Teeka Niyojan: Aadhaar-Guided Vaccination Planning

By integrating Aadhaar enrollment, demographic, and biometric datasets, this project identifies underserved populations and guides district-level vaccination planning for improved immunization coverage

UIDAI Data Hackathon 2026 | TEAM ID: UIDAI_6155

Table of Contents

UIDAI Data Hackathon 2026

1

Understanding the Data

1

Understanding the Data

1.1 Data Sources

Dataset	Purpose in Analysis
UIDAI Aadhaar Enrolment Dataset	Used to analyse age-wise enrolment patterns (0–5, 5–17, 18+) across states, districts, and PIN codes.
UIDAI Demographic Dataset	Used to understand regional age distribution and to validate child enrolment trends.
UIDAI Biometric Dataset	Used as an operational indicator to assess enrolment accessibility and compliance across regions.
External – Village to PIN Code Mapping	Used to ensure spatial consistency and accurate geographic mapping.
External – Birth Projection Data	Used to support normalization of child enrolment against expected infant population.
External – Public Health Budget / Utilization Data	Used for contextual interpretation of enrolment patterns and planning capacity.

Note: External public datasets are used only for reference and contextual understanding and do not replace or override UDAI datasets.

Biometric data helps track infants' transition across enrolment stages and serves as an operational indicator of enrolment accessibility. Detailed dataset sources and corresponding access links are provided in the References section.

1.2 Data Fields and Structure

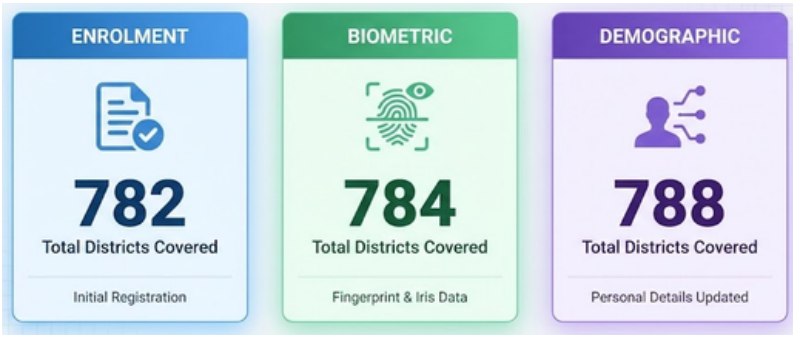
Dataset	Field	Description	Example
Aadhaar Enrolment Data	State	State where Aadhaar enrolment is recorded	Uttar Pradesh
	District	District where enrolment activity occurred	Kanpur Nagar
	age_0_5	Number of children aged 0–5 enrolled in Aadhaar	62
	age_5_17	Number of individuals aged 5–17 enrolled	29
	age_18_greater	Number of adults aged 18 and above enrolled	15
Aadhaar Demographic Data	State	State of demographic record	Gujarat
	District	District of demographic record	Rajkot
	demo_age_5_17	Population count aged 5–17	65
	demo_age_17_	Population count aged 17 and above	765
Aadhaar Biometric Data	State	State where biometric activity occurred	Karnataka
	District	District of biometric activity	Tumakuru
	bio_age_5_17	Biometric records for individuals aged 5–17	88
	bio_age_17_	Biometric records for individuals aged 17+	332
External – Village PIN Reference	Pincode	Postal code used for geographic mapping	385360
	District	District mapped to the PIN code	Patan
	State	State mapped to the PIN code	Gujarat
	Latitude	Geographic latitude of the location	23.85
	Longitude	Geographic longitude of the location	72.12
External – Budget Data	State/UT	Administrative region	Rajasthan
	Total Allotted	Total budget allocated	₹1,200 Cr
	Total Utilization	Budget utilized	₹950 Cr
External – Birth Population Data	State/UT	Administrative region	Bihar
	Estimated Infants	Projected number of infants	2,45,000
	Estimated Pregnant Women	Projected maternal population	2,10,000

