CERTIFY OR ABSTAIN: AUDITING THRESHOLDED DECISIONS UNDER DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Decisions to raise an alarm, withhold an answer, or refer a case for human review are commonly made by comparing a score against a calibrated cutoff. Claims about the resulting population operating rate are typically assessed under independence, although evaluation units may share a patient, time series, or prompt. We formulate a certify-or-abstain audit that identifies the independence conditions required for valid certification. Under r-wise independence, symmetrisation reduces the analysis of any procedure to a randomised rule based on the total count. For each null rate, its maximal certification probability is a discrete moment problem. The corresponding primal and dual solutions yield exact rational certificates of feasibility and impossibility. This formulation gives exact minimum sample requirements for randomised rules, certified deterministic gates, and a scaling law across independence orders. For independent clusters, we characterise a strong criterion with unrestricted intracluster dependence and give finite-sample moment and Bernstein audits when average intracluster correlation is bounded. The r-wise hierarchy is used as a conditional sensitivity analysis rather than as a fitted model of positive intracluster dependence. Applications to clinical referral, spacecraft telemetry, and language model reliability demonstrate how the framework separates conclusions warranted by the sampling design from those that rely on unjustified independence at the observation level.

1 INTRODUCTION

Many deployed machine learning systems act by comparing a score with a threshold. Anomaly detectors raise alarms, selective predictors withhold answers, and clinical triage models refer cases for human review, and each such system is deployed with a promise about how often it will act. While principled methods exist for selecting thresholds and controlling risk from held-out data (Angelopoulos et al., 2024; 2025; Geifman & El-Yaniv, 2017), a complementary question has received far less attention. Once the threshold is fixed, when do the held-out data actually establish that the promised rate holds?

We formalise this question as an equivalence test. Let $Y_i \in \{0, 1\}$ indicate whether the system acts on evaluation unit i, let $p = \mathbb{P}(Y_i = 1)$, and fix a tolerance region (L, U) around the nominal rate p_0. A procedure returns CERTIFY only if the data support $p \in (L, U)$ at level α, and returns ABSTAIN otherwise. Both limits matter, since a detector that never fires satisfies any upper-rate constraint while no longer serving its purpose (Berger & Hsu, 1996; Schuirmann, 1987; Wellek, 2010). The central difficulty is that standard analyses treat $Y_1, \ldots, Y_N$ as independent, an assumption that is often least plausible precisely when N is large. Forecasting windows share a telemetry channel (Hundman et al., 2018), repeated generations share a prompt (Miller, 2024), and ECG windows share a patient (Wagner et al., 2020). Intracluster correlations and design effects can diagnose this issue, but they capture only second-order dispersion and do not control the tail event on which a certification decision depends (Kish, 1965; Noonan, 2026).

We therefore study two families of weaker assumptions. The first is r-wise independence, under which every r outcomes are independent with marginal p while higher-order dependence is left unrestricted. We use this hierarchy to ask how much moment information a conclusion needs. We do not treat finite r as an empirical model of dependence induced by shared patients, prompts, or time series. Finite order does hold by construction when the j-th audited unit is $h(j)$ in a fixed population,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for a hash h drawn from a k-wise independent family (Carter & Wegman, 1979; Wegman & Carter, 1981). The N outcomes are then k-wise independent with marginal equal to the population failure rate, and N is fixed in advance. Simple tabulation hashing is three-wise but not four-wise independent (Pătraşcu & Thorup, 2012). For rules based on the total count $T = \sum_i Y_i$, it fixes the first r factorial moments of T, and the worst case error of a rule becomes a finite moment problem. Classically, such problems bound the tail of a statistic that has already been chosen (Bertsimas & Popescu, 2005; Prékopa, 1990). We turn them into a design tool, constructing the test itself so that it remains valid at every rate in the null and under every dependence structure the assumption allows. The primal solution is an explicit adversarial law and the dual is a low degree polynomial that bounds the rule on every possible count, so every certificate and every counterexample can be checked in exact arithmetic. The second family is the clustered design that motivates the problem, with independent clusters, such as patients or prompts, and dependence inside each cluster that is either arbitrary or bounded in average correlation.

Our first main result is an exact sample complexity. For a typical referral target of 5% certified at the 1% level, 936 independent units suffice, whereas 1,518 are needed under four-wise and 8,173 under pairwise independence. Certified deterministic interval gates attain the target at 943 and 1,996 units under independence and four-wise independence. The randomised optima are mathematical limits. The deterministic gates are the operational default when auxiliary randomisation is unacceptable. We prove that the requirement at even order $2k$ grows as $\alpha^{-1/k}$, which separates successive even orders as α decreases, and that odd orders can improve on the even order below them only through the range of the count. The observed proximity of orders three and two is empirical. Our second main result concerns clustered designs. When clusters are independent, dependence within them is unrestricted, and nominal retention must hold uniformly over every admissible cluster law, the number of clusters must itself reach the independent requirement, and auditing one prespecified unit per cluster attains it. When the average intracluster correlation is bounded, finite-sample moment and Bernstein certificates use every unit.

These results support reporting the audit as a sensitivity analysis. For a decision certified under independence, its *robustness order* is the lowest order of independence under which it remains certified. It is a sensitivity index over nested moment classes, not an estimate of dependence, and because every order $r \geq 2$ already assumes pairwise independence it addresses higher-order dependence only. The prespecified PTB-XL referral evaluation certifies under patient independence with robustness order six, and a secondary analysis that pools ten windows per patient certifies under arbitrary dependence within patients. Most NASA telemetry designs and the Qwen2.5 evaluation certify at the level of individual outcomes, but with only 22 or 42 independent channels and 180 independent prompts no procedure can simultaneously control level over every admissible cluster law and attain 75% retention under every nominal cluster law. Valid lower-retention procedures are not ruled out. Dependence with zero correlation inside every channel already overturns most telemetry designs that certify under independence, so a design effect of one does not justify treating outcomes as independent. In simulations with clustered dependence, cluster-t TOST, the pairs and wild cluster bootstraps and design-effect corrections certify false claims up to 3.4 times their nominal level, while every finite-sample audit controls error at a cost in clusters.

The main contributions of this work are summarised below.

- We formulate operating rate validation as a finite sample certify-or-abstain problem under r-wise independence and reduce it exactly, via symmetrisation, to a moment problem on the count (Section 2).
- We derive primal and dual certificates, checkable in exact arithmetic, that determine the minimum sample size exactly for randomised rules at orders two to five and under independence, and we report certified deterministic gates alongside (Section 3).
- We prove the scaling $N^{\star}_{2k} = \Theta(\alpha^{-1/k})$ and that odd orders act only through the range of the count (Section 3).
- We characterise clustered designs under uniform retention over arbitrary within cluster dependence, and give finite-sample moment and Bernstein audits under bounded average intracluster correlation (Section 2.3).
- We apply the audits to clinical, telemetry, and language model evaluations, compare them with Cantelli gates, clustered TOST, the cluster bootstrap and design-effect corrections, and release every exact decision as a verifiable certificate (Sections 4 and 5).

2 CERTIFICATION UNDER BOUNDED INDEPENDENCE

2.1 THE CONTRACT

Calibration produces binary outcomes $Y = (Y_1, \dots, Y_N)$ and the count $T = \sum_i Y_i$. A randomised count rule is a function $\phi : \{0, \dots, N\} \to [0, 1]$, interpreted as the conditional probability of returning CERTIFY after observing $T = t$. The composite null is $\Theta_0 = [0, L] \cup [U, 1]$. For $1 \le r \le N$, let $\mathcal{L}_r(p)$ be the set of laws on Y for which every set of at most r coordinates is mutually independent with $\mathrm{Bern}(p)$ marginals. These classes are nested, $\mathcal{L}_{r+1}(p) \subseteq \mathcal{L}_r(p)$, and $\mathcal{L}_N(p)$ is full independence.

Definition 1 (Validity and retention). *A rule is valid at order r if $\sup_{p \in \Theta_0} \sup_{P \in \mathcal{L}_r(p)} \mathbb{E}_P[\phi(T)] \le \alpha$. Its nominal retention is $\mathbb{E}_{T \sim \mathrm{Bin}(N, p_0)}[\phi(T)]$. Define $N_r^\star$ as the smallest N for which a valid order-r rule has nominal retention at least ρ.*

Validity is worst case because it is the guarantee conveyed by certification. Retention is evaluated under the intended independent design and quantifies usefulness when the system meets its target. Its range over all laws in $\mathcal{L}_r(p_0)$ is reported in Section 7.

2.2 SYMMETRISATION AND THE MOMENT PROGRAM

Write $(t)_s = t(t-1)\cdots(t-s+1)$ and $(N)_s = N(N-1)\cdots(N-s+1)$. Let $\Pi_r(p)$ contain distributions $w = (w_0, \dots, w_N)$ satisfying $\sum_t w_t (t)_s / (N)_s = p^s$ for $s = 0, \dots, r$.

Lemma 1 (Symmetrisation). *For every count rule,*

$$W_r(\phi, p) := \sup_{P \in \mathcal{L}_r(p)} \mathbb{E}_P \phi(T) = \max_{w \in \Pi_r(p)} \sum_{t=0}^{N} \phi(t) w_t.$$

Moreover, because $\mathcal{L}_r(p)$ is convex and invariant under permutations of the coordinates, any rule on the full binary vector can be symmetrised into a count rule with the same nominal retention and no larger worst case error.

The reverse construction is explicit. Draw $T \sim w$ and place T ones uniformly among the N coordinates. An extremal w has at most $r + 1$ support points, so every primal attack is short and directly checkable.

Proposition 1 (Polynomial dual). *For fixed (N, r, p, ϕ),*

$$W_r(\phi, p) = \min_{c \in \mathbb{R}^{r+1}} \Big\{ \sum_{s=0}^{r} c_s p^s \,:\, \sum_{s=0}^{r} c_s \frac{(t)_s}{(N)_s} \ge \phi(t) \;\; \forall t \in \{0, \dots, N\} \Big\}.$$

The dual polynomial is a proof, not merely an optimiser output, because its domination is checked on every integer count and its expectation is determined by the matched moments. The audit therefore has three outputs. A dual polynomial certifies validity, a moment-matched law shows that a proposed certificate is invalid, and when neither is found the result is reported as unresolved.

2.3 CLUSTERED DESIGNS

In the motivating applications the units come in clusters $g = 1, \dots, G$ of sizes m_g, such as the windows of a channel, the generations for a prompt or the recordings of a patient. Let $\mathcal{C}(p)$ contain the laws under which clusters are mutually independent and every unit is $\mathrm{Bern}(p)$, with arbitrary dependence inside each cluster, and let $\mathcal{C}_r(p) \subseteq \mathcal{C}(p)$ add r-wise independence inside each cluster. Here p is the unconditional action probability for a fresh cluster and a prespecified within cluster unit drawn according to the declared sampling scheme, not the average of fixed cluster-specific probabilities.

Proposition 2 (Clustered designs). *(a) $\mathcal{C}_r(p) \subseteq \mathcal{L}_r(p)$, so every order-$r$ certificate is valid for independent clusters with r-wise independence inside them. (b) Under $\mathcal{C}$, a procedure that is valid on the null and has retention at least ρ under every law in $\mathcal{C}(p_0)$, including the one with homogeneous clusters, exists if and only if $G \ge N_\infty^\star$. When $G \ge N_\infty^\star$, it suffices to apply the optimal independent rule to one unit per cluster, chosen before the data are seen.*

Part (b) holds because the prespecified units are independent $\mathrm{Bern}(p)$ under every law in $\mathcal{C}(p)$, whereas the law that copies one $\mathrm{Bern}(p)$ outcome to every unit of its cluster belongs to $\mathcal{C}(p)$ and reduces any procedure to a test on G independent outcomes (Appendix B). If retention were required only under full independence, $G \geq N^\star_\infty$ would remain sufficient but not necessary, since a procedure can exploit within cluster information. That gain disappears when clusters are homogeneous, which is the dependence the evaluator is guarding against.

For an intermediate model, let the clusters have a common size m, so $N = Gm$, write Y_{gi} for unit i of cluster g and $S_g = \sum_i Y_{gi}$, and define the average intracluster correlation

$$\bar{\gamma}(P, p) = \frac{1}{Gm(m-1)p(1-p)} \sum_{g=1}^{G} \sum_{i \neq j} \mathrm{Cov}_P(Y_{gi}, Y_{gj}).$$

Let $\mathcal{C}_\kappa(p)$ contain laws with independent clusters, common $\mathrm{Bern}(p)$ marginals, and $0 \leq \bar{\gamma}(P, p) \leq \kappa$, and let $V_p(\kappa) = Np(1-p)\{1 + (m-1)\kappa\}$.

