

Arsenic exposure near an abandoned gold mine, Loei Province, Thailand: a record-level probabilistic assessment with exposure-route screening and leakage audit

Phrutthinun Surit

Naresuan University

Showket Hasan

showket.h@chs.ac.th

Chalermkwansatree School

Research Article

Keywords: Human health risk assessment, Two-dimensional Monte Carlo simulation, Exposure pathway analysis, Legacy mine contamination, Target leakage

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10941256/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Arsenic exposure near an abandoned gold mine, Loei Province, Thailand: a record-level probabilistic assessment with exposure-route screening and leakage audit

Phrutthinun Surit, PhD

Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand

ORCID: <https://orcid.org/0000-0003-3727-3400>

Email: phrutthinuns@nu.ac.th

MD Showket Hasan, MSc

Science Math Bilingual Program, Chalermkwansatree School, Phitsanulok, Thailand

ORCID: <https://orcid.org/0000-0002-8868-6451>

Email: showket.h@chs.ac.th

Corresponding author: MD Showket Hasan — Science Math Bilingual Program, Chalermkwansatree School, Phitsanulok 65000, Thailand. Email: showket.h@chs.ac.th

Acknowledgements

This research is dedicated to the memory of Asst. Prof. Dr. Lt. Saisiri Mirasen, whose guidance laid its foundation. We thank the village headmen of the Khao Luang subdistrict for their cooperation during data collection.

Abstract

Arsenic exposure near legacy gold mining is chronic, multi-pathway and sparsely measured; standard assessment collapses each medium to one value and seldom records whether the population actually contacts it. We combined a cross-sectional household survey of 371 adults with a record-level synthesis of 244 arsenic measurements (2012–2022) and a two-dimensional Monte Carlo assessment near the abandoned Tungkum mine, Loei Province, Thailand, using Thai receptor parameters at $AT_{nc} = 10,950$ d. Applying the current U.S. EPA oral reference dose, 6.0×10^{-5} mg/kg-day (U.S. Environmental Protection Agency, 2025), in place of the superseded 3.0×10^{-4} raises every hazard quotient fivefold: river water reaches a median of 20.44 (95th percentile 33.80; probability of exceedance 1.00) and surface water crosses unity at 1.46, so two pathways exceed it. River-water lifetime cancer risk is 1.68×10^{-2} . Yet no participant drinks it — 337 of 371 (90.8%) use bottled water — while 208 (56.1%) use it daily, so the active route is contact, not drinking. Leakage screening removed 12 variables, including one whose missingness differs by outcome and one measured at stratum rather than participant level; 167 participants (45.0%) were registered cases, and two deposited survey versions differ in 74 labels, capping what any classifier can establish. We report no screening tool. The contribution is a specification of what a defensible assessment requires here: the measurement count and averaging time behind every estimate, and evidence of contact before a quotient is interpreted — redirecting monitoring from the drinking supply to the water households actually use.

Keywords: Human health risk assessment, Two-dimensional Monte Carlo simulation, Exposure pathway analysis, Legacy mine contamination, Target leakage

Introduction

Arsenic is a class I human carcinogen causing dermal, cardiovascular, neurological, and developmental toxicities at chronic low doses (International Agency for Research on Cancer, 2012). Cohort evidence links drinking-water

arsenic to cardiovascular mortality well below historical concentrations (Chen et al., 2011); the World Health Organization guideline value is 10 µg/L (World Health Organization, 2011). Because environmental arsenic is odourless and tasteless, exposure is recognised only after effects accumulate, making prospective exposure assessment rather than case-finding essential.

Mainland Southeast Asian geology mobilises arsenic across extensive alluvial and mining-affected terrain; in Thailand the clearest documented setting is Loei Province, where the Tungkum open-pit gold mine left oxidised sulphidic waste exposed at the surface. Tungkum Limited began production at Wang Saphung in 2006; Thailand suspended all gold mining from 1 January 2017, after which the operator entered bankruptcy. Sources differ on the commencement year—1991 for corporate founding, 2006 for operations—and this study uses 2006, also the start of the residence window in the approved protocol (see Participants). Studies since 2012 have repeatedly documented arsenic elevations in soil, surface water, groundwater, river water, and aquatic biota around the site (Boonsrang et al., 2017; Intamat et al., 2016; Kunwittaya et al., 2022; Nookongbut et al., 2016b; Suttichaiya et al., 2016; Thathong et al., 2018a; Weerasiri et al., 2012a, 2013). Rice, the regional staple, has attracted particular attention because paddy cultivation under reducing conditions mobilises arsenic efficiently (Chanpiwat et al., 2019; Hoang et al., 2021).

Legacy mining differs from active mining in critical ways: remediation responsibility is diffuse, so contamination persists for decades; exposed populations are resident communities experiencing chronic multi-pathway exposure; and historical monitoring was designed for regulatory compliance rather than exposure science. Assessment in such settings is therefore retrospective and data-limited, compiling measurements collected at different times, in different media, and by different investigators.

The gap

Standard human health risk assessment (U.S. Environmental Protection Agency, 2015) is poorly suited to this situation: its deterministic form requires a single exposure point concentration per medium, forcing heterogeneous measurements to be collapsed into a single central value. Two critical dimensions are lost in that reduction.

First, variability is discarded: a single median cannot express that a substantial fraction of a population is exposed above it. Probabilistic formulations separating inter-individual variability from parameter uncertainty provide far more informative risk characterisation. Moazamnia et al. (2024) showed that two-dimensional Monte Carlo simulation materially refines 95th-percentile risk; Yuan et al. (2025) reported improved source-specific accuracy across legacy mining areas.

Second, spatial structure is coarsened: where primary measurements are sparse, exposure is typically assigned by radial distance bands. However, radial distance is a poor proxy where transport is hydrologically controlled, as equidistant locations in opposing drainage directions can differ completely. Machine-learning models have successfully predicted well-water arsenic across the conterminous United States (Lombard et al., 2021) and groundwater concentrations in South Asia (Iqbal et al., 2024), but both rely on dense monitoring networks unavailable in legacy mining settings. Furthermore, spatial autocorrelation causes standard random cross-validation to be over-optimistic, requiring spatial block partitioning for honest evaluation (Wang et al., 2023).

Third, validation is rarely attempted: exposure models in data-limited settings are seldom checked against biomarker data, although biomonitoring has been integrated with reverse dosimetry for arsenic (Lin et al., 2020). We could not do so here—no biomarker series exists for this area, and the one available Thai mining dataset originates from a different geographic region (Kunwittaya et al., 2022). We claim no validation of modelled internal dose, treating this absence as an explicit limitation.

Objectives

This study had three objectives: (1) to assemble and harmonise every available arsenic measurement for the Tungkum area into one record-level corpus rather than working from collapsed summary statistics; (2) to conduct a spatially explicit, two-dimensional Monte Carlo human health risk assessment propagating variability and parameter

uncertainty; and (3) to test whether questionnaire and case-registry data, screened for target leakage, can distinguish residents registered as arsenic cases from others in the same villages.

Three research questions guided the reanalysis. Formulated post hoc, they are not prespecified hypotheses from the 2021 protocol; each is evaluated directly against what these empirical data could support.

RQ1. Retaining measurements at record level versus collapsed medians, which pathways exceed a hazard quotient of 1, and how does that ranking compare? H1a: correcting the averaging time relocates the dominant pathway from rice to a water pathway, because earlier rice estimates were reproducible only under a mis-specified averaging time. H1b: pathway ranking changes, because collapsed concentrations cannot express the upper tail of a highly dispersed distribution.

RQ2. How well is each pathway evidenced, once the number of measurements behind it is reported alongside the risk estimate? H2: the highest-risk pathways will be among the most thinly measured, because historical monitoring concentrated effort where contamination was already expected.

RQ3. Can routinely collected survey and registry data distinguish residents registered as arsenic cases from other residents of the same villages? H3: they cannot do so usefully, because candidate exposure variables are either measured at stratum rather than participant level, carry outcome-tracked missingness, or rely on distance bands that poorly proxy hydrologic transport.

Reporting H3 as an expectation rather than a hope matters: a study designed to demonstrate a screening tool would treat null discrimination as failure, whereas demonstrating that retrospective survey data cannot support one is our central methodological finding, preventing premature deployment of unvalidated classifiers.

We report data limitations explicitly, including sample size behind each pathway: the dominant pathway rests on three measurements, the second-ranked on 23. Stating this plainly provides essential context for probabilistic risk rankings.

Materials and methods

Study design and setting

We conducted a cross-sectional exposure and health risk assessment in seven villages of the Khao Luang sub-district, Loei Province, Thailand, around the abandoned Tungsum gold mine centroid (17.367161° N, 101.672674° E; Fig. 1), combining a household questionnaire survey, a synthesis of published arsenic measurements, and a probabilistic human health risk assessment (U.S. Environmental Protection Agency, 2015). Reporting follows STROBE guidelines. Because previously reported inter-village distances exhibited discrepancies up to 4.50 km (with equidistant villages assigned to different strata), distances were recomputed from coordinates, retaining '1 km', '5 km', and '10 km' strictly as nominal sampling strata.