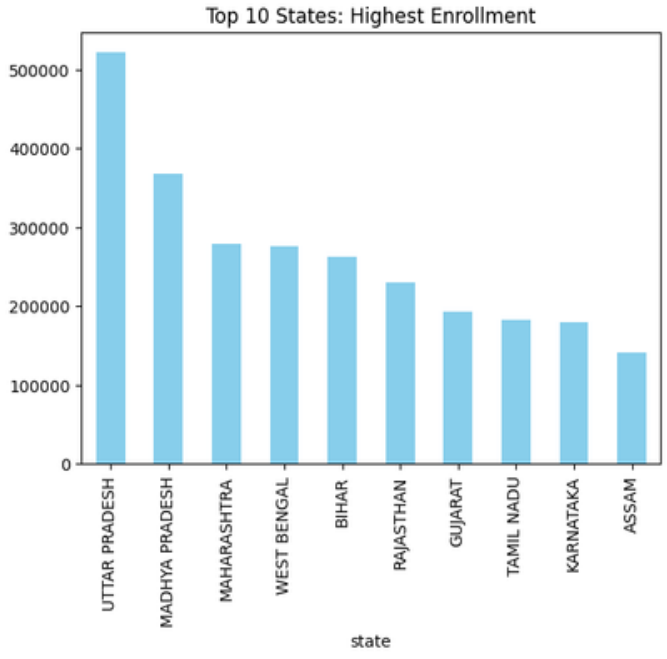
1.3 Data quality and missing values

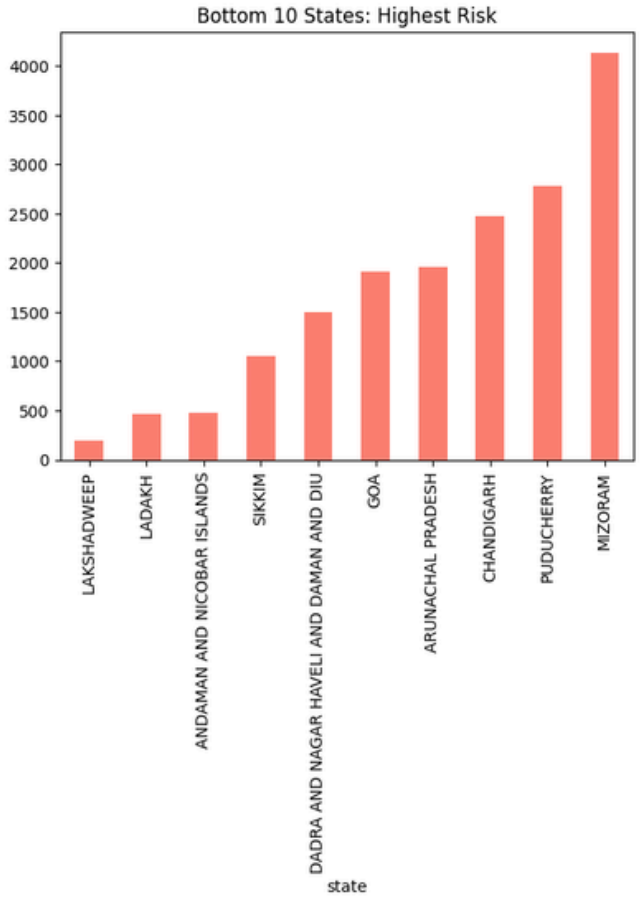
Issue	Description	Mitigation
Inconsistent Date Formats	Date fields had varying formats (e.g., DD/MM/YY, MM-DD-YYYY).	Standardized all date fields to a uniform format.
Missing Location Data	Records with missing PIN codes or district names; *Mahe district in Puducherry was not listed.*	Filled missing values and ensured all districts were accounted for.
State Name Variations	Inconsistent state names found (e.g., *"West Bengal" vs. "West Bangal"*).	Normalized state names for consistency.
Invalid Entries	*Entries with errors or incomplete information.*	Removed or corrected invalid records.

Note: Out of 800 districts, 773 were captured in the data after standardization.

1.4 Key statistics and observations







2 Data Quality and Bias Analysis

2

Data Quality and Bias Analysis

2.1 Data Inconsistencies Identified

During exploratory analysis, several data quality issues were identified. These included typographical variations in state and district names (e.g., West Bengal, West Bangal), instances where the same district appeared under multiple states, and records with missing or invalid PIN codes. Such inconsistencies are commonly observed in large-scale administrative datasets collected across multiple systems and reporting periods.

