Supplementary information: transparent record of the human–AI workflow

for "On marginal and conditional error in adaptive clinical trials"

Qingyuan Zhao
Statistical Laboratory, University of Cambridge

28 July 2026

Approval record. Qingyuan Zhao, the sole author of this supplement and the author who directed the recorded workflow, reviewed and approved it on 24 July 2026. No additional author approval is required for this supplement.

1 Purpose, scope, and relation to the manuscript

This supplement discloses how generative artificial intelligence (AI) was used to develop the manuscript and its computational materials. It expands the brief generative-AI statement in the manuscript into an auditable account of the inputs, human decisions, AI-assisted tasks, validation steps, revisions, failures, and recoveries. It is based on two exported OpenCode/OpenChamber session records supplied for this purpose, nine audio-derived manuscript-development transcripts, the versioned manuscript packages, and their source and provenance logs.

The record covers substantive manuscript work from 13 to 24 July 2026. The exported sessions show that AI-assisted drafting and simulation work began on 13 July, that the first complete package was produced on 14 July, and that the clean V8 package was generated on 24 July. The manuscript disclosure uses the same complete interval. On 28 July, V9 was prepared for arXiv and journal submission without substantive changes. This supplement expands the concise manuscript disclosure without changing the scientific content.

"Full workflow" here means a complete summarized audit trail. It does not mean disclosure of private model chain-of-thought, proprietary internal states, system prompts, or every low-level tool message. The primary records for the observable conversation are the two exported session files:

- `marginal-vs-conditional-error-rates-paper-draft-2026-07-24.md`; and
- `paper-revision-from-meeting-notes-2026-07-24.md`.

Selected initiating prompts, with one funding-application identifier redacted, are reproduced in Section 10; later directions and decisions are summarized in Sections 5 and 7. The two exports are not reproduced in this summarized supplement; they are retained by the supplement's author as non-public audit records under the filenames listed above. The exports remain the authoritative records if this summary differs from either export.

No patient-level or empirical data were used. Local administrative, costing, partner, and funding-application details were excluded. Project-relative scientific filenames are reported where needed for reproducibility; personal absolute paths are omitted.

2 Human and AI roles

2.1 Human direction and decisions

The research question and intended contribution originated with the human user: to turn a motivating adaptive-trial simulation and discussion into a short paper that would initiate debate about marginal versus selection-conditional error rates. Human contributors supplied the motivating concept note, proposal materials, preliminary R code, recorded discussions and dictated manuscript comments, coauthor annotations, and repeated version-specific instructions.

The human user and, where recorded, coauthors made consequential editorial and scientific decisions. These included the target audience and journal, the article's normative stance, the examples to retain, the final simulation setting, which derivations belonged in the paper, which reviewer suggestions to reject, the title and abstract wording, the table design, the author list and order, affiliations, literature positioning, and the user-requested responsibility wording. All listed manuscript authors approved the final manuscript before V9 submission preparation. In many rounds the AI proposed alternatives, but the user explicitly accepted or rejected them before a clean version was created.

The human contributors to the recorded workflow were not passive recipients of a one-shot AI draft. They supplied detailed oral and written revisions, challenged mathematical and interpretive language, requested new literature checks, corrected author and affiliation information, and reversed changes they considered too large or unclear. Examples are recorded in Section 7.

The 41 direct chat messages were only one input channel. A separate channel comprised nine audio-derived plain-text transcripts of recorded discussions, monologues, and dictated revision notes; under the counting rule in Section 4.2, the retained oral-transcript text was substantially longer than the direct-chat text.

2.2 AI-assisted work

AI systems were used to:

1. inspect local source materials and extract the motivating scientific argument;
2. search for, retrieve, compare, and verify scientific, journal-policy, and regulatory sources;
3. formulate mathematical definitions and derive or check Gaussian, multiple-testing, and selective-inference arguments;
4. design, write, revise, run, and validate R simulation code;
5. draft and repeatedly revise the manuscript, tables, captions, appendices, response letters, source logs, and package documentation;
6. generate tracked-change versions, accept or reject changes as instructed, compile LaTeX, and inspect logs and rendered PDFs;

7. organize specialist reviews of statistical methodology, mathematical validity, prose, bibliography, journal fit, and visual presentation; and

8. draft this supplementary workflow record.

AI systems did not qualify for authorship, did not submit the paper, did not communicate with a journal or regulator, and cannot accept responsibility for the work. Tool-based checks and model reviews were aids to human verification, not substitutes for it.

