CLINICAL STUDY DOCUMENT

Statistical Analysis Plan

EVOLVE-CXR

Official title

EvoXRare-Assisted Chest Radiograph Interpretation in Emergency and Critically Ill Adults: A Nationwide Multicenter Randomized Controlled Trial

Unique protocol ID	WHUH-CXR-AI-RCT-2026-01
Document version	English review draft 1.0
Document date	November 2026
ClinicalTrials.gov ID	To be inserted after assignment.
Sponsor	Union Hospital, Tongji Medical College, Huazhong University of Science and Technology
Principal investigator	Heshui Shi, Department of Radiology
Study coordinator	Xiaoyu Han, Department of Radiology
Funding	National Natural Science Foundation of China (82672507); Interdisciplinary Research Program of HUST (2026JCYJ026).

Study synopsis

A pragmatic, nationwide, multicenter, individually randomized, two-arm superiority trial evaluating whether EvoXRare-assisted chest radiograph interpretation reduces patient-level clinically important diagnostic misses compared with routine physician interpretation alone. Approximately 70,000 adults will be randomized 1:1 across 15–20 hospitals in China. The primary endpoint is the patient-level miss rate for prespecified clinically important abnormalities on the index chest radiograph, determined by blinded central adjudication. Key secondary outcomes address critical and rare findings, diagnostic and reporting timeliness, false-positive findings, additional imaging, workflow performance, and selected short-term clinical outcomes.

Document status and provenance

Prepared to accompany the ClinicalTrials.gov registration draft for EVOLVE-CXR. Core design, population, intervention, sample size and endpoint definitions are aligned across the registration, protocol and statistical analysis plan. Administrative fields such as final site list, local approvals, consent/waiver determinations and the assigned NCT number must be verified before activation or public posting. This working draft does not itself establish ethics approval, prospective registration, or site activation.

1 Scope, design and analysis principles

This SAP accompanies protocol WHUH-CXR-AI-RCT-2026-01. It formalizes the statistical analysis for the patient-level clinical RCT of EvoXRare-assisted chest radiograph interpretation. It does not apply to any separate retrospective model-development study, silent reader experiment or nonrandomized workflow run-in.

The trial uses individual 1:1 allocation to EvoXRare-assisted interpretation or standard interpretation, with central concealed randomization stratified by center and care setting. The target is 70,000 participants across approximately 15–20 centers. The primary analysis follows the intention-to-treat principle. The SAP should be approved and version-locked before treatment-group outcome results are unblinded to the analysis team.

1.1 Analysis populations

Intention-to-treat (ITT)	All validly randomized participants, analyzed according to assigned group regardless of AI processing failure, nonadherence, crossover, physician override or subsequent care. Primary efficacy population.
Evaluable reference-standard set	Randomized participants for whom the primary endpoint can be adjudicated under the prespecified reference-standard rules. Used for the primary binary endpoint with missing-outcome sensitivity analyses across the full ITT population.
Per-protocol supportive set	Participants without prespecified major deviations affecting assigned interpretation exposure or endpoint ascertainment. Supportive only; does not replace ITT.
Safety / workflow set	All randomized participants with available index workflow data; denominators are stated separately for each outcome.
Finding-level set	Prespecified abnormalities with an adjudicated finding-level reference standard. Analyses account for multiple findings within participant and clustering by reader/center as appropriate.

1.2 Primary estimand

Population	All randomized participants in the EVOLVE-CXR trial.
Treatment contrast	Assignment to EvoXRare-assisted interpretation versus standard interpretation.
Outcome	At least one prespecified Category A or B clinically important abnormality present on the adjudicated reference standard but absent from the initial finalized chest radiograph report.
Summary measures	Adjusted and unadjusted group risks; adjusted relative risk preferred where estimable; absolute risk difference (AI minus control); 95% confidence intervals.
Intercurrent events	Treatment-policy approach: AI failure, nonadherence, crossover, physician override and subsequent management remain part of the assigned-strategy effect.
Primary horizon	Index radiograph interpretation through completion of the prespecified reference-standard/adjudication window.

2 Endpoint derivation and coding

2.1 Primary endpoint

Code the primary event as 1 when the blinded central adjudication determines that at least one prespecified Category A or B abnormality was present on the index chest radiograph and that abnormality was not documented in the initial finalized report produced under the randomized strategy. Code 0 only when the primary endpoint is evaluable and no such miss is present. Do not code unresolved reference-standard status as 0.

