Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

District level weekly surveillance data on notifiable infectious diseases in Khyber Pakhtunkhwa Pakistan from 2023 to 2026

Sara Nawab
University of Swat
Gulalai Khan
Swat Medical College
Aamir Abbas
Vital Pakistan Trust
Junaid Alam
Khyber Medical College
Sania Ayub
University of Swat
Mohammad Ali
University of Swat
Khalid Nawab
knawab@pennstatehealth.psu.edu

Penn State Health

Data Note

Keywords: Infectious disease surveillance, Pakistan, Khyber Pakhtunkhwa, IDSR, Notifiable disease reporting, Public health data, Disease panel data

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11018824/v1

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

Additional Declarations: No competing interests reported.

Abstract

This article presents a structured, long-format panel dataset of district-week notifiable infectious disease case counts for Khyber Pakhtunkhwa (KP) province, Pakistan, covering 164 weekly epidemiological bulletins published between April 2023 and July 2026. The data were extracted from Table 4 of the National Institute of Health (NIH) Pakistan Integrated Disease Surveillance and Response (IDSR) weekly Public Health Bulletins, which reports district-level case counts for the most frequently reported notifiable conditions in KP each week. The dataset comprises 59,824 district-week-condition observations spanning 44 canonical district- and sub-division-level reporting units and 16 distinct notifiable conditions. Six conditions — acute diarrhoea (non-cholera), malaria, influenza-like illness, acute lower respiratory infection in children under five, bloody diarrhoea, and typhoid — are reported in all 164 bulletins. Ten further conditions appear only in a subset of weeks, because the non-core columns of the source table rotate according to that week's case volume: from 2023 week 32 onward the table reports the highest-ranking conditions by provincial case total rather than a fixed set, producing 106 distinct column orderings across the corpus. The dataset documents this rotating-column structure explicitly, and is accompanied by a time-scoped district-name crosswalk resolving administrative-boundary changes over the study period — including one reporting label whose underlying geography is redefined mid-series — a week-by-week column-order log, a condition presence matrix, and machine-readable flags for identified source-side data-quality issues. No comparable structured, validated dataset derived from Pakistan's national disease surveillance bulletins is currently available in the public domain.

1 Introduction

Pakistan's Integrated Disease Surveillance and Response (IDSR) system publishes a weekly Public Health Bulletin through the National Institute of Health (NIH) [4]. Each bulletin includes a province-specific table, designated Table 4 in the Khyber Pakhtunkhwa (KP) series, reporting district-level case counts for the notifiable conditions under surveillance that week. The system has been assessed for its effect on surveillance management at health-facility level [5], and is the subject of a national five-year strengthening roadmap [6]. The bulletins are published openly as PDF documents and require no registration or data-use agreement to access.

The conditions carried in Table 4 are of substantive public health importance in this setting. Cutaneous leishmaniasis is endemic across KP, with a burden concentrated in the merged tribal districts and a substantial paediatric share [7, 8]. Typhoid is endemic, and Pakistan has been the focus of international concern over extensively drug-resistant Salmonella Typhi [9]. Diarrhoeal disease remains a leading cause of childhood morbidity across South Asia and exhibits marked spatiotemporal structure that district-resolution surveillance data are well suited to characterise [10]. Weekly district-level counts for these conditions in KP have not previously been available in analysable form.

Despite the openness of the source, the content of the bulletins has not previously been assembled into a structured, analysis-ready form. The practical obstacles are the format of the source rather than its

accessibility: the reporting table is not a fixed schema, the roster of reporting districts is unstable across the period, and the case counts are distributed across 164 separate PDF documents whose filenames follow no consistent convention. As a result, the district-level detail contained in this surveillance stream has remained effectively unavailable for quantitative research.

The dataset described here was assembled to remove that obstacle, and to serve as the shared data foundation for a set of planned analyses on climate variability and enteric disease in KP. The contribution is the dataset and its documented structure, not an epidemiological finding derived from it.

The resulting resource makes four contributions. It is the first structured, machine-readable and internally validated panel built from Pakistan's national IDSR bulletins, and no equivalent currently exists for the country. It documents and corrects a previously unreported structural property of the source table, a rotating, rank-ordered column format that would cause any position-based extraction of these bulletins to misattribute case counts to the wrong conditions. It supplies an explicit, time-scoped crosswalk mapping 50 raw district and sub-division labels onto 44 canonical reporting units, including one label whose underlying geography is redefined mid-series and which a static mapping would silently conflate. And it flags every identified data-quality issue explicitly rather than retaining or removing such records silently, so that secondary users can make informed and reproducible decisions about exclusion or imputation.

2 Data description

Table 1
Specifications of the deposited dataset.