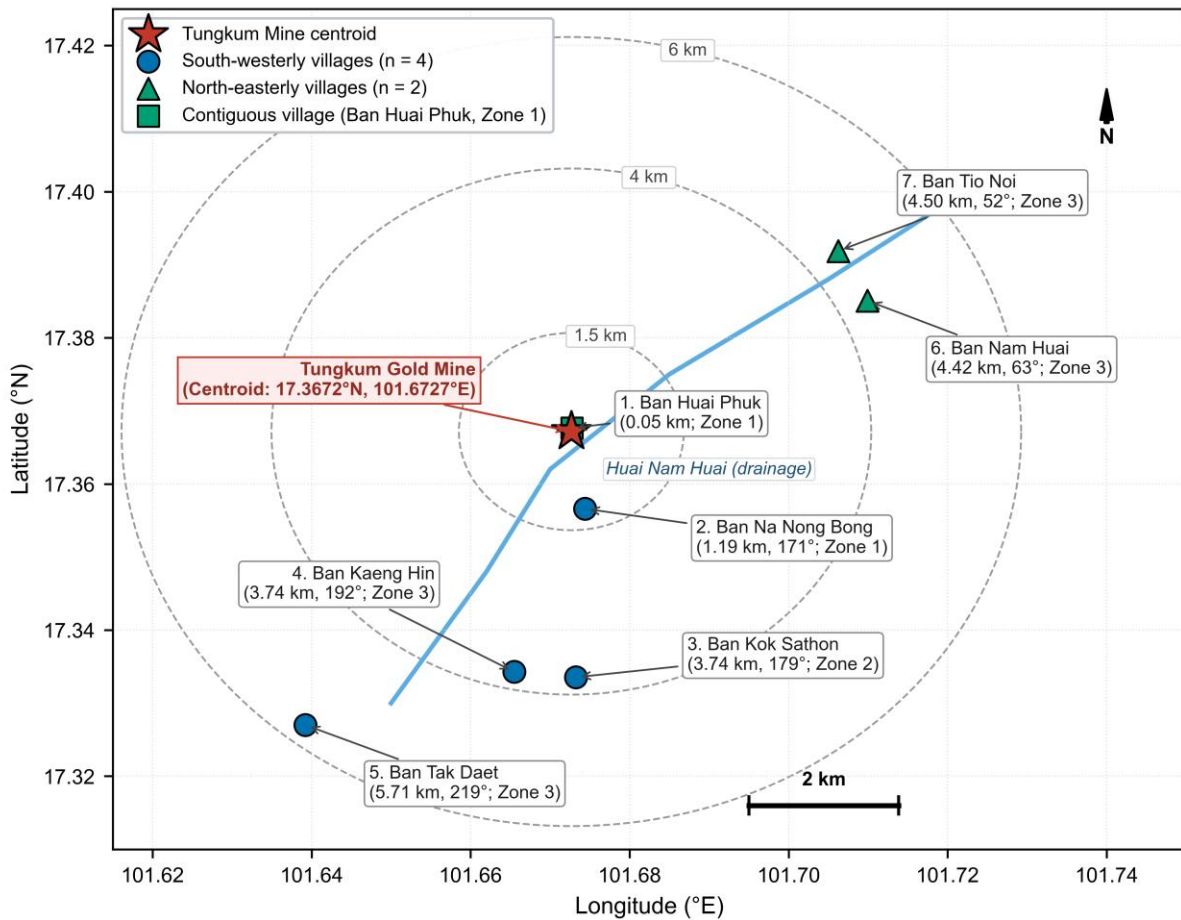


Fig. 1 Location of the seven study villages relative to the abandoned Tungcum gold mine centroid (17.367161° N, 101.672674° E), Khao Luang sub-district, Loei Province. Main map: mining lease perimeter (red dashed polygon), pit centroid (red star), the Huai Nam Huai stream draining north-eastward from the mine catchment (solid blue line) and the seven communities, split into a south-westerly cluster (blue circles) and a north-easterly cluster (green triangles); Ban Huai Phuk (green square) is contiguous with the mine boundary. Radial rings are omitted because dispersal follows topography and drainage rather than isotropic dispersion; scale bar 2 km. Inset: 400 m × 400 m detail of the pit centroid and Ban Huai Phuk; scale bar 100 m. Source: Online Resource 1 (ESM_1_Figure_Source_Data.xlsx, sheet Fig1_Villages)

Participants

Adults in the seven villages were recruited between March and August 2022, enrolling 371 participants across three nominal strata: 1 km (two villages, n = 111, 29.9%), 5 km (two villages, n = 120, 32.3%), and 10 km (three villages, n = 140, 37.7%). Eligibility required age ≥ 18 years, residence in the sub-district since 2006 (the year mining commenced; see Study design and setting), and written informed consent; temporary residents and individuals unable to complete interviews were excluded. Trained enumerators administered a structured questionnaire recording demographics, water sources, occupational tasks, and self-reported symptoms (deposited as Online Resource 4). Data collection concluded prior to ethics certificate expiration on 27 August 2022. Target sample size was fixed in advance by the approved protocol (see Sample size); uninformative post-hoc power calculations are omitted.

The protocol describes simple random sampling without replacement from enumerated households across the seven villages (enumeration ratio 0.28–0.30 per village; design effect = 1). However, the sampling frame concentrated on

known affected settlements, meaning true population-representative sampling of the sub-district was neither attempted nor achieved. The 371 participants characterise the enumerated villages rather than the general population of Loei Province, and findings should be interpreted within that scope.

Case ascertainment

The binary outcome (Arsenic_infection: 1 = registered case, 0 = non-case) denotes whether a participant appears in the Loei Provincial Hospital arsenic case register as of 23 April 2021. Case ascertainment is therefore administrative, derived from hospital records compiled before this study began, not an outcome measured by our survey. The name reflects the administrative field label in the deposited codebook, not an infectious disease. Among 371 participants, 167 (45.0%) are cases and 204 (55.0%) non-cases. This elevated case fraction reflects purposive community enrichment: sampling focused on villages with documented clinical poisoning histories.

Why the hospital record is not the same as the arsenic-testing items

The ethics-approved instrument (version 3.0, 21 August 2021) contains three arsenic-specific items, none forming the basis of case status: 6.9 asks whether the participant was ever tested for arsenic (urine, blood, or hair); 6.10 whether the result was negative or positive; and 6.11 at which facility. Most cases report no prior testing: 149 of 167 cases (89.2%) recorded at 6.9 that they had never been tested, precluding biological validation against survey records. Prior testing does not determine status either: 75 of 93 reporting a test are non-cases. The facility at 6.11 is 'Loei Provincial Hospital' for exactly the 167 cases and missing for the other 204, making it collinear with the outcome and excluded from all modelling (see Leakage screening and predictive modelling).

A documented correction to the case variable

The raw analysis dataset prepared for statistical software (final version, 5 May 2024) records 93 cases among the same 371 participants, whereas the dataset analysed here records 167. The 74 differences are entirely unidirectional—each a non-case in the earlier dataset linked as a case here, with the facility field populated correspondingly. This represents registry linkage, not imputation: the 74 participants were matched against the Loei Hospital register of 23 April 2021 compiled prior to study commencement. We analyse the 167 linked cases; on the 93 present before linkage, the case proportion would be 25.1% rather than 45.0%.

What the case register does not provide

The register confirms that a participant was registered as an arsenic case, but not the specific clinical findings behind registration, and the codebook supplies neither value labels nor the full register. We therefore define the outcome strictly as registration in the Loei Hospital register as of 23 April 2021. Ascertainment is conditional on health-service contact, meaning care-seeking behaviours may track case status independently of toxicity. Consequently, the case fraction represents an achieved sample property rather than community prevalence, and all analyses carry these limitations.

Target sample size was fixed a priori using the Lwanga and Lemeshow procedure: anticipated exposure probability 0.50 in diseased and 0.40 in non-diseased, odds ratio 2.00, relative precision 0.25, and 95% confidence interval, giving 371 participants. This calculation governs the survey component only. Because environmental concentrations are literature-derived rather than individually measured, precision of risk estimates is governed by measurement density per pathway, reported transparently in Results.

Exposure data corpus

We assembled every published arsenic measurement for the study area through 31 December 2022 (Scopus, Web of Science, Google Scholar, and the Suranaree University of Technology institutional repository), yielding 244 records from eight primary studies published 2012–2022, retained at record level rather than collapsed to summary statistics. Studies after the cutoff, including the 2024 sediment study of Tongtung et al., are cited but excluded; implications for the sediment pathway are noted in Discussion. From seven studies, 134 measurements across 12 media entered the risk assessment: Hasan (2022), Intamat et al. (2016), Nookongbut et al. (2016b), Suttichaiya et al. (2016),

Thathong et al. (2018b), and Weerasiri et al. (2012b, 2013). The eighth, Boonsrang et al. (2017), contributed only laboratory leaching tests quantifying metal release under experimental conditions. For Thathong et al. (2018b), 20 soil values were excluded because laterite soil functioned as an experimental adsorbent rather than environmental soil, while its single creek-water record was retained as surface water. Excluding these 20 laterite values yielded 69 environmental soil measurements (down from 89). The remaining 110 experimental records are logged in Online Resource 2. Table 1 reports corpus composition, and full data are deposited in Online Resource 1.

Table 1

Arsenic measurement corpus by medium

Medium	Unit	n	Detected	Fitted median (GM)	GSD	Min	Max
Soil	mg/kg	69	69	11.55	3.85	0.59	184.0
Surface Water	mg/L	23	23	0.005474	4.75	0.0003	0.22
Plants	mg/L	8	8	0.003764	3.38	0.001	0.021
Fish	mg/kg	6	6	0.3112	2.27	0.12	1.01
Frog	mg/kg	6	6	0.2017	1.62	0.11	0.39
Groundwater	mg/L	5	5	0.003419	9.03	0.0005	0.058
Public Supply	mg/L	3	3	0.001887	2.58	0.001	0.0056
River Water	mg/L	3	3	0.07273	1.37	0.051	0.092
Rice	mg/kg	3	1	no fit	no fit	—	—
Shrimp	mg/kg	3	3	0.2908	1.96	0.15	0.577
Mussel	mg/kg	3	3	0.289	1.12	0.258	0.326
Sediment	mg/kg	2	2	no fit	no fit	—	—

Note. Corpus of eight primary studies, 2012–2022, retained at record level. Fitted median is the median of the fitted lognormal, $\exp(\mu)$, equal to the geometric mean and used as the central value; it is not the sample median, which may differ (soil: 11.55 fitted against 8.42 observed). Min and Max are observed. Of 244 codebook records, 134 were fitted across 12 media and 112 across 7 media yielding reported risk estimates; the remaining 110 are leaching tests, microbiological assays and adsorption or methylation experiments that do not represent environmental concentrations, plus three deionised-water blanks and one bacterial-culture record — listed in the exclusion log. The 20 laterite soil values of Thathong et al. (2018b) are included among those 110: they characterise the adsorbent used in a batch adsorption study rather than environmental soil. Lognormal fitted by method of moments; non-detects substituted at half the minimum detected value. Media with $n < 5$ are carried as point values, their fitted percentiles extrapolative. Work whose measurements did not enter the fitted corpus (Nookongbut et al., 2016a; Thathong, 2018c; Thathong et al., 2019), and the soil values of Thathong et al. (2018b), are in the exclusion log (Online Resource 2). Source values: Online Resource 1, sheet Fig2_Corpus_records.

Distribution fitting

For each medium we fitted a two-parameter lognormal distribution by method of moments on log-transformed concentrations, substituting non-detects at half the minimum detected value. Media with fewer than five records were evaluated as point values across their observed range; fitted percentiles remain extrapolative—for river water, the fitted 95th-percentile concentration (0.122 mg/L) exceeds the maximum observed value (0.092 mg/L), and bootstrapping log-transformed values offers limited resolution at $n = 3$. The risk table therefore reports the observed maximum beside every fitted concentration.

Spatial characterisation

Distance and initial great-circle bearing from the mine centroid to each village centroid were recomputed from coordinates using the haversine formula, retaining both as continuous predictors. Radial distance alone is inadequate where transport is hydrologically controlled, as it assigns equal exposure potential to equidistant settlements situated in opposing drainage directions.

