Research Square

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

Protecting the Label: Abstention, Prespecification, and Alpha Allocation for AI Derived Subgroups in Confirmatory Trials

Luca Pani

lpani@miami.edu

University of Miami

Joseph Geraci

Queen's University

Research Article

Keywords: selective prediction, closed testing, familywise error rate, adaptive enrichment, regulatory strategy, product information

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11158567/v1

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

Additional Declarations: No competing interests reported.

Abstract

Phase 3 programs fail when a treatment that benefits some patients is tested in a population where many do not respond, and the overall effect is diluted below significance. A subgroup found after unblinding cannot rescue such a program, because no type I error was allocated to it. A subgroup can support a label only if it was defined before unblinding, by a rule an independent statistician can apply, and tested under a multiplicity strategy fixed in advance. Most classifiers produced by modern machine learning cannot be written into a protocol in that form, because membership depends on a fitted probability, distance, or similarity. We argue that abstention resolves this. A system that returns No Call where the evidence does not support an assignment can express membership as a compact logical rule over baseline values: deterministic, frozen before unblinding, and verifiable without access to the model. Model Derived Subgroups (MDS) of one to four variables, distilled from higher dimensional models, have this form. Carried into an all comers trial under weighted graphical multiple testing, an MDS gives the sponsor two paths to a label. If the overall population succeeds, the subgroup is tested at full alpha at no cost to the primary claim. If the overall population fails, a prespecified split allocation keeps the subgroup hypothesis testable, and a positive result can support a label in the population the rule defines. We set out the allocation choices, their costs, and what the rule must satisfy to define a label. *(249 words)*

1. Introduction

Late-stage programs often fail for a reason that has little to do with whether the drug works. A treatment that benefits a subset of patients is tested in a heterogeneous population, the benefit is diluted by those who do not respond, and the overall effect falls short of significance. The sponsor may then see, in the data, a group that clearly benefited. By then it is too late. A subgroup identified after unblinding carries no allocated type I error, and regulators treat it as hypothesis generating.[1, 2] A program that could have been saved is not.

The obstacle is not statistical theory. Methods for testing a subgroup hypothesis alongside an overall population hypothesis with strong control of the familywise error rate (FWER) have existed for decades and are described in regulatory guidance.[1, 2] Heterogeneity of treatment effect is likewise well characterized as a multivariable estimation problem rather than a sequence of post hoc comparisons.[3]

The obstacle is that the subgroup must be defined before the trial, by a rule that can be written into the protocol, and most classifiers produced by modern machine learning cannot be. Guidance from the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) on artificial intelligence (AI) in medicines development scales the evidentiary burden with the consequence of the model output.[4, 5, 6] For a model output that defines a hypothesis tested at a controlled alpha level, that burden includes a requirement the guidance does not spell out: an independent statistician must be able to determine, from the recorded baseline data and the written rule alone, which patients belong to the subgroup.

This paper makes three arguments. First, a confirmatory subgroup requires a classifier with five properties, and forcing an assignment for every patient defeats them. Second, abstention restores them: a classifier permitted to return No Call expresses membership as a logical predicate over observed values, and the gates that release its rules provide error control at discovery, before any rule reaches a statistical analysis plan (SAP). Third, such a rule, embedded in an all comers trial under weighted graphical multiple testing, gives the sponsor two paths to a label: one when the overall trial succeeds, and one when it does not.

We draw on the published description of one such system for the architecture of the rules it produces.[7] This is a communication on design and regulatory strategy; it reports no new analyses.

2. What a subgroup must satisfy to support a label

A subgroup hypothesis tested at a controlled alpha level in a registration trial rests on a classifier with five properties.

1. 1. Prespecified. The variables, the thresholds, the assignment logic, and the intended use are fixed in the protocol and the SAP before the trial begins.
2. 2. Frozen before unblinding. No component is reestimated, retrained, or retuned once outcome data begin to accrue.
3. 3. Deterministic. The same patient with the same baseline record receives the same assignment on every evaluation, with no dependence on seed, fold, fitting order, or software version.
4. 4. Computable from baseline covariates alone. Assignment uses only data recorded before randomization, on the case report form, with no postbaseline information.
5. 5. Independently verifiable. A statistician holding the baseline dataset and the written rule can reproduce every patient assignment without access to the model, its parameters, or its training data.