Proposition 3 (Bounded intracluster correlation). *For every $P \in \mathcal{C}_\kappa(p)$, $\mathbb{E}_P T = Np$ and*

$$(N)_2 p^2 \leq \mathbb{E}_P(T)_2 \leq (N)_2 p^2 + N(m-1)\kappa p(1-p),$$

so maximising the acceptance probability over count laws with these moments bounds it over $\mathcal{C}_\kappa(p)$. Moreover, for an observed count $NL < x < NU$,

$$\sup_{P \in \mathcal{C}_\kappa(L)} \mathbb{P}_P(T \geq x) \leq \exp\Big\{ - \frac{(x - NL)^2}{2[V_L(\kappa) + m(1-L)(x-NL)/3]} \Big\},$$

$$\sup_{P \in \mathcal{C}_\kappa(U)} \mathbb{P}_P(T \leq x) \leq \exp\Big\{ - \frac{(NU - x)^2}{2[V_U(\kappa) + mU(NU-x)/3]} \Big\}.$$

For interval gates, these boundary bounds control the complete null. At $\kappa = 1$ the Bernstein bounds hold for every law in $\mathcal{C}(p)$.

The moment relaxation keeps only two aggregate moments and discards independence across clusters. The Bernstein bound retains that independence, and at $\kappa = 1$ its variance bound $m^2 p(1-p)$ per cluster holds for every within cluster law, so a count that certifies at $\kappa = 1$ needs no assumption about dependence within clusters (Appendix B). Both certificates are rigorous, but failure of either to certify is not an impossibility result, and a cap below one must be justified externally.

3 Exact sample complexity

The robust design problem is a semi-infinite linear program, in which ϕ has $N + 1$ variables while validity ranges over a continuum of null rates and a moment polytope at every rate. The following result gives sound finite certificate forms and guarantees their existence when the relevant inequalities have strict slack.

Theorem 1 (Finite verifiable certificates under strict slack). *Fix a rational contract $(L, U, p_0, \alpha, \rho)$ and order r. (a) Any finite collection of rational null rates, rational laws in their moment polytopes, and nonnegative rational multipliers whose combined dual bound on nominal retention is below ρ certifies impossibility at the stated N. (b) Any rational rule ϕ together with a finite rational interval cover of Θ_0, on each element of which a rational degree-r polynomial dominates ϕ at every integer count and has expectation at most α, certifies feasibility. (c) Feasibility is monotone in N, since discarding a uniformly chosen observation converts a valid rule at N into one at $N + 1$ with unchanged nominal retention. A certificate of type (a) exists if the optimal nominal retention at N is strictly below ρ, and one of type (b) exists if some rule has nominal retention strictly above ρ and worst case null error strictly below α. By (c), matching certificates at $N - 1$ and N establish $N^\star_r = N$ without trusting the numerical solver used to discover them.*

The interval expectation of each dual polynomial is itself a degree-r polynomial in p. Expressing it in the Bernstein basis provides an exact certificate over an entire interval from finitely many rational coefficient inequalities (Farouki, 2012).

Theorem 2 (Primary contract). *For $(L, U, p_0, \alpha, \rho) = (0.025, 0.075, 0.05, 0.01, 0.75)$,*

$$N^\star_\infty = 936, \qquad N^\star_5 = 1498, \qquad N^\star_4 = 1518, \qquad N^\star_3 = 7958, \qquad N^\star_2 = 8173.$$

Table 1: Exact randomised sample complexity and certified deterministic gates for $p_0 = 0.05$ and retention 0.75. The primary contract is $(L, U) = (0.025, 0.075)$ with $\alpha = 0.01$, and the others change α or the region. Randomised values have matching lower and upper certificates. The deterministic entries are certified half-line gates. At order four the gate is minimal over all deterministic rules, numerically, at order five among half-line gates, under independence it is exact TOST, and at orders two and three it is minimal among half-line gates within a 150-value backward scan (Appendix E).

	independent		5-wise		4-wise		3-wise		pairwise	
Contract	rand.	det.	rand.	det.	rand.	det.	rand.	det.	rand.	det.
primary	936	943	1,498	1,993	1,518	1,996	7,958	11,929	8,173	12,223
$\alpha = 0.05$	594	598	768	1,002	774	1,003	1,589	2,598	1,648	2,656
narrow $(0.03, 0.07)$	1,449	1,482	2,277	2,996	2,311	3,016	11,737	17,641	12,002	17,977
wide $(0.02, 0.08)$	658	664	1,078	1,437	1,090	1,455	5,832	8,715	6,021	8,971

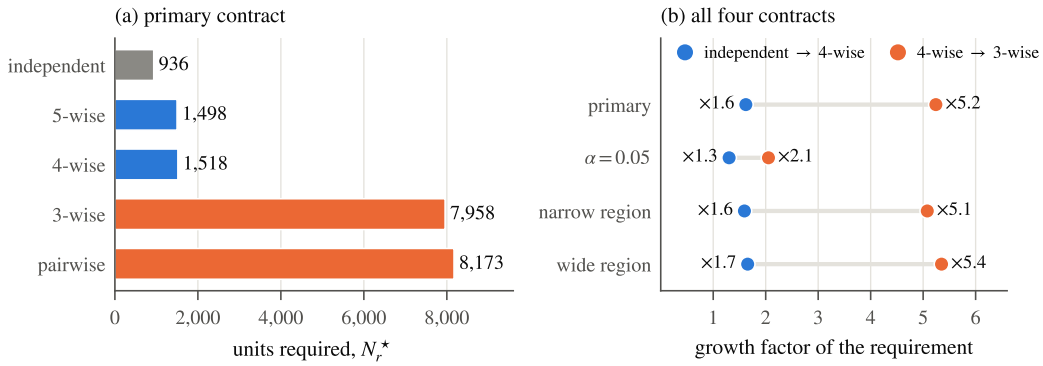


Figure 1: Exact sample sizes, each certified in both directions. (a) Units required for the primary contract under each assumption. (b) Growth of the requirement from full to four-wise independence (blue) and from four-wise to three-wise independence (orange), for the primary contract and for three further contracts with level $\alpha = 0.05$, a narrow region $(0.03, 0.07)$ and a wide region $(0.02, 0.08)$.

At $N = 1{,}517$, the exact fourth-order dual bounds the maximum retention by 0.7499325. At $N = 1{,}518$, an explicit rational rule has retention 0.7503507 and is covered on the complete null by 159 verified polynomial intervals. Orders five, three and two are closed in the same way, and so are all four orders in three further contracts (Table 1). Each upper certificate covers the entire null rather than inferring complete-null validity from boundary calculations.

The reported deterministic gates use 30 to 34% more units than the exact randomised requirement at orders four and five and 49 to 64% more at orders two and three. The gap reflects the shape of the dual. Under independence the likelihood ratio is steep and monotone, so the optimal rule randomises at two counts and a gate loses under 1%. Under a few moment constraints the dual polynomial is flat near the boundaries, so the optimal rule ramps over many counts, 196 at order two and 79 at order four (Figure 2), and a 0 or 1 rule forfeits the partial acceptances on that ramp. A mixed-integer program over all deterministic rules, not only intervals, finds none below the reported gate at order four (Appendix E). The computed jump from four-wise to three-wise independence is large at $\alpha = 0.01$ and small at $\alpha = 0.05$ (Figure 1b). Across the contracts examined, changing only the half-width rescales all sample sizes by approximately its inverse square, while changing α changes the order gap, an empirical pattern that the next result explains asymptotically.

Proposition 4 (Scaling in the error level)**.** *Fix $0 < L < p_0 < U < 1$, $\rho \in (0, 1)$ and $k \geq 1$. There are constants $0 < c \leq C$ such that $c\,\alpha^{-1/k} \leq N^\star_{2k} \leq C\,\alpha^{-1/k}$ for all sufficiently small α, while $N^\star_\infty = \Theta(\log(1/\alpha))$. Hence $N^\star_{2k}/N^\star_{2k+2} \to \infty$ as $\alpha \to 0$, and $N^\star_{2k+2} \leq N^\star_{2k+1} \leq N^\star_{2k}$.*

The upper bound applies Markov's inequality to $(T - Np)^{2k}$, whose expectation is fixed by the first $2k$ moments, and the lower bound perturbs the binomial law at the boundary by moment-preserving mass placed where a useful rule must certify (Appendix C). The computed boundary-program sizes

Table 2: Independent-data gates at $N = 943$. Size columns report the larger exact boundary error at L or U. The target is retention at least 0.75 and size at most 0.01. Error curves over the whole null are in Appendix A.6.

Rule	Acceptance counts	Retention	Independent size	4-wise size	2-wise size
Exact TOST	36 to 52	0.7539	0.00977	0.05875	0.15696
99% Clopper and Pearson	38 to 50	0.6267	0.00456	0.04058	0.13201

agree (Table 6), with local exponents $-\,d\log N_r/d\log\alpha$ of 0.93 to 1.01 at order two, 0.41 to 0.48 at order four, and 0.26 to 0.35 at order six, which straddles the asymptotic value $1/3$ over the finite grid rather than remaining uniformly below it.

Odd orders. Let $\Pi_r^{\mathbb{Z}}(p)$ be the finitely supported laws on the integers whose first r factorial moments match $\mathrm{Bin}(N,p)$, and let $N_r^{\mathbb{Z}}$ be the sample size when validity is required against them, with every rule extended by zero outside $\{0,\dots,N\}$.

Theorem 3 (Odd orders act only through the range of the count)**.** *Let $k \geq 1$. (a) For every count rule and every $p \in [0,1]$, $\sup_{w\in\Pi^{\mathbb{Z}}_{2k+1}(p)} \sum_t \phi(t)w_t = \sup_{w\in\Pi^{\mathbb{Z}}_{2k}(p)} \sum_t \phi(t)w_t$, hence $N^{\mathbb{Z}}_{2k+1} = N^{\mathbb{Z}}_{2k}$. (b) $N^{\star}_{2k+1} \leq N^{\star}_{2k} \leq N^{\mathbb{Z}}_{2k} = N^{\mathbb{Z}}_{2k+1}$. If $N^{\mathbb{Z}}_{2k} = N^{\star}_{2k}$, the saving $N^{\star}_{2k} - N^{\star}_{2k+1}$ is exactly the value, at order $2k+1$, of knowing that $0 \leq T \leq N$.*

The mechanism is the classical one on an unbounded domain, where a vanishing atom placed far away moves an odd moment at negligible cost. What the lattice setting adds is that the relaxation is tight in practice. All eight even order certificates behind Figure 1 remain valid against integer-supported laws, so $N^{\mathbb{Z}}_{2k} = N^{\star}_{2k}$ in every contract, and the odd order savings are exactly the value of the range constraint (Appendix D). A related reduction holds for the probability that k-wise independent bits are all one (Benjamini et al., 2012; Peled et al., 2011).

4 Boundary calibration and standard tests

Under independence, exact TOST is the intersection of two one-sided binomial tests, each monotone in p, so its error on the null is largest at L or U, and the rule that maximises retention subject to error control at L and U is the uniformly most powerful equivalence test (Lehmann & Romano, 2005). Boundary feasibility alone does not imply complete-null validity for an arbitrary nonmonotone rule, even under independence, and it matters in practice when validating numerical optimisation output. Numerically optimal solutions of the boundary program can put acceptance mass on counts that interior null laws reach (Appendix E), so every upper certificate is checked on the complete null. That repair costs nothing in sample size in all sixteen cells of Table 1.

Standard tests. Table 2 fixes $N = 943$, the first sample size at which exact TOST reaches the target retention, and maximises each gate's acceptance probability over count laws matching Bernoulli moments through order four or two at L and U. Each entry has matching exact primal and dual certificates, and because the boundaries belong to the null these values are rigorous lower bounds on the complete null size. A design-effect-adjusted TOST with estimated correlation zero reproduces the first row, and pairwise independent laws have correlation zero, so the correction cannot detect them. The required amount of conservatism cannot be determined without the robust analysis.

Comparison with one-sided moment bounds. Cantelli and fourth-moment Markov gates calibrated separately at L and U first reach retention 0.75 at $N = 12{,}224$ under pairwise and $N = 2{,}443$ under four-wise independence (Table 5). The exact gates need $12{,}223$ and $1{,}996$, so exactness gains nothing over Cantelli at order two and 18% at order four. Randomisation reduces the requirements by further factors of 1.50 and 1.31, and the lower certificates show that no procedure can do better, which no elementary inequality establishes.

For evaluators who need the same decision from the same data, a one-sided relaxation restores a valid boundary reduction. For $q \leq L$, superpose independent Bernoulli noise to raise the marginal to L, and for $q \geq U$, independently thin successes to lower it to U. Both preserve r-wise independence and move the count monotonically.

Table 3: Application decisions for the primary contract. "Indep." records whether the observed count triggers exact TOST. For $G \in \{22, 42, 180\}$, exact TOST on one prespecified unit per cluster has an empty gate, and Proposition 2 shows that no procedure combines uniform validity with retention 0.75 over every cluster law, although valid lower-retention procedures exist. $G \times m$ gives clusters times units per cluster (patients, channels or prompts). The pooled PTB-XL row is a secondary analysis, and the Qwen estimate of κ is 0.68. Parenthesised $r^\dagger$ are observation-level indices that the clustered design does not license.