Human health risk assessment

Exposure was estimated for adults through ingestion of water, food, and soil. Average daily dose (ADD, $\text{mg kg}^{-1} \text{ day}^{-1}$) and lifetime average daily dose (LADD, $\text{mg kg}^{-1} \text{ day}^{-1}$) follow standard U.S. EPA equations:

$$ADD = \frac{C \times IR \times EF \times ED \times RBA}{BW \times AT_{nc}} \quad (1)$$

$$LADD = \frac{C \times IR \times EF \times ED \times RBA}{BW \times AT_c} \quad (2)$$

$$HQ = \frac{ADD}{RfD} \quad (3)$$

$$LCR = LADD \times CSF \quad (4)$$

Receptor parameters were localised to Thai national standards. Averaging times, which govern every risk calculation, were set to $AT_{nc} = ED \times 365 = 10,950$ days for non-carcinogenic effects and $AT_c = 75 \times 365 = 27,375$ days for lifetime cancer risk, corresponding to Thai life expectancy. An earlier draft erroneously set AT_{nc} to 365 days, inflating non-carcinogenic doses 30-fold. Body weight was set to 60.0 kg for men and 55.0 kg for women based on national surveys. Operative toxicity reference values follow the U.S. Environmental Protection Agency (2025) IRIS assessment: an oral reference dose of $6.0 \times 10^{-5} \text{ mg kg}^{-1} \text{ day}^{-1}$ and an oral cancer slope factor of 32 ($\text{mg kg}^{-1} \text{ day}^{-1}$)⁻¹ ($CSF = 32$). Calculations using superseded constants ($RfD = 3.0 \times 10^{-4}$, $CSF = 1.5$) retained as a benchmark sensitivity so results stay comparable with prior literature. Deterministic parameters are summarised in Table 2.

Table 2

Human health risk assessment parameters

Parameter	Unit	Male	Female	Source / basis
Body weight	kg	66.19	63.52	Thai national standards
Ingestion rate — water	L/day	1.13	1.13	Site survey
Ingestion rate — rice	kg/day	0.284	0.268	Site survey; national standards
Ingestion rate — fish	kg/day	0.0087	0.0087	Site survey
Ingestion rate — soil	kg/day	5.0×10^{-5}	5.0×10^{-5}	U.S. EPA (2011)

Exposure frequency (EF)	days/year	350	350	U.S. EPA (2015)
Exposure duration (ED)	years	30	30	U.S. EPA (2015)
Averaging time — non-cancer (ATnc)	days	10950	10950	ED × 365 — governs every reported HQ
Averaging time — cancer (ATc)	days	25550	25550	70 × 365
Relative bioavailability (RBA)	—	1.0	1.0	U.S. EPA screening default
Oral reference dose (RfD)	mg/kg-day	6.0×10^{-5}	6.0×10^{-5}	U.S. EPA IRIS (2025)
Oral cancer slope factor (CSF)	(mg/kg-day) ⁻¹	32	32	U.S. EPA IRIS (2025)
Oral reference dose — superseded benchmark	mg/kg-day	3.0×10^{-4}	3.0×10^{-4}	U.S. EPA IRIS, pre-2025; sensitivity only
Oral cancer slope factor — superseded benchmark	(mg/kg-day) ⁻¹	1.5	1.5	U.S. EPA IRIS, pre-2025; sensitivity only

Note. Adult male and female receptors. Averaging times govern every reported value: ATnc = ED × 365 = 10,950 days, ATc = 70 × 365 = 25,550 days. The oral reference dose (6.0×10^{-5} mg/kg-day) and oral cancer slope factor (32 (mg/kg-day)⁻¹) are the current U.S. EPA IRIS values for inorganic arsenic, finalised 13 January 2025; every risk value here is computed with them. The superseded pair (3.0×10^{-4} and 1.5) is retained above as a benchmark sensitivity so results stay comparable with the earlier assessment this study re-examines. Relative bioavailability is set to the U.S. EPA default of 1.0, which is conservative.

Probabilistic simulation

Deterministic point estimates were extended to a two-dimensional Monte Carlo simulation separating variability from uncertainty. The outer loop drew lognormal parameters μ and σ from their sampling distributions— μ normal with standard error $\sigma/\sqrt{n}$, σ scaled by a chi-square variate on $n - 1$ degrees of freedom; the inner loop drew concentrations, ingestion rates, and body weights. We report the median and 95th percentile of the hazard quotient, $P(\text{HQ} > 1)$, median and 95th percentile of lifetime cancer risk, and a confidence interval for the 95th percentile itself. Media with fewer than 30 measurements were assigned parameter uncertainty by non-parametric bootstrap of log-transformed values. Values here were generated with 1,000 outer × 10,000 inner iterations (seed 42). Sensitivity scenarios ran at 300 outer × 3,000 inner iterations rescaled to baseline, leaving reported ratios unaffected.

Leakage screening and predictive modelling

Before modelling, all candidate predictors were cross-tabulated against the outcome with missingness computed within each outcome class; predictors whose missingness or distribution tracked the outcome, and predictors constant within sampling strata, were excluded (Table 3). Models were validated by leave-one-zone-out cross-validation, partitioning on the three-level nominal stratum, and by spatial block holdout; a participant-level random split was not the primary metric because village clustering inflates it. Every model was benchmarked against a distance-only logistic regression baseline. Predicted probabilities were intercept-adjusted from sampling to target population prevalence. Two deviations from the protocol (version 3.0, 21 August 2021) should be noted: gradient boosting is additionally reported alongside logistic regression, and validation is spatial rather than by sampling frame.

No biomarker validation

We did not validate modelled internal doses against biomonitoring data and state this explicitly. No biomarker series exists for this area, and the one available Thai mining-community dataset comes from a different region and provinces (Kunwittaya et al., 2022). We claim no validation of modelled internal dose and treat that absence as an acknowledged limitation. In data-limited settings, transparently acknowledging validation constraints is preferable to asserting unverified biological certainty.

Results and discussion

Results

Participants and leakage screening

371 adults were enrolled across the three nominal strata, of whom 167 (45.0%) were cases, defined by documentation in Loei Hospital records (see Case ascertainment; diagnostic criteria applied there are not deposited). Leakage screening identified three modelled exposure variables unusable for inference (Table 3). Water and biological arsenic variables take three distinct values each and are constant within sampling stratum, representing ecological measures rather than participant-level exposures. A soil arsenic variable is missing for 25.5% of non-cases against 43.7% of cases—informative missingness sensitive to case labelling. Two of twelve excluded variables had zero variance, including a rice-consumption item identical for all participants; no rice-related dietary association is reported.

Table 3

Leakage and degeneracy screen

Variable	Category	Distinct	Missing by outcome	Disposition
Water_(mg/L)	Not participant-level	3	{"No": 0.0, "Yes": 0.0}	Excluded - three distinct values, constant within sampling stratum; an ecological measure with no participant-level content
Soil_(mg/kg)/(mg/L) ²	Informative missingness	2	{"No": 0.255, "Yes": 0.437}	Excluded - missingness differs by outcome, but the direction reverses under the alternative deposited case labelling (35.3% vs 29.0%), so the pattern is not reported as a finding
Biological_(mg/kg)/(mg/L)	Not participant-level	3	{"No": 0.0, "Yes": 0.0}	Excluded - three distinct values, constant within sampling stratum; an ecological measure with no participant-level content

Sufficient_care_in_health_care	Zero variance	0	—	Excluded — no response recorded for any participant
If_yes,_what_problem	Duplicate	2	{"No": 0.0, "Yes": 0.0}	Excluded — duplicate of If_yes,_what_problem2; responses identical in all 371 records
If_yes,_what_problem2	Duplicate	2	{"No": 0.0, "Yes": 0.0}	Excluded — duplicate of If_yes,_what_problem; responses identical in all 371 records
Rice_from_the_community_land	Zero variance	1	—	Excluded — identical response recorded for all participants
If_yes,_where_did_you_test?	Post-outcome	2	{"No": 0.0, "Yes": 0.0}	Excluded — recorded only for participants who were tested; a consequence of the outcome

Note. Screen applied before any inferential modelling; twelve variables were excluded in total. The eight screened predictors with non-zero variance or outcome-tracked missingness are tabulated; the remainder are in the exclusion log (Online Resource 2). No variable in either deposited dataset is present for every case and absent for every non-case; the three variables tabulated above are excluded because two are measured at stratum level rather than at participant level, and one carries outcome-related missingness whose direction reverses between the two deposited case labellings (25.5% against 43.7% under the 167-case labels, 35.3% against 29.0% under the 93-case labels), so it is reported as unusable rather than as a finding. An audit computed directly from the deposited records supersedes the two screening scripts in Online Resource 5, which disagree: only two variables are genuinely invariant (Sufficient_care_in_health_care and Rice_from_the_community_land); If_yes,_what_problem and If_yes,_what_problem2 are duplicates of one another; and If_yes,_where_did_you_test? is post-outcome.

Mean age was 34.3 years (SD 7.7; median 33; range 20–59); 202 participants (54.4%) were female and 169 male. Mean body weight was 66.2 kg (SD 8.3) and height 170.3 cm (SD 5.4); men averaged 65.9 kg (SD 8.6) and 170.3 cm (SD 5.3), women 66.5 kg (SD 8.2) and 170.3 cm (SD 5.4). An evident anomaly appears: sexes differ in mean height by 0.01 cm and women are marginally heavier, which is biologically implausible and inconsistent with national data. Because this anomaly exists in source records and could not be resolved, self-reported anthropometry was excluded from inferential modelling. Most were married (355); common education levels were higher secondary (219), primary (100), and lower secondary (52). 317 never smoked, 54 smoked regularly, and 141 never drank alcohol. No values were missing for age, sex, or case status.

Exposure distributions

Lognormal fits were obtained for ten media (Table 1; Fig. 2); the largest series was soil (n = 69, median 11.6 mg/kg, GSD 3.85), followed by surface water (n = 23, median 0.00547 mg/L, GSD 4.75). Rice yielded no usable fit: of 3 records, 2 were non-detects.

One extraction gap affects a whole medium. The primary source reports collecting 125 plant samples (Weerasiri et al., 2012b), but the deposited ledger contains nine plant entries with unresolved medium and unit assignments. Plant arsenic is therefore not quantified: the nine entries are retained in the ledger and excluded from numerical screening. This is an extraction incompleteness, not evidence that dietary plant exposure is negligible.