Subject	**Public Health and Health Policy**
Specific subject area	District-level weekly notifiable infectious disease surveillance in Khyber Pakhtunkhwa province, Pakistan
Type of data	Table (CSV); Raw (case counts as printed in source) and Analysed (canonical condition and district labels, column-order and presence indicators, data-quality flags)
Data collection	Table 4 of the NIH Pakistan IDSR weekly Public Health Bulletins, the KP province district-wise surveillance table, was transcribed from 164 publicly available PDF bulletins covering 2023 week 16 and weeks 22–52, all of 2024, all of 2025, and 2026 weeks 1–28. Transcription used a generative AI document-reading tool, Claude (Anthropic), operating on each source PDF, preserving column headers, row labels and cell values exactly as printed. Transcribed data were then validated by the internal consistency checks described in Methods.
Data source location	Institution: National Institute of Health, Pakistan (publisher). Subject region: Khyber Pakhtunkhwa province (approximate centroid 34.97°N, 72.33°E). Country: Pakistan.
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/9yzvfgrhkt.1 Direct URL to data: https://data.mendeley.com/datasets/9yzvfgrhkt/1 Source bulletins are separately and publicly available from the publisher at https://www.nih.org.pk/; verified retrieval URLs for all 164 bulletins used are included in the deposited dataset.
Related research article	None.

The repository contains eleven files in a single flat directory, with no subfolders. Eight are comma-separated value files holding the panel and its supporting tables; the remaining three are a documentation file, the raw transcribed source text, and the parser that generates the CSV files from it. Every file is sorted chronologically where it carries a time dimension. Each is described individually below, in the order it appears in the repository listing.

idsr_kp_long.csv is the primary file: the full panel in long (tidy) format, with one row per district per week per condition. Fields are *year*, *epi_week*, *district_as_printed*, *district* (canonical), *unit_type* (district, subdivision, or typo), *parent_district*, *condition_as_printed*, *condition* (canonical), and *cases*. The file contains 59,824 rows covering 164 weeks, 44 canonical reporting units and 16 canonical conditions. A blank *cases* value corresponds to a printed value of “NR” (not reported) in the source and is distinct from a reported zero; 287 such cells occur across the panel.

Table 4_columns_by_week.csv records the exact column position and canonical mapping of every condition in each of the 164 bulletins, providing the evidence base for the rotating-column structure

described below.

condition_presence_matrix.csv gives a 16 × 164 condition-by-week grid indicating which conditions appeared in the table each week.

district_crosswalk.csv maps every district and sub-division label observed in the source (50 raw labels) to its canonical reporting unit (44 units), with a *unit_type* classification and, for sub-division rows, a *parent_district* reference. Two further fields, *applies_from* and *applies_to*, carry the validity window for labels whose referent changes during the study period; they are populated for the two time-scoped "SWA" entries and empty for all other labels, whose mapping is constant.

data_quality_flags.csv provides machine-readable flags for the identified duplicate source table, four non-data rows dropped from the panel, the week in which the "SWA" label is redefined, and the single week in which two overlapping South Waziristan reporting units appear simultaneously.

idsr_kp_swat.csv contains Swat district's row for every one of the 164 weeks in wide format, one column per canonical condition. It is a convenience extract: every value in it is also present in idsr_kp_long.csv.

swat_acute_diarrhoea_series.csv contains Swat's acute diarrhoea (non-cholera) time series alone, across all 164 weeks. It is likewise a convenience extract, provided because this series is the outcome variable of a planned climate-health analysis.

bulletin_links_verified.csv lists the verified retrieval URL for every bulletin used, together with the five weeks confirmed absent from the source archive.

README.md is a plain-text documentation file describing the repository contents, the field definitions of the primary file, the three properties of the data that most affect its reuse (right-censoring of rotating conditions, the mid-series redefinition of one district label, and the absence of denominators), and the full list of flagged data-quality issues. It duplicates no data.

tables_raw.txt is the complete raw transcription of Table 4 from all 164 bulletins, exactly as read from the source PDFs and before any canonicalisation, correction or sorting. Each bulletin occupies a block headed "### YEAR WEEK", followed by a pipe-delimited header row and one pipe-delimited row per district. It is included so that every published value can be traced back to the text it was derived from.

build.py is the parser that generates all eight CSV files from tables_raw.txt. It carries the condition-header normalisation map, the district crosswalk including its time-scoped entries, the dropped-cell corrections, and the build-time integrity assertion described under Validation. It requires only the Python standard library; setting its RAW and OUT paths and running it reproduces the deposited CSV files exactly.

2.1 Coverage

The panel covers 164 of the 169 epidemiological weeks in the target window: 2023 week 16 and weeks 22–52, all 52 weeks of 2024, all 52 weeks of 2025, and weeks 1–28 of 2026. The five absent weeks (2023 weeks 17–21) were confirmed by repeated inspection of the publisher's own bulletin index to be genuinely missing from the source archive, rather than a failure of retrieval.

The number of district rows reported per week ranges from 28 to 40, with a median of 37. This variation reflects genuine week-to-week differences in which districts submitted reports, together with the administrative reorganisation described under Canonicalisation, and is not an artefact of extraction. Because an undercount of rows would produce exactly this appearance, the claim was tested directly: the row counts of the bulletin at the minimum of the range (2023 week 16, 28 rows) and of the largest-discrepancy week (2023 week 41, 34 rows) were counted by hand in the source PDFs and both match the transcription exactly.

