When a “Successful” Health Intervention Stops Looking Successful

An independent reanalysis of a drinking-water arsenic information trial in India

Everion Cao

GHEAS, Waterloo, Ontario, Canada

ORCID: 0009-0005-8238-8583; correspondence: everioncao@gmail.com

Version: Discussion Draft 1.1, 22 July 2026

Comments welcome: This paper has not been submitted. We welcome criticism of the data decisions, models, and interpretation.

Abstract

Background

A randomized trial in Bihar, India reported that giving households arsenic test results and practical information reduced unsafe drinking-water use and may have improved health. We revisited the study not to declare the original authors right or wrong, but to ask a broader question: can the conclusion change when the same data are analysed with different, scientifically reasonable models?

Methods

We used the public version 1.2 data and code. An audit found several reconstructable village-code displacements and three villages containing one treatment record that could not be resolved from the released evidence. Our primary dataset repaired the deterministic errors and excluded the three conflicted villages, leaving 2,289 households in 153 villages. We first reproduced the published models. We then estimated the effect of randomized assignment without controlling for knowledge change or other variables measured after the intervention. Finally, we examined the associations of baseline water arsenic with stomach problems, depression, and skin problems.

Results

The published health result was reproducible: group assignment was associated with a 9.34-percentage-point reduction in stomach problems (p=0.005). In the model that did not condition on post-intervention knowledge, the estimate was -4.36 percentage points (95% CI -11.00 to 2.27), leaving substantial uncertainty about whether there was any effect. The individual-arm estimate changed from -4.16 to +2.83 percentage points. Stepwise modelling showed that the main shift occurred when knowledge change entered the model.

Mean arsenic in the individual arm fell from 27.4 to 24.2 µg/L, but the baseline-adjusted difference from control was -5.11 µg/L (95% CI -11.99 to 1.76). Higher baseline arsenic was associated with more stomach problems, but not with stable differences in depression or skin problems. A flexible skin model initially produced an extremely small p-value. That signal was entirely driven by one observation at 1.0 mg/L; removing it changed the joint p-value from approximately 2.0×10^-7 to 0.14.

Conclusions

The intervention may have improved arsenic knowledge and may have changed water choices for some households. The public evidence, however, does not establish a stable reduction in arsenic exposure or an arsenic-mediated improvement in health. This case shows why reproducible code is not enough: a statistically significant result can depend strongly on which variables are placed in the model.

Keywords: drinking-water arsenic; randomized trial; post-treatment adjustment; dose response; skin lesions; reproducibility

1. Why revisit this study?

Inorganic arsenic is one of the world's most important chemical drinking-water contaminants. The World Health Organization (WHO) uses 10 µg/L as a provisional guideline and recommends keeping concentrations as low as reasonably possible [2]. India uses the same 10 µg/L limit [5]. This is a management threshold, not a biological boundary separating “safe” from “harmful” exposure.

The most characteristic consequences of long-term inorganic arsenic exposure include pigmentation changes, palmoplantar keratosis, skin cancer, and increased risks affecting neurological, cardiovascular, metabolic, developmental, bladder, and lung outcomes [2-4]. Severe abdominal pain, vomiting, and diarrhoea are more typical of acute high-dose poisoning. Chronic exposure may also involve gastrointestinal symptoms, but common reports such as gastritis, constipation, or diarrhoea are not specific enough to identify arsenic as the cause.

Priyam, Salicath, and Sutter conducted a three-arm randomized trial among 2,334 households in 156 villages. Both intervention arms received arsenic education, household water testing, and disclosure of the result. Control households watched an unrelated video and were not told their test result. The paper concluded that the intervention increased the probability of consuming safe water and presented suggestive evidence that group delivery reduced stomach problems [1].

The study is a useful methodological case because it combines unusually open materials with several challenges: coding anomalies, variables measured after treatment, nonspecific health outcomes, and multiple defensible models. Our purpose was not to select the most impressive estimate. It was to trace how each analytic choice changed the answer.