Index report	The first finalized radiology report produced under the assigned interpretation strategy before any subsequent radiology addendum triggered by later information.

Clinically important abnormality	Finding included in the locked Category A or B list in the adjudication manual.
Miss	Adjudicated abnormality present on the index image but not reported in the initial finalized report at the required level of specificity defined in the manual.
Multiple abnormalities	Participant counts once for the patient-level primary endpoint, regardless of number of missed Category A/B abnormalities. Finding-level counts are analyzed separately.
Later addendum	A finding added after initial finalization remains a primary miss if the adjudication manual defines it as absent from the initial report; the addendum timing is retained for process analyses.
Indeterminate status	If evidence is insufficient to determine presence/absence or report concordance, code primary endpoint as unresolved and handle under missing-outcome procedures.

2.2 Key secondary endpoints

- Category A patient-level miss rate.
- Rare/easily missed finding sensitivity and detection rate overall and by prespecified class.
- Time from image availability to first documented recognition of a critical abnormality.
- Time from image availability to initial finalized report.
- Time from image availability or recognition to effective critical-result communication, depending on the locked operational definition.
- Patient-level false-positive clinically important finding rate.
- Radiology recall, repeat radiograph, CT, ultrasound or other additional imaging after the index interpretation.
- AI suggestion acceptance, rejection, editing and high-risk prompts that changed the final report (intervention arm descriptive analyses).
- AI technical availability, failure and latency.
- Exploratory ICU transfer, length of stay and in-hospital mortality.

3 Sample-size assumptions and event-rate verification

Selected target	70,000 randomized participants; approximately 35,000 per arm.
Primary assumptions	Control miss rate 1.5% versus EvoXRare 1.2%; absolute reduction 0.3 percentage points; relative reduction 20%.
Design parameters	Two-sided alpha 0.05; 90% power; equal allocation; approximately 10% allowance for non-evaluable primary outcome or major data incompleteness.
Approximate calculation	Approximately 62,000 evaluable participants under a two-sample proportion approximation. Inflating by approximately 10% gives about 68,900; operational target rounded to 70,000.
Sensitivity	If control is 2.0% versus 1.6%, fewer participants are needed; if 1.0% versus 0.8%, more than 90,000 may be required. These are design scenarios, not observed treatment effects.
Blinded verification	Aggregate, treatment-blinded data may be used to verify overall event prevalence and endpoint evaluability. Any sample-size amendment must not use unblinded between-group effect estimates unless a separate prespecified adaptive design amendment is approved.

4 Primary analysis

4.1 Main model

The primary comparison will estimate the effect of randomized assignment on the patient-level clinically important abnormality miss rate. The preferred model is a modified Poisson regression with log link and robust variance to estimate

an adjusted risk ratio directly. Prespecified adjustment factors are treatment group, center, care setting (emergency/resuscitation vs ICU/other critical care), and portable versus fixed radiography if this factor was used in randomization or is consistently available. Age and sex may be included as prespecified precision covariates if completeness is adequate. The final covariate set is frozen before unblinding.

Report the adjusted risk ratio, group-specific marginal standardized risks and absolute risk difference (EvoXRare minus control), each with 95% confidence intervals, together with an unadjusted comparison. A negative risk difference favors EvoXRare. Statistical significance for the primary hypothesis uses a two-sided alpha of 0.05.

4.2 Model fallback and diagnostics

Inspect convergence, sparse center strata, collinearity and implausible fitted values. If the modified Poisson model is not estimable, use a logistic regression with the same covariates and report standardized absolute risks, risk difference and odds ratio, explicitly identifying the fallback. If the full adjusted model fails, use a reduced model retaining treatment and randomization stratification factors; if that also fails, report the unadjusted contrast. Do not select the model yielding the most favorable P value.

5 Reader, center and case clustering

The primary patient-level model treats center through prespecified adjustment and uses robust variance. Because multiple physicians interpret multiple cases and individual readers may work across shifts or settings, supportive multireader/multicase analyses will assess whether the treatment effect is robust to reader-level clustering when reliable reader identifiers are available.

Primary supportive structure	Cross-classified or mixed-effects logistic/Poisson model with random intercepts for reader and center where estimable; treatment effect remains fixed.
Alternative marginal approach	Generalized estimating equations with center/reader clustering where the correlation structure and data density support estimation.
Rare finding analyses	Use participant-clustered or reader/center-aware models; for extremely sparse findings, emphasize exact or profile-likelihood confidence intervals and descriptive rates rather than unstable adjusted models.
Repeated examinations	Only one index encounter per participant enters the primary randomized analysis. Any recurrent-encounter exploratory dataset is analyzed separately and does not alter the primary denominator.