Study and endpoint	$G \times m$	x/N	Indep.	$r^\dagger$	Design-aware output
PTB-XL, position 4 (prespecified)	1877×1	101/1877	Cert.	6	Certify
PTB-XL, positions 0 to 9	1877×1	82 to 114	Cert.	4 to > 12	Sensitivity only
PTB-XL, pooled windows	1877×10	1011/18770	Cert.	n/a	Certify, any dependence
SMAP, innovation	42×30	54/1260	Cert.	(4)	Abstain at every κ
SMAP, rolling maximum	42×30	53/1260	Cert.	(4)	Abstain at every κ
MSL, innovation	22×30	31/660	Cert.	(> 8)	Abstain at every κ
MSL, rolling maximum	22×30	34/660	Abst.	n/a	Abstain at every κ
Qwen2.5-32B, GSM8K failures	180×50	436/9000	Cert.	(4)	Abstain, needs $\kappa \leq 0.11$

Proposition 5 (Composite-null deterministic gate)**.** *Let* $p_r^-(x) = W_r(\mathbf{1}\{T \geq x\}, L)$ *and* $p_r^+(x) = W_r(\mathbf{1}\{T \leq x\}, U)$*. Then* $A_r = \{x : p_r^-(x) \leq \alpha,\ p_r^+(x) \leq \alpha\}$ *is valid over the complete composite null. For the primary contract at* $r = 4$*,* $N = 1{,}996$ *gives* $A_4 = \{77, \ldots, 106\}$*, with boundary bounds* 0.009223 *and* 0.009987 *and nominal retention* 0.750197*.*

The gate should be designed at its own sample size. At $N = 1{,}518$ the order-4 gate is $\{62, \ldots, 75\}$ with retention only 0.447.

5 Applications as sensitivity analyses

Operational use. The assumption is selected from the evaluation design before outcomes are inspected. If the evaluation units are independently sampled, use the independent audit. If repeated units share a patient, prompt, or channel, use the cluster audit. With no defensible correlation cap, set $\kappa = 1$ or audit one prespecified unit per cluster. If the j-th unit is $h(j)$ for a k-wise independent hash h into a fixed population, the order-k audit applies by construction. Otherwise a robustness order is only a stress test.

Estimand and robustness order. In PTB-XL, p is the referral probability for a new patient at a prespecified window position, or at a uniformly selected position in the pooled secondary analysis. In the Qwen and NASA studies, the common marginal statements use a superpopulation interpretation in which a fresh prompt or channel and its prespecified unit are drawn according to the declared sampling scheme. For NASA this is a modelling assumption, since the released channels are a fixed set. An observed count x certifies at order r when $x \in A_r$. Since $\mathcal{L}_{r+1} \subseteq \mathcal{L}_r$, the gates are nested, $A_r \subseteq A_{r+1} \subseteq A_\infty$, and for a count certified under independence the *robustness order* $r^\dagger(x) = \min\{r \geq 2 : x \in A_r\}$ records where the decision changes along this hierarchy. For example, $r^\dagger = 6$ means the decision is valid throughout the six-wise class while an explicit five-wise null law overturns it. It neither estimates dependence nor validates six-wise independence, and positive intracluster dependence is addressed separately by Propositions 2 and 3. Certifications in Table 3 are backed by exact dual polynomials and abstentions by exact moment-matched laws.

PTB-XL. PTB-XL contains multiple ECG records per patient (Wagner et al., 2020). Folds 1 to 8 were used for model fitting, fold 9 for the referral threshold, and fold 10 for evaluation. The first eligible record of each patient was retained, giving $1{,}877$ test patients (Appendix F.1), and each record was divided into ten fixed window positions, of which position 4 was prespecified. The threshold targeted a referral rate of 0.05, and the classifier had window AUC 0.840 and patient AUC 0.870. Because the outcomes at a given position come from distinct patients, patient-level independence is the defensible assumption, and the prespecified position certifies, since its 101 referrals lie in the exact TOST gate $\{64, \ldots, 114\}$. Its robustness order is six. It abstains under five-wise independence (upper-null p-value 0.0125) and certifies under six-wise independence (0.0036). The other nine positions reuse

the same patients and have robustness orders from 4 to more than 12. Pooling all ten windows gives $1{,}011$ referrals among $18{,}770$ outcomes, with plug-in average intrapatient correlation 0.101. The two-moment relaxation certifies only up to $\kappa = 0.02$, but the Bernstein bound is 0.00249 even at $\kappa = 1$, so the pooled rate certifies under arbitrary dependence within patients, without relying on the estimated correlation.

Qwen2.5-32B. The study generated 50 answers for each of 180 GSM8K prompts (Cobbe et al., 2021) using Qwen2.5-32B-Instruct (Qwen Team, 2024), giving $9{,}000$ generations with 436 failures, a rate of 0.04844. Treating generations as independent gives the exact gate $\{261, \ldots, 617\}$, and the count certifies with robustness order four. That order is not defensible here, because every $r \geq 2$ implies zero within-prompt correlation, while the estimated correlation is 0.680. With $G = 180$ independent prompts and arbitrary dependence among the generations of a prompt, no procedure can guarantee level α and retention ρ under every cluster law. The optimal randomised rule has retention 0.034, and the exact TOST gate for one generation per prompt is empty ($c_1 = 11 > c_2 = 5$). The Bernstein certificate holds only for $\kappa \leq 0.110$, far below the plug-in correlation. Even if retention is required only under a beta-binomial nominal law at the estimated correlation rather than under the copy law, explicit rate-mixture laws show that no valid half-line gate reaches the target with fewer than 645 prompts, and even validity under beta-binomial laws alone caps the retention of every rule based on the total count at 0.062 with 180 prompts (Appendix F.4). Under the more natural one-sided contract $p < 0.075$, retention 0.75 requires 865 prompts.

NASA telemetry. Two alarm scores were computed on SMAP and MSL spacecraft telemetry (Hundman et al., 2018), the standardised absolute one-step innovation of an AR(8) forecaster and its trailing maximum over eight steps, with the forecaster, the pooled 95% threshold and the audited windows taken from successive 40%, 30% and 30% segments of each nominal training series (Appendix F.3). The fixed designs take 30 equally spaced windows from each of 42 SMAP and 22 MSL channels, and the channel is the natural cluster. At the window level, the two SMAP designs certify with robustness order four, the MSL innovation count triggers the independent gate $\{27, \ldots, 33\}$, and the MSL rolling-maximum count does not. Let channels be independent and the windows of a channel exchangeable and pairwise independent, so that every correlation is zero. Under one such law, in which almost every channel raises zero or one alarm and a rare channel raises a burst of 27, a count of at least 54 has probability 0.028 at the rate L, so the SMAP innovation certificate is not valid. Across ten stratified resamples of 30, 40 or 50 windows per channel, 93 designs certify under independence, and laws of this kind overturn 81 of them, including 27 that certify at order four (Table 10). These laws have correlation zero and so belong to $\mathcal{C}_\kappa$ for every κ, so no fixed design can certify under any cap, whatever the plug-in correlations (0.030 to 0.100). A design effect of one therefore does not validate a certificate based on independence.

Standard clustered practice. On these data, cluster-t TOST and the cluster bootstrap reach the same decisions as the Bernstein audit (Table 11, Appendix F.4). Agreement on observed data does not imply validity. Simulations with random cluster rates or serially dependent windows and 20 to 150 clusters (Table 13) show that at $p = U$ cluster-t TOST, which here coincides with CR2, the pairs and wild cluster bootstraps, and the design-effect TOST certify in up to 3.2%, 3.4%, 2.5% and 1.8% of replications, against the nominal 1%. No finite-sample audit certified at either boundary, at a cost in retention. The Bernstein audit with the true cap reaches 0.74 with 150 clusters of 50 units at correlation 0.02, while betting and empirical-Bernstein audits, which assume nothing within clusters, have zero retention in every setting, and betting first reaches 0.92 with 400 clusters. Standard practice thus gains power by giving up the finite-sample guarantee when rates are small and clusters few.

6 Related work

Equivalence testing. Exact equivalence tests, which certify only when each of two one-sided tests rejects (Berger & Hsu, 1996; Schuirmann, 1987), the exact binomial intervals of Clopper & Pearson (1934) and acceptance sampling plans control type-I error when observations are independent (Wellek, 2010). Our full-independence results recover this setting as a special case and extend it to families of dependent laws controlled over the whole composite null.

Moment problems, robust test design and limited independence. Optimisation over distributions with prescribed moments is the discrete moment problem (Boros & Prékopa, 1989; Prékopa, 1990),

and convex formulations yield sharp probability inequalities, including tight bounds under pairwise independence (Bertsimas & Popescu, 2005; Ramachandra & Natarajan, 2023). Closely related work studies the largest probability that k-wise independent bits with identical marginals are all one (Peled et al., 2011; Benjamini et al., 2012; Berend et al., 2026). Magesh et al. (2023; 2022) optimise randomised minimax tests on finite alphabets with moment constraints, so neither moment duality nor optimisation over the test is new by itself. Our arguments symmetrise arbitrary rules on N Bernoulli coordinates, control error over a continuum of null rates, certify exact minimum sample sizes at a prescribed retention, and characterise the count range at odd orders. Limited independence theory asks when low order moments suffice to control threshold functions (Diakonikolas et al., 2010; Schmidt et al., 1995), and its tail bounds anticipate our scaling law.

Robust testing and dependent data. Minimax robust tests protect against perturbations around a nominal model (Huber, 1965; Gao et al., 2018), and mixing conditions bound time-series dependence through rates that a single calibration set cannot verify (Bradley, 2005). We keep every marginal rate and leave the dependence unobserved.

Dependence corrections and distribution-free guarantees. Design effects and clustered variance estimators quantify the precision lost under dependence (Kish, 1965; Miller, 2024). Noonan (2026) shows that the relevant intracluster correlation depends on the operating point, so a single design effect cannot serve every threshold. Our unrestricted cluster result avoids estimating it, and Proposition 3 converts a declared cap into finite-sample certificates without a normal approximation. Empirical-Bernstein and betting audits of cluster rates (Maurer & Pontil, 2009; Waudby-Smith & Ramdas, 2024) are valid under arbitrary dependence within clusters. On our data the former certifies only pooled PTB-XL, and in simulation both need several hundred clusters (Appendix G). Conformal prediction and risk-control methods give finite sample guarantees under exchangeability and study how these degrade when it fails (Angelopoulos et al., 2024; Barber et al., 2023; Vovk et al., 2005). Our work complements them by fixing a binary operating rate contract and determining exactly what can be certified under each assumption.

7 Limitations and conclusion

The audit does not test whether an independence order holds. It quantifies the consequence of an assumption that the evaluator must defend from the sampling design. The r-wise results are exact sensitivity frontiers, since no application supplies a mechanism that is finite-r but not fully independent. The common marginal assumption is essential. The exact guarantees do not automatically apply to a fixed collection of units with arbitrary rates p_i, since heterogeneous marginals replace the identities $\mathbb{E}(T)_s = (N)_s p^s$ by elementary symmetric functions and destroy the one-dimensional reduction, and a shift between calibration and deployment cannot be detected from the calibration set alone. Poisson-binomial comparisons bound some one-sided tails under heterogeneous rates but do not give both validity and retention here. Proposition 3 covers bounded average correlation for independent equal-sized clusters, its Bernstein audit is valid but conservative, and a cap below one must be justified externally. Deterministic gates use 30 to 64% more units than the randomised optima, and their minimality over all deterministic rules is established only numerically, at order four. Retention is defined under independence. For the certified order-4 rule at $N = 1{,}518$ it ranges from 0.682 to 0.849 over all laws in $\Pi_4(p_0)$, and for the order-2 rule at $N = 8{,}173$ from 0.699 to 0.753 (computed in floating point). The boundary reduction for rules that certify only inside the region is a conjecture, and the small savings at odd orders are observed rather than proven.

The central lesson is operational. A large number of recorded outcomes is not automatically a large calibration sample. Once dependence is admitted, the audit states how much independence a certificate needs, counts clusters when that is all the design supports, and returns ABSTAIN whenever neither the data nor the assumption support the promised rate.

Reproducibility statement

Every certificate behind Theorem 2, Table 1 and Figure 1, the sixteen deterministic gates, the counterexamples of Table 8, every robust p-value behind Table 3 and Appendix F.1, and the gates and overturning laws of Table 10 are supplied in the anonymous supplementary artifact.

It also contains the scripts and outputs for the one-sided moment gates of Table 5, the moment and Bernstein bounds and clustered comparisons of Table 11, the simulation of Appendix G, and the empirical-Bernstein baseline. Figure 2 is generated directly from the released exact upper certificates. The top-level `README.md` maps each claim to its files. From the extracted directory, `python verify_all.py` checks the manifest and verifies moments, domination, interval coverage, error and retention in rational arithmetic using only the Python standard library. It rejects altered certificates and completes in about 40 seconds on one core. Model weights and public raw datasets are referenced rather than copied.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing, to draft sections of the paper that the authors then revised, for retrieval and discovery of related work, for research ideation and execution, including drafting analysis and verification code, and to help draft proofs of mathematical claims. We reviewed all AI-assisted work. Every cited claim was checked against its source and every proof was checked by the authors. No model output is used as evidence. Every sample size, robust p-value and counterexample described as exact rests on the released certificates, which the standalone verifier checks in exact arithmetic. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Robert M. Bell and Daniel F. McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–181, 2002.