3 Models, agents, and tools

3.1 Recorded model labels

The main drafting and revision model was OpenAI GPT-5.6 Sol, recorded by the runtime as `openai/gpt-5.6-sol`; the interface sometimes described the reasoning setting as "max" or "maximum reasoning." The session exports also record work by `openai/gpt-5.6-luna`, `openai/gpt-5.6-luna-fast`, and `opencode-go/deepseek-v4-flash` for selected implementation, review, visual, or summary tasks. These names are service/runtime labels preserved from the session metadata; they should not be interpreted as independent human reviewers.

The same principal runtime, `openai/gpt-5.6-sol`, drafted this supplement on 24 July 2026 after one consolidated question round about format, level of detail, technical disclosure, sanitization, and responsibility wording.

3.2 Specialized agents

The OpenCode/OpenChamber workflow delegated bounded tasks to agents with the following recorded role labels. Their outputs were returned to the principal AI, reconciled with other evidence, and sometimes rejected by the user.

Agent or role label	Use in the workflow
`stat-reviewer` and fast variants	Reviewed statistical guarantees, simulation estimands, Monte Carlo reporting, novelty, practical interpretation, and comparison fairness.
`math-reviewer`	Checked probability identities, closure, Dunnett calibration, inverse-normal combination, Gaussian projections, truncation events, selective pivots, and edge cases.
`proof-creator`	Produced candidate proof text and derivations, including material later rejected or removed from the manuscript.
`proofreader` and fast variants	Checked grammar, notation, redline rendering, tables, citations, package consistency, and LaTeX artifacts.
`librarian`	Mapped relevant literature and verified bibliographic, journal-policy, and regulatory metadata.
`journal-editor-fast` and adversarial reviewer	Assessed journal fit, novelty, likely reviewer objections, and weaknesses in the manuscript's premise or comparisons.
`observer`	Inspected PDFs and source documents for layout, figures, table legibility, clipping, and page breaks.

Agent or role label	Use in the workflow
`fixer`	Implemented bounded code or document changes requested by the principal agent, including simulation revisions and figure regeneration.
`explorer`	Mapped local project files and identified likely source materials.
`oracle`	Supplied senior methodological checks on selected arguments and simulation designs.

The word "independent" in some historical agent prompts meant a separate model invocation or review pass. It did not establish statistical, institutional, or intellectual independence: agents could share model families, context supplied by the principal agent, or common failure modes.

3.3 Software and external tools

The workflow used file-search and file-reading tools; web retrieval and downloads; publisher, Crossref, PubMed, arXiv, FDA, EMA, ICH, and Wiley pages; base R; LaTeX, BibLaTeX/Biber, and `latexmk`; PDF inspection; and tracked-change tools including `latexdiff`. R was used for simulation and numerical checks. LaTeX tools were used to compile the manuscript, response letters, and this supplement. No generative-image model was used. Scientific figures were generated from simulation CSV files by AI-assisted R code under human direction.

4 Inputs, source control, and exclusions

4.1 Local scientific inputs

The main local inputs were:

- a four-page concept note containing the motivating five-treatment example and conceptual framing, with programme identifiers omitted from the manuscript workflow;
- an expanded research proposal and local bibliography used for definitions and literature leads, excluding administrative sections;
- supplied R files implementing exploratory combination, selective, helper, and Gaussian simulation calculations;
- nine audio-derived transcript files, together with companion summaries where available, dated 14–16 July and 22 July 2026;
- comments supplied by Zijun Gao, including PDF annotations and a text extraction of those comments;
- downloaded papers used for close comparison, especially Takahashi et al. (2022) and Sato et al. (2026); and
- every versioned package from the first draft through `agent_files/paperv9/`, including manuscripts, response records, code, numerical outputs, figures, provenance logs, and submission-preparation materials.

The final computational record is in `agent_files/paperv9/code/simulate.R` and `agent_files/paperv9/results/`. The clean manuscript is `agent_files/paperv9/manuscript.tex`. The cumulative source inventory is `agent_files/paperv9/source_log.md`.

