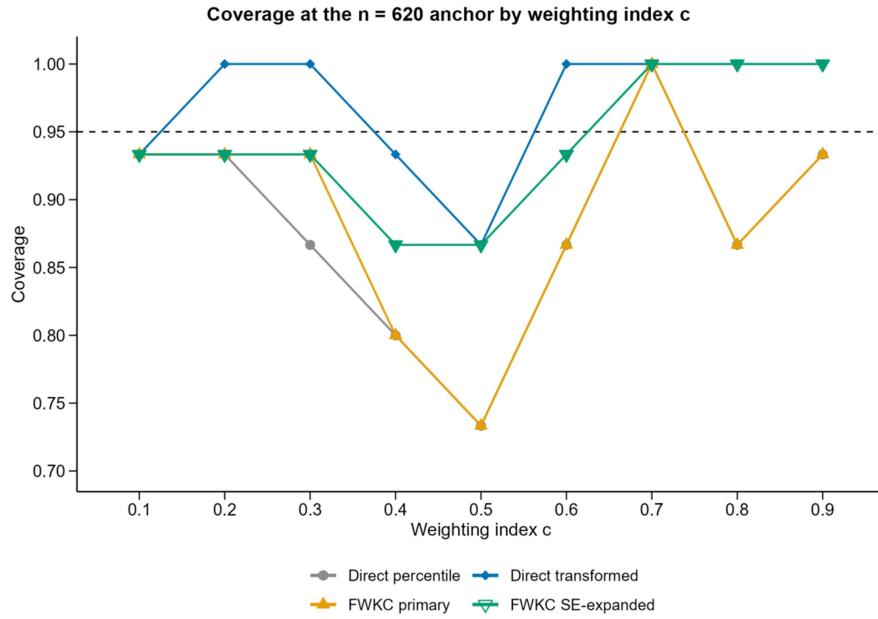


Figure S10. Coverage at the anchor sample size $n = 620$ by weighting index

$M = 15$ datasets per value of c ; values in Table S20. The expansion factor of Eq. (17) is inactive at this sample size.

S5.6 Sensitivity analyses

Each of the 34 variants changes one setting of the baseline and is evaluated on the three stress scenarios of Section 2.9 of the main text (Table S21). Figure 6 of the main text displays these results graphically.

Table S21. Coverage and RMSE for all 34 sensitivity variants

#	Factor	Variant	FWKC primary	FWKC SE-exp.	Direct transformed	Direct percentile	RMSE FWKC
1	Prior	Jeffreys (reference)	0.767	0.933	0.867	0.833	0.228
2	Prior	Uniform	0.800	0.933	0.867	0.833	0.230
3	Prior	Naïve Beta($nSe, n(1 - Se)$)	0.871	1.000	0.867	0.867	0.218
4	Generator	Beta posterior (reference)	0.767	0.933	0.867	0.833	0.228
5	Generator	Stratified binomial bootstrap	0.743	1.000	0.867	0.767	0.226
6	Pseudo-rater mode	Posterior-predictive (reference)	0.767	0.933	0.867	0.833	0.228
7	Pseudo-rater mode	Parameter-only	0.733	0.967	0.867	0.733	0.228
8	Fusion rule	Inverse variance	0.767	0.933	0.867	0.833	0.229
9	Fusion rule	Inverse SE (reference)	0.767	0.933	0.867	0.833	0.228
10	Fusion rule	Equal weights	0.767	0.933	0.867	0.833	0.228
11	Fusion rule	Bayesian only	0.800	0.933	0.867	0.833	0.230
12	Fusion rule	Random forest only	0.767	0.967	0.867	0.833	0.226
13	Transformation	$2 \times$ Fisher z (reference)	0.767	0.933	0.867	0.833	0.228
14	Transformation	Fisher z	0.867	0.933	0.867	0.867	0.228
15	Transformation	Identity	0.767	0.933	0.867	0.833	0.226
16	ML layer	Random forest (reference)	0.767	0.933	0.867	0.833	0.228
17	ML layer	Cross-fitted splines	0.833	0.967	0.867	0.800	0.228
18	ML layer	None (Bayesian only)	0.767	0.967	0.867	0.833	0.230
19	Calibration rule	No expansion	0.767	0.867	0.867	0.833	0.228
20	Calibration rule	Square-root SE expansion (reference)	0.767	0.933	0.867	0.833	0.228
21	Calibration rule	Linear SE expansion	0.767	0.967	0.867	0.833	0.228
22	Calibration rule	Adaptive (n_{eff}) expansion	0.767	1.000	0.867	0.833	0.228
23	Calibration rule	contraction rule	0.767	0.533	0.867	0.833	0.228

#	Factor	Variant	FWKC primary	FWKC SE-exp.	Direct transformed	Direct percentile	RMSE FWKC
24	n_0	75	0.767	0.933	0.867	0.833	0.228
25	n_0	100 (reference)	0.767	0.933	0.867	0.833	0.228
26	n_0	125	0.767	0.967	0.867	0.833	0.228
27	n_0	150	0.767	0.967	0.867	0.833	0.228
28	n_{mc}	250	0.800	0.967	0.867	0.767	0.227
29	n_{mc}	500	0.833	0.967	0.867	0.833	0.228
30	n_{mc}	1000	0.833	0.967	0.867	0.833	0.226
31	n_{mc}	2000	0.833	0.967	0.867	0.867	0.225
32	B	100	0.833	0.967	0.867	0.867	0.228
33	B	200	0.867	0.967	0.867	0.867	0.228
34	B	500	0.867	0.967	0.867	0.867	0.228

Each variant changes one setting of the baseline (“reference”) and is evaluated on the three stress scenarios of Section 2.9 of the main text with $M = 10$ datasets each; values are means over the three scenarios. Rows marked “reference” repeat the baseline within each factor. The direct intervals do not depend on the FWKC settings except through the shared bootstrap resamples.

S5.7 External AuditC data

Table S22. External AuditC data: estimates, interval widths and prevalence-adjusted measures by study

Study	n	Mean $\hat{\kappa}_D$	Mean $\hat{\kappa}_{FWKC}$	Max. abs. diff.	Width, percentile	Width, FWKC primary	Width, FWKC SE- exp.	PABAK	AC1
AuditC 01	895	0.437	0.432	0.009	0.154	0.154	0.168	0.754	0.846
AuditC 02	1992	0.380	0.376	0.007	0.095	0.095	0.104	0.676	0.785
AuditC 03	233	0.579	0.569	0.025	0.305	0.300	0.330	0.811	0.878
AuditC 04	393	0.419	0.411	0.014	0.180	0.180	0.196	0.588	0.700
AuditC 05	1319	0.472	0.467	0.009	0.101	0.100	0.111	0.651	0.750
AuditC 06	243	0.499	0.492	0.017	0.155	0.155	0.172	0.424	0.424
AuditC 07	603	0.502	0.496	0.014	0.122	0.121	0.136	0.629	0.724
AuditC 08	6955	0.214	0.213	0.002	0.024	0.024	0.026	0.072	0.158
AuditC 09	255	0.528	0.521	0.018	0.175	0.172	0.195	0.529	0.569