2.2 The rotating-column structure

Table 4 in the source bulletins is not a fixed-schema table. From 2023 week 32 onward, covering 153 of the 164 weeks, each week's table reports the notifiable conditions with the highest province-wide case totals for that week, ordered from highest to lowest. This is directly evidenced by the printed "Total" row, which is strictly descending in these weeks, and by the 106 distinct column orderings observed across the corpus.

In the remaining 11 weeks (2023 week 16 and weeks 22–31), the table instead used a single fixed, unranked ten-column layout that did not vary from week to week, and whose Total row is not sorted. The transition between these two regimes is a genuine change in the source bulletin's own format rather than a data error, and it is fully documented in table4_columns_by_week.csv.

The table carries ten conditions in 162 of the 164 weeks and nine in the remaining two (2023 week 52 and 2024 week 1). Six conditions appear in every one of the 164 bulletins irrespective of the rotation — acute diarrhoea (non-cholera), malaria, influenza-like illness, acute lower respiratory infection in children under five, bloody diarrhoea, and typhoid — and these six are therefore complete across the panel. A seventh, severe acute respiratory infection, is present in 163 of 164 weeks, absent only in 2026 week 25.

The remaining nine conditions rotate through the table's residual capacity, with the following observed frequencies: animal or dog bite (133 weeks), cutaneous leishmaniasis (108), measles (85), tuberculosis (78), viral hepatitis A and E (57), acute watery diarrhoea with suspected cholera (12), dengue (10), viral hepatitis B, C and D (6), and chickenpox or varicella (2).

For these rotating conditions, absence from a given week's table from 2023 week 32 onward indicates only that the condition did not rank among that week's highest by provincial case volume. **Such absence is right-censoring and must not be interpreted as a reported zero.** This distinction is material to any time-series or modelling application and is the single most important caveat for users of this dataset.

Because the pre-transition period (2023 week 16 and weeks 22–31) used a fixed layout, the ten conditions reported in that period are observed in every one of those weeks and are not censored.

2.3 Known data-quality issues

Table 4 of the 2025 week 9 bulletin is identical to that of week 8 in every district-level value and in the Total row. Both source PDFs were independently re-retrieved to confirm this: the week 9 bulletin is titled “Week 09” but reproduces week 8’s data table exactly. This most plausibly reflects a source-side republishing error rather than an extraction fault. The week is retained in the raw panel for transparency and flagged as duplicate_source_table in data_quality_flags.csv, but it should be excluded or imputed in any week-over-week or time-series analysis.

Four rows in the source bulletins (2024 weeks 36, 37, 40 and 41) contain explanatory notes recording that Kohat and/or Battagram were not reported as rows in that week, rather than case data. These rows are excluded from the panel and logged separately in data_quality_flags.csv. They represent genuine reporting gaps that the source itself documents.

One bulletin (2025 week 40) contains no printed provincial Total row.

3 Methods

3.1 Source and retrieval

Data were extracted from the weekly Integrated Disease Surveillance and Response (IDSR) Public Health Bulletins published as PDF documents by the National Institute of Health, Pakistan [4]. Two mirror hostnames serve the same bulletin paths; the www.nih.org.pk mirror was used throughout, as the alternative host returned HTTP 404 responses consistently during the retrieval period.

Bulletin filenames follow no consistent convention across the corpus — observed forms include Weekly_Report-08-2025.pdf, Weekly Report-01-2024.pdf (containing a literal space), and IDSRS Weekly Report-28-2026 NG.pdf. Retrieval therefore harvested actual hyperlinks from the publisher’s bulletin index pages rather than constructing URLs from an assumed template. The scripts used to harvest those links were written with the assistance of Claude Code (Anthropic). Every resolved URL is recorded in bulletin_links_verified.csv.

3.2 Extraction

Table 4, the KP province district-wise surveillance table, was transcribed from each bulletin with column headers, row labels and cell values preserved exactly as printed, including the literal string “NR” where a value was recorded as not reported. Because raw PDF binaries could not be bulk-downloaded in the working environment, transcription was performed using a generative AI document-reading tool, Claude (Anthropic), operating directly on each source PDF. The transcribed values were subsequently verified by the authors against the source bulletins, as described under Validation. This use of an AI tool forms part

of the research process rather than of manuscript preparation, and is reported here accordingly; its separate use in preparing the manuscript text is disclosed in the declaration at the end of this article. Values were transcribed as printed rather than inferred or reconstructed. The implications of this extraction method are addressed under Limitations.

The full transcribed source text is deposited alongside the dataset, and the parser that converts it to the published CSV files is provided, so that every step from transcribed source to final panel is independently reproducible. That parser was written with the assistance of Claude Code (Anthropic).

