A Privacy-Preserving Replication Pack for Environmental Health and Development Research: Standards and Integrity Checks for AI-Assisted Work with Confidential Data

Young Yun (youngyun@umd.edu)

Over the last decade, many researchers have improved transparency and reproducibility by sharing data and code, writing pre-analysis plans, and encouraging replications. These practices have made empirical research more credible and more useful for policy. But applied research is now at a turning point. More and more high-value work uses large administrative and health datasets that cannot be publicly shared, complex data pipelines that are hard to document and expensive to rerun, and AI-assisted tools that can speed up tasks like data cleaning, text extraction, coding, and evidence synthesis. These changes can produce faster and more detailed insights, but they also create new questions: How can we keep research trustworthy and reproducible when the underlying microdata are confidential? How do we balance privacy with transparency? And how do we manage new risks introduced by AI tools and complex computation?

These questions are especially important in environmental health and development research. Many studies use sensitive health or administrative data to estimate pollution burdens, evaluate public health programs, or measure benefits of risk-reduction interventions. These analyses often inform real decisions about regulations, budgets, and service delivery; However, data access restrictions can make it hard for others to verify results. At the same time, AI-assisted steps can add new failure points, such as unclear data provenance, unstable outputs, and hidden intermediate steps.

Privacy-Preserving Republication Pack: PPRP

The key contribution is a simple standard for what a research team should provide when the raw microdata cannot be shared. The goal is not to claim that others can perfectly reproduce every number without the original data. Instead, the goal is to make the work auditable and credible: others should be able to understand the pipeline, return key steps using a privacy-safe substitute dataset, and see how sensitive the main results and policy conclusions are to privacy protections.

In short, a PPRP is a package of materials that a project delivers alongside a paper or report. It helps reviewers, replication teams, or policy users answer basic questions:

- What data sources were used and how were variables created?
- What code produced the main tables and figures?
- What is reproducible without the confidential microdata?
- If a privacy-safe substitute dataset is used, how close are the results?
- If exact verification requires restricted access, what is the pathway for secure verification?

This approach is practical because it matches real constraints. Many institutions cannot release microdata, but they can release code, documentation, and a privacy-safe substitute dataset. They can also provide a clear path for secure verification by approved auditors or replication units.

PPRP components

The proposed PPRP has four required components, and one strongly recommended component:

(1) **Executable code and a fixed computing environment**

The pack should include end-to-end code that runs the analysis pipeline, including cleaning, variable construction, estimation, and outputs (tables and figures). It should also include a fixed computing environment to reduce reproducibility failures from version drift. This can be as simple as:

- A package list with version (e.g., R regression info, Python requirements), and
- A clear run script (what to run, in what order), and
- Random sees and deterministic settings where possible.

This element addresses a common problem in computationally intensive work: code may exist, but others cannot run it reliably because environments differ.

(2) **Codebook plus a machine-readable transformation ledger**

A standard codebook is helpful but often not enough. Many reproducibility failures come from undocumented transformations: merges, filters, recoding rules, and derived measures that change the analysis sample or variables. Therefore, the PPRP requires two documentation layers:

- Codebook: definitions of variables, units, inclusion/exclusion rules, missingness treatment, and key assumptions.
- Machine-readable transformation ledger: a structured record listing every major step that transforms raw inputs into analysis variables.

The transformation ledger is the backbone of auditability. It should record, in a consistent format:

- Data source and version (what file or table, when accessed)
- Transformation type (merge/filter/recode/derive/aggregate/geospatial join)
- Parameters (thresholds, join keys, buffer sizes, time windows)
- Outputs created (variable names, dataset names), and
- Quality checks (counts before/after, missingness changes, outlier flags)

Why machine-readable? Because it can be checked programmatically and it integrates well with AI pipelines and modern data workflows. It is also easier to reuse across projects and easier for institutions to standardize.