4.2 Recorded oral input and quantitative accounting

Beyond the direct chat prompts, the manuscript workflow used nine audio-derived plain-text transcript files supplied by Qingyuan Zhao: two files dated 14 July, three dated 15 July, three untimestamped dictated revision-note files dated 16 July, and one monologue dated 22 July. One 14 July recording is a discussion between Qingyuan Zhao and Zijun Gao; the remaining files principally record Qingyuan Zhao's oral comments and instructions. The transcripts include broad framing and literature questions, line-by-line manuscript criticism, proposed wording, simulation choices, table and figure instructions, and explicit acceptance or rejection of AI suggestions. The AI was repeatedly instructed to address every substantive transcript comment, and versioned response letters mapped those comments to manuscript changes.

Six transcript files contain timestamps whose final positions sum to approximately 3 hours 41 minutes. The three 16 July files are untimestamped, so the total duration of the underlying recordings is longer but cannot be reconstructed exactly from the retained text.

Because the retained input is partly in Chinese, English/alphanumeric words and Chinese characters were counted separately rather than combined into a single multilingual word count. English/alphanumeric words were counted as sequences of Latin letters or numerals, allowing internal apostrophes; Chinese characters were counted individually. Each plain-text transcript file was counted once after removing timestamps and speaker labels, and companion summaries and duplicate JSON representations were excluded. Under this convention, the nine audio-derived transcripts contain 10,689 English/alphanumeric words and 24,833 Chinese characters. The 41 top-level direct human messages in the two exported sessions contain another 2,184 English/alphanumeric words and 502 Chinese characters after appended AI-to-subagent transcripts were excluded. The combined retained chat and oral-transcript record therefore contains 12,873 English/alphanumeric words and 25,335 Chinese characters. These are text counts, not model-token counts, and they exclude the separately supplied concept note, proposal materials, R code, transcript summaries, and coauthor annotations.

4.3 Online inputs

The AI searched or checked primary articles on adaptive treatment selection, combination testing, Dunnett procedures, selective inference, adaptive enrichment, randomization inference, adaptive experiments, econometrics, and bandits. Bibliographic metadata were checked against DOI, publisher, Crossref, PubMed, or arXiv records. Regulatory quotations were checked against the FDA adaptive-design guidance and the ICH E20 Step 2b draft. Journal and AI-disclosure requirements were checked against Wiley and *Statistics in Medicine* pages. Complete scientific citations appear in the V9 bibliography; the source log records key DOI and URL provenance.

4.4 Excluded material

No patient data were available or used. Budget, partner, grant-administration, and other confidential application material did not contribute to the manuscript. Absolute local paths and account-specific

information are omitted here. The workflow did not use AI to fabricate or alter empirical observations. The numerical results arose from synthetic Gaussian draws generated by the documented simulation.

5 Chronological workflow

The chronology below consolidates the two session exports and the versioned source logs. Dates refer to the recorded work period, not necessarily the date printed on every intermediate PDF.

Date	Stage	Human direction	AI-assisted work and outcome
13–14 July	Initial draft / Paper V1	Turn the motivating simulation and discussion into a short paper; ask one question round, then work autonomously.	Located and extracted the concept note and proposal; mapped literature and journal requirements; reconstructed a Gaussian design because complete original code was initially unavailable; wrote base-R simulation code; drafted and compiled the first manuscript package. Added statistical, mathematical, editorial, adversarial, and visual checks.
14 July	Initial computational revision	Continue after a reviewer produced no usable output.	Detected that the first independent-arm reconstruction did not reproduce the desired power ordering. Added shared-control and correlation-aware benchmark calculations. A smoke test accidentally overwrote canonical CSVs; the AI restored the fixed-seed two-million-repetition outputs and re-ran package gates.
14–15 July	Paper V2	Follow two recorded meetings, use supplied R functions, compare the requested procedures, and save a second version.	Audited the supplied combination, Gaussian selective, pooled, and stage-2-only methods. Rewrote the simulation on the participant-mean scale. Found that an exploratory Cauchy local-intersection calculation was mildly anti-conservative under shared-control dependence and replaced it in the manuscript's closed test with Gaussian Dunnett local tests. Added proofs, provenance, and a response-oriented preparation record.
15–16 July	Papers V3, V3.1, V3.2	Address all substantial comments from three transcript sets; produce a point-by-point response; then implement three further 16 July transcripts.	Reorganized the abstract and Introduction, clarified marginal and selection-conditional FWER, checked citations, and expanded technical derivations. Ran a disclosed effect grid and selected $\theta_1 = 0.25$ because it maximized the estimated selective-minus-combination joint-power difference on that grid; retained $\theta_1 = 0.30$ as sensitivity. Multiple reviewers found and corrected proof, notation, regulatory, and reporting issues.