3.3 Canonicalisation

Column-header spelling and ordering vary between bulletins. A header-normalisation function mapped every observed header variant — differing in case, punctuation, spacing and hyphenation, for example “AD (Non-Cholera)”, “AD (noncholera)” and “AD (NonCholera)” — onto one of 16 canonical condition names. Two single-week header artefacts (“AD (NonCholeraA)” and “AD (NonCholeraI)”, in 2025 weeks 12 and 13) were identified as the same condition occupying the same table position in adjacent weeks and mapped accordingly.

A parallel crosswalk mapped 50 raw district and sub-division labels onto 44 canonical reporting units. This resolves three distinct issues. First, two evident data-entry errors in the source (“Abbottaba” for Abbottabad, appearing in one week, and “Hangue” for Hangu, in two) were harmonised and flagged with unit_type = typo. Second, six sub-division-level reporting units (SD Bannu, SD DI Khan, SD Kohat, SD Lakki, SD Peshawar and SD Tank) appear in some weeks in place of, or alongside, their parent districts. These are retained as distinct units with unit_type = subdivision and an explicit parent_district reference, rather than merged into the parent district, because a sub-division and its parent district are not the same geography and merging them would double-count or misattribute cases.

Third, and least tractable, the source uses four labelling conventions for South Waziristan across different periods — full names, the legacy agency abbreviations “SWA” and “SWU”, and a short form (“SW (L)”, “SW(U)”). Three of these map onto canonical units without ambiguity. “SWA”, however, does not denote a fixed geography, and establishing what it denotes required evidence from the panel itself rather than from the labels.

From 2023 week 34 to 2024 week 49, “SWA” is the only South Waziristan row present in Table 4, and its all-condition weekly total runs between roughly 350 and 450 cases. Measured over the same epidemiological weeks one year later, when both halves report separately, South Waziristan Lower and Upper sum to approximately 548 cases per week, of which Upper alone contributes approximately 158. The magnitude of the “SWA” series is therefore consistent with the undivided district and inconsistent with either half. From 2024 week 50, a distinct “South Waziristan (Lower)” row begins reporting at 286–342 cases per week and the “SWA” series falls abruptly to 49–61, indicating that “SWA” has become the residual, that is, the Upper portion. That reduced series then continues without discontinuity into the “SWU” label from 2025 week 1 (2024 week 52: 61 cases; 2025 week 1: 50 cases). Decisively, in 2024

weeks 49–52 the labels “SWA” and “South Waziristan (Lower)” appear in the same tables carrying different values, which excludes the possibility that they denote the same unit.

“SWA” is accordingly resolved by week rather than by label: to a distinct canonical unit, South Waziristan (undivided), through 2024 week 49, and to South Waziristan Upper from 2024 week 50 onward. The crosswalk records this through applies_from and applies_to fields. One week, 2024 week 49, resists clean assignment: “SWA” still reports at undivided-district magnitude (340 cases) while a small “South Waziristan (Lower)” row (42 cases) also appears, so the two rows overlap geographically. That week is coded as undivided and flagged in data_quality_flags.csv rather than silently resolved. A build-time assertion confirms that no two source rows within a week collapse onto the same canonical district and condition.

3.4 Validation

Four internal consistency checks were applied to the transcribed corpus.

First, each week’s printed Total row was tested for descending order. In 153 of 164 weeks the row is strictly descending, consistent with a rank-ordered top-conditions design; the 11 non-descending weeks are precisely 2023 week 16 and weeks 22–31, which independently share a single fixed column layout. The two findings corroborate each other and together establish the format transition as a genuine property of the source rather than an anomaly of extraction.

Second, every data row’s cell count was checked against the column-header count for its own week. No mismatches were found in any of the 164 bulletins.

Third, the suspected duplicate week was verified by independently re-retrieving both the 2025 week 8 and week 9 source PDFs and comparing their tables directly, confirming that the duplication is present in the published source.

Fourth, and as an independent check on transcription accuracy, the sum of district rows was compared against the printed provincial Total row for every column in every week carrying such a row. This test located a specific and otherwise undetectable class of transcription error, described below. After correction, 137 of the 163 weeks with a printed Total (84.0%) reconcile exactly on every column.

The error arises where a district row in the source contains a genuinely blank cell. The document-reading tool returns only the populated values, left-aligned, so every value after the gap is assigned to the column heading one position to its left, and a spurious value is padded onto the end of the row. The digits themselves are read correctly; only their column assignment is wrong. Such a row is undetectable by inspection because it appears complete and internally plausible.

Candidate rows were identified by testing every non-reconciling row against a single-parameter hypothesis — that one blank cell was dropped at position k — and retaining only those for which a single value of k causes every column in that week to reconcile exactly. Four rows in the corpus satisfy this

test: Mansehra in 2026 week 26, and Malakand in 2023 weeks 16, 22 and 26. A week in which only one column fails to reconcile, and then by a small amount, was deliberately not corrected on this basis, because a lone small discrepancy can be absorbed by many different transformations and cannot be distinguished from arithmetic inconsistency in the bulletin; 2024 week 50, whose single affected column differs by one case, is left uncorrected for this reason.