Itai Benjamini, Ori Gurel-Gurevich, and Ron Peled. On K-wise independent distributions and Boolean functions, 2012.

Daniel Berend, Philip A. Ernst, Aryeh Kontorovich, and Rishi Kumar. Exact expressions for the maximal probability that all k-wise independent bits are 1. *Journal of Theoretical Probability*, 39 (3):58, 2026. doi: 10.1007/s10959-026-01487-4.

Roger L. Berger and Dennis D. Boos. P values maximised over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

Endre Boros and András Prékopa. Closed form two-sided bounds for probabilities that at least r and exactly r out of n events occur. *Mathematics of Operations Research*, 14(2):317–342, 1989. doi: 10.1287/moor.14.2.317.

Richard C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

J. Lawrence Carter and Mark N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18(2):143–154, 1979.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010. doi: 10.1137/100783030.

Rida T. Farouki. The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419, 2012. doi: 10.1016/j.cagd.2012.03.001.

Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, 2018.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Peter J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Söderström. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 387–395, 2018.

Leslie Kish. *Survey Sampling*. Wiley, 1965.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

Akshayaa Magesh, Zhongchang Sun, Venugopal V. Veeravalli, and Shaofeng Zou. Robust multi-hypothesis testing with moment-constrained uncertainty sets. *arXiv preprint arXiv:2210.12869*, 2022.

Akshayaa Magesh, Zhongchang Sun, Venugopal V. Veeravalli, and Shaofeng Zou. Robust hypothesis testing with moment constrained uncertainty sets. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10096843.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024.

Adam Noonan. The exceedance design effect: Effective sample size for thresholds under clustering, 2026. Preprint, arXiv:2608.21262.

Mihai Pătraşcu and Mikkel Thorup. The power of simple tabulation hashing. *Journal of the ACM*, 59 (3):14:1–14:50, 2012.

Ron Peled, Ariel Yadin, and Amir Yehudayoff. The maximal probability that k-wise independent bits are all 1. *Random Structures & Algorithms*, 38(4):502–525, 2011.

András Prékopa. The discrete moment problem and linear programming. *Discrete Applied Mathematics*, 27(3):235–254, 1990. doi: 10.1016/0166-218X(90)90068-N.

Qwen Team. Qwen2.5 technical report, 2024. arXiv:2412.15115.

Arjun Kodagehalli Ramachandra and Karthik Natarajan. Tight probability bounds with pairwise independence. *SIAM Journal on Discrete Mathematics*, 37(2):516–555, 2023. doi: 10.1137/21M1408294.

Jeanette P. Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff–Hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2):223–250, 1995.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Matthew D. Webb. Reworking wild bootstrap-based inference for clustered errors. *Canadian Journal of Economics*, 56(3):839–858, 2023.

Mark N. Wegman and J. Lawrence Carter. New hash functions and their use in authentication and set equality. *Journal of Computer and System Sciences*, 22(3):265–279, 1981.

Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, 2nd edition, 2010.

A Proofs and certificate format

A.1 Symmetrisation

For any $P \in \mathcal{L}_r(p)$,

$$\mathbb{E}_P(T)_s = \sum_{i_1,\dots,i_s \text{ distinct}} \mathbb{P}_P(Y_{i_1} = \cdots = Y_{i_s} = 1) = (N)_s p^s,$$

so the count law belongs to $\Pi_r(p)$. Conversely, for $w \in \Pi_r(p)$, draw $T \sim w$ and choose a uniformly random T-subset of coordinates to set to one. Any specified $s \le r$ coordinates are all one with probability $\mathbb{E}[(T)_s/(N)_s] = p^s$, and the principle of inclusion and exclusion gives their product Bernoulli law. For an arbitrary rule ψ on $\{0,1\}^N$, let $\bar{\phi}(t)$ be its average over the $\binom{N}{t}$ arrangements with t ones. Under the nominal binomial law the arrangement is uniform given T, so retention is unchanged. For validity, fix $P \in \mathcal{L}_r(p)$ and let P^σ be its image under a coordinate permutation σ. Because $\mathcal{L}_r(p)$ is invariant under permutations, every P^σ is in $\mathcal{L}_r(p)$, and because it is convex, the average $\bar{P}$ of the P^σ is in $\mathcal{L}_r(p)$ as well. Then $\mathbb{E}_P\bar{\phi}(T) = \mathbb{E}_{\bar{P}}\psi \le \sup_{Q\in\mathcal{L}_r(p)} \mathbb{E}_Q\psi$, so the symmetrised rule has no larger worst case error. This proves Lemma 1.

A.2 Duality

For fixed p, the primal is

$$\max_{w\ge 0} \sum_t \phi(t) w_t \quad \text{s.t.} \quad \sum_t w_t \frac{(t)_s}{(N)_s} = p^s, \qquad s = 0,\dots,r.$$

It is feasible because the binomial count law is feasible, and it is bounded because $0 \le \phi \le 1$. Ordinary linear-programming duality yields the polynomial dual in Proposition 1. A vertex of the primal has at most $r+1$ nonzero weights.

A.3 SAMPLE AUGMENTATION

Suppose ϕ is valid on N outcomes. On $N+1$ outcomes, discard a uniformly selected coordinate and apply ϕ to the remaining count. Conditional on total t, the new rule is

$$\phi^+(t) = \frac{t}{N+1}\,\phi(t-1) + \Big(1 - \frac{t}{N+1}\Big)\phi(t).$$

Every N-coordinate marginal of a law in $\mathcal{L}_r(p)$ remains in $\mathcal{L}_r(p)$, so validity is preserved. Under the nominal independent law, the remaining count is $\mathrm{Bin}(N, p_0)$, so retention is unchanged.

A.4 EXACT LOWER AND UPPER CERTIFICATES

A lower certificate contains rational null rates p_k, rational laws $w_k \in \Pi_r(p_k)$, and rational multipliers $\lambda_k \geq 0$. Every valid rule satisfies

$$\mathbb{E}_{p_0}\phi(T) \leq \sum_t \Big(b_{N,p_0}(t) - \sum_k \lambda_k w_k(t)\Big)_+ + \alpha \sum_k \lambda_k.$$

If the right-hand side is below ρ, no valid rule can attain the target retention.

An upper certificate contains rational values of $\phi(t)$ and intervals that tile $[0, L] \cup [U, 1]$. Each interval has a rational polynomial dual that dominates ϕ at all $N+1$ counts. After mapping the interval to $[0, 1]$, nonpositive Bernstein coefficients of the difference between the polynomial expectation and α certify the error bound throughout that interval. All sixteen pairs of certificates are listed in Table 4.

Existence under strict slack. For the lower certificate, let $K = \{\phi \in [0,1]^{N+1} : \mathbb{E}_{p_0}\phi(T) \geq \rho\}$. If the optimal valid retention is strictly below ρ, then every $\phi \in K$ violates the error constraint for some null rate and some law in the corresponding moment polytope. These strict violations form an open cover of the compact set K, so a finite subcover suffices. The rates and laws may be replaced by nearby rational feasible points while preserving them. For an interior rate, first mix the law with the binomial law to enter the relative interior of its rational moment polytope, and then approximate the rate and the resulting feasible point rationally. At rates zero and one the feasible law is already rational. The resulting finite rational linear program has optimum below ρ, and its finite-dimensional dual supplies the nonnegative rational multipliers in Theorem 1(a).

For the upper certificate, start from a rule whose retention exceeds ρ and whose worst case null error is below α, both with strict slack. A sufficiently close rational approximation to the rule preserves both inequalities. At each $p \in \Theta_0$, strong duality for the finite moment program gives a degree-r majorant. Approximate its coefficients rationally and increase its constant term by a rational amount so that domination holds exactly at all $N+1$ counts. The strict error slack is preserved. Its expected value is a polynomial in p, so continuity gives a neighbourhood on which it remains below α. Compactness of Θ_0 yields a finite subcover with rational endpoints. After sufficiently fine rational subdivision, the Bernstein coefficients certify the strict polynomial inequality on every interval. This proves the two existence statements in Theorem 1. At equality, an exact certificate may still exist, as it does in our computations, but the theorem does not assert completeness.

How certificates are found. (1) Solve the boundary program by dualising the inner supremum at L and U, which gives one linear program in ϕ and the dual coefficients. (2) Find the smallest N at which it reaches ρ. Its dual at $N-1$ gives the lower certificate after exact re-solution of each law on its support. (3) At that N, break ties by minimising total certification mass, set ϕ to zero on counts whose nominal probability is below 10^{-15} (which can only lower every error), and shrink the level to $(1-\kappa)\alpha$ with $\kappa \leq 2 \times 10^{-5}$. Scan the worst case error over Θ_0, add any violating local maximiser as a further null rate, and re-solve until no violation remains. (4) Cover Θ_0 adaptively. On each interval, take the dual polynomial at an endpoint, repair domination exactly, and bisect the interval if its Bernstein bound exceeds α. If bisection fails, add the local maximiser of the error on that interval to the constraint set and return to (3). A rule produced by the boundary program is therefore never accepted as a complete null certificate on boundary evidence alone. Interior rates were needed in four cells ($\alpha = 0.05$ at orders two, three and five, and the wide region at order five), and the certified sample size was unchanged in all of them.

Under independence the optimal rule is the uniformly most powerful equivalence test, valid on the whole null (Lehmann & Romano, 2005). Its requirement is $N^\star_\infty = 936$, 594, 1,449 and 658 for the four contracts, with exact binomial retentions below 0.75 at $N^\star_\infty - 1$.

Table 4: Exact certificates behind Table 1 and Figure 1. The lower column gives the bound above at $N_r^\star - 1$, using boundary laws only. The upper columns give the exact nominal retention of the certified rule at $N_r^\star$ and the number of polynomial intervals that cover Θ_0.

Contract	r	$N_r^\star$	Lower bound at $N_r^\star - 1$	Retention at $N_r^\star$	Intervals
primary	5	1498	0.749811	0.750335	159
primary	4	1518	0.749933	0.750351	159
primary	3	7958	0.749921	0.750003	163
primary	2	8173	0.749987	0.750017	167
$\alpha = 0.05$	5	768	0.749380	0.750222	159
$\alpha = 0.05$	4	774	0.749511	0.750030	159
$\alpha = 0.05$	3	1589	0.749724	0.750305	161
$\alpha = 0.05$	2	1648	0.749652	0.750012	181
narrow $(0.03, 0.07)$	5	2277	0.749953	0.750292	162
narrow $(0.03, 0.07)$	4	2311	0.749943	0.750219	162
narrow $(0.03, 0.07)$	3	11737	0.749978	0.750005	170
narrow $(0.03, 0.07)$	2	12002	0.749999	0.750061	174
wide $(0.02, 0.08)$	5	1078	0.749805	0.750245	158
wide $(0.02, 0.08)$	4	1090	0.749956	0.750276	156
wide $(0.02, 0.08)$	3	5832	0.749987	0.750001	157
wide $(0.02, 0.08)$	2	6021	0.749875	0.750075	164

A.5 Shape and use of the randomised optima

Figure 2 plots the released primary-contract rules rather than only their objective values. The pairwise and three-wise rules have positive acceptance probability only when $t/N \in (L, U)$. The four-wise and five-wise rules each have 21 positive entries outside that empirical-rate interval. Their largest such probabilities are 0.00387 and 0.00350, and their contributions to nominal retention are only 5.31×10^{-8} and 4.07×10^{-8}. These entries are nevertheless included in the complete-null certificates and are not justified by boundary calculations alone.

The randomised rules establish the information-theoretic limits in Theorem 2. They are not the default operational recommendation. If a randomised rule is used, its auxiliary uniform draw must be independent of the evaluation outcomes, generated once after the data are locked, and recorded with the audit. Otherwise the deterministic gate of Proposition 5 should be used.

A.6 Independent-data baselines

The first deterministic exact TOST interval for the primary contract appears at $N = 943$, with gate $36 \le T \le 52$, boundary errors 0.009443 and 0.009769, and retention 0.753902. Requiring the 99% two-sided interval of Clopper & Pearson (1934) to lie inside $(0.025, 0.075)$ first reaches the retention target at $N = 1{,}111$. At $N = 943$, this interval rule accepts the counts $\{38, \dots, 50\}$ and has retention 0.626679. For either gate and $r \in \{2, 4\}$, Table 2 is obtained by solving $\max_{w \in \Pi_r(p)} \sum_{t=c_1}^{c_2} w_t$ at $p \in \{L, U\}$. Every reported value has a feasible moment matching law and a rationally repaired dual polynomial checked at all 944 integer counts. For TOST, the fourth-order lower-boundary witness has support $\{17, 18, 25, 26, 36\}$ and probability 0.0467656, and the upper-boundary witness has support $\{52, 67, 68, 82, 83\}$ and probability 0.0587492. The pairwise values are 0.1287679 and 0.1569590. For the interval rule the corresponding maxima are 0.0405799 at order four and 0.1320131 at order two. The curves in Figure 3 solve the same program on a grid of null rates, and for both gates and both orders the worst case over the grid is attained at L or U. A cluster bootstrap is not assigned an exact r-wise size, because its rejection probability depends on the empirical partition, the resampling scheme and the full within cluster law, none of which is determined by global moment constraints.