Date	Stage	Human direction	AI-assisted work and outcome
22 July	Papers V4 and V4.1	Use a new monologue to simplify the paper, introduce one- and two-treatment continuation examples, and mark all changes; then roll back the numerical design to $N = 100$ and $\theta = (0.25, 0, 0, 0, -0.02)$.	Rebuilt the scientific narrative and simulation around top-one and unordered top-two selection with shared controls. Initially implemented a different numerical setting requested in the monologue, then regenerated all outputs after the explicit rollback. Added exact Dunnett and selective-Gaussian calculations and updated regulatory quotations.
22–23 July	Papers V5 and V5.1	Accept previous changes, streamline Sections 2–4, retain only new red/blue marks, rename Settings as Examples, improve tables and figures, and verify the package.	Created tracked revisions, updated table layouts and figures, reran the final simulation, and checked that the V5.1 outputs matched V5. The workflow repeatedly removed temporary clean sources because the user requested that only the marked manuscript be retained at those stages.
23 July	Papers V5.2–V5.4	Refine the selective-Gaussian explanation and tables; briefly move a derivation to Appendix A, then remove it; restore descriptive table headings and move FWER notation to captions.	Derived the top-one/top-two truncation law and checked it in a separate AI review pass. Added the derivation in V5.2, replaced it in V5.3 by a concise zero-correlation/joint-Gaussian explanation at the user's direction, and revised table headings and captions through V5.4.
23 July	Papers V6 and V6.1	Accept all changes into a clean paper; remove the preparation-record appendix; replace the AI byline with human authors; incorporate Zijun Gao's comments and improve tables.	Produced a clean manuscript, removed the earlier transparent-preparation appendix from the paper, set affiliations, revised wording and table structure, and moved Discussion material. The user's and one coauthor's recorded reviews determined the retained text at this stage.
23 July	Paper V7	Obtain statistical, mathematical, and proofreading reviews; implement a numbered set of proposed final revisions; make Zijun Gao first author and add Ting Ye as middle author.	Generated 13 revision groups, including stronger proof text, Monte Carlo reporting, literature additions, and submission statements. Specialist agents reviewed the mathematics and statistics. This deliberately expansive V7 served as a review vehicle rather than an automatically accepted final version.

Date	Stage	Human direction	AI-assisted work and outcome
23–24 July	Papers V7.1 and V7.2	Accept only specified V7 changes; reject overlarge proof and framing changes; delete Appendix A; characterize Takahashi and Sato carefully; revise the power language.	Implemented the user's itemized accept/reject decisions. Read the downloaded Takahashi and Sato papers, compared Sato with Freidling et al., Sampson and Sill, and Li et al., and added qualified citations in the Introduction, power discussion, and planning paragraph. Repaired tracked-change rendering and compiled cleanly.
24 July	Paper V8	Accept V7.2, create a clean manuscript, standardize multi-author citations, and add "full" to the responsibility statement.	Created `agent_files/paperv8/`, removed change tracking and temporary sources, set `uniquelist=false`, rebuilt the manuscript, updated the complete AI-use interval, and retained the final simulation artifacts unchanged. The final manuscript and responsibility wording were subsequently approved by all listed authors.
24 July	This supplement	Create a standalone TeX/PDF record; summarize the complete audit trail; name models and agents; sanitize paths; and initially mark author review as pending.	Inspected both exported sessions, V8, prior transparent-preparation appendices, and source logs; asked one consolidated question round; drafted, compiled, and arranged a separate proofreading pass. After the resulting corrections, Qingyuan Zhao approved the supplement and confirmed that he is its sole author.
28 July	Paper V9	Prepare the approved manuscript for arXiv and *Statistics in Medicine* without substantive changes.	Preserved the V8 scientific source, updated approval status, curated public reproducibility files, and created separate platform-specific submission bundles.

6 How AI affected the scientific content

6.1 Research question and framing

The central research question was supplied by the human user. AI systems helped translate it into formal definitions, the total-probability decomposition, a taxonomy separating selection-conditional error from the classical conditional-error function, and a discussion of clinical-trial interpretation. The final normative position—that selection-conditional risk is often relevant but is not an established regulatory requirement—was repeatedly directed and refined by the user.