Study	n	Mean $\hat{\kappa}_D$	Mean $\hat{\kappa}_{FWKC}$	Max. abs. diff.	Width, percentile	Width, FWKC primary	Width, FWKC SE-exp.	PABAK	AC1
AuditC 10	3551	0.283	0.281	0.005	0.077	0.077	0.082	0.650	0.775
AuditC 11	626	0.544	0.540	0.008	0.119	0.120	0.133	0.581	0.628
AuditC 12	600	0.474	0.470	0.010	0.140	0.140	0.153	0.647	0.748
AuditC 13	112	0.627	0.611	0.026	0.235	0.236	0.263	0.607	0.622
AuditC 14	555	0.483	0.479	0.012	0.152	0.150	0.160	0.499	0.519

Means and widths are averaged over the nine values of c ; Max. abs. diff. is the largest absolute difference between the two estimates over c . $B = 49$. PABAK and AC1 do not depend on c (Proposition 8).

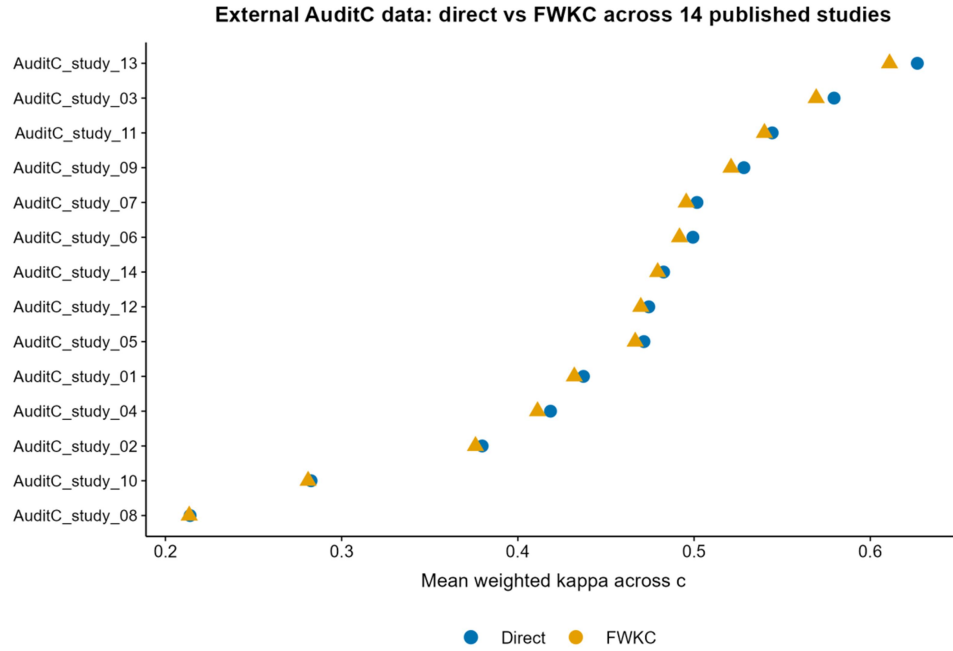
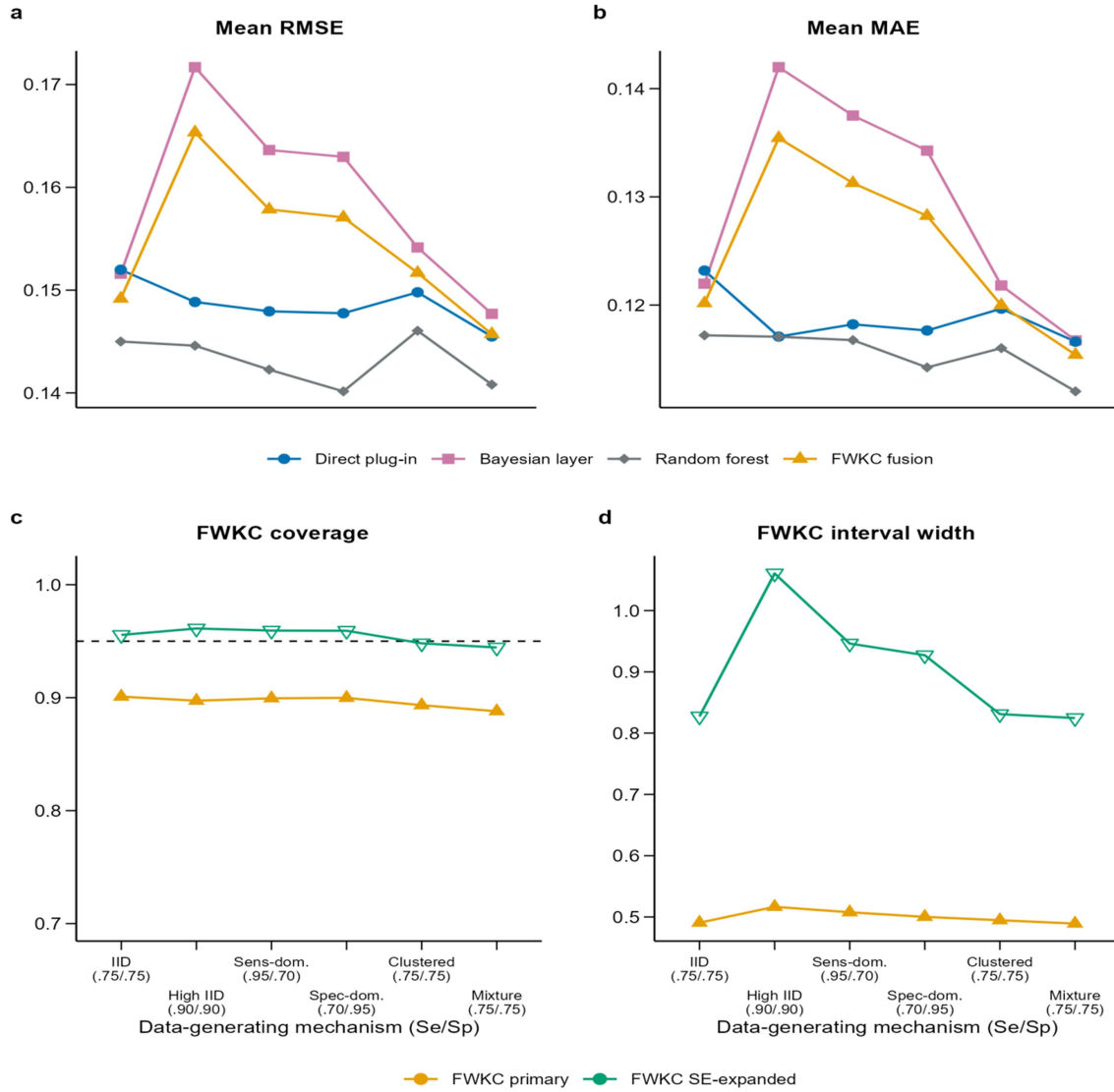


Figure S11. External AuditC data: direct and FWKC estimates in 14 studies

Mean weighted kappa over the nine values of c for each study. The two estimates differ by at most 0.026 in any study- c combination.