Classifiers with a closed form, such as a logistic regression with published coefficients and a cut point, or a published scoring rule, can satisfy all five. The machine learning classifiers now proposed for subgroup discovery generally cannot, and the failure is structural. A model obliged to assign every patient must place a boundary somewhere, and for an ensemble, a neural network, or a kernel method that boundary is computed by the fitted model: a predicted probability against a cut point, a distance to a centroid, a similarity to a prototype. Membership then depends on the fitted object. It cannot be verified from the trial data alone, because reproducing it requires rerunning the model. It cannot be stated in a protocol in a form an investigator or reviewer can apply, because it describes what a fitted model says about the patient rather than the patient. And near the boundary it is unstable: in the small samples typical of discovery, a borderline assignment is set by parameter estimates that would change on resampling.

This is the practical content of the preference for intrinsically interpretable models where the stakes are high.[8] Reporting frameworks such as TRIPOD + AI and CONSORT-AI ask what must be described.[9, 10] The confirmatory setting asks what must be recomputable.

Selective prediction, in which a model may withhold an assignment, has been studied mainly as a way to improve accuracy on the cases a model does classify,[11] and conformal procedures supply coverage guarantees for prediction sets without distributional assumptions.[12] The property we draw on is different: permitting abstention removes the need to assign patients the rule does not cover, which is what allows membership to be written as a predicate over recorded values and keeps unstable borderline assignments out of the subgroup. The literature on testing a prespecified subgroup assumes such a classifier exists.[13] Our argument concerns how one is obtained.

3. Model Derived Subgroups: abstention as prespecification

3.1 From higher dimensional models to a compact rule

For clarity, a Model Derived Subgroup (MDS) is a compact rule that characterizes patients predicted to show, collectively, a significant separation on the endpoint, between treatment arms or, where a single treatment is studied, in response. An MDS consists of one to four variables, each with a defined interval or ordinal range. These variables are learned in a preceding step that discovers synergistic combinations of variables through higher dimensional models. When a model is reported with more than four variables, it is one of these higher dimensional models that is being shared.[7] What is prespecified in the SAP is the MDS, not the higher dimensional model, and its compactness is what makes it verifiable: a rule of at most four interval conditions can be printed in a protocol, read by an investigator, and recomputed by hand.

3.2 The assignment procedure

Assignment is deterministic and conjunctive. A patient joins a subgroup only if every variable in the bundle is observed and lies inside its interval, with no distance, similarity, or probability threshold applied at assignment time.[7, 14] A patient whose observed values do not satisfy any advanced rule receives No Call. Abstention also operates earlier than assignment: the higher dimensional models from which the synergistic variable combinations are learned themselves abstain during learning, in that each candidate configuration is scored on the patients it covers and leaves the remainder unassigned rather than partitioning the whole cohort.

Membership is therefore a Boolean function of recorded baseline values and a fixed table of intervals. An independent statistician holding the baseline data and the written rule reproduces every assignment exactly, without the model, its training data, or its software, and the rule returns the same answer on any cohort to which it is applied.

None of this survives if the classifier is required to assign every patient. A rule of the form "every variable observed and in interval" has no natural extension to patients who fail it; extending it requires a nearest match, which requires a metric, which reintroduces the model. Abstention is therefore not a property added alongside prespecifiability. It is the property that produces it.

3.3 The advancement gates as error control at the discovery stage

The second requirement is that the candidate rule was not selected from a large space by chance. In the published system a candidate is advanced only if it covers at least 15% of the available training population, satisfies a prespecified treatment effect or response enrichment requirement, reaches nominal significance, meets an effect size criterion, and achieves an empirical p value below 0.05 against a random set null built by running the complete search on randomly assembled variable sets of matching dimensionality under the same optimization budget.[7] The learning objective seeks the largest subpopulation in which treatment and control separate on the endpoint, because coverage sets the size of the trial needed to test the subgroup (Section 4.3).

The random set null is the substantive element: it asks whether the observed separation exceeds what the same adaptive search obtains from arbitrary variable combinations given equivalent freedom. These gates are screening criteria calibrated to the search itself.[7] Formal multiplicity control belongs to the confirmatory stage. The gates determine which rules are eligible to be written into an SAP; the confirmatory alpha is then spent on a single prespecified rule, and the discovery search, which is not repeated, does not enter the confirmatory analysis.