2. How did we handle the data and models?

2.1 Public materials and trial design

Baseline interviews took place from December 2019 to February 2020, with follow-up approximately one year later. The public Edmond version 1.2 archive contains panel, wide, and health files together with Stata code [1,6]. Field-kit arsenic results were recorded at 0, 10, 25, 50, 100, 200, 300, 500, or 1,000 µg/L. Health outcomes were reported by households rather than clinically assessed.

2.2 What did the data audit find?

Before examining any GHEAS effect estimate, we compared village, treatment, and baseline information across the released files. Twenty-two displaced village codes could be reconstructed uniquely from repeated public information. After those repairs, villages 2016, 2032, and 4022 each retained one record assigned differently from the other 14 households in the same village. Nothing in the archive established which value was correct.

We therefore did not guess and did not ask the original investigators to fill the gap for us. The primary analysis excluded all three villages. Three sensitivity datasets tested whether other reasonable decisions would change the conclusion.

Scenario	Data decision	Purpose	Health rows/villages
D0	Repair 22 displacements; exclude 3 conflicted villages	Primary	2,289 / 153

Scenario	Data decision	Purpose	Health rows/villages
D1	Repair displacements; exclude only 3 conflicted records	Sensitivity	2,331 / 156
D2	Repair displacements; assign 3 records to the village majority	Sensitivity	2,334 / 156
D3	Preserve the released coding	Published-model reproduction	2,334 / 156

The complete-case primary health models used 2,232 D0 records. Every transformation was recorded in a ledger, and the source archive remained unchanged.

2.3 What did the health measures actually represent?

Only three health outcomes were directly available for analysis: any household stomach problem, a household depression indicator, and any household skin problem. Stomach problems combined gastritis, constipation, and diarrhoea. Depression was derived from PHQ-9. Skin problems combined irritation, melanosis, and hyperkeratosis [1].

These measures are not equally informative about arsenic toxicity. Depression cannot stand in for peripheral neuropathy, numbness, cognition, or child neurodevelopment. Skin is more closely aligned with chronic arsenic toxicity, but the indicator combines nonspecific irritation with the more characteristic findings of melanosis and hyperkeratosis, and no clinical examination was performed. Stomach problems were common but especially nonspecific.

2.4 Why compare two kinds of model?

The first question in a randomized trial is straightforward: what was the overall consequence of being assigned to the intervention? Our primary model therefore included randomized assignment and randomization blocks, but not knowledge change, arsenic conversations, water-source choice, or other events that occurred after assignment.

The published health model adjusted for change in arsenic knowledge. Yet increasing knowledge was one of the intervention's intended mechanisms. Once knowledge is held constant, the question changes from “What was the total effect of the intervention?” to “Among people with the same knowledge change, was there any additional intervention effect?” Those are different estimands, and conditioning on a post-treatment variable can introduce bias [8].

We fitted linear probability models so that coefficients could be read as percentage-point differences. Uncertainty was clustered by village. Randomization inference, multiplicity correction, leave-one-village-out analysis, and all four data scenarios were used to assess stability.

The arsenic-health analyses were observational associations. They did not become causal merely because the records came from a randomized trial. We compared continuous concentration, the 10 µg/L threshold, three concentration categories, and a flexible natural-spline model.

3. What did we find?

3.1 How much did water arsenic change?

Mean baseline arsenic was approximately 27 µg/L in every arm, and the median was 10 µg/L. In the primary sample, 34.3% of households exceeded 10 µg/L and 18.8% were at or above 50 µg/L.

Arm	Baseline mean µg/L	Endline mean µg/L	Within-household mean change µg/L	>10 µg/L baseline→endline	≥50 µg/L baseline→endline
Control	26.6	27.1	+0.08	27.3%→30.1%	15.3%→15.0%
Individual	27.4	24.2	-3.47	39.5%→36.1%	19.6%→15.8%
Group	26.8	26.7	+0.02	35.7%→36.4%	21.4%→16.8%
Overall	26.9	26.0	—	34.3%→34.3%	18.8%→15.9%

The individual arm appeared to improve, but its baseline-adjusted difference from control was -5.11 µg/L (95% CI -11.99 to 1.76; p=0.144). The interval is compatible with a meaningful reduction, but also with little or no reduction. The group-arm contrast was -0.37 µg/L (p=0.909). At endline, mean arsenic in the individual arm remained about 2.4 times the guideline, and 36.1% of households still exceeded it.