One-sided moment inequalities. For a lower acceptance cutoff c_1, Cantelli's inequality gives

$$\mathbb{P}_L(T \ge c_1) \le \frac{NL(1-L)}{NL(1-L) + (c_1 - NL)^2},$$

and the analogous bound at U controls $\mathbb{P}_U(T \le c_2)$. Under four-wise independence, the fourth central moment equals

$$\mu_4(q) = 3\{Nq(1-q)\}^2 + Nq(1-q)\{1 - 6q(1-q)\},$$

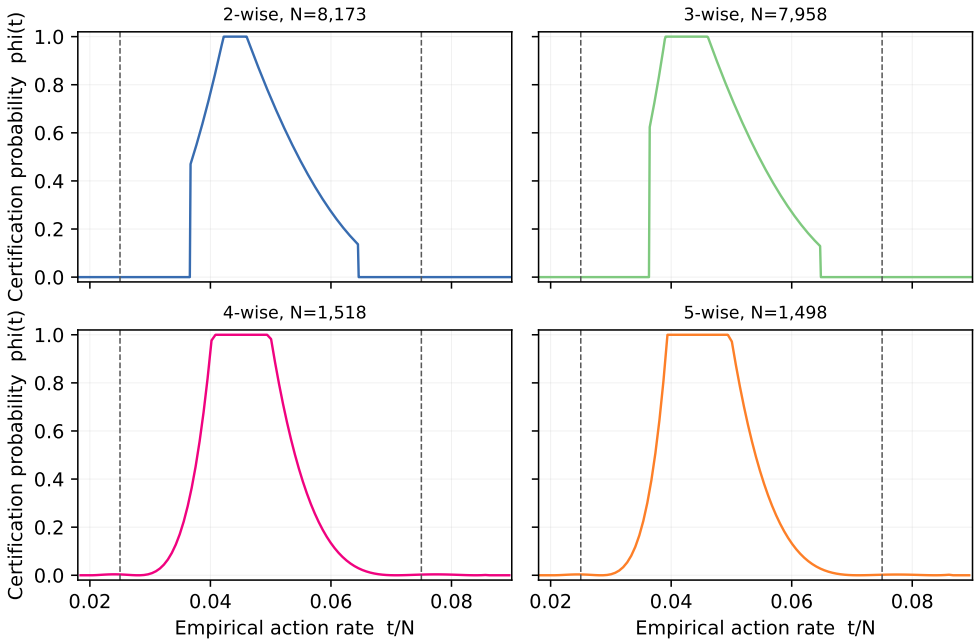


Figure 2: Released randomised rules $\phi(t)$ at the exact minimum sample size for the primary contract. Dashed lines mark $L = 0.025$ and $U = 0.075$. The figure is generated directly from the exact rational upper certificates.

so Markov's inequality gives the corresponding one-sided bound $\mu_4(q)/a^4$. The coupling in Proposition 5 extends each boundary calculation to its complete null half-line. Exhaustive integer scans give the first Cantelli gate at $N = 12{,}224$, with counts 478 to 627, retention 0.751693, and exact boundary bounds $7449/750493$ and $7067/706934$. At $N = 12{,}223$ its retention is 0.739156. The first fourth-moment Markov gate occurs at $N = 2{,}443$, with counts 94 to 129, retention 0.751893, and exact boundary bounds $9121089523/1002817363707$ and $73619221123/7377631400907$. At $N = 2{,}442$ its retention is 0.723846.

Table 5: One-sided moment gates for the primary contract, each calibrated separately at L and U. "Reduction" compares the exact deterministic gate with the corresponding elementary gate.

Assumption	Bound	Bound N	Exact gate	Reduction	Exact randomised
pairwise	Cantelli	12,224	12,223	$< 0.01\%$	8,173
four-wise	fourth-moment Markov	2,443	1,996	18.3%	1,518

B CLUSTERED DESIGNS

Proof of Proposition 2. (a) Let $P \in \mathcal{C}_r(p)$ and let S be a set of at most r units. Its intersection with each cluster has at most r units, which are independent $\mathrm{Bern}(p)$ by assumption, and the clusters are independent, so the units of S are independent $\mathrm{Bern}(p)$. Hence $P \in \mathcal{L}_r(p)$.

(b) *Sufficiency.* Fix one unit u_g in each cluster before the data are seen, for example the first recorded window or the first generation. Under any $P \in \mathcal{C}(p)$ the outcomes $Y_{u_1}, \dots, Y_{u_G}$ are independent, because they lie in different clusters, and each is $\mathrm{Bern}(p)$. Applying a rule that is valid for G independent outcomes therefore gives validity over the null, and its retention under any $P \in \mathcal{C}(p_0)$ equals its retention under $\mathrm{Bern}(p_0)^{\otimes G}$. If $G \geq N^{\star}_{\infty}$, Theorem 1(c) applied to full independence provides such a rule with retention at least ρ.

Necessity. Let ψ be any procedure, randomised or not and using the cluster labels in any way, that is valid over the null under $\mathcal{C}$ and has retention at least ρ under every law in $\mathcal{C}(p_0)$. For $q \in [0, 1]$ let

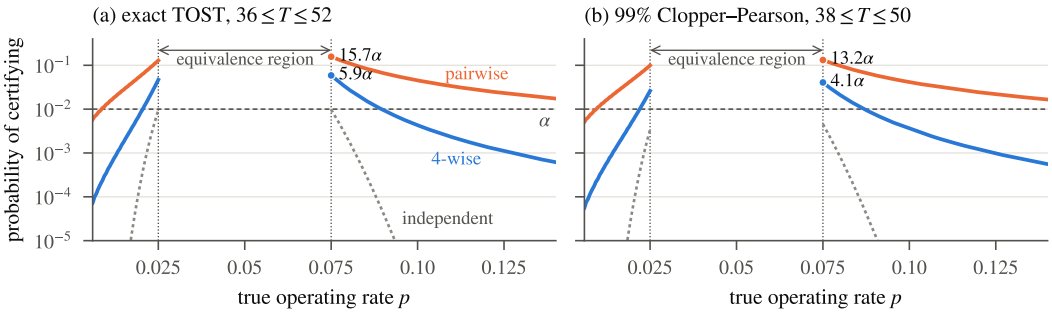


Figure 3: Probability of certifying outside the equivalence region for the two gates of Table 2 at $N = 943$, under independence (dotted) and in the worst case under four-wise and pairwise independence. Any value above α disproves validity, and for these gates the worst case is at the boundary.

P_q^{copy} draw $B_1, \dots, B_G$ independent $\mathrm{Bern}(q)$ and set every unit of cluster g equal to B_g. This law belongs to $\mathcal{C}(q)$. Define $\tau(b) = \mathbb{E}[\psi(Y) \mid B = b]$, a test on G bits. Validity of ψ under P_q^{copy} for $q \in \Theta_0$ says that τ is valid for G independent $\mathrm{Bern}(q)$ outcomes, and retention under $P_{p_0}^{\mathrm{copy}}$ says that τ has retention at least ρ. By Lemma 1 with $r = G$, τ can be replaced by a count rule with the same properties, so $G \geq N_\infty^\star$. For the primary contract, exact TOST on the prespecified units is a deterministic procedure at $G = 943$. □

The construction discards all but one unit per cluster. This is unavoidable in the worst case over within cluster laws, since the copying law makes the other units uninformative. When the evaluator is willing to assume r-wise independence within clusters, part (a) applies and every unit counts.

Proof of Proposition 3. Write $T = \sum_{g,i} Y_{gi}$. Pairs in different clusters are independent and have joint mean p^2. The definition of $\bar{\gamma}$ gives

$$\mathbb{E}_P(T)_2 = \sum_{(g,i)\neq(h,j)} \mathbb{E}_P(Y_{gi}Y_{hj}) = (N)_2 p^2 + Gm(m-1)\bar{\gamma}(P,p)p(1-p),$$

which equals $(N)_2 p^2 + N(m-1)\bar{\gamma}(P,p)p(1-p)$. The stated moment interval follows from $0 \leq \bar{\gamma} \leq \kappa$. Every count law induced by $\mathcal{C}_\kappa(p)$ is therefore feasible for the relaxed moment program, whose optimum bounds the acceptance probability over the cluster class.

For the sharper bound, the cluster sums S_g are independent, $0 \leq S_g \leq m$, and

$$\mathrm{Var}_P(T) = Np(1-p)\{1 + (m-1)\bar{\gamma}(P,p)\} \leq V_p(\kappa).$$

At $p = L$, the independent mean-zero variables $S_g - mL$ are bounded above by $m(1-L)$, and one-sided Bernstein applied to their sum gives the first bound. At $p = U$, apply the same inequality to $mU - S_g$, which is bounded above by mU. For any law with $\mathrm{Bern}(p)$ marginals, $\mathrm{Var}(S_g) \leq (\sum_i \mathrm{sd}\, Y_{gi})^2 = m^2 p(1-p)$, which equals the per-cluster variance bound at $\kappa = 1$, so at $\kappa = 1$ the Bernstein bounds hold for every law in $\mathcal{C}(p)$, including those with negative correlation.

It remains to extend the boundary bounds to the complete null. If $q \leq L$, independently set each zero to one with probability $(L-q)/(1-q)$. The resulting vector has marginal L, its count is no smaller, cluster independence is preserved, and its intracluster correlations are multiplied by $q(1-L)/\{L(1-q)\} \leq 1$. If $q \geq U$, retain each success independently with probability U/q. The count is no larger and correlations are multiplied by $U(1-q)/\{q(1-U)\} \leq 1$. Both couplings map $\mathcal{C}_\kappa(q)$ into $\mathcal{C}_\kappa(L)$ or $\mathcal{C}_\kappa(U)$, and $\mathcal{C}(q)$ into $\mathcal{C}(L)$ or $\mathcal{C}(U)$. Since an interval gate is contained in the corresponding half-line on each side, the two boundary certificates prove validity over the complete null. □

C Scaling in the error level

Proof of Proposition 4. By Lemma 1 it suffices to consider count rules. Let $\delta = \min(p_0 - L, U - p_0)$.

Upper bound. Let $\phi(T) = \mathbf{1}\{|T - Np_0| < N\delta/2\}$. For $q \in \Theta_0$ we have $|q - p_0| \geq \delta$, so certification implies $|T - Nq| > N\delta/2$. For $P \in \mathcal{L}_{2k}(q)$, $(T - Nq)^{2k}$ is a polynomial of degree $2k$ in T, so its expectation equals the central moment $\mu_{2k}(N, q)$ of $\mathrm{Bin}(N, q)$, and $\mu_{2k}(N, q) \leq C_k N^k$ for all $q \in [0, 1]$ by the inequality of Marcinkiewicz and Zygmund. Markov's inequality gives $\mathbb{E}_P \phi(T) \leq C_k 4^k \delta^{-2k} N^{-k} \leq \alpha$ once $N \geq (C_k 4^k/\alpha)^{1/k} \delta^{-2}$, and Chebyshev's inequality gives retention at least $1 - 4p_0(1 - p_0)/(N\delta^2) \geq \rho$ once $N \geq 4p_0(1 - p_0)/((1 - \rho)\delta^2)$. Hence $N^\star_{2k} = O(\alpha^{-1/k})$. Under independence Hoeffding's inequality replaces Markov's and gives $N^\star_\infty = O(\log(1/\alpha))$. Conversely, for any rule valid at L under independence, $\alpha \geq \mathbb{E}_L \phi \geq \min_t\{b_{N,L}(t)/b_{N,p_0}(t)\}\, \mathbb{E}_{p_0} \phi \geq \rho\, m^N$ with $m = \min\{L/p_0, (1 - L)/(1 - p_0)\} \in (0, 1)$, so $N^\star_\infty \geq \log(\rho/\alpha)/\log(1/m)$. This is where $L > 0$ is needed, since with $L = 0$ and $U = 1$ the null contains only degenerate laws.