4. Alpha allocation: two paths to a label

4.1 Closed testing and graphical procedures

The closed testing principle provides strong FWER control: a hypothesis may be rejected at level alpha only if it and every intersection hypothesis containing it are rejected at level alpha by their own local tests.[15] Weighted graphical procedures realize this principle for the small families that arise in confirmatory trials.[16, 17] Hypotheses are nodes carrying initial weights that sum to one, and each is tested at alpha times its weight. When a hypothesis is rejected, its weight passes along directed edges to the hypotheses not yet rejected, whose levels rise accordingly. FWER is controlled in the strong sense throughout, and FDA guidance on multiple endpoints describes these methods for registration settings.[18]

4.2 A two node graph for a prespecified subgroup

Let H1 be the null hypothesis for the overall randomized population and H2 the null hypothesis for the called subgroup, with initial weights w1 and w2 summing to one and edges in both directions carrying the full weight of a rejected node to the other. Figure 1 shows the graph for the three configurations in Table 1. The choice among them is a choice about what the subgroup is for.

Table 1
Alpha allocation configurations for a two node graph

Configuration	w1 / w2	Level at which the primary claim is tested	Level available to the subgroup hypothesis	When a sponsor chooses it
Hierarchical (fixed sequence)	1.00 / 0.00	Full alpha; no reduction	Full alpha if H1 is rejected; not tested otherwise	Complement expected to respond much as the called subgroup does; the subgroup claim is upside, not insurance
Recycled split, conservative	0.80 / 0.20	0.80 alpha; a real and permanent reduction	0.20 alpha if H1 fails; full alpha if H1 is rejected	Complement expected to show a smaller or null effect; the subgroup keeps a path to a label open if the overall claim fails
Balanced	0.50 / 0.50	0.50 alpha; a substantial reduction	0.50 alpha if H1 fails; full alpha if H1 is rejected	Complement expected to show no effect or signal reversal; the subgroup is the likely basis of the label

H1, overall population hypothesis; H2, called subgroup hypothesis; w1 and w2, initial node weights summing to one. Recycling transfers weight to hypotheses not yet rejected. It does not restore the level at which an already tested hypothesis was evaluated, so the reduction in the primary claim's level under a split is not recovered. Because the subgroup is nested in the overall population, Bonferroni based graphs are conservative in this setting; weighted Simes and parametric versions recover part of the loss.[17]

Under a recycled split the subgroup is insurance. With w1 equal to 0.8, H1 is tested at 0.8 alpha and H2 at 0.2 alpha from the outset. If the overall claim fails, H2 is still tested, and a positive result is a confirmatory finding in a prespecified population: a basis for a label that the program would otherwise not have. If H1 succeeds, its alpha recycles and H2 is tested at full alpha. The premium is real and should be stated plainly. The level at which the primary claim is evaluated is permanently reduced, with a corresponding loss of power, because recycling moves weight forward to hypotheses not yet rejected and never restores the level at which an already tested hypothesis was evaluated. What the sponsor buys with that premium is a second route to approval.

Under hierarchical testing the subgroup is upside. With w1 equal to one and w2 equal to zero, H1 is tested at full alpha and the presence of H2 costs the primary claim nothing. If H1 is rejected, H2 is tested at full alpha, and a positive result sharpens the label and the clinical message. If H1 fails, however, H2 is never tested, so this configuration does not rescue a failing program. A sponsor whose principal concern is the risk of an overall failure should not choose it.

Because the called subgroup is nested within the overall population, the two test statistics are positively correlated, and Bonferroni based graphs are conservative. Weighted Simes and parametric versions of the same graphs recover part of the loss,[17] and methods developed for the overall population and subgroup problem offer further refinement.[13, 19, 20, 21]

4.3 Sizing the subgroup hypothesis

The size of the called subgroup determines the power available to H2 and nothing else. Power for H2 depends on the expected subgroup fraction, which the rule supplies as its coverage, and on the treatment effect within the called subgroup, which is a design assumption the sponsor states explicitly. The separation observed at discovery motivates H2 as a hypothesis; the randomized confirmatory trial estimates the comparative effect it tests. Coverage enters sizing directly: a trial that needs a given number of participants in the H2 analysis set must randomize that number divided by the coverage. At a coverage of 30%, that is roughly 3.3 times the subgroup sample, with every additional participant contributing to H1. This is why the discovery objective seeks the largest subpopulation in which treatment and control separate: every gain in coverage reduces the trial needed to test the subgroup.