3.2 Could we reproduce the published unsafe-water result?

Yes. With the released coding and published model, the individual assignment-by-endline estimate was -7.28 percentage points (p=0.044); the group estimate was -2.36 points (p=0.445). These nearly match the table values of -7.3 and -2.4 [1].

The evidence was nevertheless borderline. Without post-treatment knowledge and conversation variables, the individual estimate was -6.00 points (p=0.060). Adding knowledge shifted it to -7.33 (p=0.043). In the structurally corrected D0 sample, the complete published model produced -7.11 (p=0.055). The original article also reported that the individual result did not remain significant after Holm-Bonferroni correction.

A fair summary is therefore that individual delivery may have reduced the proportion above 10 µg/L, but the evidence was weak and sensitive to modelling and data decisions. The group arm showed no clear improvement in measured arsenic.

3.3 Why did the stomach-health conclusion change?

Outcome and model	Individual, percentage points (95% CI)	Group, percentage points (95% CI)	Interpretation
Stomach, published D3 model	-4.16 (-11.29, 2.97)	-9.34 (-15.85, -2.84)	Reproduced; adjusts for post-treatment variables
Stomach, GHEAS D0 primary	+2.83 (-3.65, 9.30)	-4.36 (-11.00, 2.27)	Randomized total-effect estimate
Depression, GHEAS D0 primary	-2.89 (-10.88, 5.10)	-0.08 (-7.83, 7.68)	No clear effect
Skin, GHEAS D0 primary	+1.08 (-4.07, 6.23)	-2.24 (-6.84, 2.36)	No clear effect

None of the six health comparisons excluded a zero effect after Holm correction. Randomization-inference p-values for the primary stomach model were 0.419 for individual assignment and 0.208 for group assignment.