Bayesian, random-forest and FWKC robustness summary



S5.8 Complete scenario-level results of the primary grid

Table S23. Complete results of the 243 primary scenarios

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
30	0.1	0.1	0.147	0.775	0.850	0.783	0.783	0.783	0.725	0.118	0.132	0.136	0.124
30	0.1	0.2	0.162	0.867	0.917	0.892	0.892	0.850	0.808	0.114	0.126	0.129	0.121
30	0.1	0.3	0.179	0.850	0.908	0.892	0.883	0.858	0.792	0.114	0.124	0.127	0.116
30	0.1	0.4	0.201	0.867	0.933	0.883	0.883	0.850	0.767	0.127	0.134	0.136	0.123
30	0.1	0.5	0.229	0.908	0.967	0.958	0.950	0.933	0.825	0.128	0.129	0.130	0.126
30	0.1	0.6	0.267	0.817	0.917	0.867	0.825	0.833	0.733	0.178	0.173	0.175	0.165
30	0.1	0.7	0.318	0.883	0.942	0.933	0.908	0.900	0.875	0.155	0.150	0.152	0.133
30	0.1	0.8	0.395	0.792	0.875	0.850	0.825	0.808	0.700	0.231	0.227	0.230	0.197
30	0.1	0.9	0.520	0.866	0.950	0.883	0.808	0.867	0.667	0.251	0.239	0.241	0.227
30	0.2	0.1	0.255	0.900	1.000	0.950	0.933	0.892	0.900	0.136	0.156	0.160	0.151
30	0.2	0.2	0.275	0.892	0.983	0.933	0.917	0.892	0.867	0.127	0.133	0.136	0.132
30	0.2	0.3	0.299	0.883	0.975	0.925	0.925	0.892	0.867	0.143	0.152	0.155	0.148
30	0.2	0.4	0.327	0.850	0.983	0.900	0.883	0.850	0.842	0.164	0.170	0.173	0.166
30	0.2	0.5	0.360	0.892	0.992	0.950	0.917	0.908	0.883	0.140	0.146	0.151	0.143
30	0.2	0.6	0.401	0.917	0.983	0.942	0.917	0.900	0.850	0.166	0.169	0.172	0.166
30	0.2	0.7	0.453	0.917	1.000	0.958	0.942	0.933	0.850	0.162	0.166	0.169	0.163
30	0.2	0.8	0.520	0.883	0.983	0.933	0.900	0.900	0.800	0.186	0.183	0.185	0.180
30	0.2	0.9	0.611	0.875	1.000	0.850	0.833	0.867	0.717	0.205	0.207	0.209	0.204

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
30	0.3	0.1	0.338	0.900	0.992	0.958	0.933	0.892	0.875	0.142	0.166	0.172	0.157
30	0.3	0.2	0.359	0.900	0.992	0.892	0.908	0.875	0.842	0.137	0.150	0.155	0.146
30	0.3	0.3	0.382	0.908	0.992	0.950	0.933	0.892	0.833	0.150	0.165	0.169	0.158
30	0.3	0.4	0.410	0.908	0.992	0.917	0.917	0.900	0.842	0.151	0.159	0.164	0.155
30	0.3	0.5	0.441	0.882	1.000	0.925	0.942	0.892	0.892	0.153	0.164	0.169	0.159
30	0.3	0.6	0.477	0.842	1.000	0.942	0.908	0.883	0.792	0.162	0.171	0.176	0.165
30	0.3	0.7	0.520	0.867	1.000	0.933	0.892	0.875	0.850	0.174	0.180	0.184	0.174
30	0.3	0.8	0.572	0.842	0.992	0.892	0.892	0.842	0.817	0.177	0.179	0.182	0.176
30	0.3	0.9	0.635	0.883	1.000	0.875	0.883	0.867	0.817	0.194	0.199	0.202	0.195
30	0.4	0.1	0.402	0.867	0.983	0.942	0.925	0.875	0.833	0.151	0.171	0.179	0.162
30	0.4	0.2	0.421	0.875	1.000	0.942	0.892	0.875	0.842	0.160	0.179	0.185	0.169
30	0.4	0.3	0.442	0.900	1.000	0.975	0.950	0.900	0.892	0.140	0.154	0.159	0.148
30	0.4	0.4	0.466	0.858	1.000	0.958	0.942	0.850	0.817	0.146	0.158	0.164	0.152
30	0.4	0.5	0.491	0.900	0.992	0.975	0.950	0.900	0.892	0.147	0.159	0.164	0.153
30	0.4	0.6	0.520	0.867	1.000	0.942	0.892	0.867	0.792	0.166	0.177	0.183	0.170
30	0.4	0.7	0.553	0.817	1.000	0.917	0.900	0.817	0.867	0.170	0.178	0.183	0.172
30	0.4	0.8	0.590	0.900	1.000	0.942	0.908	0.892	0.867	0.154	0.156	0.160	0.152
30	0.4	0.9	0.632	0.833	0.992	0.808	0.775	0.783	0.700	0.208	0.206	0.209	0.201
30	0.5	0.1	0.453	0.933	1.000	0.967	0.950	0.925	0.900	0.146	0.171	0.179	0.158
30	0.5	0.2	0.468	0.858	0.992	0.942	0.900	0.858	0.808	0.165	0.194	0.202	0.178
30	0.5	0.3	0.484	0.925	1.000	0.958	0.958	0.925	0.858	0.151	0.168	0.175	0.158
30	0.5	0.4	0.502	0.839	0.992	0.933	0.892	0.825	0.800	0.175	0.192	0.199	0.179