Lower bound. Let ϕ be valid at order $2k$ with retention at least ρ at sample size N, and write b and g for the $\mathrm{Bin}(N, L)$ and $\mathrm{Bin}(N, p_0)$ laws, $\sigma^2 = NL(1 - L)$ and $h = \lfloor \sigma/(2k + 1) \rfloor$. By Theorem 1(c) we may assume $N \geq N_0$, where N_0 is large enough for the steps below. For each shift $j = 0, \ldots, h - 1$ let $X_j = \{x_0 + j + ih : i = 0, \ldots, 2k\}$ with $x_0 = \lceil NL - \sigma \rceil$, so every X_j lies in the window $W = \{x : |x - NL| \leq \sigma\}$ and every $x \in W$ belongs to at most $2k + 1$ of them. Let $\nu^{(j)}$ be the signed measure on X_j whose factorial moments of order at most $2k$ equal those of g. Its weights are $\nu^{(j)}_x = \mathbb{E}_g \ell_{j,x}(T)$, where $\ell_{j,x}$ are the Lagrange basis polynomials of degree $2k$ on X_j. Since $|\ell_{j,x}(t)| \leq c_k((|t - NL| + \sigma)/h)^{2k}$ and $\mathbb{E}_g |T - NL|^{2k} \leq CN^{2k}$, we get $|\nu^{(j)}_x| \leq C' N^k$. The average $\nu = h^{-1} \sum_j \nu^{(j)}$ has the same moments and satisfies $|\nu_x| \leq (2k+1) C' N^k/h \leq C'' N^{k-1/2}$ on W. By the local limit theorem $b(x) \geq c/\sqrt{N}$ on W, so $|\nu_x| \leq K b(x) N^k$ for a constant K. With $\varepsilon = 1/(2KN^k)$, the measure $w = b + \varepsilon(g - \nu)$ is nonnegative, has the factorial moments of b up to order $2k$, and so belongs to $\Pi_{2k}(L)$. Validity at w and at b gives

$$\alpha \;\geq\; \sum_t \phi(t) w_t \;\geq\; \varepsilon \sum_t \phi(t) g_t - \varepsilon \sum_{x \in W} \phi(x) |\nu_x| \;\geq\; \varepsilon\rho - \varepsilon K N^k \sum_x \phi(x) b(x) \;\geq\; \frac{\rho}{2KN^k} - \frac{\alpha}{2},$$

so $N^k \geq \rho/(3K\alpha)$. For α small this exceeds N_0^k, which justifies the assumption $N \geq N_0$ and gives $N^\star_{2k} = \Omega(\alpha^{-1/k})$. The ordering of odd orders follows from $\mathcal{L}_{2k+2} \subseteq \mathcal{L}_{2k+1} \subseteq \mathcal{L}_{2k}$. □

Table 6 reports the sample sizes of the boundary program over a grid of settings. They are lower bounds on $N^\star_r$ computed in floating point, and they coincide with the certified values in all sixteen cells of Table 4. The observed jump from order four to order three is about 1.4 at $\alpha = 0.1$, 2 at 0.05, 5 at 0.01 and 8 at 0.005, and the same ordering holds for $p_0 = 0.2$, an asymmetric region, and $\rho \in \{0.5, 0.9\}$.

Table 6: Boundary-program sample sizes (floating point) for $\alpha = 0.01$, retention $\rho = 0.75$, $p_0 = 0.05$ and region $(0.025, 0.075)$ unless stated. Proposition 4 gives exponent $1/k$ at order $2k$ as $\alpha \to 0$. On this finite grid the order-6 local estimates range from 0.26 to 0.35 and therefore straddle $1/3$. Entries marked n/c were not computed.

Setting	independent	6-wise	5-wise	4-wise	3-wise	pairwise
$\alpha = 0.001$	1,422	2,131	4,342	4,392	n/c	n/c
$\alpha = 0.002$	1,273	1,755	3,122	3,156	n/c	n/c
$\alpha = 0.005$	1,080	1,371	2,040	2,066	16,034	16,471
$\alpha = 0.01$	936	1,144	1,498	1,518	7,958	8,173
$\alpha = 0.02$	788	954	1,116	1,129	3,953	4,061
$\alpha = 0.05$	594	735	768	774	1,589	1,648
$\alpha = 0.1$	445	575	579	579	828	865
$p_0 = 0.2$, region $(0.15, 0.25)$	772	931	1,166	1,196	5,293	5,770
region $(0.03, 0.08)$	968	1,163	1,411	1,468	6,076	6,862
$\rho = 0.5$	672	826	1,066	1,096	5,220	5,382
$\rho = 0.9$	1,261	1,549	1,997	2,011	9,978	10,222
narrow $(0.03, 0.07)$	1,449	1,764	2,277	2,311	11,737	12,002
wide $(0.02, 0.08)$	658	806	1,078	1,090	5,832	6,021

D ODD ORDERS AND THE RANGE OF THE COUNT

Proof of Theorem 3. Throughout, rules vanish outside $\{0,\dots,N\}$, and $\Pi_r^{\mathbb{Z}}(p)$ contains the finitely supported laws w on $\mathbb{Z}$ with $\sum_x w_x(x)_s = (N)_s p^s$ for $0 \le s \le r$.

(a) Since $\Pi_{2k+1}^{\mathbb{Z}}(p) \subseteq \Pi_{2k}^{\mathbb{Z}}(p)$, only $\ge$ needs proof. Write S_r for the supremum over $\Pi_r^{\mathbb{Z}}(p)$. Fix ϕ, $p \in (0,1)$, $\eta > 0$ and $w \in \Pi_{2k}^{\mathbb{Z}}(p)$ with $\sum \phi w \ge S_{2k} - \eta$. For $\theta \in (0,1)$ let $w_\theta = (1-\theta)w + \theta\, b$ with $b = \mathrm{Bin}(N,p)$. Then $w_\theta \in \Pi_{2k}^{\mathbb{Z}}(p)$, $\sum \phi w_\theta \ge (1-\theta)(S_{2k} - \eta)$, and $w_\theta(x) \ge \theta\, b(x) > 0$ for $x \in \{0,\dots,2k\}$ (assume $N \ge 2k$). Let $D = \sum_x w_\theta(x)(x)_{2k+1} - (N)_{2k+1}p^{2k+1}$. If $D = 0$ we are done. Otherwise pick an integer z with $(z)_{2k+1}$ of sign opposite to D (a large positive z if $D < 0$, or a large negative z if $D > 0$, since $(z)_{2k+1}$ is then a product of $2k+1$ negative factors), put $\varepsilon = -D/(z)_{2k+1} > 0$, and let ν be the signed measure on $\{0,\dots,2k\}$ with $\sum_x \nu_x(x)_s = \varepsilon(z)_s$ for $0 \le s \le 2k$. This system is triangular with diagonal entries $s!$, because $(x)_s = 0$ for $0 \le x < s$, so ν exists, is unique, and satisfies $\max_x |\nu_x| \le C_k\, \varepsilon \max_{s \le 2k} |(z)_s| = O(|D|/|z|)$. The measure $w' = w_\theta - \nu + \varepsilon\delta_z$ has the factorial moments of w_θ up to order $2k$, and, because $(x)_{2k+1} = 0$ for $0 \le x \le 2k$, its moment of order $2k+1$ is $(N)_{2k+1}p^{2k+1} + D + \varepsilon(z)_{2k+1} = (N)_{2k+1}p^{2k+1}$. The constraint for $s = 0$ gives $\sum_x \nu_x = \varepsilon$, so w' has total mass one. For $|z|$ large, $|\nu_x| < \theta\, b(x)$, so $w' \ge 0$ and $w' \in \Pi_{2k+1}^{\mathbb{Z}}(p)$. Since $\phi(z) = 0$ and $0 \le \phi \le 1$, $\left|\sum \phi w' - \sum \phi w_\theta\right| \le \sum_x |\nu_x| = O(|D|/|z|)$. Letting $|z| \to \infty$ and then $\theta, \eta \to 0$ gives $S_{2k+1} \ge S_{2k}$.

The remaining cases are degenerate. At $p = 0$, $(x)_2 \ge 0$ on $\mathbb{Z}$ with zeros only at $\{0,1\}$, so the first two constraints force $\Pi_{2k}^{\mathbb{Z}}(0) = \{\delta_0\}$. At $p = 1$ the first two moments give variance zero, so $\Pi_{2k}^{\mathbb{Z}}(1) = \{\delta_N\}$. Both laws satisfy every higher constraint. If $N < 2k$, then $(x)_{2k} \ge 0$ on $\mathbb{Z}$ with zeros exactly at $\{0,\dots,2k-1\}$ and $(N)_{2k} = 0$, so every $w \in \Pi_{2k}^{\mathbb{Z}}(p)$ lives on $\{0,\dots,2k-1\}$, where the order-$(2k+1)$ constraint holds automatically. Validity depends on the class only through these suprema, so the two classes admit the same valid rules and $N_{2k+1}^{\mathbb{Z}} = N_{2k}^{\mathbb{Z}}$.

(b) $\mathcal{L}_{2k+1}(p) \subseteq \mathcal{L}_{2k}(p)$ gives $N_{2k+1}^{\star} \le N_{2k}^{\star}$. Every law on $\{0,\dots,N\}$ is a law on $\mathbb{Z}$, so $\Pi_{2k}(p) \subseteq \Pi_{2k}^{\mathbb{Z}}(p)$, and a rule valid against the larger class is valid against the smaller, so $N_{2k}^{\star} \le N_{2k}^{\mathbb{Z}}$. The last equality is (a). □

Relation to earlier work. The perturbation is the classical device of moving an odd moment with a vanishing far-away atom. For the probability that N bits which are k-wise independent with marginal p are all one, Benjamini et al. (2012) show that the maximum for odd k equals p times the maximum for $N-1$ bits and order $k-1$ (see also Peled et al., 2011). That reduction concerns one event at the edge of the range, where the range constraint binds by definition. Theorem 3 concerns every count rule and isolates the range constraint as the only channel through which odd orders act. The computational side below shows that this relaxation is tight for the even orders in every contract we certified.

Computational side. All eight even order certificates of Table 4 dominate their rule at every integer, not only on $\{0,\dots,N\}$ (an option of the verifier checks this exactly), so $N_{2k}^{\mathbb{Z}} = N_{2k}^{\star}$ in every contract and every odd order saving in Figure 1 is exactly the value of $0 \le T \le N$. Every odd order certificate uses that constraint. Table 7 shows the saving shrinking as the constraint is relaxed. It reports the optimal retention of the boundary program when count laws may use the integers in $[-M, N+M]$. The even order value does not change with M at these N. At $N = 1{,}498$ the gap falls from 0.0093 ($M = 0$) to 0.0023 ($M = 1{,}500$), and at $N = 7{,}958$, where $M \ll N$, only from 0.0183 to 0.0153. Both ends of the range contribute. At $N = 1{,}498$ and $M = 1{,}500$, relaxing only the lower end gives 0.745706 and only the upper end 0.748034.

Table 7: Optimal retention of the boundary program for the primary contract when count laws may use the integers in $[-M, N+M]$. The last column is the even order, which does not depend on M.

	$M=0$	20	100	400	1500	even order
$N = 1498$, $r = 5$ vs. 4	0.7503	0.7489	0.7472	0.7452	0.7434	0.7411
$N = 7958$, $r = 3$ vs. 2	0.7500	0.7500	0.7498	0.7491	0.7469	0.7317

E BOUNDARY CALIBRATION

Under independence,

$$\frac{d}{dp}\mathbb{P}_p(c_1 \le T \le c_2) = N\{b_{N-1,p}(c_1-1) - b_{N-1,p}(c_2)\}.$$

The ratio of the two binomial terms is a positive constant times $\{p/(1-p)\}^{c_2-c_1+1}$, so the derivative changes sign once. This proves unimodality for deterministic interval rules such as exact TOST, and hence the usual boundary reduction for them.

Where boundary feasibility breaks. Let ϕ be feasible at the boundaries with dual polynomials π_L, π_U. Any rule with $\phi \le \phi' \le \min(1, \pi_L, \pi_U)$ is dominated by the same polynomials and therefore remains feasible at the boundaries. Since every $b_{N,p_0}(t)$ is strictly positive, any strict increase in ϕ strictly increases nominal retention, so distinct rules in this range cannot both be exact maximisers. In our numerical solutions the increases occur on counts with extremely small nominal probability, and interior null laws can place substantial mass on those counts. Table 8 shows two such rules. Their exact nominal retentions are 0.7503560510, differing by less than 4×10^{-14} and within 8×10^{-7} of the numerical optimum, yet explicit rational five-atom laws at interior null rates make them certify far more often than α. This is a finite-precision failure mode for boundary optimisation, not a claim that the exact optimiser has slack. Interior maxima of size functions are familiar from exact unconditional tests with nuisance parameters (Berger & Boos, 1994), and here the nuisance is the dependence structure.

Table 8: Exact interior null counterexamples at $N = 1{,}518$ under four-wise independence. The first two rules are feasible at the boundaries and numerically nearly optimal outputs of the boundary program. The last row is certified over the complete null.

Rule	Null side	Marginal rate p	$\mathbb{P}(\text{CERTIFY})$	Multiple of α
LP-vertex boundary rule	Lower	0.01653	0.0451	4.5
	Upper	0.09770	0.1082	10.8
Maximal boundary rule	Lower	0.01375	0.2208	22.1
	Upper	0.10125	0.6944	69.4
Certified complete null rule	Both	every $p \in \Theta_0$	≤ 0.0100	≤ 1

Shapes of the rules in Table 8. The LP-vertex rule certifies with positive probability at 43 counts outside $(LN, UN) = (37.95, 113.85)$, from $T = 23$ (probability 0.065) to $T = 145$ (probability 0.161). Its lower witness places 0.686 of its mass on $T = 23$ and its upper witness 0.673 on $T = 145$. The maximal rule certifies with probability essentially one (0.999999, after the scaling by $1 - 10^{-6}$ used for certification) for $T \le 11$ and near $T = N$, and with positive probability at $1{,}439$ counts outside the region. Neither is a rule anyone should deploy. They show why numerical solutions of a boundary program still require a complete null certificate.