The correction to 2026 week 26 was confirmed directly against the source PDF. Mansehra's row there carries nine numeric values against ten column headings, the cell beneath "Animal / Dog Bite" is empty, and the value 217 is printed beneath "TB" — the position the Total-row arithmetic predicts, and not beneath "ALRI < 5 years" where the uncorrected transcription had placed it. This single correction moves 217 cases between two conditions. The four corrected rows are recorded individually in data_quality_flags.csv.

Twenty-six weeks still contain at least one column that does not reconcile. These residual discrepancies are predominantly small: the median worst-case discrepancy in an affected week is 3 cases and 19 of the 26 differ by no more than 10 cases in any column. Two weeks contain a column differing by more than 100 cases (2023 week 41 and 2026 week 4); the former is addressed immediately below and the latter, in which cutaneous leishmaniasis is short by 214 cases, remains undiagnosed. The largest residual discrepancy, 2023 week 41, was examined directly against its source PDF. Two candidate explanations were tested and both were excluded: the bulletin prints 34 district rows, exactly the number transcribed, so no row was dropped; and the row for Upper Kurram — the reporting unit whose typical weekly profile most closely matches the shortfall — carries the same near-zero values in the source as in the panel, so it was not misread. The district rows printed in that bulletin therefore do not sum to the provincial Total printed alongside them, and the inconsistency is internal to the published source rather than introduced by transcription. For the remaining weeks the printed Total row is itself part of the source and was not used to construct the panel, so this test cannot by itself distinguish residual transcription error from arithmetic inconsistency in the published bulletin; both possibilities remain open and the affected weeks are listed so that users can weight them accordingly. That at least one such discrepancy has been positively attributed to the source, and that every row inspected against its PDF was otherwise faithful apart from the dropped-cell cases described above, suggests these residuals are not predominantly extraction errors.

4 Limitations

Several limitations apply to this dataset. The first three are inherent to the source; the remainder concern the extraction and should be weighed by any user of the data.

No denominators. The bulletins report no per-district count of reporting health facilities and no catchment population. All values in this dataset are therefore raw case counts, not incidence rates, and reporting completeness cannot be observed from the data itself. Apparent between-district differences

may reflect differences in reporting effort as much as differences in disease burden, and no adjustment for this is possible from the source alone.

Partial condition coverage. Each bulletin's narrative text discusses approximately 30 notifiable conditions, while Table 4 reports only the nine or ten that reach district-level tabulation. The remainder are published only at national or provincial level and lie outside the scope of this dataset.

Right-censoring of rotating conditions. As described under Data Description, absence of a rotating condition from a given week's table from 2023 week 32 onward indicates non-inclusion in that week's ranked columns, not a zero count. Analyses treating such absences as zeros will be biased.

Extraction method. All 164 tables were transcribed using a generative AI document-reading tool, Claude (Anthropic), rather than by programmatic parsing of the PDF byte stream. The structural findings reported here are robust to this choice, because each was derived from consistency checks across the transcribed corpus or from independent re-retrieval of source documents: the descending-Total test, the cell-count test, the direct re-fetch confirming the duplicated week, and the district-to-total reconciliation. Individual cell values, however, have not been systematically verified against the source PDFs by an independent method. The reconciliation check reported under Validation both quantifies this risk and demonstrates it: it identified four rows in which values had been assigned to the wrong condition, one of which was then confirmed against the source PDF and affected 217 cases. That check constrains but does not eliminate the possibility of further errors, since it can only detect discrepancies that disturb a column total. Users requiring exact counts for a specific district-week are advised to confirm those values against the corresponding source bulletin, whose URL is provided in bulletin_links_verified.csv.

Discontinuity in the South Waziristan series. Although the referent of the "SWA" label is resolved as described under Methods, the underlying reporting practice still changes mid-corpus. No separate Lower and Upper series exist before 2024 week 50; a single undivided series exists instead, and it is published here as its own canonical unit rather than being apportioned between the two successor districts, which the source does not support. Users requiring a continuous South Waziristan series across the full study period must either analyse the undivided unit and the sum of its two successors as a single aggregate, or restrict analysis to 2024 week 50 onward. Week 49 of 2024 should be excluded from either approach. All other district mappings, including the treatment of "NWA" as North Waziristan, are constant across the corpus and unambiguous in the source.

Residual total-row discrepancies. After the dropped-cell corrections described under Validation, 26 of 163 weeks still contain at least one column in which the district rows do not sum to the printed provincial total. Most are small — 19 differ by no more than 10 cases — but 2023 week 41 and 2026 week 4 each contain a column differing by more than 100 cases. The origin of these residual discrepancies, whether transcription or the source itself, is not resolved, and the affected weeks are identified in data_quality_flags.csv so that users may exclude them where exact agreement with the published provincial totals matters.

Declarations

Ethics approval and consent to participate

Not applicable. This work did not involve human participants, animal subjects, or the collection of any primary data. The dataset is derived entirely from previously published, publicly available aggregate surveillance bulletins issued by a national public health authority. The source bulletins report case counts aggregated to district level and contain no individual-level or personally identifiable information. No ethical approval was therefore required.