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
30	0.5	0.5	0.520	0.867	0.992	0.933	0.925	0.858	0.842	0.159	0.173	0.180	0.164
30	0.5	0.6	0.541	0.842	1.000	0.958	0.925	0.858	0.858	0.154	0.167	0.174	0.158
30	0.5	0.7	0.563	0.892	1.000	0.917	0.908	0.867	0.833	0.169	0.183	0.188	0.176
30	0.5	0.8	0.586	0.925	0.992	0.942	0.917	0.925	0.858	0.157	0.161	0.166	0.155
30	0.5	0.9	0.612	0.908	1.000	0.892	0.900	0.900	0.842	0.167	0.171	0.175	0.163
30	0.6	0.1	0.492	0.858	1.000	0.942	0.908	0.867	0.883	0.174	0.195	0.204	0.181
30	0.6	0.2	0.501	0.867	1.000	0.950	0.900	0.850	0.833	0.172	0.202	0.212	0.184
30	0.6	0.3	0.511	0.873	1.000	0.933	0.892	0.842	0.825	0.171	0.193	0.200	0.179
30	0.6	0.4	0.520	0.875	1.000	0.942	0.883	0.858	0.817	0.174	0.197	0.205	0.182
30	0.6	0.5	0.531	0.900	0.992	0.975	0.925	0.892	0.867	0.155	0.174	0.181	0.162
30	0.6	0.6	0.541	0.857	1.000	0.958	0.908	0.883	0.842	0.170	0.189	0.196	0.177
30	0.6	0.7	0.552	0.883	0.983	0.958	0.875	0.858	0.833	0.162	0.181	0.188	0.168
30	0.6	0.8	0.563	0.858	1.000	0.908	0.900	0.867	0.800	0.174	0.190	0.197	0.180
30	0.6	0.9	0.575	0.850	1.000	0.892	0.875	0.875	0.792	0.183	0.185	0.190	0.177
30	0.7	0.1	0.522	0.883	0.992	0.925	0.908	0.892	0.800	0.195	0.211	0.218	0.195
30	0.7	0.2	0.521	0.883	0.983	0.908	0.917	0.875	0.842	0.181	0.199	0.207	0.187
30	0.7	0.3	0.520	0.849	1.000	0.917	0.850	0.808	0.808	0.181	0.205	0.214	0.188
30	0.7	0.4	0.520	0.874	1.000	0.967	0.942	0.850	0.833	0.170	0.191	0.199	0.178
30	0.7	0.5	0.519	0.916	1.000	0.992	0.958	0.908	0.867	0.162	0.188	0.198	0.169
30	0.7	0.6	0.519	0.908	0.992	0.975	0.917	0.875	0.842	0.165	0.191	0.200	0.173
30	0.7	0.7	0.518	0.899	0.992	0.958	0.908	0.892	0.875	0.176	0.193	0.200	0.180
30	0.7	0.8	0.518	0.849	0.983	0.924	0.899	0.825	0.833	0.198	0.212	0.218	0.194

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
30	0.7	0.9	0.517	0.915	1.000	0.950	0.925	0.917	0.892	0.166	0.186	0.194	0.168
30	0.8	0.1	0.537	0.821	0.966	0.825	0.783	0.792	0.708	0.259	0.258	0.264	0.242
30	0.8	0.2	0.520	0.875	0.950	0.849	0.832	0.858	0.717	0.245	0.248	0.257	0.226
30	0.8	0.3	0.505	0.899	0.992	0.942	0.892	0.892	0.800	0.196	0.201	0.209	0.187
30	0.8	0.4	0.490	0.948	0.991	0.958	0.942	0.925	0.833	0.182	0.211	0.220	0.187
30	0.8	0.5	0.476	0.858	0.967	0.850	0.867	0.817	0.792	0.217	0.230	0.236	0.216
30	0.8	0.6	0.462	0.915	0.975	0.958	0.883	0.908	0.808	0.192	0.225	0.236	0.201
30	0.8	0.7	0.450	0.847	0.992	0.941	0.933	0.858	0.842	0.188	0.223	0.233	0.198
30	0.8	0.8	0.438	0.916	0.992	0.892	0.900	0.900	0.842	0.192	0.224	0.235	0.199
30	0.8	0.9	0.427	0.880	0.966	0.908	0.917	0.892	0.792	0.211	0.244	0.256	0.219
30	0.9	0.1	0.520	0.849	0.941	0.761	0.701	0.775	0.608	0.310	0.269	0.275	0.242
30	0.9	0.2	0.469	0.849	0.958	0.838	0.786	0.790	0.630	0.276	0.272	0.281	0.238
30	0.9	0.3	0.427	0.857	0.924	0.913	0.861	0.842	0.767	0.238	0.256	0.268	0.216
30	0.9	0.4	0.392	0.863	0.923	0.888	0.776	0.832	0.681	0.251	0.274	0.286	0.236
30	0.9	0.5	0.362	0.842	0.917	0.838	0.795	0.775	0.700	0.247	0.276	0.289	0.233
30	0.9	0.6	0.337	0.852	0.904	0.875	0.848	0.775	0.700	0.239	0.273	0.284	0.235
30	0.9	0.7	0.314	0.849	0.916	0.929	0.911	0.839	0.780	0.196	0.254	0.271	0.206
30	0.9	0.8	0.295	0.866	0.950	0.922	0.904	0.857	0.832	0.197	0.278	0.297	0.232
30	0.9	0.9	0.278	0.923	0.966	0.891	0.874	0.842	0.758	0.201	0.262	0.278	0.222
50	0.1	0.1	0.147	0.908	0.967	0.925	0.925	0.917	0.867	0.092	0.093	0.090	0.094
50	0.1	0.2	0.162	0.858	0.933	0.892	0.892	0.842	0.817	0.104	0.105	0.103	0.107
50	0.1	0.3	0.179	0.892	0.958	0.925	0.917	0.900	0.875	0.088	0.086	0.083	0.087

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
50	0.1	0.4	0.201	0.892	0.950	0.908	0.908	0.892	0.883	0.100	0.096	0.095	0.097
50	0.1	0.5	0.229	0.875	0.942	0.900	0.900	0.858	0.858	0.116	0.112	0.110	0.114
50	0.1	0.6	0.267	0.792	0.917	0.892	0.867	0.808	0.775	0.136	0.134	0.133	0.133
50	0.1	0.7	0.318	0.883	0.975	0.942	0.925	0.917	0.867	0.136	0.136	0.137	0.134
50	0.1	0.8	0.395	0.817	0.933	0.892	0.900	0.867	0.808	0.177	0.171	0.171	0.171
50	0.1	0.9	0.520	0.825	0.975	0.900	0.900	0.892	0.775	0.187	0.186	0.187	0.184
50	0.2	0.1	0.255	0.933	0.992	0.950	0.950	0.933	0.908	0.098	0.103	0.103	0.104
50	0.2	0.2	0.275	0.850	0.975	0.908	0.908	0.842	0.808	0.115	0.119	0.119	0.120
50	0.2	0.3	0.299	0.892	0.967	0.933	0.908	0.883	0.858	0.107	0.108	0.108	0.109
50	0.2	0.4	0.327	0.900	1.000	0.967	0.967	0.892	0.900	0.106	0.108	0.108	0.108
50	0.2	0.5	0.360	0.833	0.967	0.917	0.908	0.842	0.858	0.128	0.127	0.127	0.128
50	0.2	0.6	0.401	0.900	0.983	0.958	0.967	0.925	0.908	0.111	0.112	0.113	0.112
50	0.2	0.7	0.453	0.883	0.983	0.958	0.942	0.883	0.875	0.130	0.131	0.133	0.131
50	0.2	0.8	0.520	0.867	0.992	0.950	0.958	0.900	0.875	0.145	0.150	0.153	0.148
50	0.2	0.9	0.611	0.858	1.000	0.883	0.858	0.867	0.792	0.159	0.155	0.156	0.155
50	0.3	0.1	0.338	0.917	0.992	0.975	0.967	0.925	0.900	0.100	0.102	0.102	0.104
50	0.3	0.2	0.359	0.917	0.983	0.958	0.958	0.925	0.892	0.108	0.107	0.107	0.109
50	0.3	0.3	0.382	0.900	0.975	0.933	0.933	0.908	0.875	0.107	0.109	0.112	0.108
50	0.3	0.4	0.410	0.925	0.975	0.950	0.950	0.917	0.900	0.112	0.111	0.113	0.111
50	0.3	0.5	0.441	0.842	0.975	0.917	0.908	0.850	0.850	0.128	0.129	0.131	0.128
50	0.3	0.6	0.477	0.883	1.000	0.967	0.958	0.892	0.867	0.117	0.117	0.120	0.117
50	0.3	0.7	0.520	0.858	0.992	0.933	0.933	0.875	0.858	0.130	0.133	0.135	0.132