If AI is used, the ledger should also log AI-related details. For example, if a model is used to classify text fields or code open-ended responses, the ledger should capture:

- tool/model and version,
- Prompt template or labeling rules,
- Validation approach (human review, inter-coder checks, spot checks), and
- How human edits were applied.

This directly supports the call's focus on provenance and explainability in AI-assisted analysis.

(3) A privacy-safe substitute dataset

When microdata cannot be shared, the pack should include a dataset that allows others to rerun key steps. This substitute dataset is not expected to be identical to the true confidential dataset. Instead, it should allow:

- Pipeline testing (does the code run?),
- Method checking (does the estimator behave as described?), and
- Sensitivity assessment (how much do key results change under privacy-safe data?).

Depending on the study, the substitute dataset may be:

- Aggregated data with suppressed small cells and clear rules for aggregation,
- Synthetic microdata generated using a disclosed method, or
- A hybrid approach (public components + aggregated confidential components).

The point is to provide an input that supports verification while respecting privacy.

(4) Secure verification pathway

Some claims cannot be verified without the original microdata. The PPRP should therefore include a clear pathway for secure verification, such as:

- Restricted access through a secure enclave,
- A data use agreement (DUA) process for qualified verifiers, or
- Verification by an internal or third-party replication unit that can access the data and report results.

This pack clarifies how verification can happen safely.

(5) A short replication summary report

The PPRP works best with a short report that tells readers what is reproducible and what is not. This report should include the integrity checks described next. It should also clearly state:

- Which outputs should match exactly (e.g., figures produced from public data),
- Which outputs should match approximately (e.g., results from substitute data), and
- Which outputs require restricted verification.

Integrity Checks: the key novelty

The most distinctive part of this proposal is a small set of integrity checks that make the privacy-replicability trade-off measurable. Too often, studies simply say data are confidential, and the reader cannot tell what means for trust and reproducibility. Integrity checks provide a practical solution: they quantify how privacy protections affect estimates and policy conclusions.

The integrity checks have three parts:

(1) **Estimate and inference stability**

This check asks 'are the key estimates stable when we rerun the analysis using the substitute dataset or privacy-preserving variants?' It includes:

- Changes in core coefficients or treatment effects (Δβ),
- Changes in standard errors and confidence intervals, and
- Threshold stability for inference, including whether results cross a pre-specified significance threshold (e.g., whether p-value stay below or above α such as 0.05)

The purpose is not to over-focus on p-values. Rather, many research and policy workflows treat statistically significant vs. not as a decision input. If privacy protection changes a result from significant to not significant, that is important to report clearly.

(2) **Decision stability (policy-relevant reproducibility)**

Decision stability asks: do the main policy conclusions change under the privacy-safe substitute dataset or under reasonable privacy-driven perturbations? Many policy users care more about conclusions than about exact coefficients. Examples include:

- Whether an intervention is still recommended under plausible cost assumptions,
- Whether the same high-risk groups remain the priority, or
- Whether the ranking of options stays the same.

Decision stability is the bridge between technical reproducibility and real-world use.

(3) **Provenance completeness**

This check evaluates whether the workflow has enough documentation to audit:

- Data source inventory and versions,
- Completeness of the transformation ledger, and
- AI usage disclosure.

This part addresses bias, provenance, and explainability.

Cases: Environmental Health Issues

To show that this approach is realistic, I describe two specific environmental health cases where the PPRP and Integrity Checks can be applied in a feasible way:

(1) **Indoor air risk-reduction evaluation using a policy-relevant survey with sensitive covariates**

Context: Indoor air interventions (e.g., ventilation, filtration, and germicidal technologies) protection, equity effects, and cost-effectiveness. Many studies use stated-preference surveys to estimate willingness to pay for risk reduction and to understand public acceptance.

Confidentiality considerations: Even if surveys do not include direct identifiers, detailed covariates (income bands, household composition, health-related factors, location proxies) can create re-identification risk. Vendor contracts may also restrict sharing microdata.