Consent for publication

Not applicable. The manuscript contains no data from any individual person.

Availability of data and materials

The dataset described in this article is publicly available in the Mendeley Data repository under DOI 10.17632/9yzvfgrhkt.1 (https://data.mendeley.com/datasets/9yzvfgrhkt/1), released under a CC BY 4.0 licence. The deposit comprises eleven files, including the raw transcribed source text and the parser that generates the published files from it. The source bulletins from which the dataset is derived are publicly available from the National Institute of Health, Pakistan, at https://www.nih.org.pk/, and the verified retrieval URL for each of the 164 bulletins used is included in the deposited dataset.

Competing interests

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Clinical Trial Number

not applicable

Author contributions

Sara Nawab: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. Gulalai Khan: Validation, Visualization, Writing – original draft, Writing – review & editing. Aamir Abbas: Software, Formal analysis, Writing – review & editing. Junaid Alam: Writing – original draft, Writing – review & editing. Sania Ayub: Validation, Writing – original draft, Writing – review & editing. Mohammad Ali: Supervision, Project administration, Validation, Writing – review & editing. Khalid Nawab: Conceptualization, Supervision, Project administration, Writing – review & editing.

All authors read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Use of generative AI and AI-assisted technologies

The manuscript was written by the authors. During its preparation the authors used Claude (Anthropic) in order to improve the language and readability of the text they had written. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article. Uses of AI tools forming part of the research process rather than of manuscript preparation are described in Methods: Claude (Anthropic) was used to transcribe Table 4 from the source PDF bulletins, and Claude Code (Anthropic) was used to assist in writing the link-harvesting scripts and the parser. No AI tool is listed as an author, and the authors accept full responsibility for the accuracy and integrity of the work.

Acknowledgements

The authors thank the National Institute of Health, Pakistan, for publishing the Integrated Disease Surveillance and Response weekly bulletins openly, which made this dataset possible.

References

1. Kaur G, Ghoshal S, Marbate R, Malviya N, Kaur A, SB V, Srivastava AK, Singh M. EpiClim: Weekly district-wise all-India multi-epidemics climate-health dataset for accelerated GeoHealth research, arXiv:2501.18602 [physics.soc-ph] (2025). https://doi.org/10.48550/arXiv.2501.18602.
2. Hossain N, Madaniyazi L, Ng CFS, Nasrin D, Seposo XT, Chua PLC, et al. Short-term associations of diarrhoeal diseases in children with temperature and precipitation in seven low- and middle-income countries from Sub-Saharan Africa and South Asia in the Global Enteric Multicenter Study. PLoS Negl Trop Dis. 2024;18(10):e0011834. https://doi.org/10.1371/journal.pntd.0011834.
3. Sharif F, Shahzad L, Batool M. The association between climatic factors and waterborne infectious outbreaks with a focus on vulnerability in Pakistan. Int J Environ Health Res. 2024;34(9):3299–316. https://doi.org/10.1080/09603123.2024.2302040.

4. National Institute of Health, Pakistan. Integrated Disease Surveillance and Response (IDSR) weekly public health bulletins. https://www.nih.org.pk/ (accessed 6 August 2026).
5. Haq ZU, Fazid S, Hussain B, Khan MF, Betanni A, Behrawar B, Afridi S. Assessment of the impact of integrated disease surveillance and response system on surveillance management at healthcare facilities in Pakistan, East Mediterr. Health J. 2024;30(2):109–15. https://doi.org/10.26719/emhj.24.026.
6. National Institute of Health, Ministry of National Health Services. Regulations and Coordination, Pakistan, IDSR Roadmap for Pakistan 2024–2028: A 'System Strengthening' and 'Collaborative Surveillance' Approach, Islamabad, 2024.
7. Ullah W, Yen TY, Niaz S, Nasreen N, Tsai YF, Rodriguez-Vivas RI, Khan A, Tsai KH. Distribution and risk of cutaneous leishmaniasis in Khyber Pakhtunkhwa, Pakistan, Trop. Med Infect Dis. 2023;8(2):128. https://doi.org/10.3390/tropicalmed8020128.
8. Lu C, Khan K, Khan F, Shah SU, Jamal M, Badshah N. Epidemiology of cutaneous leishmaniasis in children of Khyber Pakhtunkhwa, Pakistan, Trop. Med. Int Health. 2024;29(7):633–46. https://doi.org/10.1111/tmi.14005.
9. Rasheed MK, Hasan SS, Babar ZU, Ahmed SI. Extensively drug-resistant typhoid fever in Pakistan. Lancet Infect Dis. 2019;19(3):242–3. https://doi.org/10.1016/S1473-3099(19)30051-9.
10. Badr HS, Colston JM, Nguyen NH, Chen YT, Burnett E, Ali SA, et al. Spatiotemporal variation in risk of Shigella infection in childhood: a global risk mapping and prediction model, Lancet Glob. Health. 2023;11(3):e373–84. https://doi.org/10.1016/S2214-109X(22)00549-6.