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
50	0.3	0.8	0.572	0.867	1.000	0.917	0.900	0.867	0.850	0.135	0.134	0.137	0.133
50	0.3	0.9	0.635	0.858	0.983	0.900	0.875	0.867	0.808	0.139	0.131	0.133	0.130
50	0.4	0.1	0.402	0.883	1.000	0.950	0.925	0.900	0.883	0.113	0.115	0.116	0.115
50	0.4	0.2	0.421	0.942	1.000	0.983	0.983	0.942	0.917	0.104	0.104	0.105	0.104
50	0.4	0.3	0.442	0.883	0.992	0.925	0.900	0.883	0.850	0.123	0.123	0.125	0.123
50	0.4	0.4	0.466	0.925	0.983	0.958	0.950	0.900	0.892	0.112	0.115	0.118	0.114
50	0.4	0.5	0.491	0.933	1.000	0.975	0.975	0.917	0.925	0.102	0.103	0.105	0.103
50	0.4	0.6	0.520	0.850	0.975	0.925	0.925	0.858	0.867	0.127	0.127	0.130	0.126
50	0.4	0.7	0.553	0.858	1.000	0.942	0.925	0.867	0.842	0.124	0.124	0.126	0.124
50	0.4	0.8	0.590	0.900	0.983	0.950	0.933	0.900	0.850	0.127	0.128	0.132	0.126
50	0.4	0.9	0.632	0.892	0.992	0.917	0.875	0.867	0.800	0.145	0.140	0.142	0.139
50	0.5	0.1	0.453	0.875	0.975	0.950	0.917	0.858	0.850	0.120	0.123	0.125	0.123
50	0.5	0.2	0.468	0.900	0.983	0.967	0.950	0.883	0.875	0.117	0.117	0.119	0.117
50	0.5	0.3	0.484	0.833	1.000	0.875	0.883	0.842	0.833	0.133	0.135	0.138	0.133
50	0.5	0.4	0.502	0.875	0.983	0.942	0.900	0.892	0.875	0.129	0.132	0.134	0.131
50	0.5	0.5	0.520	0.875	0.983	0.975	0.933	0.875	0.875	0.116	0.117	0.119	0.116
50	0.5	0.6	0.541	0.883	1.000	0.967	0.950	0.900	0.833	0.119	0.118	0.120	0.118
50	0.5	0.7	0.563	0.875	0.975	0.942	0.950	0.867	0.867	0.123	0.127	0.131	0.125
50	0.5	0.8	0.586	0.900	1.000	0.942	0.942	0.883	0.842	0.120	0.118	0.120	0.117
50	0.5	0.9	0.612	0.883	0.992	0.933	0.867	0.867	0.800	0.140	0.136	0.138	0.135
50	0.6	0.1	0.492	0.817	0.992	0.925	0.917	0.808	0.767	0.147	0.149	0.153	0.147
50	0.6	0.2	0.501	0.892	0.983	0.942	0.917	0.900	0.842	0.135	0.140	0.145	0.137

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
50	0.6	0.3	0.511	0.883	0.992	0.950	0.925	0.883	0.850	0.128	0.129	0.133	0.128
50	0.6	0.4	0.520	0.858	0.983	0.933	0.942	0.867	0.833	0.137	0.137	0.140	0.135
50	0.6	0.5	0.531	0.858	0.992	0.958	0.917	0.867	0.858	0.120	0.121	0.124	0.119
50	0.6	0.6	0.541	0.883	0.950	0.917	0.900	0.883	0.825	0.141	0.143	0.146	0.142
50	0.6	0.7	0.552	0.867	0.983	0.958	0.925	0.900	0.875	0.132	0.135	0.139	0.132
50	0.6	0.8	0.563	0.967	1.000	0.967	0.942	0.967	0.892	0.111	0.110	0.114	0.108
50	0.6	0.9	0.575	0.908	0.992	0.967	0.925	0.908	0.892	0.120	0.119	0.121	0.117
50	0.7	0.1	0.522	0.858	0.992	0.950	0.900	0.833	0.800	0.154	0.156	0.159	0.153
50	0.7	0.2	0.521	0.875	0.983	0.967	0.925	0.858	0.850	0.144	0.144	0.147	0.142
50	0.7	0.3	0.520	0.850	0.967	0.933	0.892	0.833	0.850	0.145	0.150	0.153	0.147
50	0.7	0.4	0.520	0.900	1.000	0.975	0.958	0.892	0.908	0.120	0.124	0.127	0.120
50	0.7	0.5	0.519	0.900	0.992	0.967	0.967	0.917	0.908	0.132	0.135	0.139	0.133
50	0.7	0.6	0.519	0.874	0.983	0.942	0.925	0.883	0.858	0.140	0.147	0.151	0.144
50	0.7	0.7	0.518	0.892	1.000	0.925	0.908	0.883	0.842	0.141	0.141	0.144	0.139
50	0.7	0.8	0.518	0.942	0.992	0.942	0.925	0.917	0.900	0.135	0.134	0.138	0.131
50	0.7	0.9	0.517	0.892	0.992	0.950	0.925	0.875	0.875	0.146	0.150	0.153	0.148
50	0.8	0.1	0.537	0.892	0.983	0.933	0.925	0.883	0.875	0.171	0.169	0.173	0.164
50	0.8	0.2	0.520	0.892	0.975	0.942	0.908	0.883	0.850	0.159	0.158	0.160	0.155
50	0.8	0.3	0.505	0.883	0.950	0.900	0.883	0.883	0.850	0.167	0.167	0.170	0.163
50	0.8	0.4	0.490	0.883	0.975	0.933	0.933	0.892	0.875	0.149	0.149	0.153	0.145
50	0.8	0.5	0.476	0.916	0.992	0.942	0.908	0.908	0.858	0.149	0.152	0.155	0.149
50	0.8	0.6	0.462	0.867	0.950	0.900	0.900	0.850	0.850	0.158	0.161	0.165	0.158