How PPRP applies:

- Code and environment: scripts for cleaning and constructing WTP variables, estimating double-bounded models, and generating the main tables and figures.
- Codebook + transformation ledger: clear documentation of skip logic, variable coding rules, sample exclusions, and construction of weights and strata, recorded in a machine-readable ledger.
- Substitute dataset: either an aggregated dataset by coarse strata (e.g., region x age band x education band) or a synthetic microdata file created from a disclosed procedure, sufficient to rerun key estimations and demonstrate the pipeline.
- Secure verification: a defined route for restricted verification.
- Integrity checks
    - Estimate/inference stability: changes in WTP estimates, confidence intervals, and whether key findings stay below/above significance thresholds;
    - Decision stability policy metric: whether the main policy conclusion remains unchanged;
    - Provenance completeness: full documentation of how variables were built, including any AI-assisted coding of open-ended responses.

: This is realistic even without publicly releasing survey microdata.

(2) Linking environmental exposures to health outcomes using restricted administrative health data and public exposure measures

Context: A common environmental health task is to link exposure to health outcomes, often to support targeted action and environmental justice priorities.

Confidentiality constraint: Administrative health data often contain sensitive information and fine geographic identifiers. Even aggregated health outcomes may require suppression for small counts. Exposure data may be public, but the linkage between health records and location can be restricted.

How PPRP applies:

- Code and environment: code to build exposure metrics, align time windows, assign exposures to geographic units, run models, and create maps/plots.
- Codebook + transformation ledger: precise logging of geospatial joins, time alignment rules, filters, and any imputation or smoothing steps.
- Substitute dataset: one feasible substitute is a public exposure dataset combined with aggregated health outcome rates at a geographic level that meets privacy rules. Another option is synthetic linkage keys that preserve distributions without revealing identities.
- Secure verification: a two-run verification design
  - A public run using the substitute dataset to verify pipeline integrity and provide approximate replication,
  - A restricted run by an approved verifier to confirm that the main claims hold using the true microdata.
- Integrity Checks:
  - estimate/inference stability: direction & rough size of exposure/outcome relationships and whether core findings cross inference thresholds;
  - Decision stability policy metric: whether the identification of priority geographies and vulnerable groups for environmental health action remains consistent across substitute vs. restricted runs;
  - Provenance completeness: clear tracking of exposure data versions, geospatial processing, and any AI-assisted steps.

: This example reflects privacy-replicability trade-offs, data integrity for sensitive data, and reproducibility standards for computationally intensive workflows.

How this can be adopted in practices

The PPRP is meant to be adopted without large new infrastructure. It can be used by:

- Research teams, as a project deliverable at dissemination time;
- Data stewards, as a standard checklist for what can be shared and how verification works;
- Institutions and training programs, as a teaching module for confidentiality-safe open science and AI governance.

The key is standardization. If many projects use a similar pack and the same Integrity Checks, the community can compare practices, improve norms, and reduce friction for reviewers and policymakers.

Limits and open questions

This approach has clear limits. Substitute datasets are approximations and will not replicate every result perfectly. Some verification will still require restricted access. AI logging adds work. But these are unavoidable realities. The advantages of the PPRP and Integrity Checks in that they make the trade-offs explicit and measurable rather than hidden.

An open question is how to define stability thresholds that are appropriate across context. Another question is how to standardize AI logging across tools and institutions.

Conclusion

As confidential data, AI-assisted analysis, and complex computation become central to policy-relevant research, the field needs a practical definition of transparency and reproducibility that works under real privacy constraints. The proposed PPRP and Integrity Checks provide an adoptable standard to make workflows auditable and conclusions more trustworthy without requiring public release of sensitive microdata. Grounded in feasible environmental health cases, this approach offers a clear path for researchers, data stewards, and institutions to balance privacy, security, and credible open science in development and public health settings.

Keywords: open science; reproducibility; confidential data; privacy-preserving replication; environmental health; administrative health data; synthetic data, aggregated data; audit trail; machine-readable transformation ledger; AI-assisted analysis; integrity checks.