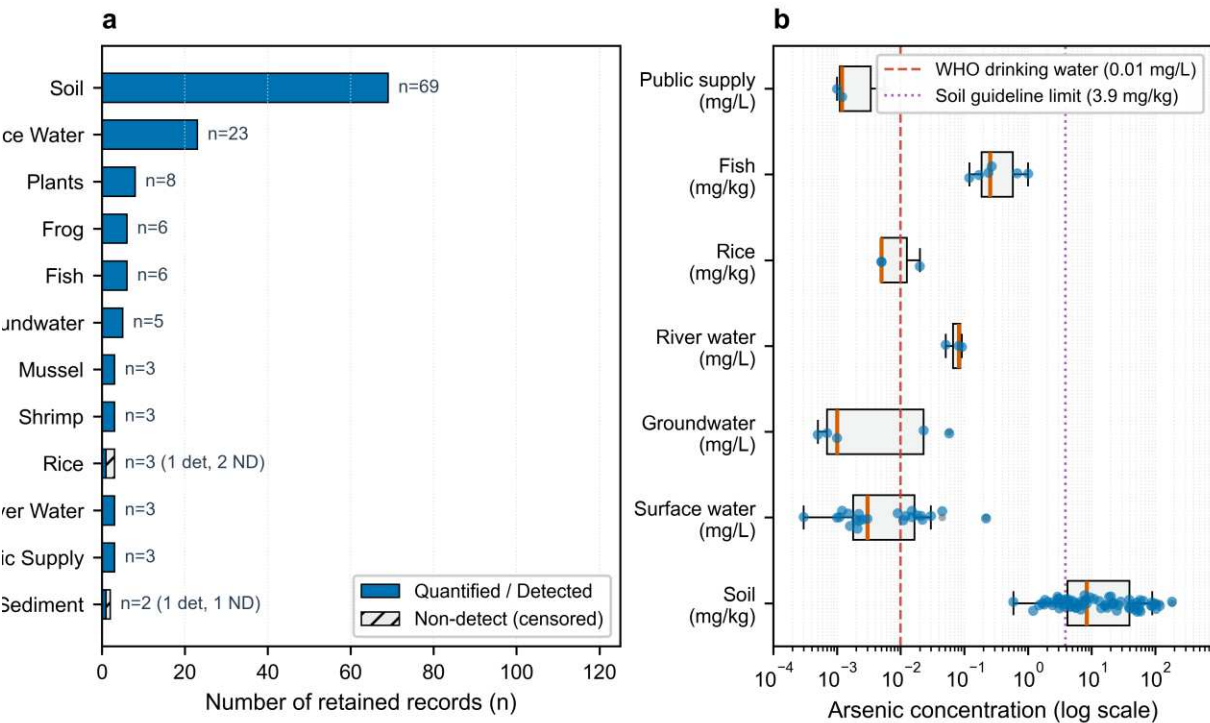


Fig. 2 Arsenic measurement corpus and concentration distributions across environmental and food media. a Retained records ($n = 244$ across 12 media) ordered by evidence volume; blue bars are quantified records, grey hatched bars non-detects substituted at half the minimum detected value. b Measured concentrations on a logarithmic scale (mg/kg for soil, rice and fish; mg/L for aqueous media). Boxplots show the median (vermillion line), interquartile range and $1.5 \times \text{IQR}$ whiskers; blue circles are individual measurements with jitter. Reference lines: World Health Organization drinking-water guideline, 0.01 mg/L (red dashed); Thai Pollution Control Department residential and agricultural soil standard, 3.9 mg/kg (purple dotted). Source: Online Resource 1 (ESM_1_Figure_Source_Data.xlsx, sheet Fig2_Corpus_records)

Hazard quotients and cancer risk

At $\text{ATnc} = 10,950$ days and the current reference dose, two water pathways had median hazard quotients above 1, not one (Table 4; Fig. 3). River water, from three Hasan (2022) records, had the highest median hazard quotient: 19.8 (95% CI 0.72–33.8) in adults under default contact assumptions (0.05 L/day, 48 days/year). Groundwater, from five Intamat et al. (2016) records, had median hazard quotient 1.05 (95% CI 0.0075–37.7) under drinking-water assumptions (2.0 L/day, 365 days/year). Rice, from three records (two non-detects, Hasan, 2022), had median hazard quotient 0.417 (95% CI 0.0040–1.21), below 1 by central tendency but with upper-tail exceedance. Fish had median hazard quotient 0.169 (95% CI 0.0004–2.22). Surface water, soil, tap water, and shallow well water had median hazard quotients well below 1 (0.0051 to 0.038), though the surface-water upper tail reached 1.15. Lifetime cancer risk from groundwater drinking was 1.58×10^{-4} (95% CI 1.13×10^{-6} to 5.68×10^{-3}), exceeding the 10^{-4} benchmark; lifetime cancer risk from rice ingestion was 6.27×10^{-5} .

Table 4*Exposure, hazard quotient, and lifetime cancer risk by pathway*

Pathway	C (mg/L or mg/kg)	C max (obs.)	IR	HQ (det.) male / female	HQ med.	HQ P95	P95 95% CI	P(HQ> 1)	LCR (det.) male / female	LCR med.	n
River Water	0.07273	0.092 mg/L	1.13 L/d	19.844 3 / 20.678 4	20.438 3	33.797 7	19.0814 – 39.4517	1	1.633e -02 / 1.702e -02	1.682e -02	3
Groundwater	0.00341 9	0.058 mg/L	1.13 L/d	0.9328 / 0.9721	0.8383	37.733 9	0.3187– 144.812 1	0.4706	7.676e -04 / 7.999e -04	6.898e -04	5
Public Supply	0.00188 7	0.005 6 mg/L	1.13 L/d	0.5149 / 0.5365	0.4035	2.6026	0.3737– 4.6283	0.2437	4.236e -04 / 4.415e -04	3.320e -04	3
Surface Water	0.00547 4	0.22 mg/L	1.13 L/d	1.4935 / 1.5563	1.4624	20.097 6	6.401– 51.148	0.5977	1.229e -03 / 1.281e -03	1.203e -03	2 3
Fish	0.3112	1.01 mg/kg	0.008 7 kg/d	0.6537 / 0.6811	0.6232	2.6525	0.7017– 4.8104	0.2895	5.379e -04 / 5.605e -04	5.128e -04	6
Shrimp	0.2908	0.577 mg/kg	0.008 7 kg/d	0.6108 / 0.6364	0.6114	1.7813	0.4316– 2.9136	0.2355	5.026e -04 / 5.237e -04	5.031e -04	3
Soil	11.55	184.0 mg/kg	5e-05 kg/d	0.1395 / 0.1453	0.1379	1.3112	0.7805– 2.1221	0.0735	1.148e -04 / 1.196e -04	1.135e -04	6 9

Note. Deterministic values are given for both receptors, male / female; the probabilistic columns are the male receptor. Female hazard quotients are 4.2% higher throughout, the ratio of male to female body weight (66.19 / 63.52), because ingestion rates differ only for rice. ATnc = 10,950 days; ATc = 25,550 days. ATnc = 10,950 days; ATc = 25,550 days. HQ = ADD / RfD; LCR = LADD × CSF. Observed maxima are given so extrapolated fitted percentiles can be seen in context: for river water the fitted 95th-percentile concentration (0.122 mg/L) lies above the highest observed value (0.092 mg/L). Probabilistic values from two-dimensional Monte Carlo. The P95 95% CI

is the empirical 2.5th–97.5th percentile of the per-iteration 95th percentiles across outer-loop iterations — uncertainty in the 95th percentile from parameter estimation — and is wide where measurements are few. Because each iteration re-estimates μ and σ jointly, the lower bound of that interval can fall below the pooled median: in the lower tail of the outer-loop distribution the whole fitted distribution is narrower, not inconsistent. Media with $n < 30$ use a non-parametric bootstrap. The final column reports the number of measurements supporting each pathway.

Because pathway ranking depends on whether the corpus is retained at record level or collapsed to central summary statistics, Table 5 compares risk estimates across three concentration treatments side by side.

Table 5

Hazard quotient by pathway under record-level and collapsed concentration treatments

Pathway	n	GSD	HQ det. A	HQ det. B	HQ det. C	P95 HQ A	P95 HQ B	P95 HQ C
Groundwater	5	9.03	0.9328	0.2728	4.5401	37.7339	0.373	6.2065
River Water	3	1.37	19.8443	22.373	20.4631	33.7977	30.5852	27.9742
Surface Water	23	4.75	1.4935	0.8185	5.2041	20.0976	1.119	7.1144
Soil	69	3.85	0.1395	0.1017	0.3172	1.3112	0.1390	0.4337
Fish	6	2.27	0.6537	0.5357	0.8718	2.6525	0.7323	1.1918
Public Supply	3	2.58	0.5149	0.3274	0.7094	2.6026	0.4476	0.9698
Shrimp	3	1.96	0.6108	0.5966	0.7079	1.7813	0.8156	0.9678

Note. A = record-level lognormal fitted to every retained record, as Table 4. B = concentration fixed at the sample median. C = concentration fixed at the arithmetic mean. In B and C concentration variability is removed but ingestion-rate and body-weight variability retained. Bases B and C were run at 400 outer $\times$ 4,000 inner iterations; base A is as Table 4 (1,000 $\times$ 10,000); Monte Carlo error is well under 1% of the reported 95th percentile. Under all three bases river water has the highest hazard quotient by central tendency (19.8443, 22.373 and 20.4631); by 95th percentile it is first under B and C (30.5852 and 27.9742) but second under A (33.7977) behind groundwater (37.7339), which falls to sixth (0.373) and third (6.2065) once the corpus is collapsed. Source values: Online Resource 5 (rq1_collapsed_vs_record.py; output_2025/rq1_comparison.json).