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
50	0.8	0.7	0.450	0.883	0.950	0.925	0.908	0.900	0.808	0.165	0.168	0.171	0.164
50	0.8	0.8	0.438	0.917	0.983	0.925	0.875	0.917	0.817	0.153	0.159	0.163	0.155
50	0.8	0.9	0.427	0.850	0.958	0.917	0.875	0.858	0.808	0.153	0.163	0.168	0.158
50	0.9	0.1	0.520	0.875	0.967	0.866	0.840	0.875	0.750	0.236	0.214	0.217	0.202
50	0.9	0.2	0.469	0.867	0.933	0.867	0.833	0.850	0.733	0.237	0.217	0.220	0.212
50	0.9	0.3	0.427	0.883	0.942	0.916	0.908	0.883	0.800	0.200	0.196	0.202	0.184
50	0.9	0.4	0.392	0.875	0.942	0.925	0.925	0.867	0.808	0.177	0.175	0.179	0.170
50	0.9	0.5	0.362	0.883	0.933	0.917	0.917	0.892	0.808	0.165	0.169	0.174	0.162
50	0.9	0.6	0.337	0.900	0.975	0.950	0.925	0.892	0.817	0.157	0.168	0.173	0.159
50	0.9	0.7	0.314	0.917	0.967	0.942	0.950	0.892	0.858	0.148	0.161	0.167	0.150
50	0.9	0.8	0.295	0.875	0.942	0.916	0.874	0.875	0.825	0.168	0.194	0.201	0.180
50	0.9	0.9	0.278	0.900	0.958	0.933	0.933	0.917	0.842	0.151	0.165	0.169	0.158
100	0.1	0.1	0.147	0.867	0.900	0.933	0.933	0.858	0.875	0.053	0.055	0.054	0.055
100	0.1	0.2	0.162	0.900	0.908	0.933	0.933	0.900	0.892	0.066	0.066	0.064	0.067
100	0.1	0.3	0.179	0.875	0.950	0.950	0.950	0.892	0.933	0.066	0.065	0.062	0.067
100	0.1	0.4	0.201	0.875	0.908	0.942	0.933	0.892	0.875	0.072	0.070	0.068	0.071
100	0.1	0.5	0.229	0.917	0.917	0.967	0.967	0.925	0.867	0.074	0.073	0.074	0.074
100	0.1	0.6	0.267	0.883	0.925	0.942	0.942	0.892	0.900	0.089	0.088	0.088	0.089
100	0.1	0.7	0.318	0.875	0.925	0.942	0.950	0.900	0.892	0.094	0.095	0.097	0.095
100	0.1	0.8	0.395	0.892	0.933	0.958	0.925	0.883	0.850	0.111	0.110	0.111	0.110
100	0.1	0.9	0.520	0.850	0.908	0.917	0.925	0.858	0.767	0.130	0.127	0.129	0.127
100	0.2	0.1	0.255	0.925	0.958	0.983	0.983	0.942	0.900	0.062	0.061	0.061	0.063

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
100	0.2	0.2	0.275	0.883	0.925	0.933	0.917	0.858	0.842	0.077	0.076	0.075	0.077
100	0.2	0.3	0.299	0.917	0.958	0.958	0.958	0.925	0.925	0.068	0.069	0.070	0.069
100	0.2	0.4	0.327	0.883	0.917	0.917	0.917	0.892	0.867	0.080	0.079	0.079	0.081
100	0.2	0.5	0.360	0.875	0.942	0.967	0.958	0.875	0.842	0.086	0.084	0.084	0.085
100	0.2	0.6	0.401	0.900	0.933	0.942	0.942	0.900	0.917	0.084	0.085	0.089	0.085
100	0.2	0.7	0.453	0.858	0.908	0.942	0.942	0.883	0.850	0.094	0.095	0.098	0.095
100	0.2	0.8	0.520	0.858	0.867	0.933	0.950	0.817	0.800	0.109	0.112	0.115	0.110
100	0.2	0.9	0.611	0.867	0.942	0.950	0.925	0.875	0.875	0.109	0.110	0.115	0.107
100	0.3	0.1	0.338	0.883	0.900	0.900	0.917	0.883	0.875	0.077	0.076	0.077	0.077
100	0.3	0.2	0.359	0.883	0.925	0.942	0.933	0.883	0.867	0.083	0.083	0.084	0.084
100	0.3	0.3	0.382	0.892	0.900	0.933	0.933	0.917	0.883	0.082	0.083	0.085	0.083
100	0.3	0.4	0.410	0.908	0.958	0.975	0.975	0.917	0.892	0.072	0.071	0.073	0.072
100	0.3	0.5	0.441	0.917	0.967	0.967	0.967	0.908	0.908	0.079	0.079	0.081	0.080
100	0.3	0.6	0.477	0.892	0.942	0.950	0.958	0.917	0.892	0.086	0.088	0.093	0.088
100	0.3	0.7	0.520	0.875	0.908	0.925	0.917	0.875	0.858	0.094	0.097	0.101	0.096
100	0.3	0.8	0.572	0.858	0.942	0.917	0.908	0.833	0.842	0.096	0.095	0.098	0.094
100	0.3	0.9	0.635	0.842	0.892	0.925	0.908	0.842	0.767	0.110	0.110	0.113	0.109
100	0.4	0.1	0.402	0.842	0.942	0.950	0.925	0.850	0.900	0.085	0.084	0.084	0.085
100	0.4	0.2	0.421	0.875	0.925	0.950	0.933	0.892	0.858	0.087	0.085	0.085	0.087
100	0.4	0.3	0.442	0.917	0.950	0.950	0.925	0.900	0.892	0.078	0.078	0.081	0.078
100	0.4	0.4	0.466	0.933	0.950	0.958	0.950	0.942	0.908	0.081	0.080	0.084	0.080
100	0.4	0.5	0.491	0.883	0.925	0.950	0.925	0.908	0.850	0.089	0.089	0.092	0.089

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
100	0.4	0.6	0.520	0.908	0.967	0.975	0.967	0.925	0.917	0.073	0.073	0.078	0.073
100	0.4	0.7	0.553	0.925	0.958	0.958	0.942	0.917	0.892	0.078	0.077	0.080	0.077
100	0.4	0.8	0.590	0.908	0.908	0.925	0.942	0.900	0.883	0.087	0.088	0.091	0.087
100	0.4	0.9	0.632	0.850	0.950	0.950	0.933	0.867	0.892	0.098	0.100	0.103	0.098
100	0.5	0.1	0.453	0.875	0.950	0.958	0.933	0.883	0.900	0.085	0.084	0.083	0.086
100	0.5	0.2	0.468	0.867	0.942	0.933	0.925	0.858	0.883	0.091	0.090	0.094	0.090
100	0.5	0.3	0.484	0.958	0.983	0.992	0.975	0.950	0.900	0.078	0.076	0.080	0.076
100	0.5	0.4	0.502	0.875	0.942	0.933	0.958	0.875	0.883	0.082	0.082	0.083	0.082
100	0.5	0.5	0.520	0.900	0.942	0.950	0.958	0.908	0.900	0.077	0.080	0.087	0.078
100	0.5	0.6	0.541	0.858	0.908	0.917	0.908	0.867	0.875	0.089	0.090	0.095	0.088
100	0.5	0.7	0.563	0.883	0.958	0.967	0.967	0.883	0.925	0.079	0.081	0.087	0.080
100	0.5	0.8	0.586	0.900	0.958	0.942	0.933	0.875	0.867	0.094	0.094	0.097	0.093
100	0.5	0.9	0.612	0.925	0.958	0.958	0.950	0.933	0.908	0.088	0.088	0.091	0.088
100	0.6	0.1	0.492	0.900	0.967	0.958	0.958	0.917	0.867	0.091	0.090	0.091	0.091
100	0.6	0.2	0.501	0.875	0.933	0.942	0.933	0.883	0.900	0.090	0.090	0.094	0.089
100	0.6	0.3	0.511	0.883	0.950	0.975	0.975	0.892	0.942	0.082	0.083	0.089	0.081
100	0.6	0.4	0.520	0.900	0.967	0.975	0.933	0.892	0.892	0.084	0.083	0.086	0.083
100	0.6	0.5	0.531	0.850	0.917	0.942	0.917	0.833	0.875	0.092	0.092	0.094	0.092
100	0.6	0.6	0.541	0.908	0.958	0.975	0.967	0.933	0.925	0.088	0.090	0.094	0.090
100	0.6	0.7	0.552	0.933	0.975	0.983	0.958	0.950	0.858	0.082	0.081	0.082	0.082
100	0.6	0.8	0.563	0.908	0.925	0.942	0.917	0.892	0.850	0.099	0.097	0.098	0.098
100	0.6	0.9	0.575	0.942	0.958	0.967	0.967	0.942	0.925	0.091	0.089	0.093	0.089