6.2 Simulation design and code

The original motivating table did not provide enough information to reconstruct every parameter or algorithm. The first AI-assisted session therefore built a transparent Gaussian reconstruction and discovered that the desired power ordering depended on covariance and information assumptions. After the user supplied additional R code, AI systems audited the supplied methods and replaced the manuscript's exploratory Cauchy local intersection calculation with Dunnett-calibrated local tests because the former showed a reproducible finite-level global-null excess under shared-control dependence.

The final V9 simulation program, unchanged from V8, was largely written or rewritten by AI under human direction. It generates sufficient Gaussian sample means directly, compares a closed equal-weight inverse-normal combination test, a selective Gaussian test (with Holm adjustment for two selections), and a stage-2-only test (with closed step-down Dunnett for two selections), and reports marginal and selection-conditional FWER and power. The final design uses five treatments, $N = 100$ per observed group and stage, a principal vector $(0.25, 0, 0, 0, -0.02)$, a grid over θ_1, and a separate global null. The principal and global-null runs use two million repetitions; the other grid values use 200,000. The reproducibility command is

```
Rscript code/simulate.R 2000000 20260722 200000
```

The post-hoc selection of $\theta_1 = 0.25$ during an earlier version was AI-executed from a disclosed effect grid and was reported in the versioned source record. The user later explicitly chose the same setting when rolling V4 back to $N = 100$ and $(0.25, 0, 0, 0, -0.02)$. This history matters because the principal configuration was not selected independently of the observed simulated power contrast.

6.3 Mathematical arguments

AI systems formulated, derived, or checked several mathematical components: strong marginal FWER under closure, adaptive inverse-normal combination, shared-control Dunnett probabilities, the Gaussian projection used for the selective test, the top-one and unordered top-two selection inequalities, the truncation threshold, boundary-null uniformity and composite-null super-uniformity, Holm adjustment under dependence, and the stage-2-only Dunnett argument.

The rigor of these arguments evolved. Early reviewers accepted concise validity statements; later reviewers identified missing monotonicity, conditioning, latent-variable, and composite-null details. V7 contained a long AI-generated proof appendix intended to close those gaps. The user rejected that expansion as disproportionate to the short note and directed its deletion. The final V9

manuscript therefore contains the concise mathematical presentation retained at the user's direction and approved by all listed authors, rather than the most expansive AI-generated proof text. The rejected proof material remains part of the session record, not part of V9.

6.4 Literature and citation positioning

AI systems carried out a broad literature search and metadata audit, including adaptive treatment selection, conditional error functions, selective inference, randomization inference, adaptive enrichment, econometrics, and bandits. The AI identified Takahashi et al. (2022) and Sato et al. (2026) as close clinical-trial antecedents. At the user's request, the downloaded papers were read more closely. The final wording distinguishes Sato et al.'s design-specific, selection-conditioned stage-1 randomization procedure from the more general selective-randomization framework of Freidling et al. and from the paper's own general contrast between marginal and selection-conditional FWER.

Citation metadata, titles, DOI records, and regulatory quotations were repeatedly checked, but AI discovery and verification can still err. The user specifically challenged the Sato positioning and the unusual rendered Friede citation, leading to further analysis and correction. The V9 bibliography and source log, rather than an unverified model-generated list, are the final reference record.

6.5 Writing and editing

Most manuscript prose was initially generated or substantially rewritten by AI. Human revision was extensive: the user supplied sentence-level wording, reorganized the Introduction and Discussion, controlled the amount of mathematical detail, accepted or rejected reviewer proposals, and iteratively accepted revisions and authorized the generation of successive clean versions. AI systems also generated response letters, source logs, abstract counts, table/caption revisions, tracked changes, and package documentation. Earlier preparation-record appendices were AI-drafted and later removed from the main manuscript at the user's direction; the present supplement restores a fuller record outside the article.

7 Evidence of human control and rejected AI suggestions

The following examples show that AI outputs were treated as proposals rather than final authority:

- The user rejected changing the title to include "selection-conditional," retaining "On marginal and conditional error in adaptive clinical trials."
- The user rejected much of the large V7 proof and framing expansion, including the new technical appendix, the FDA 0.025 discussion, tie-breaking language, and several broad assumption paragraphs.
- The user required the V5.2 selective-Gaussian derivation to be moved to an appendix, then in V5.3 required that appendix to be deleted and the main text reduced to the zero-correlation and joint-Gaussian explanation.
- The user rejected simplified FWER table headers in V5.3, restored descriptive headers in V5.4, and then made additional row/column-label corrections in V6.1.