5. Design: all comers, and no participant excluded by the rule

The design that follows enrolls all comers. The rule labels participants; it does not screen them.

No Call and subgroup negative participants are enrolled, randomized, and analyzed. Every randomized participant contributes to H1, whatever their rule status. Rule status is determined from baseline data before randomization, and randomization is stratified by it, so the called and not called strata are balanced across arms. The rule variables are mandated baseline assessments. Where a required value is nonetheless missing, the participant receives No Call, remains in the trial and in H1, and is excluded only from the H2 analysis set. Missingness can affect the subgroup analysis set; it cannot affect access to the trial.

What remains estimable. This design preserves the overall treatment effect, the treatment effect within the called subgroup, the treatment by subgroup interaction, safety in both strata, and the effect among participants the rule does not call. A design that enrolled only the called subgroup would forfeit four of these five. Retaining them is what allows a regulator to judge where the benefit lies: in the overall population, or in the population the rule defines.

Enrichment only is the relevant comparison. Against an all comers trial without a subgroup hypothesis, adding a prespecified H2 costs nothing under hierarchical testing. Against an enrichment only trial, the all comers design is larger, randomizing roughly 3.3 participants for each one in the H2 analysis set at 30% coverage. The additional participants buy the overall claim, the four estimands that enrichment only forfeits, and both paths to a label. Whether that is worth the cost depends on the complement.

The complement determines the allocation. Participants the rule does not call enter H1. Where they are expected to respond much as the called stratum does, they add information to H1 and hierarchical testing is appropriate. Where they are expected to show little or no effect, they dilute H1, and weight should move toward H2 through a recycled split. Where randomized discovery data show signal reversal in the complement, with control outperforming treatment, the complement works against H1; the allocation should then place substantial weight on H2, or the design should limit enrollment of the complement as described below. When discovery data are randomized, the effect in the complement is

estimated alongside the effect in the called subgroup, so this choice is made from evidence rather than preference. Table 1 sets out the correspondence.

Adaptive enrichment offers a middle course.[22, 23, 24] The trial opens to all comers, and at a prespecified interim analysis enrollment of the complement stops if it meets a prespecified futility criterion. Every estimand remains available up to the interim, and the design then recovers much of the efficiency of enrichment. The interim, the decision rule, and the combination of evidence across stages are fixed in advance, and methods that control the type I error of such designs are established. Any restriction then rests on randomized evidence that the complement does not benefit, not on the rule alone.

Representativeness. The concern that AI derived subgroups will narrow trial populations and encode underrepresentation in eligibility criteria has been raised in the governance literature, including by one of us.[25, 26] The design set out here answers that concern directly, because the rule excludes no one. It obtains the subgroup claim while preserving the evidence needed to characterize benefit and harm in the population that will receive the product.

Change control. The frozen rule is identified in the SAP by classifier version. Any revision, to an interval boundary or to a variable, produces a different rule whose coverage must be characterized again before confirmatory use. This is documentation of the same kind as a database lock, not additional study.

6. Labeling: two paths to a label

The design yields a label on either of two paths, and both depend on the prespecification set out above.

When the overall claim succeeds, the label follows the overall population. The subgroup result, tested at full alpha after recycling, can inform the description of clinical studies and the clinical message without narrowing the indication. A label should not narrow merely because a subgroup responds better; the question is whether the evidence supports an acceptable balance of benefit and risk outside it, and an all comers design is what makes that question answerable.

When the overall claim fails, a prespecified subgroup can save the program. A positive result on H2, tested at its allocated alpha, is a confirmatory finding in a population defined before unblinding by a rule any statistician can apply, and it can support an indication in that population. Abstention strengthens that case. Because membership is a predicate over recorded values, a reviewer can reconstruct the analysis population patient for patient from the submitted data, and because the rule declines to assign borderline patients, the population in which H2 succeeded is not an artifact of where a fitted model happened to draw its boundary. It does not guarantee one. Regulators will weigh the consistency of the subgroup result, the plausibility of the mechanism, and the effect and safety in the complement, and may ask for confirmation.[1, 2] But it is the only route within the trial itself by which a program that misses its overall endpoint can still yield a labeled product, and it exists only if the subgroup and its alpha were prespecified. A subgroup found after unblinding offers no such route.