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
100	0.7	0.1	0.522	0.850	0.908	0.917	0.933	0.875	0.850	0.106	0.106	0.109	0.105
100	0.7	0.2	0.521	0.875	0.908	0.933	0.917	0.883	0.842	0.108	0.106	0.106	0.106
100	0.7	0.3	0.520	0.900	0.933	0.967	0.958	0.908	0.883	0.094	0.094	0.097	0.094
100	0.7	0.4	0.520	0.867	0.908	0.917	0.900	0.842	0.825	0.103	0.102	0.104	0.102
100	0.7	0.5	0.519	0.850	0.933	0.958	0.933	0.842	0.892	0.093	0.092	0.095	0.093
100	0.7	0.6	0.519	0.883	0.950	0.950	0.917	0.900	0.892	0.099	0.100	0.103	0.099
100	0.7	0.7	0.518	0.900	0.950	0.958	0.958	0.908	0.867	0.091	0.091	0.094	0.092
100	0.7	0.8	0.518	0.883	0.933	0.933	0.908	0.900	0.850	0.109	0.108	0.108	0.109
100	0.7	0.9	0.517	0.867	0.942	0.942	0.925	0.883	0.883	0.108	0.108	0.108	0.109
100	0.8	0.1	0.537	0.900	0.933	0.950	0.942	0.900	0.883	0.109	0.108	0.110	0.107
100	0.8	0.2	0.520	0.883	0.942	0.942	0.925	0.867	0.850	0.112	0.108	0.111	0.108
100	0.8	0.3	0.505	0.858	0.917	0.933	0.917	0.875	0.842	0.114	0.111	0.113	0.110
100	0.8	0.4	0.490	0.875	0.950	0.942	0.942	0.892	0.867	0.111	0.111	0.114	0.111
100	0.8	0.5	0.476	0.950	0.967	0.975	0.975	0.908	0.908	0.100	0.100	0.100	0.100
100	0.8	0.6	0.462	0.842	0.908	0.925	0.908	0.858	0.850	0.117	0.115	0.116	0.117
100	0.8	0.7	0.450	0.908	0.958	0.958	0.933	0.900	0.850	0.108	0.106	0.105	0.108
100	0.8	0.8	0.438	0.883	0.958	0.950	0.908	0.875	0.858	0.112	0.111	0.110	0.113
100	0.8	0.9	0.427	0.833	0.900	0.925	0.925	0.850	0.858	0.115	0.116	0.116	0.117
100	0.9	0.1	0.520	0.875	0.892	0.900	0.883	0.867	0.833	0.162	0.151	0.152	0.151
100	0.9	0.2	0.469	0.892	0.908	0.917	0.908	0.900	0.833	0.144	0.137	0.139	0.136
100	0.9	0.3	0.427	0.883	0.925	0.942	0.925	0.875	0.850	0.134	0.130	0.130	0.131
100	0.9	0.4	0.392	0.858	0.917	0.925	0.925	0.875	0.867	0.128	0.125	0.126	0.124

n	p	c	$\kappa(c)$	FWKC primary	FWKC SE- exp.	Transformed	Wald	Percentile	Basic	RMSE direct	RMSE FWKC	RMSE Bayes.	RMSE RF
100	0.9	0.5	0.362	0.867	0.925	0.925	0.917	0.883	0.867	0.122	0.119	0.118	0.120
100	0.9	0.6	0.337	0.875	0.942	0.933	0.925	0.875	0.875	0.119	0.114	0.113	0.116
100	0.9	0.7	0.314	0.842	0.925	0.933	0.925	0.850	0.867	0.113	0.109	0.109	0.110
100	0.9	0.8	0.295	0.883	0.917	0.917	0.908	0.883	0.900	0.100	0.099	0.099	0.100
100	0.9	0.9	0.278	0.875	0.942	0.942	0.917	0.875	0.875	0.111	0.112	0.108	0.115

$Se = 0.854$, $Sp = 0.667$; $M = 120$ independent datasets per scenario (119 in a few scenarios with a failed replicate). Coverage columns are for nominal 95% intervals; the Monte Carlo standard error of a coverage value near 0.95 is 0.020.

S5.9 Scope and limitations of the supplementary evidence

The primary 243-scenario grid covers the full prevalence and weighting-index ranges used in the manuscript but a single operating point for sensitivity and specificity. The alternative-mechanism tier covers the full $3 \times 3 \times 3$ grid of n , p and c (162 scenarios, $M=40$), so it provides robustness evidence across the design rather than at one point within it; the residual gap relative to the primary grid is precision ($M=40$ rather than $M=120$) and a lower FWKC fitting-success rate in the most demanding cells, which lowers the minimum SE-expanded coverage in this tier to 0.675 once failed fits are counted as non-coverage (S5.4). Only the clustered mechanism induces dependence between subjects; the mixture mechanism is observationally equivalent to the matched independent mechanism (Proposition 22).

The sensitivity tier comprises 34 prespecified variants evaluated on three stress scenarios with $M = 10$. It can identify large changes or clear monotone patterns, but it cannot rank settings whose coverage differs by a few hundredths. The AuditC analysis is a cross-domain check of numerical stability; it documents successful fitting and agreement between the FWKC and direct estimates, but it is not a coverage study, because the true agreement parameter is unknown in real data.

The FWKC point estimator is not generally more accurate than the direct estimator: its mean RMSE is higher at $n = 30$ (0.192 against 0.177), similar at $n = 50$ (0.137 against 0.136) and slightly lower at $n = 100$ (0.094 against 0.095), and the $n = 620$ bridge shows no practical difference (0.0466 against 0.0471). No claim is made for outcomes with more than two categories. The framework is computationally expensive relative to the transformed closed-form interval, and the analysis has not yet been independently replicated. These limits are part of the interpretation of the results rather than exceptions to it.

