

Google Search Anomalies Before IDSP-Recorded Dengue Outbreak Weeks in India: A Retrospective Multistate Proof-of-Concept Study, 2009–2022

Supplementary Information (Online Resource 1)

Contents

Supplementary Methods.....	2
SM1. Cohort provenance.....	2
SM2. Search dictionary.....	2
SM3. Event-centred extraction and quality control.....	2
SM4. Endpoint construction and corrected conditional timing	2
SM5. Exploratory predictive archive and reproducibility boundary	2
Supplementary Tables.....	3
Supplementary Table S1. Detailed cohort reconciliation.....	3
Supplementary Table S2. Locked event records by disease.....	4
Supplementary Table S3. Search dictionary composition by disease and language.....	5
Supplementary Table S4. Dictionary and extraction roles.....	6
Supplementary Table S5. Event-centred extraction completion.....	7
Supplementary Table S6. Quality of event-centred weekly series.....	8
Supplementary Table S7. Archived versus corrected timing derivation.....	9
Supplementary Table S8. Positive-week eligibility and control matching	10
Supplementary Table S9. Selection-bias audit for matched versus full dengue-positive weeks.....	11
Supplementary Table S10. Interpretive boundary for archived feature-set summaries.....	12
Supplementary Figures	13
Supplementary Fig. S1. State-disease search-series reliability scores.....	13
Supplementary Fig. S2. Archived all-disease relative-week anomaly profiles	14

Supplementary Methods

SM1. Cohort provenance

The Google Trends workflow received the 8,751-record harmonized EpiClim table rather than the 8,985-row upstream source. Three epidemiological week-53 records whose week-start dates could not be resolved were removed, followed by one repeated event row, producing a locked 8,747-record event panel. The primary-query event summary contained 8,717 linked records. Counts and disease-wise composition are provided in Supplementary Tables S1 and S2.

SM2. Search dictionary

The development dictionary included formal disease names, symptoms, diagnostic terms, treatment terms, prevention terms, and media/control terms in English, Hindi, and Romanized Hindi. Duplicate normalized queries within disease were flagged rather than automatically deleted, and low-specificity or high-ambiguity terms were retained for manual review. The dictionary was used for development and quality-control work; the primary dengue endpoint used only the query dengue. Composition and analytic roles are shown in Supplementary Tables S3 and S4.

SM3. Event-centred extraction and quality control

Event-centred Google Trends requests were run at Indian state or union-territory geography for the configured event window. Extraction used cached raw responses and resume-safe execution. Request completion is summarized in Supplementary Table S5. Series were subsequently classified by the executed quality rules using temporal granularity, observation coverage, zero rate, and variability. The final quality distribution is shown in Supplementary Table S6, and the corresponding constructed state-disease reliability scores are shown in Supplementary Fig. S1.

SM4. Endpoint construction and corrected conditional timing

The historical reference window was relative weeks -12 through -5 and the detection window was weeks -4 through -1. A detected event required at least one event-baseline robust z score of 2 or higher in the detection window. The archived `first_anomaly_ge_2_week` field searched the full pre-outbreak interval and could therefore return a reference-window week. It was not used for the revised conditional timing estimate.

The corrected derivation first restricted archived weekly anomaly flags to weeks -4 through -1 and then selected the earliest qualifying week. This yielded all 303 pandemic-excluded detected event rows, with a median conditional timing of 4 weeks and an interquartile range of 3-4 weeks. The contrast with the archived unrestricted-crossing summary is shown in Supplementary Table S7. Because week -4 was the earliest permitted observation in the detection window, the corrected distribution remains boundary-censored.

SM5. Exploratory predictive archive and reproducibility boundary

The exploratory predictive layer used matched case-control construction rather than a complete monitored state-week surveillance panel. Positive-week eligibility, matching yield, and control exhaustion are summarized in Supplementary Table S8. Matched dengue-positive weeks differed materially from the full positive pool, as shown in Supplementary Table S9. The feature-set summaries therefore remain descriptive within their constructed samples and are not interpretable as incremental discrimination attributable to Google Trends; this boundary is summarized in Supplementary Table S10. The core machine-readable files required to reproduce the event, lag, quality-control, and exploratory analyses are listed in Supplementary Table S11.

Supplementary Tables

Supplementary Table S1. Detailed cohort reconciliation

Stage	n / change	Comment
Upstream published source	8,985	Contextual source count
Input to Google Trends workflow	8,751	Harmonized cohort received by the workflow
Invalid mandatory week records	-3	Assam cholera; Gujarat cholera; Tamil Nadu acute diarrhoeal disease
After event-week validation	8,748	Valid week-start date
Repeated event row	-1	Odisha, Ganjam, acute diarrhoeal disease; week of 17 June 2013
Locked event panel	8,747	Outbreak-report rows
Primary-query summaries	8,717	30 locked event records did not link to a primary-query summary

Footnote. Counts refer to event-report rows unless otherwise indicated. Abbreviation: n, number of records or stated change in record count.