Rules that certify only inside the region. We added the constraint $\phi(t) = 0$ unless $L < t/N < U$ to the boundary program. At $N = 1{,}518$ and $r = 4$ its numerical optimum agrees with the unrestricted boundary program to the reported precision (0.7503567). To search for interior failures within this class, for each of seventeen interior rates ($0.16L$ to $0.99L$ and $1.01U$ to $2.7U$) we alternated between computing the worst case law at that rate and re-optimising the rule to maximise its probability under that law, subject to boundary validity, the support restriction and retention at least 0.75. In six cells (the primary contract at orders four and five, $\alpha = 0.05$ at orders four and five, and the narrow and wide regions at order four), the largest error found was 0.95α, always at the rate closest to a boundary. This supports, but does not prove, a boundary reduction for rules that certify only inside the region.

All deterministic rules. To test minimality beyond half-line gates, we maximised nominal retention over every 0 or 1 rule on $\{0, \dots, N\}$ subject to validity at L and U, a necessary condition for complete-null validity. Validity at each boundary is written through the polynomial dual of Proposition 1, which gives a mixed-integer linear program, solved with HiGHS. At order four, for every N from $1{,}518$ to $1{,}995$, the optimum stayed below 0.75, reaching at most 0.7453 at $N = 1{,}980$, so the gate at $1{,}996$ is the smallest deterministic design. At order five the scan covered $N = 1{,}498$ to $1{,}778$, with optima at most 0.659. At orders two and three we solved the program at nine sample sizes between $N_r^\star$ and the reported gate. Every optimum was an interval, and retention rose steadily towards the target. At

$N = 11{,}928$ the order-three solver returned an interval with retention 0.7517 that exceeds α at U by 7×10^{-7} in exact arithmetic, a tolerance artefact, so candidates near the threshold are rechecked exactly.

One-sided repair. Take $q \le L$ and $Y \in \mathcal{L}_r(q)$. Let V_i be independent $\mathrm{Bern}((L-q)/(1-q))$ variables independent of Y and set $Z_i = \max(Y_i, V_i)$. Then $Z \in \mathcal{L}_r(L)$ and $\sum_i Z_i \ge \sum_i Y_i$. Hence $\mathbb{P}_q(c_1 \le T \le c_2) \le W_r(\mathbf{1}\{T \ge c_1\}, L)$. For $q \ge U$, independently retain each success with probability U/q. The thinned vector belongs to $\mathcal{L}_r(U)$ and has no larger count, giving the analogous upper-null inequality. These couplings control half-lines, not interval probabilities, so they do not contradict the interior null counterexamples. The deterministic constructions of Table 1 scan every N from $N_r^\star$ for $r \in \{4, 5\}$. For $r \in \{2, 3\}$ they bisect on N and then scan only the 150 values below the result. Thus global minimality within the half-line family is established for independence and orders four and five, but not for orders two and three. At the reported N and at $N - 1$, the best interval valid at both boundaries (searched over all intervals for $r \in \{4, 5\}$, and over intervals whose lower end is within 12 counts of the gate for $r \in \{2, 3\}$) was the half-line gate itself, so the relaxation cost nothing there. For orders four and five in the primary, $\alpha = 0.05$ and wide contracts we also searched, at every N from $N_r^\star$ to the reported value minus one, all intervals valid at both boundaries whose lower end lies within 12 counts of the gate. None reached retention 0.75, the largest being 0.74998 for the primary contract at order five and $N = 1{,}977$. The endpoints of all sixteen gates are certified in exact arithmetic, and their nominal retentions are computed exactly.

F Application Details

F.1 PTB-XL

The frozen split uses PTB-XL folds 1 to 8 for model fitting, fold 9 for the threshold, and fold 10 for evaluation. A record was eligible if it carried at least one diagnostic SCP statement, and for each patient the eligible record with the lowest `ecg_id` was retained. This removed 27 of the 1,904 patients in fold 10. Each record contained 12 leads sampled at 100 Hz and was truncated to ten equal contiguous windows. Each window produced 96 features. For each lead, these were the mean, standard deviation, root mean square, minimum, maximum, mean absolute first difference, and the 0.25 and 0.75 quantiles. A standardised logistic regression with class weighting used $C = 1$, the L-BFGS solver, at most 1,000 iterations and seed 20260922. The label was the presence of a diagnostic code other than `NORM`. With predicted abnormality probability s, uncertainty was $1 - |2s - 1|$. The sample sizes were 14,823 training patients, 1,917 threshold patients, and 1,877 test patients. The empirical cutoff from fold 9 was 0.9403569. It targeted a 5% referral rate and yielded 0.049974 over its 19,170 windows. The exact TOST gate for 1,877 independent outcomes is $\{64, \ldots, 114\}$, and position 7 lies on its edge, with $\mathbb{P}_U(T \le 114) = 0.0091$. Table 9 reports the robustness order of every position. Every decision is certified. At the orders reported below the robustness order an exact law in Π_r shows that the p-value exceeds α, and at the robustness order a repaired dual polynomial shows that it does not. Where only one side is certified we report a bound. Abstention at the few intermediate orders where our numerical bounds did not close (position 7 at orders 7, 11 and above 12) follows from nesting, since abstention at order r implies abstention at every lower order.

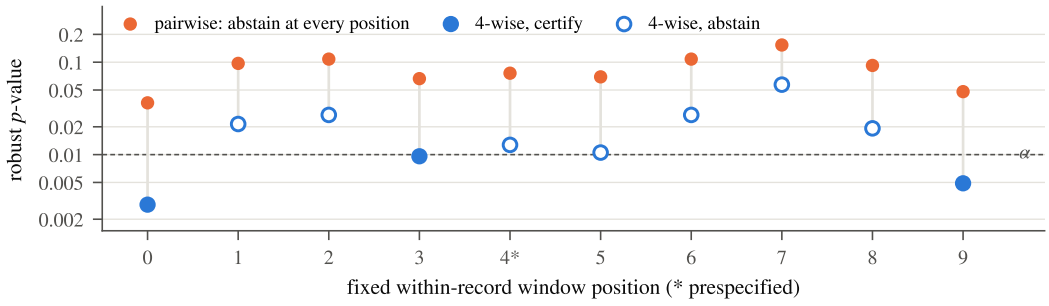


Figure 4: PTB-XL robust p-values across the ten fixed within-record positions under four-wise independence (blue, filled if it certifies and open if it abstains) and under pairwise independence (orange).

Table 9: PTB-XL positions. "Independent" is the larger exact TOST p-value. The last two columns give the larger robust p-value at the order just below the robustness order $r^\dagger$ and at $r^\dagger$. The ten positions reuse the same patients and are not independent evaluations. Position 4 was prespecified.

Position	Referrals x	Independent	$r^\dagger$	p-value at $r^\dagger - 1$	p-value at $r^\dagger$
0	82	0.000002	4	0.0352	0.0029
1	106	0.000917	6	0.0212	0.0082
2	108	0.001718	8	≥ 0.0108	0.0082
3	98	0.000051	4	0.0649	0.0096
4	101	0.000161	6	0.0125	0.0036
5	99	0.000075	6	0.0103	0.0027
6	108	0.001718	8	≥ 0.0108	0.0082
7	114	0.009105	> 12	≥ 0.0288 (at $r = 12$)	n/a
8	105	0.000660	6	0.0190	0.0069
9	90	0.000001	4	0.0468	0.0049

F.2 QWEN2.5-32B

The GSM8K test split was shuffled with seed 20260920. Shuffled positions 0 through 19 were used for the pilot and excluded. Positions 20 through 199 formed the locked 180-prompt set. Each prompt appended "Solve the problem. End with exactly: `FINAL: <number>`". Qwen2.5-32B-Instruct was sampled 50 times per prompt with temperature 0.7, at most 128 new tokens and a deterministic SHA-256-derived seed for each prompt and replicate combination. No top-p or top-k override was supplied. The numeric scorer took the last number after `FINAL:`, or the last number on the final line if the marker was absent, removed commas, parsed decimals or fractions exactly, and compared it with the number following #### in GSM8K. Missing or unparsable answers were failures. The public model repository was recorded but its exact commit was not pinned, so this is a locked empirical audit rather than an exactly reproducible benchmark.

The file contains 180 prompts and 50 generations per prompt, with no duplicate rows. Of 9,000 generations, 436 were marked failures. The prompt failure-count histogram included 151 prompts with zero failures and one prompt with 50 failures. The estimated pairwise correlation within prompt was 0.67987, the usual design effect 34.31, and $9{,}000/34.31 = 262.29$. At the observation level the exact gates are $\{261, \ldots, 617\}$ under independence, $\{281, \ldots, 581\}$ under four-wise independence, and $\{373, \ldots, 426\}$ under pairwise independence, and the robust p-values are 0.0108 at order two, 0.0105 at order three and 0.0002 at order four. These are mathematical statements about 9,000 binary outcomes and are not licensed by the design. At the prompt level, with one prespecified generation per prompt, the exact TOST gate for 180 outcomes is empty ($c_1 = 11 > c_2 = 5$) and the optimal randomised rule has retention 0.0337. For the one-sided contract that certifies $p < 0.075$ at level 0.01, the best randomised rule with 180 prompts has retention 0.144, and retention 0.75 needs 865 prompts.

F.3 NASA TELEMETRY

Protocol. For every SMAP and MSL channel of the public release of Hundman et al. (2018), the first telemetry dimension of the nominal training series is standardised, and a ridge regularised AR(8) model (penalty 10^{-4}) is fitted to its first 40%. The innovation score of a time step is the absolute single step prediction error divided by the residual standard deviation on the fitting segment, and the rolling maximum score is its trailing maximum over eight steps. The alarm threshold of each score is the empirical 95% quantile of its values on the segment from 40% to 70%, pooled over the channels of a spacecraft. Audited windows are drawn from the final 30% segment, and a window fails when its score exceeds the threshold. The fixed designs of Table 3 take 30 approximately equally spaced windows from each channel's chronologically ordered audit segment of n windows, at positions $\mathrm{round}(j(n-1)/29)$ for $j = 0, \ldots, 29$. The repeated campaign draws, for $m \in \{30, 40, 50\}$ and ten seeds, one window uniformly from each of m contiguous temporal bins of every channel's audit segment, and both scores are evaluated on the same windows. This gives 120 designs, of which 93 certify under independence. For the fixed designs, the $N = 660$ MSL independence gate has nominal retention only 0.4256, so the MSL innovation count triggers a valid gate that fails the planning target.

Exact design-level analysis. For G channels of m windows, whether a count certifies under independence or at order four, and whether a cluster law overturns it, depends only on (G, m) and the count, so Table 10 reports these sets for every design. The attacks take the channel count K to follow a law w on at most three points with $\mathbb{E}K = mq$ and $\mathbb{E}K(K-1) = m(m-1)q^2$, $q \in \{L, U\}$, place the failures of a channel uniformly at random given K, and make channels independent. Every pair of windows is then independent with $\mathrm{Bern}(q)$ marginals, so the law belongs to $\mathcal{C}_2(q)$ of Proposition 2, to $\mathcal{L}_2(q)$, and, having average intracluster correlation zero, to $\mathcal{C}_\kappa(q)$ for every $\kappa \geq 0$. We enumerate every vertex of this moment polytope and compute the exact tail of the G-fold convolution at each count of the independence gate. The lower boundary witnesses place mass on $\{a, a+1, k\}$ with $a = \lceil mL \rceil - 1$, so almost every channel raises a or $a+1$ alarms, and a rare channel raises a burst of k between 14 and 50. The fixed designs are overturned in this way. The counts 31 (MSL innovation), 54 and 53 (SMAP) have p-values 0.034, 0.028 and 0.031 under such laws, although the two SMAP counts certify at order four. A count with no overturning vertex is recorded as unresolved. The tail is a polynomial of degree G in w, so an interior law might still overturn it.

Table 10: NASA designs at the primary contract. Gates and overturned sets are exact and verified from `certs/nasa_designs.json`. The last four columns count the 20 designs per row (ten seeds, two scores) by where their observed counts fell.

Design $G \times m$	N	Indep. gate	Order-4 gate	Overturned in $\mathcal{C}_2$	Cert.	Overturned	Order 4	Both
MSL 22×30	660	27 to 33	empty	27 to 33	7	7	0	0
MSL 22×40	880	34 to 47	empty	34 to 47	16	16	0	0
MSL 22×50	1100	41 to 62	48 to 50	41 to 62	17	17	4	4
SMAP 42×30	1260	46 to 72	53 to 59	46 to 61, 72	18	16	7	7
SMAP 42×40	1680	59 to 101	67 to 85	59 to 80, 100, 101	17	12	11	9
SMAP 42×50	2100	71 to 129	80 to 112	71 to 100, 128, 129	18	13	12	7
Total					93	81	34	27

The observed count of each of the 120 designs is released with the campaign script, which regenerates it from the public telemetry. For every design, the supplementary checker confirms that its independence decision, its status and its attack p-value follow from the exact sets, and it recomputes the four count columns of Table 10 from them.