S6. Relation to the supplementary file of the original submission

The original supplementary file documented the analysis of the original submission, in which coverage was computed from bootstrap resamples of one simulated dataset per scenario (Proposition 23(b)), pseudo-raters were drawn from the distribution of Proposition 10, and intervals were contracted rather than expanded

(Proposition 18(b)). That analysis has been withdrawn, and none of its numerical results is reused here. Every mathematical component that remains valid has been retained, corrected where necessary and placed in the section listed in Table S24; the remaining items are withdrawn for the reason given.

Table S24. Disposition of each item of the original supplementary file

Original item	Status	Now in	Reason or change
S1 Simulation design and validation framework	Replaced	S4	Coverage now from independent datasets (Proposition 23)
S1.3 Small-sample shrinkage factor	Withdrawn and replaced	S3.8	The contraction rule under-covers by construction (Proposition 18(b)); replaced by the expansion of Eq. (17)
S2 Empirical anchoring and notation	Retained, corrected	S1.1, S2.1, Table S2	Notation unified with the main text
S3 Implications for methodological development	Withdrawn	—	Claims not supported by the revised analysis
S4 Hierarchical pooling with tightened priors for $n \leq n_0$	Replaced	S3.4	Priors are fixed; the model is crossed; the sampler is exact (Proposition 14)
S5 Diagnostic performance measures	Retained	S1.1, Table S5	—
S6 Bootstrap as inferential backbone	Retained	S3.7, S4.5	Full-pipeline bootstrap is the primary interval
S7 Pseudo-rater construction	Replaced	S3.1–S3.2	Jeffreys posteriors replace $\text{Beta}(n\widehat{Se}, n(1 - \widehat{Se}))$ (Proposition 10)
S8 Bayesian partial pooling on a transformed scale	Retained, corrected	S3.3–S3.4	Crossed model with draw effect; exact Gibbs sampler; CmdStan cross-check

Original item	Status	Now in	Reason or change
S9 Hybrid Bayesian–random-forest fusion	Retained, extended	S3.5–S3.6	Forest fully specified; fusion analysed for correlated components
S10 Delta-covariance uncertainty propagation	Retained as diagnostic only	S3.9	Measures posterior-predictive dispersion (Proposition 11), not sampling variance
S11 DGM from real-data-anchored parameters	Replaced	S4.2	Six further mechanisms and an anchor bridge added
S12 Dependence-aware comparative inference	Retained, extended	S4.5	Nonparametric paired tests primary; z secondary
S13 Scenario grid	Extended	S4.1	81 → 243 scenarios
S14 Stage-wise empirical results (Stages 1–5)	Withdrawn	S5	Computed from bootstrap resamples of one dataset per scenario
Appendix A limit-case summary (A.1–A.4 of the first part)	Withdrawn	S3.2	Referred to a shrinkage operator that no longer exists; replaced by Proposition 12
Appendix A.1 Delta-covariance standard error	Retained	S3.9 (Proposition 19)	—
Appendix A.2 Optimality of fusion weights	Retained, extended	S3.6 (Proposition 15)	The original derivation assumed uncorrelated components; observed correlation is 0.99
Appendix A.3 Consistency of the shrinkage factor	Retained for the expansion factor	S3.8 (Proposition 17)	Continuity and large-sample identity carry over
Appendix A.4 Variance of the dependent difference	Retained	S4.5 (Proposition 24)	—
Appendix B Tables S1–S4	Withdrawn	S5	Not reproducible from the final code; replaced by Tables S23, S19, S21
Appendix C Reproducibility protocol	Replaced	Table S13; GitHub release v1.1.1-reviewer	Seed, settings and software environment updated

Original item	Status	Now in	Reason or change
Appendix D.1 Sensitivity to n_0	Replaced	S3.8, Table S21	Replaced by the prespecified n_0 variants
Appendix D.2 Ridge regularisation bounds	Retained, corrected	S3.9	$\epsilon = 10^{-10}$
Appendix D.3 Adaptive priors	Withdrawn	S3.4	Priors are no longer sample-size dependent
Appendix D.4, E.5, F.5 Ethics and data governance	Moved	Declarations of the main text	—
Appendix D.5, E.1 Convergence (HMC, adapt_delta)	Replaced	S3.4	Primary sampler is Gibbs; HMC used only as cross-check (Table S10)
Appendix E.2–E.3 Condition number; clipping	Retained, corrected	S3.3, S3.9	$\epsilon_t = 10^{-8}$ for truncation
Appendix E.4, G.3 Computational complexity	Retained, corrected	S3.11 (Proposition 20)	—
Appendix F.1–F.2 Power and sample-size planning	Withdrawn	—	The planning layer is not part of the revised framework
Appendix F.3 Delta-method bound	Replaced	S2.2 (Proposition 5)	Exact closed form
Appendix F.4 Singularity condition	Retained	S3.9 (Proposition 19(b))	—
Appendix F.6, G.4 Boundary behaviour	Replaced	S3.1, S3.3	Jeffreys posteriors are proper with zero cells
Appendix G.1 Bias bound under Beta misspecification	Replaced	S3.2 (Proposition 12)	Large-sample behaviour derived for the actual pipeline
Appendix G.2 Monte Carlo convergence	Retained	S3.2	$O(n_{mc}^{-1/2})$ error
Appendix H PABAK comparison	Retained, extended	S2.4 (Proposition 8)	AC1 added; Table H.1 and the coverage comparison withdrawn because PABAK is a different estimand

Section and appendix labels are those of the original supplementary file.

The following statements of the original file describe the earlier analysis and are superseded by the present file: the random seed (20260912, not 2025); the Bayesian engine (an exact blocked Gibbs sampler with four chains, cross-checked by CmdStan HMC, not HMC with two chains and `adapt_delta = 0.99`); the priors (fixed $\mu \sim N(0, 2^2)$ and $IG(2, 1)$ variance priors, not sample-size-dependent priors); the convergence criterion ($\hat{R} \leq 1.05$ on the larger of the classical and rank-normalised statistics); the truncation constant (10^{-8} , not 10^{-6}); the ridge constant (10^{-10} , not 10^{-8}); the Monte Carlo and bootstrap sizes (Table S13). One statement of the original file was incorrect even for the earlier analysis: the original interval adjustment multiplied a κ -scale Wald half-width by $\min(1, n/n_0)$, not by $\min\{1, (n/n_0)^{1/2}\}$ as its Appendix A.3 stated (its Section S14.19 described yet another form).

Additional references

- [A1] Fisher RA. On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*. 1921;1:3–32.
- [A2] Jeffreys H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A*. 1946;186(1007):453–461.
- [A3] Liu JS, Wong WH, Kong A. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*. 1994;81(1):27–40.
- [A4] Kish L. *Survey Sampling*. New York: Wiley; 1965.