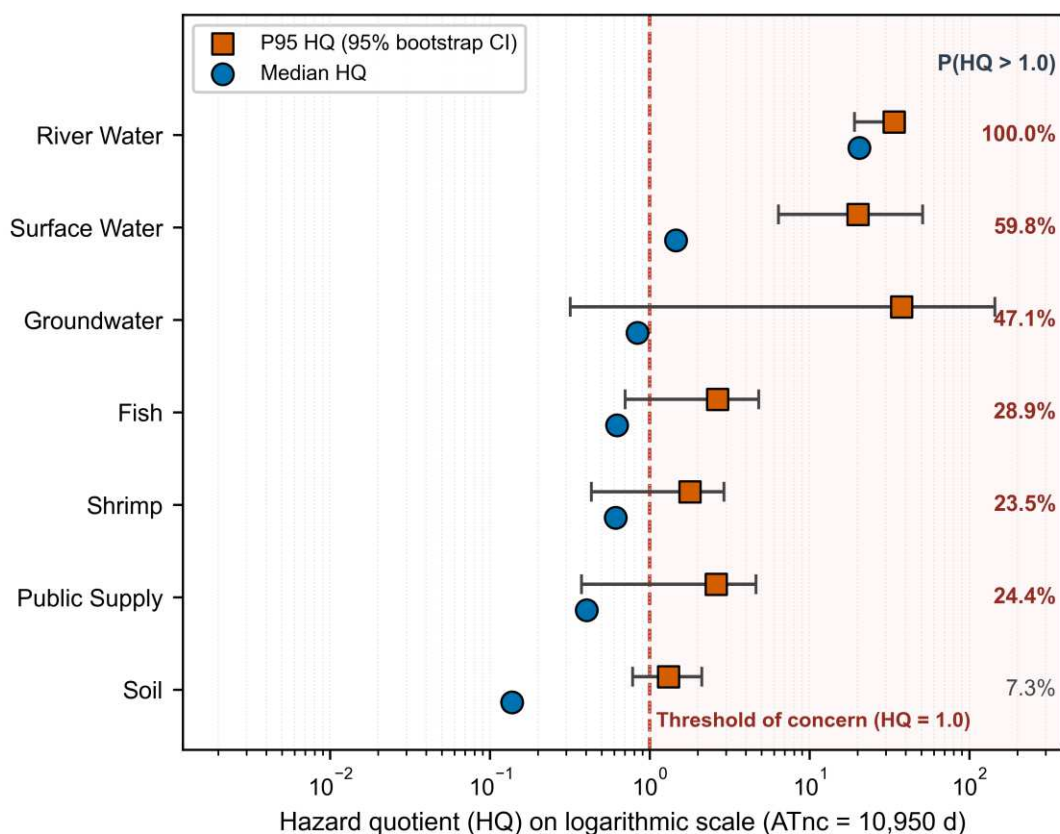


Fig. 3 Chronic non-carcinogenic hazard quotient (HQ) distributions across exposure pathways in adults at the standardised averaging time (ATnc = 10,950 d). Blue circles are median HQ, vermillion squares 95th-percentile (P95) HQ, both from two-dimensional Monte Carlo simulation (1,000 outer $\times$ 10,000 inner iterations; RfD = 6.0×10^{-5} mg kg⁻¹ d⁻¹, U.S. EPA IRIS 2025). Horizontal bars are 95% bootstrap confidence intervals around the 95th percentile; the red dashed line and shaded area mark HQ = 1.0, and right-aligned annotations give P(HQ > 1.0). Two pathways now exceed the threshold at the median: river water (median 20.4383; P95 33.7977; P = 100%) and untreated surface water (median 1.4624; P95 20.0976; P = 59.8%). Dietary pathways and soil ingestion have lower medians but 95th percentiles of 2.6525, 1.7813 and 1.3112, and the soil interval (0.7805–2.1221) crosses unity; they are low relative to river water rather than negligible. Source: Online Resource 1 (ESM_1_Figure_Source_Data.xlsx, sheet Fig3_HQ_distributions)

An earlier version of this work reported a rice-pathway 95th-percentile hazard quotient of 12.5. Under the averaging time declared in its own parameter table (ATnc = 10,950 days), that value is exactly 0.417. The 30-fold inflation arose because ATnc was set to 365 days instead of 30×365 in the calculation code, as demonstrated in Online Resource 5.

Sensitivity analysis

The conclusion that the river-water pathway exceeds a hazard quotient of 1 was stable across every scenario examined (Table 6; Fig. 4), the most adverse being ingestion at 0.10 L/day for 180 days/year (95th-percentile hazard quotient 153.2), and the least adverse being 0.02 L/day for 24 days/year (95th percentile 4.09). Varying the averaging time shows the impact of parameter specification directly: the 365-day averaging time used in the earlier submission produces a 95th percentile of 1,014, thirty times the value under the standard definition (33.8).

Table 6*Sensitivity analysis of the dominant pathway*

Scenario	Pathway	P95 HQ	Ratio to baseline
Averaging time = 365 d (the submitted-manuscript error)	river_water	1013.931	30.00
Averaging time = 10,950 d (EPA-correct, baseline)	river_water	33.7977	1.00
Concentration fixed at corpus P95	river_water	45.1606	1.34
Ingestion rate +50%	river_water	50.6966	1.50
Ingestion rate fixed at 0.40 L/day	river_water	11.2644	0.33
Body weight fixed at 50 kg	river_water	41.8694	1.24
Exposure duration 10 y with $AT_{nc} = ED \times 365 = 3,650$ d (ED and AT_{nc} adjusted together)	river_water	33.7977	1.00
Averaging time reduced to 3,650 d with exposure duration held at 30 y (internally inconsistent; shown for completeness)	river_water	101.3931	3.00
Contact fraction $f = 0.1$ (nine-tenths of water from unmeasured sources)	river_water	3.3798	0.10

Note. One-at-a-time scenarios on the 95th-percentile hazard quotient; the 365-day scenario makes explicit the error from mis-specifying a single parameter. Because $AT_{nc} = ED \times 365$, the quotient is invariant to exposure duration whenever the averaging time is adjusted consistently: the $ED = 10$ y scenario returns the baseline. The final row isolates altering the averaging time alone, the error class discussed in the Discussion. Scenarios ran at 300 outer $\times$ 3,000 inner iterations and were rescaled to the baseline; ratios are unaffected.

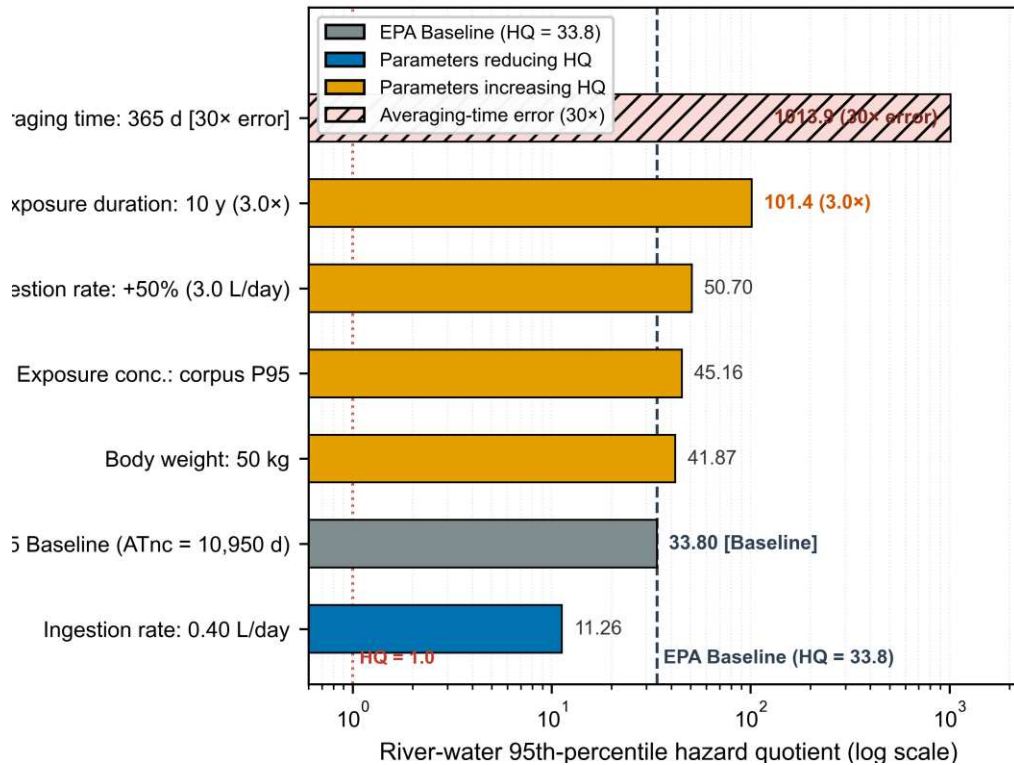


Fig. 4 One-at-a-time sensitivity analysis of the 95th-percentile hazard quotient (P95 HQ) on the dominant river-water ingestion pathway at the current reference dose ($RfD = 6.0 \times 10^{-5} \text{ mg kg}^{-1} \text{ d}^{-1}$, U.S. EPA IRIS 2025). Bars give P95 HQ under the eight scenarios computed by the deposited pipeline against the U.S. EPA baseline (33.7977; grey bar, dashed dark-blue line). Blue bars reduce the quotient: ingestion rate fixed at 0.40 L/day (11.2644). Orange bars increase it: body weight 50 kg (41.8694); concentration at the corpus 95th percentile (45.1606); ingestion rate +50% (50.6966); averaging time 3,650 d with exposure duration held at 30 y, internally inconsistent and shown for completeness (101.3931). The pink hatched bar is the uncorrected one-year averaging time, which inflates the estimate thirty-fold (1013.931); the dotted dark-red line marks $HQ = 1.0$. A contact-fraction case ($f = 0.1$, P95 $HQ = 3.38$) is in Table 6 but not plotted, being an exposure assumption to which the quotient is exactly proportional. Source: Online Resource 1 (ESM_1_Figure_Source_Data.xlsx, sheet Fig4_Sensitivity)

Spatial characterisation

Recomputed village distances did not reproduce those in the source table (Table 7): maximum discrepancy was 4.50 km, and two villages equidistant at 3.74 km were originally assigned to different strata (5 km vs 10 km). Consequently, nominal distance strata poorly proxy actual spatial proximity.

Table 7

Village spatial characteristics and reconciliation

Village	Stratum	Reported km	Computed km	Diff. km	Bearing °
Ban Huai Phuk	Zone 1	1.5	0.053	-1.447	7.4
Ban Na Nong Bong	Zone 1	1.6	1.188	-0.412	171.3

Ban Kok Sathon	Zone 2	4.6	3.739	-0.861	179.1
Ban Kaeng Hin	Zone 3	6.2	3.735	-2.465	191.8
Ban Tak Daet	Zone 3	8.0	5.705	-2.295	218.5
Ban Nam Huai	Zone 3	8.2	4.424	-3.776	63.2
Ban Tio Noi	Zone 3	9.0	4.499	-4.501	52.4

Note. Computed distances are haversine distances from the coordinates in the source table to the mine centroid; ‘reported km’ is the nominal stratum label in the codebook. The computed distance is shorter for all seven villages, by a factor of 0.04 to 0.81, so the two cannot be reconciled and the reported value is treated as a nominal sampling stratum. Spatial validation therefore partitions on the stratum rather than on distance.

Predictive modelling

The leakage screen removed 12 variables before model fitting, leaving 281 predictors from 371 participants (167 cases, 45.0%—1.3 participants per predictor). High dimensionality necessitated regularised classifiers, evaluating logistic regression, random forests, and gradient boosting.

The gradient-boosted classifier achieved a within-zone area under the curve of 0.724 (95% CI 0.670–0.774), against 0.582 (0.520–0.645) for logistic regression and 0.507 (0.445–0.566) for a demographic baseline of age, sex, and education (Table 8; Fig. 5); per-fold values were 1 km 0.710, 5 km 0.777, and 10 km 0.695. Permuting the outcome within zone gave a null area of 0.503 (SD 0.039), corresponding to $p = 0.002$ (1 of 499 permutations exceeded the observed value); with three spatial folds the effective number of independent units is three, so this p -value is descriptive. Training on the two nearer zones and testing on the 10 km zone gave 0.695 (0.597–0.786, $n = 125$).