Figures

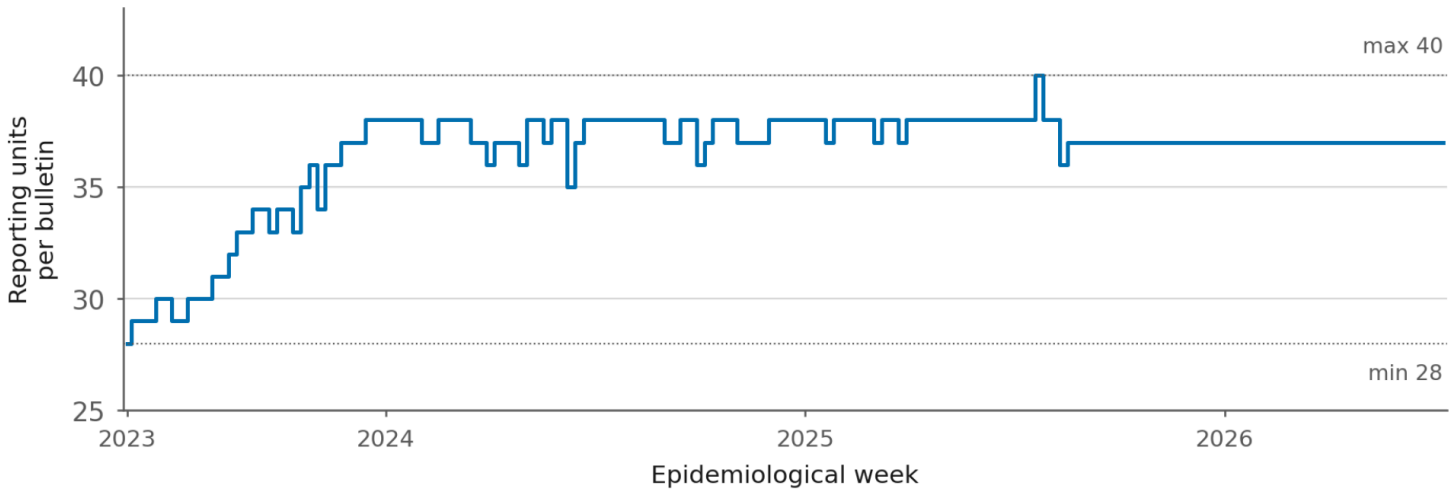


Figure 1

Number of distinct canonical reporting units appearing in Table 4 in each bulletin. The count is not fixed: it ranges from 28 to 40 across the corpus, rising through 2023 as districts enter the table and varying

thereafter as sub-division-level rows appear in place of, or alongside, their parent districts. This instability is what the district crosswalk resolves.