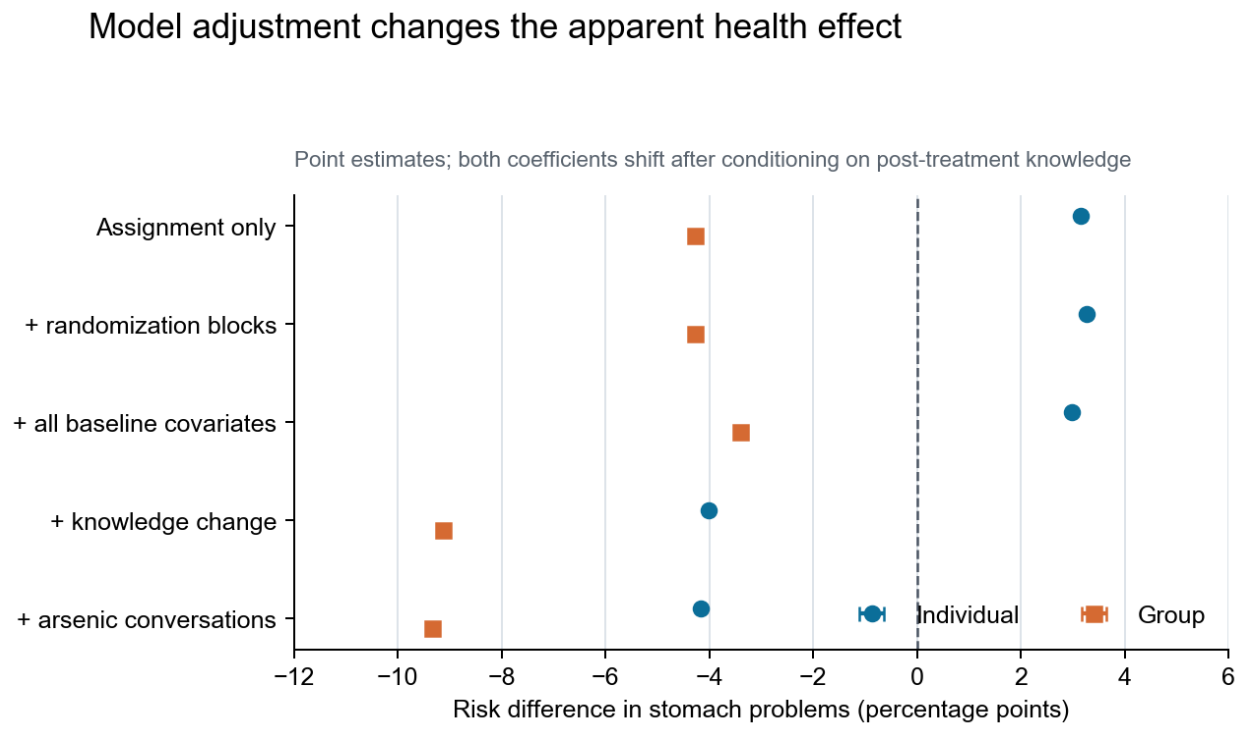

Figure 1. Post-treatment knowledge adjustment and estimated stomach effects.

Figure 1 shows where the result changed. After adding all baseline covariates, the individual and group stomach estimates were still +2.99 and -3.41 percentage points. Adding post-intervention knowledge change moved them immediately to -4.02 and -9.13. Adding arsenic conversations produced only a small further change. The significant group result therefore appeared mainly after conditioning on knowledge change, not after correcting baseline imbalance.

3.4 Was water arsenic related to depression or skin problems?

Health outcome	Per 10 µg/L, percentage points (95% CI)	>10 µg/L, percentage points (95% CI)	>10-<50 vs ≤10 µg/L	≥50 vs ≤10 µg/L
Stomach problems	+0.43 (0.09, 0.78)	+5.27 (0.86, 9.68)	—	—
Depression	+0.03 (-0.46, 0.52)	+3.48 (-2.04, 9.00)	+3.72 (-2.60, 10.04)	+3.26 (-3.68, 10.19)
Skin problems	+0.07 (-0.35, 0.49)	-0.74 (-5.11, 3.64)	-0.50 (-6.13, 5.13)	-0.96 (-6.00, 4.09)

Each 10 µg/L higher baseline arsenic was associated with about 0.43 percentage points more stomach problems (p=0.015). Depression and skin estimates, however, crossed zero whether exposure was treated continuously, divided at 10 µg/L, or grouped into three ranges. Both linear estimates could also change sign when villages were removed one at a time.