- The user replaced an AI-proposed final Discussion paragraph with supplied wording and directed where literature paragraphs should move.
- The user challenged the phrase "not power-matched." The AI agreed that it was inappropriate for procedures targeting different guarantees, and the sentence was replaced.
- The user asked for a more careful account of Sato et al. before permitting manuscript edits. This led to a separate comparison with Freidling et al., Sampson and Sill, and Li et al.
- The user set author order and affiliations: Zijun Gao first, Ting Ye in the middle, and Qingyuan Zhao last.
- The user requested that the authors take "full responsibility" in V8.

These decisions do not by themselves validate every retained claim, but they demonstrate material human control over the scientific and editorial outcome.

8 Verification, review loops, and recoveries

8.1 Computational checks

The final R program contains implementation diagnostics for Dunnett interpolation, the two-arm critical value, marginal FWER, singleton and setwise selection-conditional calibration, and marginal–conditional decompositions. All rows in the final validation file passed. These diagnostics support code correctness for the simulated configurations; they are not mathematical proofs of strong control over all parameter values. Analytical arguments provide the claimed guarantees under the manuscript's Gaussian assumptions.

Final package checks included fixed-seed reruns, comparison of CSV outputs across versions, abstract counts, citation and cross-reference checks, LaTeX warning and overflow checks, and PDF inspection. In separate AI review passes, statistical and mathematical agents recomputed or spot-checked reported quantities and Monte Carlo standard errors.

8.2 Review loops

Review findings were not accepted automatically. The principal AI reconciled mathematical, statistical, editorial, and user feedback; the user then selected the changes to retain. Some reviewers initially found no issue where later reviewers found proof gaps or reporting problems. This inconsistency is part of the reason that the workflow used repeated review and executable checks rather than relying on a single model judgment.

8.3 Recorded failures and recoveries

The session exports record several failures or limitations:

1. An early statistical-review invocation returned no useful output; the principal agent continued with other verification and later re-ran reviews.

2. A smoke test wrote to the normal results directory and overwrote canonical CSV files. The AI recognized this, restored the intended fixed-seed two-million-repetition outputs, and repeated the final gates.

3. Some specialist agents could not write their requested report files because of permission ordering. Their findings were returned inline instead.

4. Shell policy blocked one cleanup command; cleanup was completed using allowed file operations.

5. `latexdiff` produced nested-markup failures, joined words, malformed equations, and table artifacts in several versions. The AI manually repaired tracked rendering and repeatedly recompiled and proofread it.

6. One provider request exceeded an attachment-size limit, causing media to be removed from that context. Subsequent checks relied on smaller or source-based inputs.

7. The main AI occasionally offered language later judged inaccurate or unhelpful, including "power-matched," overlarge proof expansions, and imprecise citation characterizations. The user or later reviewers corrected these.

These incidents illustrate why the existence of an AI review log is not itself evidence of correctness. Reproducible code, primary sources, mathematical checking, and author review remain necessary.

9 Version and artifact record

The workflow intentionally retained multiple versions because the user requested tracked changes and selective acceptance. The main stages were `paper/` or Paper V1, followed by `paperv2`, `paperv3`, `paperv3.1`, `paperv3.2`, `paperv4`, `paperv4.1`, `paperv5`, `paperv5.1`, `paperv5.2`, `paperv5.3`, `paperv5.4`, `paperv6`, `paperv6.1`, `paperv7`, `paperv7.1`, `paperv7.2`, `paperv8`, and `paperv9`. Some versions deliberately omitted clean TeX files or response letters in accordance with user instructions. Temporary comparison sources were removed before finalizing those packages.

The deliverable V9 package contains:

- the clean LaTeX manuscript and PDF;
- the final BibLaTeX bibliography;
- base-R simulation code;
- CSV outputs, validation records, and PDF/TIFF figures;
- an abstract word-count record; and
- README and source/provenance documentation;
- an arXiv source and ancillary-file bundle; and
- *Statistics in Medicine* submission files and a reproducibility archive.

V8 accepted the V7.2 tracked changes and changed only citation-list disambiguation, the final responsibility wording, and the corrected 13–24 July 2026 AI-use interval. V9 preserves that approved scientific manuscript and adds only submission formatting, administrative metadata, workflow-approval status, and reproducibility packaging. Its numerical inputs, algorithms, estimates, CSV outputs, validation records, and figures are inherited unchanged from the verified V5.1 computational run.