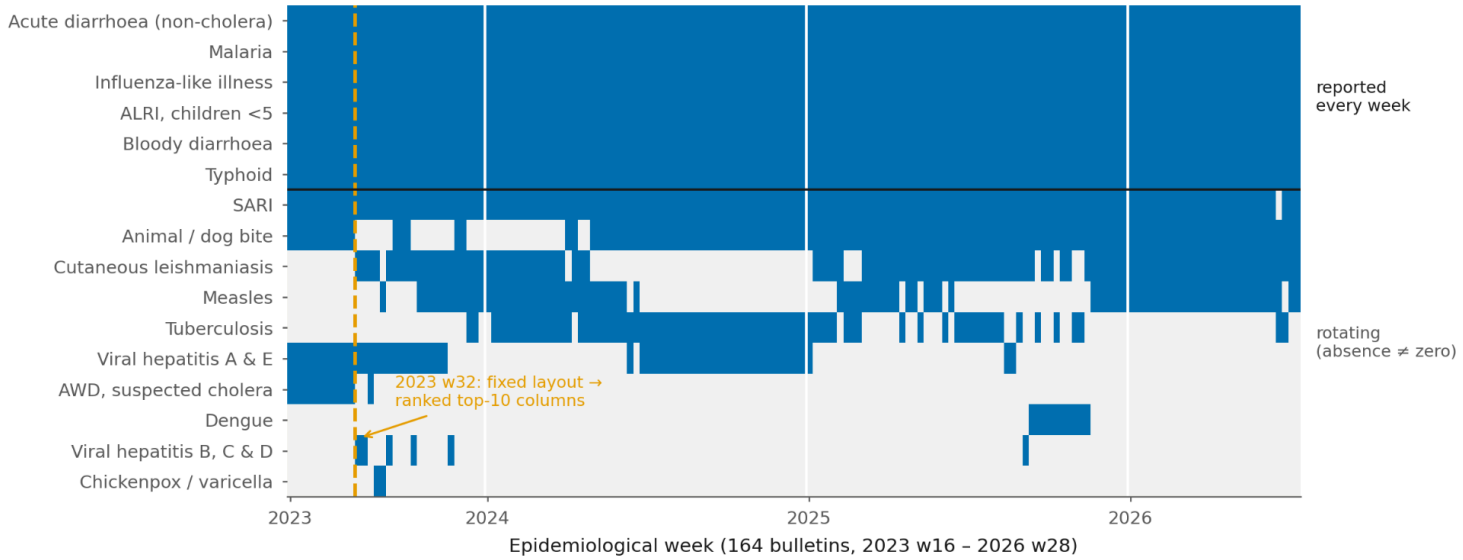


Figure 2

Presence of each notifiable condition in Table 4, by bulletin week (16 conditions × 164 weeks). Filled cells indicate that the condition appeared as a column in that week's table; empty cells indicate that it did not. The six conditions above the rule are present in every bulletin and are complete across the panel. For the ten below it, an empty cell means the condition did not rank among that week's highest by provincial case volume, and is right-censoring rather than a reported zero. The dashed line marks 2023 week 32, where the source changes from a fixed ten-column layout to ranked top-ten columns.

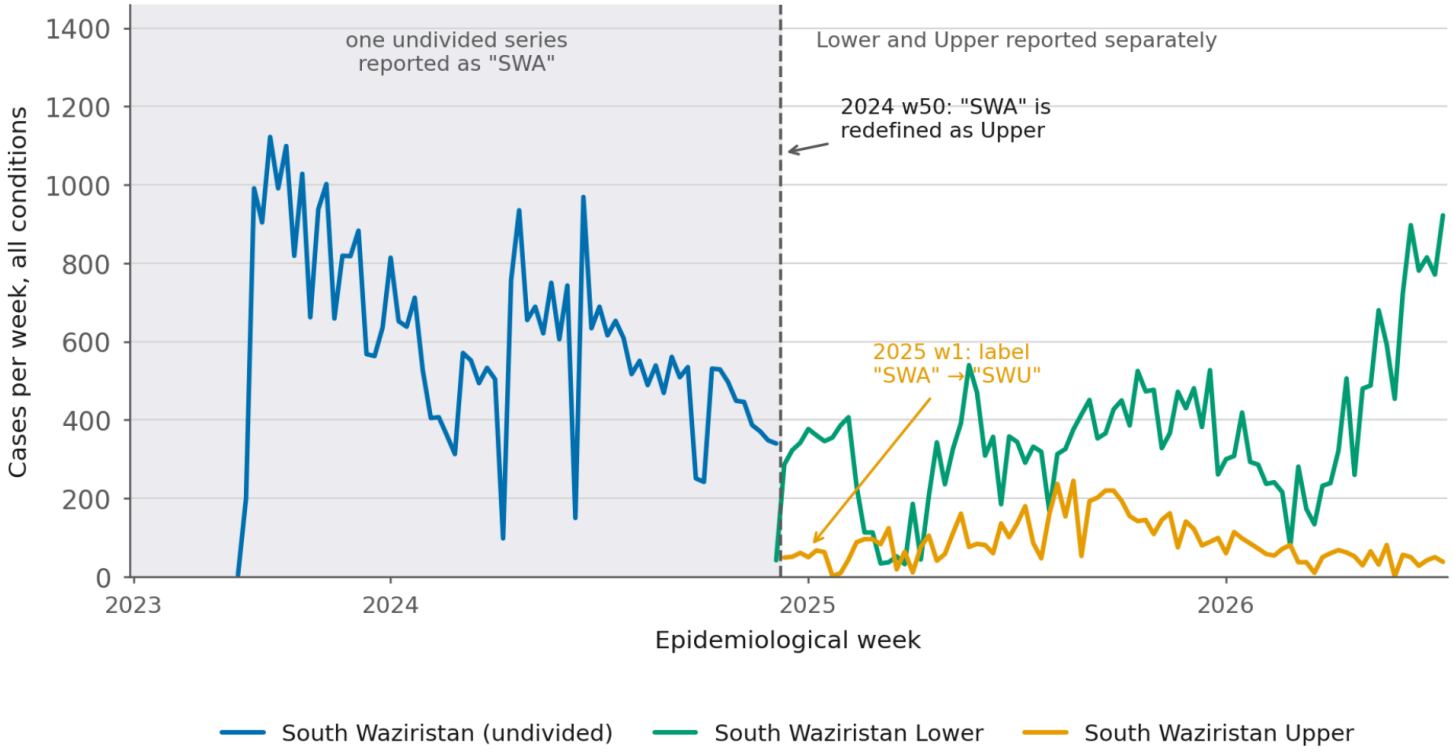


Figure 3

All-condition weekly case totals for the South Waziristan reporting units, showing why the “SWA” label cannot be mapped to a fixed geography. Through 2024 week 49 (shaded) a single series is reported under “SWA” at 350–1130 cases per week. From 2024 week 50 a distinct Lower series begins and the “SWA” series drops abruptly to 49–61 cases per week, identifying it thereafter as the Upper portion; that reduced series continues without discontinuity when the label changes to “SWU” at 2025 week 1. Mapping “SWA” to a single district for the whole period, as a static crosswalk would, conflates the undivided district with one of its successors.