F.4 Positive dependence and clustered practice

Table 11 applies Proposition 3 and two conventional clustered procedures to every application with more than one unit per cluster. The empirical-Bernstein interval of Maurer & Pontil (2009) on cluster rates certifies pooled PTB-XL, with interval $[0.0399, 0.0679]$, and abstains for Qwen and every NASA design, whose upper limits range from 0.161 to 0.697. The aggregate moment program and Cantelli give the same classifications on the tested grid because both discard independence across clusters. Bernstein retains it. This changes the finite-sample conclusion for pooled PTB-XL, which certifies at $\kappa = 1$ and hence under arbitrary dependence within patients. Qwen certifies only for $\kappa \leq 0.110$, below its plug-in correlation 0.680. For NASA the abstentions are exact at every κ, because the zero-correlation witnesses of Table 10 belong to every $\mathcal{C}_\kappa$. Conventional clustered procedures reach the same observed decisions, but they carry no finite-sample guarantee (Appendix G).

Random-effects clusters. Between a correlation cap and arbitrary dependence lies the random-effects model $\mathcal{M}_\kappa(p)$, in which clusters are independent and cluster g has a latent rate π_g with $\mathbb{E}\pi_g = p$ and $\mathrm{Var}\,\pi_g \leq \kappa p(1-p)$, its units being independent given π_g. It constrains every moment of a cluster sum, since S_g is a mixture of $\mathrm{Bin}(m, \pi)$ laws, and $\mathcal{M}_\kappa(p) \subseteq \mathcal{C}_\kappa(p)$. Superposition and thinning preserve it, so half-line gates again need only the boundaries. Two-point rate laws give exact lower bounds on the worst case tail by convolution, and hence necessary cluster counts for any valid half-line gate. For sufficiency, the law $K^\star$ whose stop-loss transform is $t \mapsto \sup_\pi \mathbb{E}(S_g - t)_+$ dominates every cluster sum in the increasing convex order, so $\mathbb{P}(T \geq c) \leq \min_{a<c} \mathbb{E}(S^\star - a)_+/(c-a)$ for the sum $S^\star$ of G copies, and symmetrically at U. Table 12 reports both counts, with retention under the beta-binomial law at p_0 with correlation κ, beside the design-effect rule $\lceil 943\{1 + (m-1)\kappa\}/m \rceil$. The rule is close for $m = 10$ but undercounts for $m = 50$ and small κ, where rare bursts dominate the tail. For Qwen ($m = 50$, $\kappa = 0.68$), the necessary count is 645 prompts, so 180 prompts support no valid half-line gate even when retention is measured under the beta-binomial law rather than the

Table 11: Positive-dependence and clustered-practice comparisons under the primary contract. The plug-in $\hat{\bar{\gamma}}$ is descriptive. The LP and Bernstein columns give the largest κ on the tested grid certified by the corresponding finite-sample bound (the exact Bernstein threshold for Qwen is 0.110). "None" means that the bound did not certify at $\kappa = 0$, and for NASA abstention is exact at every κ. The last two columns are 98% intervals, with C for certify and A for abstain. Cluster-t TOST is asymptotic and the cluster bootstrap is a percentile interval based on $50{,}000$ resamples.

Endpoint	$\hat{\bar{\gamma}}$	LP cap	Bernstein cap	Cluster-t CI	Bootstrap CI
PTB-XL pooled windows	.101	.02	1.00	$[.0486, .0592]$ C	$[.0486, .0594]$ C
Qwen GSM8K failures	.680	None	.10	$[.0172, .0797]$ A	$[.0210, .0823]$ A
SMAP innovation	.055	None	None	$[.0203, .0654]$ A	$[.0238, .0667]$ A
SMAP rolling maximum	.054	None	None	$[.0199, .0642]$ A	$[.0230, .0651]$ A
MSL innovation	.030	None	None	$[.0179, .0760]$ A	$[.0227, .0742]$ A
MSL rolling maximum	.100	None	None	$[.0078, .0952]$ A	$[.0182, .0955]$ A

copy law. Even if validity is required only under the beta-binomial laws at L and U, the exact gate is empty ($c_1 = 454 > c_2 = 360$), no rule based on T, randomised or not, retains with probability above 0.062 (a linear program over all such rules). This law is not invariant under permutations of units across clusters, so Lemma 1 does not reduce every procedure to T here, and the one-sided gate $T \leq 360$ retains with probability 0.24. These counts are computed in floating point on a rate grid. The necessary counts come from explicit laws in the class, and the sufficient counts are approximate because the grid supremum can underestimate the envelope.

Table 12: Clusters needed by a valid half-line gate under the random-effects model, primary contract. "Necessary" uses explicit two-point rate laws and is a lower bound. "Sufficient" uses the convex-order bound with its envelope taken as a supremum over a rate grid, which can underestimate it, so these counts are approximate. At $\kappa = 0$ the model is full independence, and at $\kappa = 1$ Proposition 2 gives 943 with one unit per cluster.

		Clusters G		
m	κ	necessary	sufficient (approx.)	design-effect rule
10	0.01	108	129	103
10	0.05	142	186	137
10	0.2	273	387	265
50	0.01	41	53	29
50	0.05	76	119	66
50	0.2	213	350	204
50	0.68	645	878	648

G Simulation of clustered procedures

Each replication draws G independent clusters of m units at a common marginal rate $p \in \{L, U, p_0\}$. In the beta-binomial settings each cluster has a rate drawn from a beta law with mean p and intracluster correlation κ, and its units are independent given that rate. In the AR(1) settings each cluster is a stationary Gaussian AR(1) series with coefficient 0.9, thresholded at its $1 - p$ quantile, so consecutive windows are strongly dependent. Its average intracluster correlation is at most 0.217 at L, U and p_0, which we use as its cap. The procedures are exact TOST on all Gm units treated as independent, cluster-t TOST (the 98% interval $\hat{p} \pm t_{G-1,0.99}\,\widehat{\mathrm{se}}$ with the cluster-robust standard error), the design-effect TOST (exact TOST on the effective size $Gm/\{1 + (m-1)\hat{\kappa}\}$ with the ANOVA estimate $\hat{\kappa}$), the cluster bootstrap percentile 98% interval with 400 resamples, the Bernstein audit of Proposition 3 at the true cap and at $\kappa = 1$, and exact TOST on one prespecified unit per cluster. Table 13 reports $4{,}000$ replications per setting with seed 20260925, so a rate near 0.01 has Monte Carlo standard error about 0.0016. At $p = L$ every clustered procedure stayed at or below 0.0092, while exact TOST on all units reached 0.17. The prespecified-unit test never certified, since its exact TOST gate is empty at these numbers of clusters. An empirical-Bernstein audit of the cluster rates (Maurer & Pontil, 2009),

with level α on each side, made no error and never certified at p_0 in any setting, because its range term $7\log(2/\alpha)/\{3(G-1)\}$ alone exceeds the half-width 0.025 unless G is above about 500.

The table adds three procedures run on separate draws with the same design and seeds 20260926 and 20260927. The restricted wild cluster bootstrap uses Webb six-point weights and 399 draws (Cameron et al., 2008; Webb, 2023). For equal cluster sizes, CR2 with Bell and McCaffrey degrees of freedom coincides with the cluster-t row (Bell & McCaffrey, 2002). The betting audit tests each boundary by a capital process on the cluster rates with growth-rate-adaptive bets, a prior pseudo-observation at p_0 and Ville's inequality (Waudby-Smith & Ramdas, 2024), and the empirical-Bernstein audit uses the bound of Maurer & Pontil (2009) on cluster rates. Both are valid under arbitrary dependence within clusters, and both pay for it. A cluster rate can equal one, so a bet against the upper null can grow by a factor of at most $1/(1-U)$ per cluster, and the empirical-Bernstein range term $7\log(2/\alpha)/\{3(G-1)\}$ exceeds the half-width 0.025 unless $G > 495$. In the $\kappa = 0.02$, $m = 50$ design, betting retention is 0.05 with 300 clusters, 0.92 with 400 and 1.00 with 500.

Table 13: Probability of certifying at $p = U$, whose nominal bound is 0.01, and retention at p_0, reported as error / retention from $4{,}000$ replications. The Bernstein audit uses the true cap, and at $\kappa = 1$ it also never certified. The AR(1) cap is its largest average intracluster correlation over L, U and p_0.

	beta-binomial				AR(1), coefficient 0.9	
Cap κ	0.02	0.02	0.1	0.02	0.217	0.217
$G \times m$	20×50	50×50	50×50	150×50	42×30	150×30
Independent TOST	.033 / .64	.052 / .98	.185 / .84	.044 / 1.00	.177 / .50	.193 / .96
Cluster-t (CR2)	.014 / .18	.018 / .90	.032 / .12	.012 / 1.00	.000 / .00	.026 / .49
Design effect	.014 / .27	.013 / .88	.018 / .15	.009 / 1.00	.000 / .00	.014 / .47
Pairs bootstrap	.025 / .34	.024 / .91	.034 / .25	.014 / 1.00	.001 / .00	.025 / .56
Wild bootstrap	.018 / .19	.017 / .88	.021 / .12	.015 / 1.00	.000 / .00	.025 / .49
Bernstein, true cap	.000 / .00	.000 / .00	.000 / .00	.000 / .74	.000 / .00	.000 / .00
Empirical Bernstein	.000 / .00	.000 / .00	.000 / .00	.000 / .00	.000 / .00	.000 / .00
Betting	.000 / .00	.000 / .00	.000 / .00	.000 / .00	.000 / .00	.000 / .00
One unit per cluster	.000 / .00	.000 / .00	.000 / .00	.000 / .00	.000 / .00	.000 / .00

H COMPUTATION

All timings are on one core of a cloud virtual machine. Verifying all 32 sample-size certificates with the standalone verifier takes about 22 seconds, the 194 robust p-value certificates of the applications less than one second, the NASA design certificates of Table 10 half a second, and the sixteen deterministic gates about 14 seconds. At the certified sample size, constructing the lower certificate takes 0.5 seconds at order four ($N = 1{,}517$) and 1.3 seconds at order two ($N = 8{,}172$), and the complete null upper certificate 2.8 and 8.7 seconds. The bisection over N and the cutting-plane iterations multiply this by a small factor. Exact robust p-values, with lower and upper bounds that agreed to within 1% in our tests, take about 0.3 seconds at $N = 10^4$ and 3 seconds at $N = 10^5$ at orders two and four, using a centred polynomial basis and cutting planes with exact violation checks. The cost is dominated by exact evaluation at the $N+1$ counts and grows linearly in N. Floating-point solutions of these moment programs are fragile at small p, large N and high order. At orders above about ten the dual solutions often fail to repair, and the corresponding entries in the supplementary material are reported as unresolved rather than certified.

I CLAIM-STATUS LEDGER

- Theorem 2 and the randomised columns of Table 1 rest on exact lower certificates at $N_r^\star - 1$ and exact complete null upper certificates at $N_r^\star$ (Table 4).
- Figure 2 is a direct rendering of the rational $\phi(t)$ values in the four primary-contract upper certificates. The randomised optima are used as exact lower limits on sample requirements. The deterministic gates are the operational alternative.

- In the deterministic columns of Table 1, every reported gate is certified valid in exact arithmetic (Proposition 5) with exact nominal retention at least 0.75. The order-4 entry is minimal over all deterministic rules by a mixed-integer scan of every smaller N (floating point). The order-5 entry is minimal among half-line gates. The order-2 and order-3 entries are minimal among half-line gates within the stated 150-value backward scan. Mixed-integer checks at other sample sizes found interval optima throughout.
- Propositions 2, 3 and 4 and Theorem 3 are proven. That odd orders save little is observed in all four contracts. Table 6 reports floating-point lower bounds.
- In Table 8, every rule and every interior null witness is released and checked by the verifier. The boundary reduction for rules that certify only inside the region is a conjecture supported by the search above.
- In Tables 3 and 9, every decision is certified, by dual polynomials for certification and by exact laws for abstention. Intermediate orders that did not close numerically follow by nesting.
- In Table 10, every gate and every overturning law is certified exactly, and the status of each of the 120 released designs is checked against them. The observed counts are regenerated from public telemetry by the released script. Unresolved counts are not claimed either way.
- Every entry of Table 2 has matching exact lower and upper bounds in the archive.
- Table 5 reports sufficient deterministic constructions, not optimality claims. Its boundary bounds are exact rational numbers, its first sample sizes come from exhaustive integer scans, and its retentions are computed from the binomial distribution.
- Table 11 reports the largest certified value on a grid, not an exact threshold and not an estimate of correlation. "None" records failure of the bound to certify, except for NASA, where the zero-correlation witnesses make abstention exact. The clustered TOST and bootstrap columns are conventional comparisons, not certificates.
- Appendix G is a Monte Carlo study with $4{,}000$ replications per setting. Zeros in Table 13 mean no certifications were observed in those replications. They are not proofs that retention is exactly zero.