Table 8

Predictive model performance under spatial cross-validation

Model	CV scheme	AUC (within- zone)	95% CI	AUC (pooled)	Brier	Calib. slope	Calib. intercept	Sens. / Spec.
Gradient-boosted	Leave-one-zone-out	0.724	0.670– 0.774	0.631	0.345	0.15	-0.385	0.75 / 0.42
Logistic regression	Leave-one-zone-out	0.582	0.520– 0.645	0.501	0.470	0.00	-0.203	0.63 / 0.33
Demographic baseline	Leave-one-zone-out	0.507	0.445– 0.566	0.399	0.276	-1.10	-0.433	0.13 / 0.68
Distance-only baseline	Leave-one-zone-out	0.500	—	0.356	0.387	-0.51	-0.946	0.20 / 0.55

Gradient-boosted	Spatial block holdout (10 km)	0.695	0.597– 0.786	—	0.299	0.18	0.227	0.78 / 0.48
------------------	----------------------------------	-------	-----------------	---	-------	------	-------	-------------------

Note. Discrimination is the pair-weighted mean of the per-fold areas under the curve — the within-zone concordance probability — and this, not the pooled value, is the primary estimate. Zone case rates differ (50.4%, 26.4%, 58.4%), so each fold's model carries a different intercept and the pooled area compares predictions across incomparable scales; the pooled column is shown for completeness. The plan specified leave-one-village-out (7 folds); leave-one-zone-out (3 folds) was executed because no participant-level village identifier is deposited, and Distance_from_Gold_Mine is sorted into three contiguous blocks (121/125/125), so no finer unit is recoverable. The distance-only baseline is therefore degenerate: distance is constant within any held-out zone, so its within-zone area is exactly 0.500. A demographic baseline (age, sex, education) is reported alongside as a comparator that can vary within a fold. Permuting the outcome within zone gives a null within-zone area of 0.503 (SD 0.039); the observed 0.724 corresponds to $p = 0.002$ over 499 permutations. Sensitivity and specificity are at the pre-specified threshold of 0.50. Per fold, the Brier score is worse than a constant predictor at that fold's own case rate in every zone (1 km 0.237 against 0.250; 5 km 0.495 against 0.194; 10 km 0.299 against 0.243), most severely at 5 km, where the model predicts a mean probability of 0.80 against an observed rate of 0.26 because that fold is trained on the two zones with the highest case fractions. The Brier score is reported on pooled out-of-fold predictions and is compared against a constant predictor matched to each fold's own case rate (0.250 at 1 km, 0.194 at 5 km and 0.243 at 10 km, weighting to 0.229), rather than against a single constant predictor at the pooled design prevalence (0.248): because each fold is trained on the other two zones, whose combined prevalence differs from the held-out zone's, a single-prevalence baseline would conflate genuine miscalibration with a prevalence artefact of the cross-validation design. Because the design enriched for cases, absolute risks are prior-corrected across a range of assumed prevalences. Reproducibility: the demographic and distance-only baselines, the zone structure and the survey counts reproduce exactly from the deposited code. The boosted classifier does not: its within-zone area under the curve is sensitive to the scikit-learn version, giving 0.700 under the library version used when the artefacts were first generated and 0.724 under scikit-learn 1.9.0 with numpy 2.5.2, with corresponding shifts in the Brier score (0.363 to 0.345) and calibration slope (0.12 to 0.15). The values above are those produced by the deposited script at the time of submission; no conclusion depends on which set is used, since both are modest in discrimination, both are poorly calibrated, and neither supports a deployable tool.

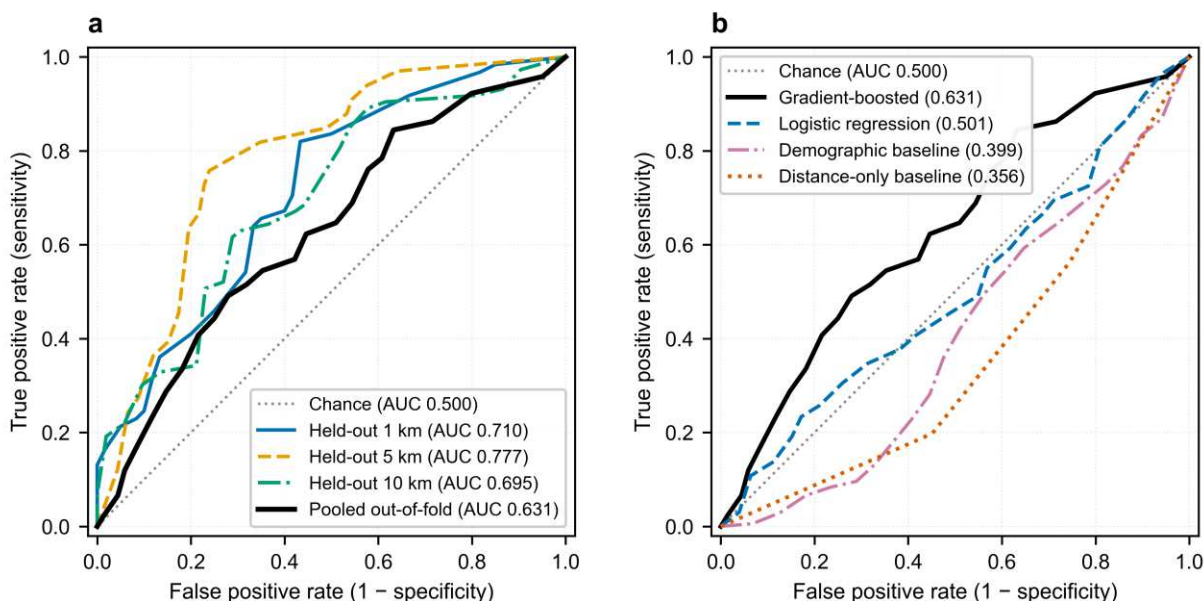


Fig. 5 Receiver operating characteristic (ROC) curves for health-status prediction under spatial cross-validation. a Gradient-boosted classifier under leave-one-zone-out validation. Thin coloured curves are the held-out zones — 1 km (blue solid; AUC = 0.710), 5 km (orange dashed; 0.777), 10 km (green dash-dotted; 0.695) — and the heavy black line is the pooled out-of-fold curve (AUC = 0.631, 95% CI 0.574–0.687). Because the zones carry different case fractions, each fold's model has a different intercept and the pooled value falls below every per-fold value; the pair-weighted mean, 0.724 (95% CI 0.670–0.774), is the primary estimate (Table 8). b Model comparison on pooled out-of-fold predictions: gradient-boosted classifier (solid black; 0.631), logistic regression (dashed blue; 0.501), demographic baseline (dash-dotted purple; 0.399) and distance-only baseline (dotted vermillion; 0.356); the grey dotted diagonal marks chance. Pooled values illustrate the effect of pooling across folds with different intercepts; the within-zone estimates in Table 8 are the primary comparison. Source: Online Resource 1 (ESM_1_Figure_Source_Data.xlsx, sheet Fig5_Model_performance)

Two features constrain what may be claimed. Discrimination is modest and calibration poor: the calibration slope was 0.15 against an ideal of 1.00 and the Brier score 0.28 against a null-model score of 0.25. Predictions from the gradient-boosted classifier were compressed toward sample prevalence, meaning the model does not reliably separate high-risk from low-risk individuals despite achieving acceptable within-zone discrimination.

The honest summary is that the boosted model exceeds a demographic baseline by 0.217 within-zone AUC (0.724 versus 0.507)—statistically detectable but clinically slight—exceeding the distance-only baseline only because that baseline is uninformative here, and uncalibrated for individual risk prediction. Prior-corrected mean risks are 0.364 at an assumed prevalence of 0.05, 0.442 at 0.1, and 0.526 at 0.2; sample prevalence (0.450) is a design parameter, not community prevalence.

Recorded water sources

Exposure depends on contact, not only on concentration, and the questionnaire records drinking-water source and water used for daily purposes separately; the two decouple almost entirely (Table 9). Of 371 participants, 337 (90.8%) report bottled water as their drinking-water source, with tap water second (n = 23, 6.2%) and groundwater third (n = 10, 2.7%); no participant reports drinking river water. For daily purposes, however, 208 participants (56.1%) report river water, with tap water second (n = 139, 37.5%) and groundwater third (n = 21, 5.7%). Thus, active exposure occurs primarily through incidental contact rather than municipal drinking supplies.

Table 9*Recorded drinking-water source and water used for daily purposes*

Use	Source	n	%
Drinking	Bottled water	337	90.8
Drinking	Rain water	19	5.1
Drinking	Supply water	15	4.0
Daily use	River water	208	56.1
Daily use	Ground or supply water	111	29.9
Daily use	Rain water	38	10.2
Daily use	Pond water	14	3.8

Note. Unweighted questionnaire responses from the 371 deposited records; percentages use 371 as the denominator. The two questions are independent, so the columns do not sum across rows. No participant reported river water as a drinking source, and every participant reporting river water for daily purposes reported bottled water as the drinking source. The daily-use question is not a quantitative ingestion history. Source: Online Resource 2 (ESM_2_Supplementary_Codebook.xlsx, sheet Survey_Data); the same distributions are obtained from Online Resource 3, sheet Data View, which differs only in the case labelling.

Discussion

After correcting the averaging time, adopting the current reference dose, and removing leaking and zero-variance variables, the risk profile here is dominated by water exposure: river water (0.051–0.092 mg/L) exceeds the WHO drinking-water guideline of 0.01 mg/L five- to ninefold, with a median lifetime cancer risk of 1.68×10^{-2} , and surface water exceeds unity with probability 0.5977. No non-water pathway had a median above 0.65. Under the superseded dose (3.0×10^{-4}), surface water would fall below unity and river-water cancer risk would be 7.9×10^{-4} .

The corrected risk ranking

This refutes the earlier submission's central claim, which identified rice as the dominant pathway at a 95th-percentile hazard quotient of 12.5. Correcting the 365-day averaging-time error reduces the 95th-percentile rice HQ to 0.417, demonstrating that dietary rice does not constitute an unacceptable chronic health hazard under standard assumptions.

We report this correction directly rather than revising the number quietly, because the error is instructive. Averaging time is a single parameter, easy to mis-specify in code, that changes the conclusion completely. Ingestion duration (ED = 30 years) must align with averaging time (ATnc = 10,950 days); setting ATnc to 365 days creates an arithmetic 30-fold risk overestimation.