6 Multiplicity and hierarchy

The primary endpoint is the single confirmatory efficacy endpoint. Key secondary endpoints will be reported with effect estimates and 95% confidence intervals. Unless the final protocol explicitly adopts a fixed hierarchical sequence before unblinding, secondary P values are interpreted descriptively/exploratorily and are not used to claim multiplicity-controlled confirmatory superiority.

If a hierarchy is adopted, the recommended order is: (1) primary clinically important abnormality miss rate; (2) Category A critical finding miss rate; (3) report turnaround time or a single prespecified workflow endpoint. Any such sequence must be approved before unblinding and documented identically in protocol, SAP and registry when required.

7 Missing data and indeterminate reference standards

7.1 Missing primary outcome

Missing or indeterminate reference-standard status is not treated as absence of disease. Summarize the proportion and reason for non-evaluable primary outcomes by randomized group and center. The main model is fit to participants with an adjudicated primary endpoint, while the randomized ITT population defines the target estimand and missing-outcome sensitivity analyses assess robustness.

If primary-outcome missingness exceeds a trivial level, use multiple imputation under a missing-at-random assumption with treatment group, center, care setting, age, sex, portable/fixed acquisition, report timing, follow-up availability, and other prespecified predictors of endpoint observability and abnormality status. Use at least 50 imputations; increase to 100 if the fraction of missing information is material. Pool estimates using Rubin's rules.

7.2 Sensitivity to missing-not-at-random

Conduct tipping-point or delta-adjustment analyses that vary the assumed odds of a primary miss among unresolved participants separately by randomized group. Also present best/worst plausible assignment bounds and a complete-adjudication analysis. Preserve all known endpoint states; do not overwrite an observed event through imputation.

7.3 Missing covariates

Do not create an "unknown" category for continuous variables solely to avoid missingness. Use multiple imputation for incomplete prespecified covariates when required for the adjusted model, or omit nonessential precision covariates if missingness materially threatens model stability; treatment and randomization stratification variables should be obtained from the randomization system and therefore are expected to be complete.

8 Secondary outcome analyses

Category A miss rate	Analyze similarly to the primary binary endpoint. If events are sparse, use a reduced adjustment set or marginal risk difference with robust/exact confidence intervals.
Rare/easily missed finding sensitivity	Report numerator/denominator for each prespecified finding class. Use finding-level clustered models or participant-resampled bootstrap confidence intervals; do not pool heterogeneous findings without presenting class-specific results.
False-positive important finding rate	Patient-level binary analysis with the same core adjustment factors as the primary endpoint.
Report turnaround time	Summarize median, IQR and selected percentiles. Compare groups using quantile regression or accelerated failure-time/survival methods as prespecified after inspecting the data-generating structure without treatment labels.
Critical recognition / notification time	Analyze among participants with adjudicated critical findings and relevant timestamps. Report medians and time-to-event curves where appropriate; clearly label the conditional denominator.
Additional imaging / recall	Binary patient-level outcomes analyzed with modified Poisson or logistic models; counts may use negative-binomial or Poisson models with robust variance if overdispersion is acceptable.
Length of stay	Exploratory; summarize distribution and use robust rank/quantile or count/time models as appropriate. Account for in-hospital death in interpretation.
In-hospital mortality / ICU transfer	Exploratory binary outcomes with risk differences and 95% confidence intervals; the trial is not powered to establish mortality benefit.
Technical performance	Descriptive rates of successful processing, failure mode, latency, report availability and actual viewing; intervention arm primary, with site-level summaries.

9 Prespecified subgroup and heterogeneity analyses

Subgroup analyses are exploratory unless otherwise stated. Estimate treatment-by-subgroup interactions in the primary model rather than comparing significance separately within each subgroup. Report subgroup-specific adjusted risks and treatment effects with 95% confidence intervals.

- Care setting: emergency/resuscitation versus ICU/other critical care.
- Portable versus fixed chest radiography.
- Reader experience category (for example junior versus senior) where a harmonized definition is available across centers.
- Presence versus absence of any Category A finding on adjudication.

- Rare/easily missed finding present versus other clinically important finding categories, recognizing post-baseline/reference-standard conditioning and interpreting cautiously.
- Center volume category or geographic region if prespecified before unblinding and sufficiently populated.