Supplementary Table S2. Locked event records by disease

Disease	Event rows	States / UTs	Districts	Years
Acute diarrhoeal disease	5,193	35	690	2009-2022
Chikungunya	729	21	195	2009-2022
Cholera	662	25	208	2009-2022
Dengue	1,621	34	345	2009-2022
Malaria	542	24	199	2009-2022

Footnote. These are locked event-panel records before search-series quality restriction. Abbreviation: UTs, union territories.

Supplementary Table S3. Search dictionary composition by disease and language

Group	Terms	English	Hindi	Romanized Hindi
Acute diarrhoeal disease	22	18	2	2
Chikungunya	19	14	2	3
Cholera	21	16	3	2
Dengue	25	19	3	3
Malaria	23	17	3	3
Generic controls	2	2	0	0
Total	112	86	13	13

Footnote. Ninety-eight terms were active for full extraction; the focused extraction used 22 terms. The primary endpoint used only the query dengue. No abbreviations are used in this table.

Supplementary Table S4. Dictionary and extraction roles

Layer	Size	Role
Development dictionary	112 entries	Ontology, language, and query-family development
Full extraction list	98 entries	Terms marked active for full extraction
Focused extraction	22 terms x 37 geographies x 3 historical windows	Anchor-based development and quality control; 1,038 successful requests
Primary event endpoint	1 query: dengue	State-level event-centred weekly series

Footnote. The primary endpoint did not combine the broader multilingual development dictionary. No abbreviations are used in this table.

Supplementary Table S5. Event-centred extraction completion

Outcome	Count	Rate / comment
Planned requests	1,332	100.0%
Weekly successful	1,125	84.5%
Valid no-data	171	12.8%
Retry-exhausted	36	2.7%
Terminal completion	1,296	97.3%
Collapsed weekly rows	60,627	249 series

Footnote. Terminal completion comprises weekly-successful and valid no-data requests. No abbreviations are used in this table.

Supplementary Table S6. Quality of 249 event-centred weekly series

Quality class	Series	Analytic role
HIGH	6	Strict cohort
MODERATE	12	Strict cohort
EXPLORATORY_ONLY	44	Expanded analyses only
UNUSABLE	187	Excluded from analytic cohorts

Footnote. The strict primary cohort required HIGH or MODERATE series quality plus minimum pre-outbreak coverage. No abbreviations are used in this table.

Supplementary Table S7. Archived versus corrected timing derivation

Derivation	Analytic period	n	Median	IQR	Interpretation
Archived unrestricted crossing	All years	356	4.5 weeks	3-5	First crossing anywhere in weeks -12 to -1; reference-window crossings can appear as lead
Correct window-first derivation	Pandemic excluded	303	4 weeks	3-4	Restrict to weeks -4 to -1 first, then identify the earliest qualifying crossing

Footnote. The corrected derivation uses archived weekly anomaly flags and does not refit a model. It remains conditional on detection and boundary-censored at week -4. Abbreviations: IQR, interquartile range; n, number of detected event rows.

Supplementary Table S8. Positive-week eligibility and control matching

Stage	n	Interpretation
Positive-week audit rows	6,025	Includes 281 rows without a series or outside the search range
Rows with history eligibility assessed	5,744	2,691 eligible; 3,053 ineligible
Full matches	345	Three controls obtained
Partial matches	44	One or two controls obtained
Controls exhausted	2,301	No matched controls remained
No control candidates	1	No candidate in the state-disease stratum
Primary matched dengue matrix	335	87 positive weeks; 248 control weeks

Footnote. Matching results document the constructed exploratory case-control archive and do not represent a complete monitored surveillance panel. Abbreviation: n, number of audit rows, matched positive weeks, or matrix observations as specified by row.