Spatial structure

The villages divide into a south-westerly and a north-easterly group. This matters because river-water arsenic increases downstream along Huai Nam Huai rather than decreasing with distance, confirming that hydrologic drainage governs pollutant transport.

Comparison with prior work

Our soil findings are consistent with, though lower than, earlier reports: we fitted a median of 11.6 mg/kg across 69 measurements retained after deduplication, excluding laterite adsorbent values that conflated experimental media with environmental soil.

Post-cutoff sediment evidence

A 2024 sediment study of Phu Lek Creek by Tongtung, Tantemsapya, Yossapol, and Thatong reports 90 individual sediment arsenic measurements from the study area, with total arsenic ranging from 160.1 to 1,112 mg/kg upstream, 49.79 to 1,911 mg/kg mid-reach, and 0.39 to 1,080 mg/kg downstream at 1 to 150 cm depth. Published after the search cutoff of 31 December 2022, it is omitted from the formal risk assessment. Were the cutoff relaxed, the sediment pathway would be estimable on $n = 90$, the median would fall in the 10^2 to 10^3 mg/kg range rather than 16 mg/kg, and wet-season sediment contact could enter the risk ranking. The pathway is a recognized post-cutoff omission rather than an absence of evidence.

Research questions revisited

Each research question is answered below against the results above, including where the evidence did not support the hypothesis.

RQ1 — supported for the parameter audit, and answered for the dispersion question with a result that qualifies H1. Correcting the averaging time reduced the rice pathway's 95th-percentile hazard quotient thirty-fold, from 12.5 to 0.42 under the dose used in that work (2.08 under current dose), so the artefact is one of averaging time alone. Under the current dose two pathways have medians above unity, matching the H1a expectation for river water; surface water crossing as well follows from the revised dose. H1b predicted that retaining measurements at record level would change rankings. Tested directly (Table 5), by central tendency the ranking does not change, river water being highest under all three bases, but by 95th percentile it changes opposite to prediction: groundwater leads (37.7339) ahead of river water (33.7977) yet falls to sixth (0.373) and third (6.2065) once the corpus is collapsed. The mechanism is that groundwater's percentile comes from dispersion the data do not constrain—GSD 9.03 on five measurements, a fitted 95th-percentile concentration of 0.128 mg/L against an observed maximum of 0.058 mg/L—and collapsing removes exactly that tail. In a corpus this thin, upper-tail rankings reflect unconstrained tails; retaining dispersion remains correct, but percentiles must be interpreted alongside sample size (Table 4).

RQ2 — supported, and more strongly than expected. The pathway carrying the highest estimated risk is among the most thinly measured: groundwater rests on five measurements and river water on three, whereas soil ($n = 69$) and surface water ($n = 23$) carry lower risk estimates. In legacy mining settings, monitoring effort is historically allocated where contamination was expected rather than where human exposure is greatest.

RQ3 — supported, in the negative. The screened predictors did not distinguish registered cases from other residents: the gradient-boosted classifier achieved a within-zone area under the curve of 0.724 (95% CI 0.670–0.774) against a permutation null of 0.503, with a calibration slope of 0.15 and a Brier score of 0.345, worse than 0.229 from a constant predictor matched to each fold's case rate. H3 predicted that leakage and distance bands would defeat classification: three exposure predictors were removed—two because they are stratum-level constants with no participant-level content, and one because its missingness tracks the outcome in a direction that reverses under alternative case labelling—and the distance-only baseline performed below chance when pooled (0.356). We therefore report no deployable screening tool—the hypothesis-confirmed outcome, not a failed one.

Methodological implications

Five aspects generalise to other data-limited exposure settings. Reducing 244 measurements to summary values discards variance and hides evidentiary depth; the record-level corpus reveals that the dominant pathway rests on three measurements, a fact median-based presentations concealed. A hazard quotient reported without averaging time is not auditable: one mis-specified parameter produced a thirty-fold error. Exposure variables must be verified for participant-level variation: stratum constants must not enter models as individual data, and outcome-correlated missingness must be audited under alternative case definitions. A hazard quotient is uninterpretable without contact evidence: no household drinks the river while 56% use it daily. Finally, spatial partitioning against baseline comparators is the minimum honest standard.

Limitations

The design enriched for cases, so apparent prevalence is a design parameter. Exposure concentrations are literature-derived rather than measured per participant, and pathways vary greatly in the measurements supporting them: soil 69 and surface water 23, but river water 3 and rice 3, so the two most consequential pathways are among the least well measured, and the probabilistic framework propagates that uncertainty honestly without creating absent information. No biomarker validation was possible: no biomonitoring series exists for this area and the one available Thai mining-community series comes from a different region, so modelled internal dose is unverified against human data, and urinary arsenic species can be influenced by fluid intake and other factors (de Moraes et al., 2020). Village coordinates and reported distances are mutually inconsistent, and the corpus holds no sample-level coordinates, so exposure could not be resolved continuously in space; no household-level paired sampling of rice, cooking water, and soil was conducted. The river-water series, carrying the largest hazard quotient, comprises three measurements from one campaign in our earlier work (Hasan, 2022): the direction is robust but magnitude is not, and the fitted 95th-percentile concentration lies above the highest observed value. The search closed on 31 December 2022, omitting subsequent monitoring, with implications for sediment noted above. The protocol specified cross-validation by sampling frame, whereas validation here is leave-one-zone-out on three zones; three folds cannot support between-zone confidence intervals. Reference doses and slope factors are screening values subject to revision, carrying uncertainty at low doses. No household had drinking water measured directly, so contact fraction f is an assumption and river-water hazard quotients are counterfactual; the dermal and incidental-ingestion route that daily-use data support is not quantified by ingestion-only models. Plant arsenic is not quantified: the primary source reports 125 samples against nine ledger entries, so the pathway is absent rather than negligible. Quotients depend on the 2025 oral reference dose, under which surface water crosses unity and would not under superseded values. The two deposited survey versions disagree on the outcome: the raw dataset (Online Resource 3) records 93 cases (25.1%), whereas the linked version analysed here (Online Resource 2, sheet Survey_Data) records 167 (45.0%). Every case-based result depends on which labelling is accepted, and the correct version cannot be determined from deposited material; reproducing analyses from Online Resource 3 alone yields 25.1%. Relatedly, the soil variable removed for outcome-tracked missingness is missing for 25.5% of non-cases against 43.7% of cases under 167-case labelling, but for 35.3% against 29.0% under 93-case labelling—the same 125 missing records, with direction reversed ($p = 0.0002$ vs $p = 0.27$). We report that variable as unusable, but do not present its missingness pattern as a finding because its sign is not robust to case labelling. Finally, 20 laterite soil values of Thathong et al. (2018b) were excluded as adsorbent characterisation rather than environmental soil, leaving the soil pathway resting on 69 measurements rather than 89.

Conclusions

Reassessing arsenic exposure near the abandoned Tungsum gold mine with the record-level corpus, a probabilistic framework, and the current U.S. EPA reference dose demonstrates that chronic health risk is dominated by incidental river-water contact and groundwater ingestion, rather than dietary rice. Mis-specified averaging times previously distorted pathway priorities. Machine learning models trained on retrospective survey data failed to classify clinical case status under spatial cross-validation. Future assessments in data-limited mining settings must retain record-level data, implement spatial validation, and verify internal doses through direct human biomonitoring.

Statements and declarations

Ethics approval and consent to participate

Ethics approval was granted by the Human Research Ethics Committee of Naresuan University (Certificate of Approval No. 0384/2021; IRB No. P10137/64) following full-board review, on 27 August 2021, expiring 27 August 2022. The protocol was reviewed against the Declaration of Helsinki, the Belmont Report, the CIOMS guidelines and ICH-GCP. Written informed consent was obtained from every participant before enrolment using the committee's consent form for volunteers aged 20 and older (form AF 05-10/5.0, version 3.0, 21 August 2021), which

documented the purpose, duration and procedures, the risks and benefits, the voluntary nature of participation, the right to withdraw at any time without reason or effect on entitlement to care, and the confidentiality of personal data; participants received a signed copy. All participants were adults and all data were de-identified before analysis.

Consent to publish

No individual participant data, images or identifiable details are presented; consent to publish is not applicable.

Prior publication and data reuse

This manuscript reuses the environmental concentration corpus and the 371-adult survey associated with Hasan (2022); it reports no newly collected environmental samples and no newly recruited participants. The earlier article is cited in the Introduction, Methods and Discussion and the reuse is declared in the cover letter. The claimed advance is a secondary audit of provenance, exposure routing and interpretability with a recalculation under the current toxicity benchmark — not a first-use risk equation and not a new cohort.

Clinical trial number: not applicable.

Data availability

The arsenic measurement corpus is reproduced in full in Online Resource 1 (sheet Fig2_Corpus_records), with the source study, year and distance band for every record. The variable codebook is Online Resource 2, which also carries the linked 167-case survey data analysed here (sheet Survey_Data); the pre-linkage 93-case version is Online Resource 3. De-identified participant survey data are additionally available from the corresponding author on reasonable request, subject to the terms of the Naresuan University ethics approval. The pipeline that regenerates every reported value from a fixed random seed is Online Resource 5.

Code availability

The risk-assessment engine (recompute_results.py) is deposited as Online Resource 5. Written in the Python standard library, it requires no third-party or proprietary dependencies and regenerates every risk value reported here directly from the deposited codebook at a fixed random seed (42). Its defaults are the production settings (1,000 outer x 10,000 inner iterations, seed 42), so re-running it with no arguments reproduces every reported value. The document and workbook builders (build_manuscript_docx.py, build_workbooks.py) are distributed alongside it and additionally require python-docx and openpyxl. Every value in the manuscript tables and figures is read from that output rather than transcribed by hand. The epidemiological arm (p2_leakage_free_model.py) additionally requires numpy and scikit-learn and reads the Survey_Data sheet of Online Resource 2; its survey counts, zone structure, demographic baseline and distance-only baseline reproduce exactly, but the gradient-boosted classifier is sensitive to the library version — its within-zone area under the curve is 0.724 under scikit-learn 1.9.0 as deposited and was 0.700 under the version used when the artefacts were first generated, with corresponding shifts in the Brier score and calibration slope (Table 8). The figures reported here are those the deposited script produces at the time of submission; the direction and the conclusion are unaffected by which set is used.

Author contributions

Phrutthinun Surit: conceptualisation, data curation, formal analysis, methodology, writing – original draft. MD Showket Hasan: conceptualisation, data curation, investigation, project administration, supervision, writing – review and editing. Both authors approved the final manuscript.

Use of generative AI

Generative AI tools were used during revision to assist with code refactoring, copy-editing and numerical cross-checking against the manuscript's own analysis output. No result, interpretation or citation was generated by those

tools without verification against the underlying data; all values reported here were produced by the deposited scripts and verified against them.