No subgroup will be promoted to a separate primary claim merely because its P value is smaller. Sparse interaction strata may be collapsed only according to prespecified rules or presented descriptively.

10 Protocol deviations, crossover and per-protocol analyses

The ITT analysis retains all validly randomized participants according to assignment. Prespecified major deviations include use of an unapproved model version, control clinician access to EvoXRare before initial report finalization, preventable failure to provide assigned AI after successful technical processing, randomization of an ineligible participant whose index exam could not meaningfully enter the trial, and material absence of required source data caused by trial-process failure.

A supportive per-protocol analysis excludes participants with major deviations judged to compromise assigned strategy exposure or endpoint ascertainment before treatment-group outcome comparison. Exclusion decisions are made using coded groups and documented rules. As-treated comparisons are descriptive because treatment receipt is post-randomization.

11 Safety and harm analyses

Potential AI-related harm includes a missed critical abnormality after an erroneous or falsely reassuring AI suggestion, inappropriate escalation after a false-positive AI prompt, clinically relevant delay attributable to the AI workflow, or other serious diagnostic harm judged plausibly related to trial participation. Events are reviewed according to the monitoring charter and summarized by randomized group.

For suspected serious AI-related harm, report the event narrative category, adjudicated abnormality, AI output, physician action, relevant timestamps and clinical consequence. The causal relationship between AI output and clinical outcome is adjudicated separately from whether the chest radiograph finding itself was missed.

12 Data quality and derivation rules

- Use a unique participant ID linked to center and randomization timestamp.
- Store the initial finalized report separately from later addenda or amendments.
- Preserve raw image availability, report finalization, critical recognition and communication timestamps; do not replace missing times with arbitrary midpoint values.
- Archive AI software version, processing status, report output, access/view logs and physician edit history when available.
- Adjudication decisions and supporting evidence are versioned and locked before treatment labels are released to the primary analysis team.
- All derived variables are reproducible from a controlled data dictionary and analysis code.

13 Statistical reporting

Participant flow	Screened, randomized, assigned, received allocated workflow, primary endpoint evaluable, analyzed, and major protocol deviations by group and center.
Baseline characteristics	Age, sex, care setting, portable/fixed acquisition, reader experience where available, center and other prespecified variables. No significance testing of randomization balance.
Primary outcome	Observed events and denominators, adjusted/unadjusted relative and absolute effects, 95% CIs, model diagnostics and missing-outcome sensitivity analyses.

Secondary outcomes	Effect estimates with 95% CIs, clearly stated denominators and time origins; identify conditional analyses.
Finding-level results	Category A/B and rare/easily missed findings with verification coverage and clustered uncertainty estimates.
Harms and implementation	AI failures, crossover, overrides, serious suspected AI-related harms, additional imaging and workflow metrics.
Subgroups	Forest plot of prespecified interaction estimates; exploratory labeling.

14 Analysis timing, software and reproducibility

The primary analysis begins only after the final participant has completed the prespecified endpoint ascertainment window, central adjudication is sufficiently complete, protocol deviations are classified, the database is locked, and the SAP has been approved. Treatment codes remain masked to programmers preparing derivations and shells when feasible.

Archive statistical software and package versions, random seeds where applicable, data dictionaries, analysis scripts, derived datasets, endpoint decisions, adjudication exports, database-lock dates and final output. Any post-lock deviation from the SAP is dated, justified and reported transparently in the primary publication.

15 Planned tables and figures

- CONSORT participant-flow diagram.
- Baseline characteristics by randomized group.
- Primary endpoint table with adjusted and unadjusted effects.
- Category A and rare/easily missed finding performance table.
- Workflow and time-to-recognition/reporting table.
- False-positive, additional imaging and recall outcomes table.
- AI technical performance and physician override table.
- Serious suspected AI-related harm and protocol deviation table.
- Prespecified subgroup forest plot.
- Center-level descriptive funnel or caterpillar display as exploratory quality assessment, not a center-ranking inferential test.

16 Methodological references

1. Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159:702-706.
2. International Council for Harmonisation. E9(R1): Addendum on Estimands and Sensitivity Analysis in Clinical Trials. 2019.
3. Liu X, et al. CONSORT-AI extension. Nat Med. 2020;26:1364-1374.
4. Cruz Rivera S, et al. SPIRIT-AI extension. Nat Med. 2020;26:1351-1363.