Supplementary Table S9. Selection-bias audit for matched versus full dengue-positive weeks

Variable	Matched n	Full n	Matched mean	Full mean	SMD	KS p
Calendar year	87	939	2011.4	2015.7	-1.805	6.5×10^{-40}
Baseline search median	87	939	3.93	14.09	-0.467	7.3×10^{-8}
Early-window z mean	87	939	4.43	2.38	0.185	0.0022
Early-window z maximum	87	939	8.70	5.24	0.164	0.00012

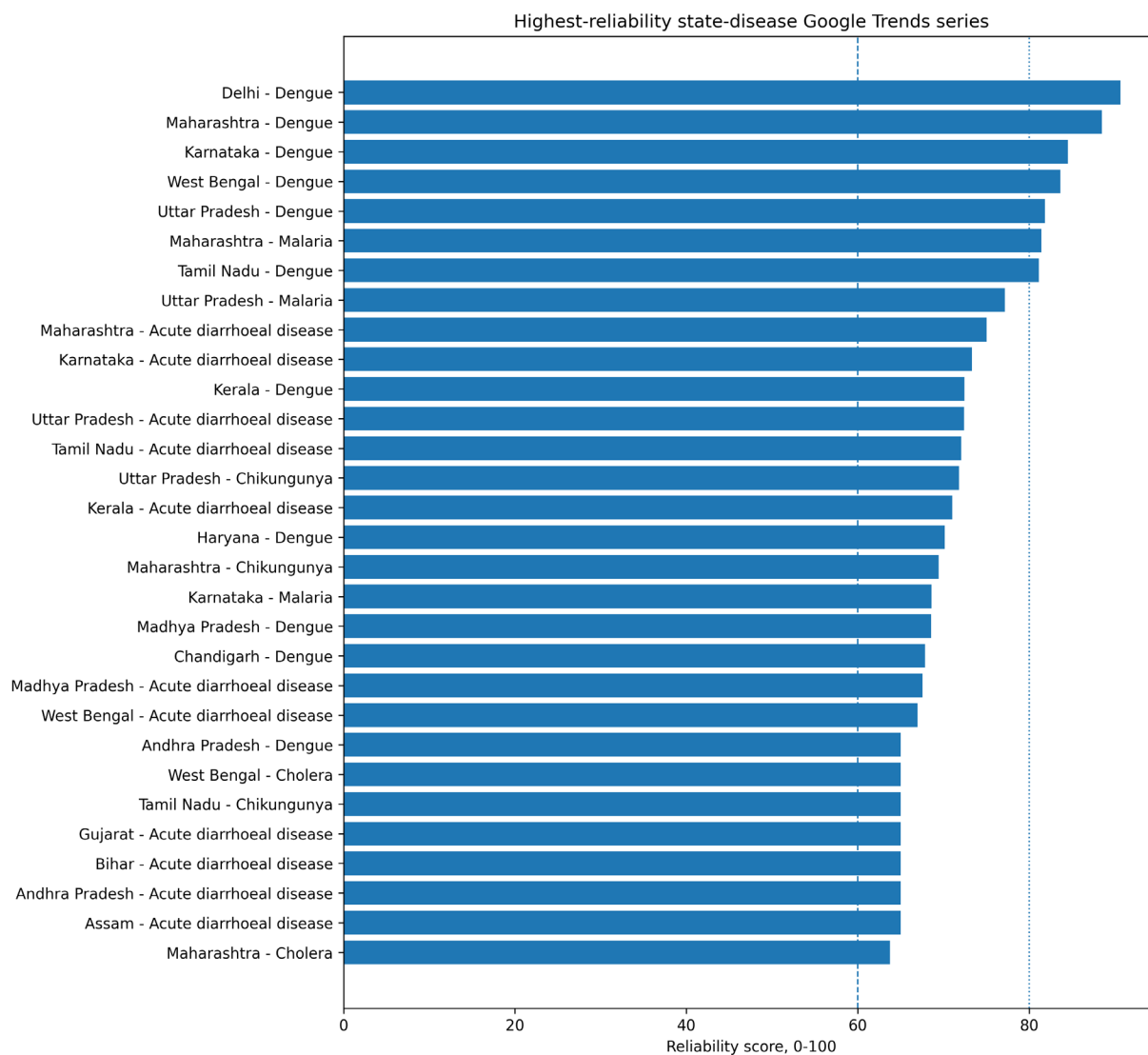
Footnote. Matched positives were substantially earlier and somewhat more anomalous than the full positive pool. Abbreviations: KS, Kolmogorov-Smimov; SMD, standardized mean difference; n, number of dengue-positive weeks; p, nominal probability value.