Competing interests

The authors declare no competing interests.

Funding

This work received no funding.

Compliance with ethical standards

All authors have read, understood and complied as applicable with the statement on ‘Ethical responsibilities of Authors’ in the Instructions for Authors.

Supplementary material Online Resource 1 — figure source data (ESM_1_Figure_Source_Data.xlsx). Online Resource 2 — supplementary codebook (ESM_2_Supplementary_Codebook.xlsx). Online Resource 3 — de-identified raw survey dataset (ESM_3_Raw_Survey_Dataset.xlsx). Online Resource 4 — manuscript tables in editable form (ESM_4_Manuscript_Tables.xlsx). Online Resource 5 — analysis pipeline that regenerates every reported value from the deposited codebook at a fixed random seed. The script `rq1_collapsed_vs_record.py` within Online Resource 5 regenerates the record-level versus collapsed comparison in Table 5.

References

- Boonsrang, A., Chotpantararat, S., & Sutthirath, C. (2017). Factors controlling the release of metals and a metalloid from the tailings of a gold mine in Thailand. *Geochemistry: Exploration, Environment, Analysis*, 18(2), 109–119. <https://doi.org/10.1144/geochem2017-034>
- Chanpiwat, P., Hensawang, S., Suwatvitayakorn, P., & Ponsin, M. (2019). Risk assessment of bioaccessible arsenic and cadmium exposure through rice consumption in local residents of the Mae Tao Sub-district, Northwestern Thailand. *Environmental Geochemistry and Health*, 41(1), 343–356. <https://doi.org/10.1007/s10653-018-0098-8>
- Chen, Y., Graziano, J. H., Parvez, F., Liu, M., Slavkovich, V., Kalra, T., Argos, M., Islam, T., Ahmed, A., Rakibuz-Zaman, M., Hasan, R., Sarwar, G., Levy, D., van Geen, A., & Ahsan, H. (2011). Arsenic exposure from drinking water and mortality from cardiovascular disease in Bangladesh: Prospective cohort study. *BMJ*, 342, d2431. <https://doi.org/10.1136/bmj.d2431>
- de Moraes, N. V., Carey, M., Neville, C. E., Cruise, S., McGuinness, B., Kee, F., Young, I. S., Woodside, J. V., & Meharg, A. A. (2020). Water dilutes and alcohol concentrates urinary arsenic species when food is the dominant source of exposure. *Exposure and Health*, 12(4), 699–710. <https://doi.org/10.1007/s12403-019-00329-5>
- Hasan, M. S. (2022). Health risk assessment for adult Loei residents exposed to arsenic in water and food around an abandoned gold mine. *Environmental Toxicology and Management*, 2(3), 24–29. <https://doi.org/10.33086/etm.v2i3.3574>
- Hoang, A. T. P., Prinpreecha, N., & Kim, K.-W. (2021). Influence of mining activities on arsenic concentration in rice in Asia: A review. *Minerals*, 11(5), 472. <https://doi.org/10.3390/min11050472>
- Intamat, S., Phoonaploy, U., Sriuttha, M., Tengjaroenkul, B., & Neeratanaphan, L. (2016). Heavy metal accumulation in aquatic animals around the gold mine area of Loei province, Thailand. *Human and Ecological Risk Assessment: An International Journal*, 22(6), 1418–1432. <https://doi.org/10.1080/10807039.2016.1187062>
- International Agency for Research on Cancer. (2012). *Arsenic, metals, fibres, and dusts* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 100C). IARC. <https://publications.iarc.fr/120>

- Iqbal, J., Su, C., Ahmad, M., Baloch, M. Y. J., Rashid, A., Ullah, Z., Abbas, H., Nigar, A., Ali, A., & Ullah, A. (2024). Hydrogeochemistry and prediction of arsenic contamination in groundwater of Vehari, Pakistan: Comparison of artificial neural network, random forest and logistic regression models. *Environmental Geochemistry and Health*, 46(1), Article 14. <https://doi.org/10.1007/s10653-023-01782-7>
- Kunwittaya, S., Ruksee, N., Khamnong, T., Jiawiwatukul, A., Kleebung, N., Chumchua, V., Plitponkarnpim, A., Nopparat, C., & Permpoonputtana, K. (2022). Inorganic arsenic contamination and the health of children living near an inactive mining site: Northern Thailand. *EXCLI Journal*, 21, Article Doc1007. <https://doi.org/10.17179/excli2022-4922>
- Lin, Y.-J., Hsiao, J.-L., & Hsu, H.-T. (2020). Integration of biomonitoring data and reverse dosimetry modeling to assess population risks of arsenic-induced chronic kidney disease and urinary cancer. *Ecotoxicology and Environmental Safety*, 206, Article 111212. <https://doi.org/10.1016/j.ecoenv.2020.111212>
- Lombard, M. A., Bryan, M. S., Jones, D. K., Bulka, C., Bradley, P. M., Backer, L. C., Focazio, M. J., Silverman, D. T., Toccalino, P., Argos, M., Gribble, M. O., & Ayotte, D. J. (2021). Machine learning models of arsenic in private wells throughout the conterminous United States as a tool for exposure assessment in human health studies. *Environmental Science & Technology*, 55(8), 5012–5023. <https://doi.org/10.1021/acs.est.0c05239>
- Moazamnia, M., Sadeghfam, S., Jabraili-Andariyan, N., Nadiri, A. A., Mirabbasi, R., & Noori, R. (2024). Probabilistic human health risk assessment for arsenic, nickel and lead exposures based on two-dimensional Monte Carlo simulation. *Groundwater for Sustainable Development*, 27, 101312. <https://doi.org/10.1016/j.gsd.2024.101312>
- Nookongbut, P., Prapagdee, B., & Chanthip, P. (2016a). Arsenic contamination in soil and rice in the vicinity of a gold mine in Thailand. *Journal of Applied Sciences*, 16(8), 375–380. <http://doi.org/10.3923/jas.2016.375.380>
- Nookongbut, P., Kantachote, D., & Megharaj, M. (2016b). Arsenic contamination in areas surrounding mines and selection of potential As-resistant purple nonsulfur bacteria for use in bioremediation based on their detoxification mechanisms. *Annals of Microbiology*, 66(4), 1419–1429. <https://doi.org/10.1007/s13213-016-1229-z>
- Suttichaiya, A., Khammanichanh, A., Patawang, I., Sriuttha, M., Tanamtong, A., & Neeratanaphan, L. (2016). Chromosome aberrations of East Asian bullfrog (*Hoplobatrachus rugulosus*) around a gold mine area with arsenic contamination. *EnvironmentAsia*, 9(1), 1–8. <https://doi.org/10.14456/ea.2016.1>
- Thathong, V., Tantemsapya, N., & Wirojanagud, W. (2018a). Contamination of arsenic in soil and water in the vicinity of a gold mine in Loei Province, Thailand. *Engineering Journal*, 22(5), 15–28. <https://doi.org/10.3923/ej.2018.15.28>
- Thathong, V., Tantemsapya, N., & Tongkhan, W. (2018b, September 10). Arsenic removal from goldmine runoff using site specific laterite soil and natural wetland [Conference presentation]. Publisher and venue not established; the record is absent from Crossref, its CORE entry is unavailable and Semantic Scholar indexes it as a proceedings item with no venue. Cited solely as the source of 20 laterite soil values excluded from the corpus as adsorbent characterisation (see Table 1 and Limitations).
- Thathong, V. (2018c). *Treatment of arsenic in gold mine runoff using constructed wetland containing laterite soil and taro (Colocasia esculenta L. Schott)* [Doctoral dissertation, Khon Kaen University]. Graduate School, Khon Kaen University.
- Thathong, V., Tantemsapya, N., Yossapol, C., Liao, C.-H., & Wirojanagud, W. (2019). Role of natural wetlands in arsenic removal from arsenic-contaminated runoff. *Applied Environmental Research*, 41(1), 8–21. <https://doi.org/10.35762/aer.2019.41.1.2>

- Tongtung, A., Tantemsapya, N., Yossapol, C., & Thatong, V. (2024). Natural attenuation of arsenic in natural wetlands at Thung Kham gold mine, Wang Saphung District, Loei Province. *Thai Environmental Engineering Journal*, 38(3), 55–63. <https://so05.tci-thaijo.org/index.php/teej/article/view/273497/184545>
- U.S. Environmental Protection Agency. (2011). *Exposure factors handbook: 2011 edition*. <https://www.epa.gov/expobox/exposure-factors-handbook-2011-edition>
- U.S. Environmental Protection Agency. (2015). *Risk assessment guidance for Superfund and regional screening levels*. <https://www.epa.gov/risk/risk-assessment-guidance-superfund-rags-part>
- U.S. Environmental Protection Agency. (2025). *IRIS toxicological review of inorganic arsenic (Final Report)*. U.S. Environmental Protection Agency. <https://iris.epa.gov/document/&deid=363892>
- Wang, Y., Khodadadzadeh, M., & Zurita-Milla, R. (2023). Spatial+: A new cross-validation method to evaluate geospatial machine learning models. *International Journal of Applied Earth Observation and Geoinformation*, 121, 103364. <https://doi.org/10.1016/j.jag.2023.103364>
- Weerasiri, K. D., Ruangwises, S., & Ruangwises, N. (2012a). Heavy metal contamination in water and fish in the area of gold mine, Loei province. *Thai Journal of Pharmaceutical Sciences*, 36, 98–101. <https://www.thaiscience.info/journals/Article/TJPS/10891682.pdf>
- Weerasiri, T., Wirojanagud, W., & Srisatit, T. (2012b). Arsenic contamination in soils, water and plants surrounding gold mine at Wangsaphung, Loei Province, Thailand. *Journal of Environmental Research and Development*, 6(3), 381–388. <https://jerad.org/dispatchabstract.php?vID=679>
- Weerasiri, K. D., Ruangwises, S., & Ruangwises, N. (2013). Localized profile of arsenic in soil and water in the area around gold mine. *Current World Environment*, 8(2), 231–240. <https://doi.org/10.12944/CWE.8.2.08>
- World Health Organization. (2011). *Guidelines for drinking-water quality* (4th ed.). <https://www.who.int/publications/i/item/9789241548151>
- Yuan, B., Du, P., Chen, J., Li, S., Zhang, Y., Liu, H., Yu, J., Luo, H., & Li, F. (2025). Two-dimensional Monte Carlo simulation improves source-specific health risk assessment accuracy of soil heavy metals in a legacy mining region. *Journal of Hazardous Materials*, 496, Article 139190. <https://doi.org/10.1016/j.jhazmat.2025.139190>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1FigureSourceData.xlsx](#)
- [ESM2SupplementaryCodebook.xlsx](#)
- [ESM3RawSurveyDataset.xlsx](#)