The rule must be writable into the label. For the second path to be open, the population the rule defines must be one a prescribing clinician can identify. This is where abstention matters most. A classifier that assigns every patient defines its population by a model score, and a model score cannot be printed in an indication. An abstaining rule defines its population by explicit intervals on baseline variables, which can. An MDS is expressed as one to four variables with explicit intervals, which is the right form for an indication statement, but each variable must also be measurable in routine care: by an analytically validated assay, a standard laboratory value, or an instrument with established operating characteristics. Where a rule variable is a molecular marker, the route to product information runs through a companion diagnostic. Where it is a single item drawn from a research instrument administered under trial conditions, it may define an analysis population soundly and still not be writable into a label. The measurability of each rule variable should therefore be assessed at the design stage and recorded in the prespecification package, so that the path to a subgroup label is known to be open before the trial begins.

7. Conclusion

Programs fail when benefit concentrated in some patients is diluted across all of them, and a subgroup found after the fact cannot rescue them. A subgroup prespecified with its own alpha can. What has stood in the way is that machine learning classifiers which assign every patient cannot be written into a protocol in verifiable form. A classifier permitted to abstain can: membership becomes a compact rule over baseline values that any statistician can recompute, released only after passing gates that provide error control at discovery.

With such a rule in hand, the remaining machinery is standard. A two node weighted graph carries the overall and subgroup hypotheses. Under hierarchical testing the subgroup is upside at no cost; under a recycled split it is insurance, bought with a stated reduction in the level of the primary claim, and it keeps a path to a label open if the overall trial fails. The expected behavior of the complement determines which to choose. Enrollment stays all comers, and the label follows the evidence: in the overall population when it succeeds, and in the population the rule defines when only the subgroup does. Abstention is what makes the subgroup prespecifiable, and prespecification is what makes it capable of saving a program.

Declarations

Conflict of interest disclosures.

Dr Pani reported serving as Chief Innovation and Regulatory Officer of NetraMark; serving during the 36 months preceding submission as a consultant or advisory board member for AbbVie, AiDvance, Boston Consulting Group, GH Research, LB Pharmaceuticals, Magdalena BioSciences, Napo Pharmaceuticals EU, Rain Scientific, and Relmada Therapeutics; and owning shares or options in AiDvance, Adapt, Enzeta,

NetraMark, Rain Scientific, and Relmada Therapeutics. Dr Geraci reported being founder and Chief Scientific and Technical Officer of NetraMark Corp and a significant shareholder of NetraMark Holdings Inc. No other disclosures were reported.

Role of the funder/sponsor.

No funder or sponsor had any role in the preparation, review, or approval of the manuscript or the decision to submit it for publication.

Funding.

No specific funding was received for this work.

Author Contribution

Concept and design: Pani, Geraci. Interpretation: Pani, Geraci. Drafting of the manuscript: Pani, Geraci. Critical revision of the manuscript for important intellectual content: Pani, Geraci. Both authors approved the final manuscript and agree to be accountable for all aspects of the work.

Acknowledgment.

A generative AI language assistant was used under the authors' direction for structural editing and language refinement. The authors conceived the analysis, verified all factual statements and references against primary sources, and take full responsibility for the content.

Data availability.

No data were generated or analyzed for this work.

References

1. US Food and Drug Administration. Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products. Guidance for Industry; March 2019.
2. European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials. EMA/CHMP/539146/2013. January 31, 2019.
3. Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020;172(1):35–45. 10.7326/M18-3667.

4. US Food and Drug Administration. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. Draft Guidance for Industry and Other Interested Parties; January 2025.
5. European Medicines Agency. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. September 9, 2024.
6. US Food and Drug Administration; European Medicines Agency. Guiding principles of good AI practice in drug development. January 14, 2026.
7. Geraci J, Qorri B, Cumbaa C, et al. Interpretable subgroup discovery with abstention in small, heterogeneous clinical trials: a retrospective multi-dataset study. AI. 2026;7(9):349. 10.3390/ai7090349.
8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. 10.1038/s42256-019-0048-x.
9. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
10. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364–74. 10.1038/s41591-020-1034-x.
11. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. Adv Neural Inf Process Syst. 2017;30:4878–87.
12. Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. Conformal risk control. In: International Conference on Learning Representations; 2024.
13. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. Stat Med. 2007;26(19):3535–49. 10.1002/sim.2825.
14. Geraci J, Qorri B, Tsay M, et al. Explainable AI-driven precision clinical trial enrichment: demonstration of the NetraAI platform with a phase II depression trial. npj Digit Med. 2025;8:749. 10.1038/s41746-025-02143-7.
15. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika. 1976;63(3):655–60. 10.1093/biomet/63.3.655.
16. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Stat Med. 2009;28(4):586–604. 10.1002/sim.3495.
17. Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. Biom J. 2011;53(6):894–913. 10.1002/bimj.201000239.
18. US Food and Drug Administration. Multiple Endpoints in Clinical Trials: Guidance for Industry. October 2022.

19. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. Stat Med. 2009;28(1):3–23. 10.1002/sim.3461.
20. Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. Stat Med. 2014;33(4):693–713. 10.1002/sim.5974.
21. Millen BA, Dmitrienko A, Ruberg S, Shen L. A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. Drug Inf J. 2012;46(6):647–56. 10.1177/0092861512454116.
22. US Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. November; 2019.
23. Wang SJ, Hung HMJ, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. Biom J. 2009;51(2):358–74. 10.1002/bimj.200900003.
24. Simon N, Simon R. Adaptive enrichment designs for clinical trials. Biostatistics. 2013;14(4):613–25. 10.1093/biostatistics/kxt010.
25. Blom JMC, Staunton C, Tascedda S, Slabbert N, Pani L, Labuschaigne M. Embedding equity in clinical research governance. Nat Med. 2026;32(4):1185–7. 10.1038/s41591-026-04232-w.
26. Blom JMC, Pani L. Embedded transparency in artificial intelligence: a prerequisite for equity and representation in AI-enabled clinical trials. npj Digit Med. 2026;9:531. 10.1038/s41746-026-02987-7.

Tables

Table 2
Prespecification package for a confirmatory subgroup hypothesis

Item	What is documented	Why it is required
Frozen rule definition	Every variable in the bundle with its value interval and endpoint convention	This is the object the protocol carries; it must be readable and applicable without the model
Classifier version and change control	Version identifier, date of freeze, and the procedure governing any revision	Establishes that the rule predates unblinding and fixes what a revision obliges the sponsor to redo
Coverage in the training population	Proportion of the development cohort meeting the rule	Determines the expected subgroup fraction and therefore the power available to H2
Assignment determinism statement	Explicit confirmation that no distance, similarity, or probability threshold is applied at assignment	Distinguishes a protocol criterion from a model output
Independent recomputation procedure	The steps by which a third party reproduces every assignment from baseline data and the written rule	Satisfies the verifiability requirement that a confirmatory claim rests on
Hypothesis structure and alpha allocation	The graph, initial weights, edge weights, and the resulting local levels	Fixes the multiplicity strategy before results are available
Handling of No Call and subgroup negative participants	Their retention, randomization, stratification, and analysis population membership	Establishes that enrollment is unrestricted and that the overall analysis is complete
Estimands	Overall treatment effect, called subgroup effect, and treatment by subgroup interaction, each with its population and intercurrent event strategy	These answer different questions and must be distinguished in advance
Measurability assessment	For each rule variable, the instrument, administrator, and setting required to obtain it in routine care	Establishes, before the trial, that the rule can define a labeled population if only the subgroup claim succeeds

Table 3
Decision gates for a prespecified abstaining classifier

Gate	What must be established	How it is documented	What it enables
Discovery	That the candidate rule survived prespecified advancement criteria, including minimum coverage, an effect size criterion, and calibration against an empirical random set null	Discovery report with the gates, the random set procedure, and the resulting rule with intervals	A compact conjunctive rule eligible to be written into a protocol
Trial design	That the rule is frozen, deterministic, computable from baseline covariates, and independently recomputable; that enrollment is all comers and randomization is stratified by rule status; that alpha is allocated across H1 and H2	Protocol and SAP, incorporating the prespecification package of Table 2	A confirmatory subgroup hypothesis tested at a known level, with the overall effect, subgroup effect, interaction, and safety all estimable
Labeling	That the evidence supports the labeled population, overall or in the population the rule defines; that each rule variable can be measured in routine care	Clinical overview and benefit and risk assessment, with the measurability assessment carried forward	A label in the overall population, or, if only the subgroup claim succeeds, a label in the population the rule defines