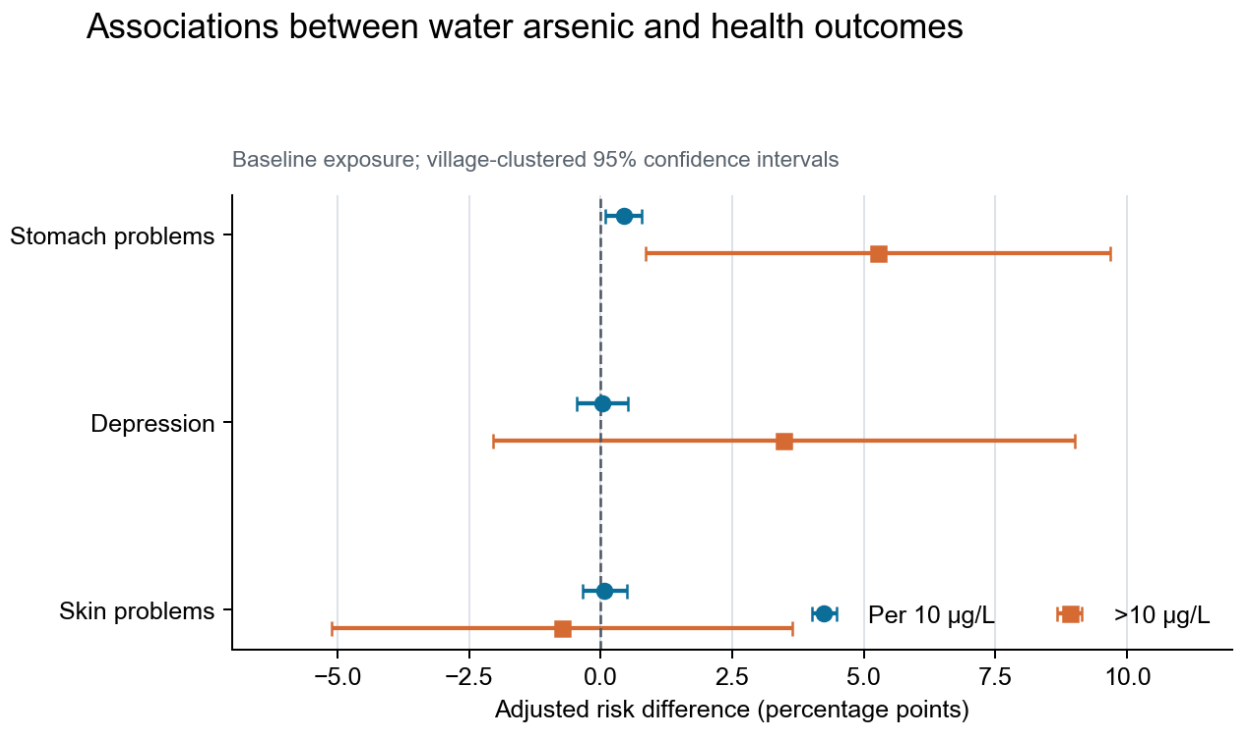

Figure 2. Adjusted associations of baseline water arsenic with three household health outcomes.

The skin spline initially yielded p≈2.0×10^-7, which looked striking. Exact-level inspection showed that the 1.0 mg/L level contained only one observation, and that household reported a skin problem. Removing that record changed the joint p-value to 0.14; restricting the highest concentration to 0.3 mg/L gave p=0.746. This was not persuasive evidence of a biological curve. It was an artifact of a sparse extreme tail combined with a flexible model.

3.5 Why did the three health outcomes not tell one coherent story?

Stomach problems were reported by 72.0% of the complete sample, compared with 14.6% for skin problems. The stomach measure probably captured a wide range of household symptoms, many unrelated to arsenic. It was positively associated with baseline arsenic, but there was no corresponding skin pattern, no neurological outcome for triangulation, and no stable chain in which the intervention reduced arsenic and lower arsenic improved health.

Skin was more relevant biologically but poorly timed for this trial. WHO notes that characteristic chronic arsenical skin changes generally take years to emerge [2]. With approximately one year of follow-up, even a real modest reduction in exposure would not be expected to reverse established pigmentation or keratosis quickly. A null skin result therefore does not show that arsenic is harmless or that the intervention could never work. It shows that the available outcome and follow-up period were not well suited to detecting chronic skin benefit.

4. What do the findings mean?

4.1 Reproducible does not mean self-interpreting

The main published coefficients were highly reproducible, so the disagreement is not about arithmetic. It concerns the question answered by the model, the variables placed in it, and whether the measured outcomes represent arsenic-related disease.

If the aim is to estimate the total effect of randomized assignment, a change caused by the intervention should not be treated as an ordinary background covariate. A model that conditions on knowledge may still be useful for exploring mechanisms, but it cannot replace the primary intention-to-treat estimate. In this case, that single modelling step reversed the health coefficients and generated the significant result.

4.2 Stomach symptoms cannot carry the arsenic-health claim alone

In a chronic exposure distribution concentrated around 10-50 µg/L, general gastritis, constipation, and diarrhoea do not identify arsenic as the cause. Microbial contamination, diet, nutrition, hygiene, infection, and reporting behaviour can all affect these symptoms. The original paper itself noted that stomach improvement might reflect reduced bacterial contamination when households boiled surface water, rather than reduced arsenic [1].

A stronger test of arsenic-related health improvement would need repeated laboratory water measurements, biomarkers such as urinary arsenic, clinical skin assessment, peripheral neurological examination, baseline health information, and longer follow-up.