Supplementary Table S10. Interpretive boundary for archived feature-set summaries

Dimension	Finding
Observation sets	Selected feature-set summaries used different observation sets
Sample sizes	Sample sizes differed across summaries
Models	Model families differed across feature sets and strata
Validation designs	Expanding temporal, grouped-state, and leave-state-out results were mixed
Calibration	Identity, Platt, and isotonic calibration outputs were mixed
Permissible interpretation	Exploratory model performance within its own constructed sample
Prohibited interpretation	Incremental predictive value attributable to Google Trends

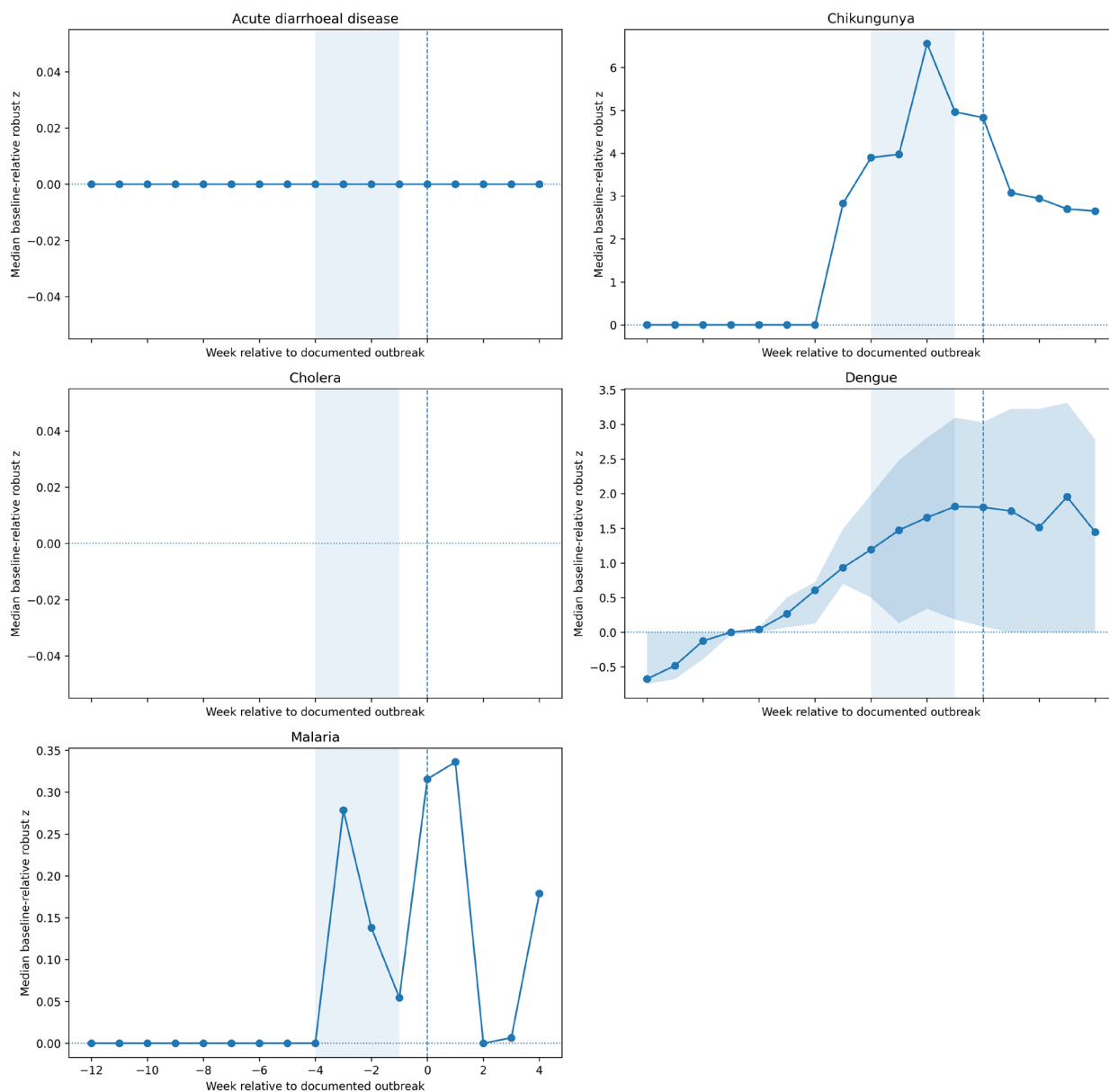
Footnote. Average-precision values from non-identical feature-set summaries must not be subtracted to claim incremental discrimination. No abbreviations are used in this table.

Supplementary Figures



Supplementary Fig. S1 Constructed state-disease Google Trends search-series reliability scores. The score summarizes series coverage, sparsity, and stability and does not measure outbreak-detection accuracy

Footnote. GT, Google Trends. The reliability score is a constructed quality-control metric on a 0-100 scale.



Supplementary Fig. S2 Archived all-disease relative-week anomaly profiles. Dengue is the only strict disease with broad state support; the chikungunya panel represents one strict event and cholera has no strict eligible event. The panels are retained for provenance and should not be interpreted as comparative disease performance

Footnote. No abbreviations are used in the figure. Relative week 0 corresponds to the EpiClim-recorded outbreak week; negative values precede that week.