Figures

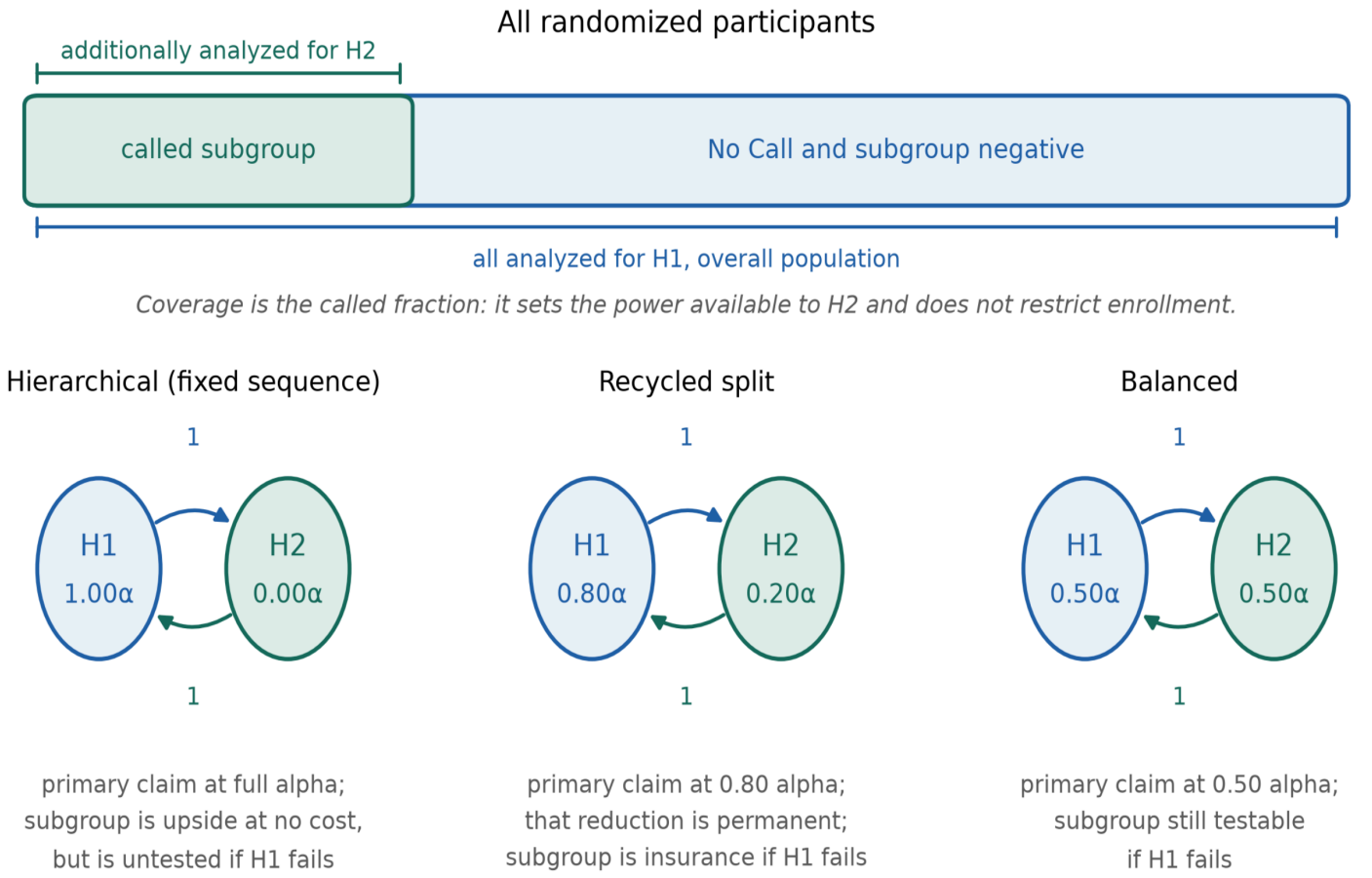


Figure 1

Two node weighted graph for a prespecified subgroup hypothesis, shown for the three allocations of Table 1. H1 is the overall population hypothesis and H2 the called subgroup hypothesis; node labels give the initial weight and the resulting local level. Directed edges carry the full weight of a rejected node to the node not yet rejected. The population panel shows that enrollment is all comers: participants receiving No Call and those who are subgroup negative are randomized and analyzed within the overall population supporting H1, and only the called stratum additionally supports H2.