10 Selected initiating prompts and record pointers

The prompts below are reproduced with one limited redaction of a funding-application identifier because they defined major workflow transitions. Redactions are shown in square brackets. The full non-public exports contain all later messages, including detailed acceptance and rejection decisions.

Original paper-development prompt

```
We would like to turn the simulation example and discussion in the [redacted funding application]
to a short paper (about 5-6 pages) to be submitted to Statistics in Medicine or Biometrics. The
main purpose is to briefly review the literature, introduce the post-selection angle, and, most
importantly, initiate the discussion on marginal vs conditional error rates in the biostatistics
and clinical trial community. You can ask me one round of questions in the beginning (as many
questions as you want), then you will work autonomously to draft that manuscript until it is ready
for submission. The manuscript should be saved to the `paper/` folder.
```

Paper V2 prompt

```
Follow the instructions and discussion in the meeting recordings in `meeting notes/` (there are
two meetings; use `transcript.txt` and `summary.md` in the subfolder). Draft a 2nd version of the
paper and save it to `paperv2/`. You have one chance to ask me as many questions as you like, then
you will work autonomously until you obtain a paper that matches all the requirements specified in
the meeting transcripts.
```

Paper V3 prompt

```
Revise the V2 manuscript and save it to `agent_files/paperv3`. This time include a point-by-point
response letter to ALL the substantial queries/comments I've made in the audio transcripts in
`2026-07-15_*`. Again you can ask me exactly one round of questions (as many as you like), you
will then work autonomously.
```

Paper V4 prompt

```
Find the transcript from my monologue today in this folder and use it to revise the paper. Make
sure that all points are addressed and all changes are highlighted. Save it to
`agent_files/paperv4`. You can ask me one round of questions.
```

Final clean-version prompt

```
Okay, make that change. All other changes can also be accepted. Make a clean `paperv8`, no need to
run any proofreading.
```

Prompt for this supplement

```
Attached are transcripts of two other OpenCode/OpenChamber sessions that produce the manuscript in
`agent_files/paperv8`. Write a supplementary document disclosing the full workflow and how AI has
been used in writing this paper and this supplement that you will write (some previous versions of
the paper have this in Appendix). You can ask me one round of questions before writing this
document autonomously.
```

The user's answers in the permitted question round requested: a standalone TeX/PDF supplement in the V8 package; a complete summarized audit rather than full transcript reproduction; explicit naming of models, agents, tools, failures, and recoveries; sanitized source paths and grant details; and wording that author review was pending at the drafting stage. Qingyuan Zhao subsequently reviewed and approved the revised supplement on 24 July 2026.

11 AI use in preparing this supplement

OpenAI GPT-5.6 Sol (runtime identifier `openai/gpt-5.6-sol`, maximum reasoning setting) was used on 24 July 2026, under human direction, to inspect the two exported session records, read V8 and earlier provenance/preparation records, ask one consolidated question round, organize the chronology, draft this document, compile it, and request a separate proofreading review. The supplement contains no new statistical analysis and does not change the manuscript's scientific content, code, or numerical results; preparation of the supplement prompted the date correction in the manuscript's AI disclosure.

Qingyuan Zhao, the sole author of this supplement and the author who directed the recorded workflow, reviewed and approved it on 24 July 2026 after the corrections arising from the separate AI proofreading pass. No approval from the manuscript's other authors is required for this separately authored supplement. Qingyuan Zhao, not the AI systems, takes full responsibility for this supplement. Responsibility for the manuscript itself remains with the manuscript's listed human authors, as stated in its disclosure.

12 Limitations of this record

This record is necessarily a summary. Session exports may omit media removed because of provider limits, tool-internal metadata, and ephemeral temporary files. It reports observable prompts, responses, tool actions, and retained artifacts, not private model reasoning. The audio-derived transcripts may contain transcription or speaker-attribution errors. The reported English/alphanumeric-word and Chinese-character counts document only the size of the retained direct human chat messages and the text of the nine transcripts; they do not by themselves establish speaker identity, authorship of the manuscript, or the correctness of any statement. Some model and agent labels are platform-specific aliases. Web pages and publication metadata can change after the recorded access dates. Review agents were AI systems and should not be described as independent human peer review. Finally, a complete process record cannot establish that every scientific statement is correct; that requires author verification and, ultimately, external peer review.