4.3 Lessons for data analysis

This case suggests a practical sequence: verify units and randomization before estimating effects; separate randomized total effects from mechanism models; require evidence that the contaminant actually declined before attributing health improvement to it; choose outcomes with appropriate specificity and latency; inspect extreme observations whenever a flexible curve produces a dramatic result; and never allow $p<0.05$ to upgrade a claim automatically.

5. Strengths and limitations

This reanalysis preserved the source files, recorded each repair, compared four data decisions, and reported results that both supported and contradicted the published interpretation. It reproduced the original models and added randomization inference, multiplicity correction, leave-one-village-out checks, and an explicit label for analyses proposed after earlier results were visible.

Important limitations remain. Public health records lacked household identifiers, and the true value of three assignment conflicts could not be recovered. Field-kit measurements were discrete and had no laboratory replication, batch, or operator information. Health outcomes were household-reported, with no baseline health measures or individual linkage. Shared water sources and information spillovers between villages could not be assessed. The extended depression and skin analyses were exploratory and results-visible, not an independent confirmatory study.

6. Conclusions

The public numerical results of this Indian arsenic-information trial were highly reproducible, but the health interpretation was not robust. Models that did not condition on post-treatment variables found no clear improvement in stomach, depression, or skin outcomes. Individual delivery may have reduced the proportion of households above 10 µg/L, but the evidence was weak; the continuous concentration reduction was also uncertain.

The most defensible conclusion is that the intervention may have improved knowledge and changed some water choices, but the public evidence does not establish a stable reduction in arsenic or an arsenic-mediated improvement in health. This does not make the original study unimportant. Its greatest value may be that its openness allows us to see exactly where evidence ends and interpretation begins.

Data, code, and transparency

The investigators released version 1.2 data through Edmond under CC BY 4.0, DOI 10.17617/3.LPX8BT [6]. GHEAS retains the protocol, anomaly rules, transformation ledger, data scenarios, model inventory, and results. The source ZIP was not modified, and the investigators were not contacted to complete or reinterpret public records.

Everion Cao directed the research question, analysis principles, and final judgement. OpenAI Codex, referred to in the project as Jace, assisted with code, execution, and drafting under a rule that prohibited AI from strengthening causal claims. Generative AI was not used to impute missing outcomes or decide the true value of conflicted records.

Competing interests

The author declares no competing interests.

References

[1] Priyam S, Salicath D, Sutter M. An information intervention in India reduces consumption of arsenic-contaminated drinking water and has small effects on health. PNAS Nexus. 2026;5(4):pgag093. https://doi.org/10.1093/pnasnexus/pgag093

[2] World Health Organization. Arsenic: fact sheet. 7 December 2022. https://www.who.int/news-room/fact-sheets/detail/arsenic

[3] World Health Organization. Exposure to arsenic: a major public health concern. 2019. WHO/CED/PHE/EPE/19.4.1. https://www.who.int/publications/i/item/WHO-CED-PHE-EPE-19.4.1

[4] United States Environmental Protection Agency. IRIS assessment: inorganic arsenic, CASRN 7440-38-2. Updated 13 January 2025. https://iris.epa.gov/ChemicalLanding/&substance_nmbr=278

[5] Government of India, Ministry of Jal Shakti. Monitoring of contamination of ground water; Annex I, IS 10500:2012 drinking-water limits. 21 July 2022. https://www.pib.gov.in/Pressreleaseshare.aspx?PRID=1843385

[6] Priyam S, Salicath D, Sutter M. Replication data for “An information-intervention in India reduces consumption of arsenic-contaminated drinking water and has small effects on health.” Edmond, version 1.2. 2026. https://doi.org/10.17617/3.LPX8BT

[7] George CM, Zheng Y, Graziano JH, et al. Evaluation of an arsenic test kit for rapid well screening in Bangladesh. Environmental Science & Technology. 2012;46:11213-11219.

[8] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009;20(4):488-495. https://doi.org/10.1097/EDE.0b013e3181a819a1

[9] Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6(2):65-70.

[10] Martin A, Rief W, Klaiberg A, Braehler E. Validity of the brief Patient Health Questionnaire mood scale (PHQ-9) in the general population. General Hospital Psychiatry. 2006;28:71-77.