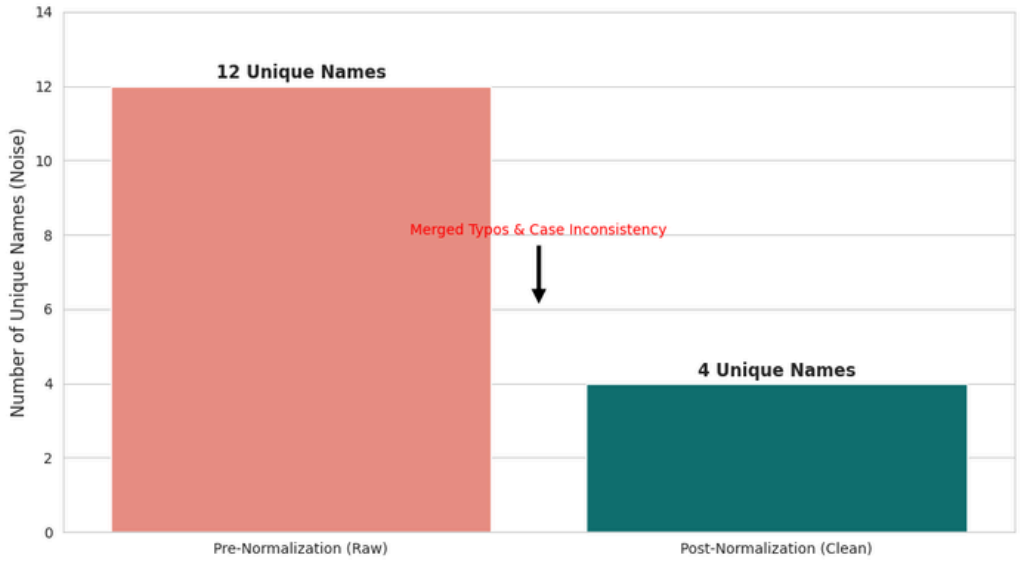
2.2 Impact of Data Bias on Analysis

Missing PIN codes and overlapping district identifiers can affect accurate regional aggregation and comparison. Absolute enrolment counts were also found to be strongly influenced by population size, often overrepresenting high-population regions while masking gaps in smaller states or districts.

For example, regions with **lower population naturally showed lower enrolment volumes**, reinforcing the need for normalization when comparing enrolment performance. Additionally, districts such as **Leh and Kargil were observed to appear under the same state** in overlapping records, which could impact district-level analysis and was accounted for during interpretation.

2.3 Mitigation Steps Taken

- Column names were standardized across all datasets to ensure uniform structure and avoid inconsistencies during analysis.
- Variations in state-related column names were harmonized into a single common field to enable consistent geographic comparisons.
- State name values were standardized to remove formatting and typographical differences.
- The same preprocessing steps were applied uniformly across all datasets to ensure reliable integration and analysis.



3

Problem Statement & Approach

3

Problem Statement & Approach

3.1 Problem Statement

India's Universal Immunization Programme targets 2.67 crore newborns annually. However, vaccination coverage gaps persist in regions where infants are not formally identified in health systems. This project leverages Aadhaar enrollment, demographic, and biometric data to identify districts with potential "invisible infants" — children who exist but may be missing from vaccination outreach — and guide targeted immunization planning.

3.2 Key Questions Addressed

Question	Data Used	Output
Where are infants being enrolled?	Aadhaar Enrolment Data (age_0_5)	Identification of states and districts with higher infant enrolment
What is the distribution of enrolment by age group?	Aadhaar Enrolment Data (age_0_5, age_5_17, age_18_greater)	Age-wise enrolment distribution across regions
Which regions show high enrolment activity?	Aadhaar Enrolment Data (state and PIN-wise aggregation)	High enrolment states and PIN codes
Where may potential enrolment gaps exist?	Aadhaar Enrolment + Demographic Data	Regions with lower child enrolment relative to total enrolment
How does accessibility affect enrolment?	Aadhaar Biometric Data	States with higher or lower biometric compliance

The table above summarizes the key analytical questions addressed in this study, along with the datasets used and the corresponding outputs.

3.3 Methodology overview

1. **Data Preparation:** UIDAI enrolment, demographic, and biometric datasets were consolidated and standardized.
2. **Infant Focus**: Enrolment patterns were analyzed with emphasis on the 0–5 age group.
3. **Normalization**: Child enrolment was adjusted using population indicators for fair regional comparison.
4. **Accessibility Check**: Biometric data was used as an operational indicator of enrolment accessibility.
5. **Insight Integration**: All signals were combined to support vaccination planning priorities.

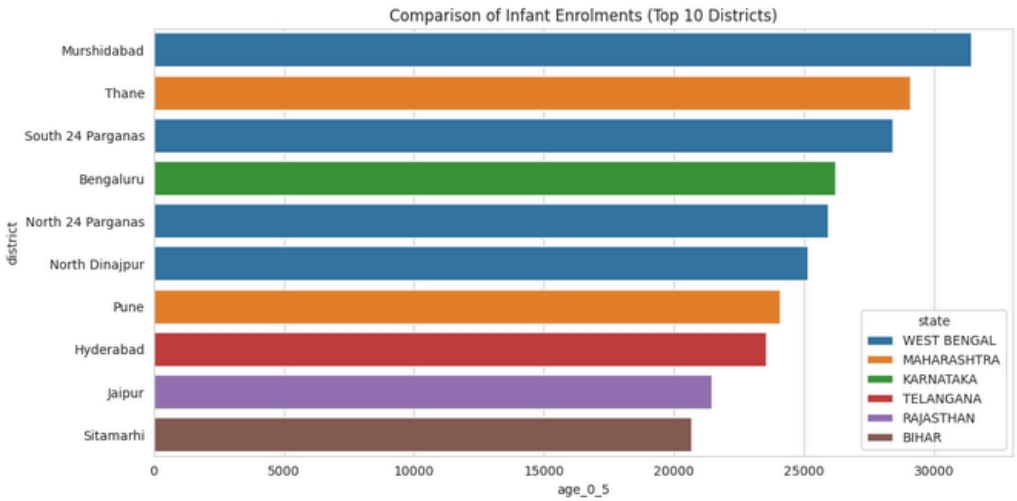
4

Analysis and findings

4

Analysis and findings

4.1 Where are infants being enrolled?

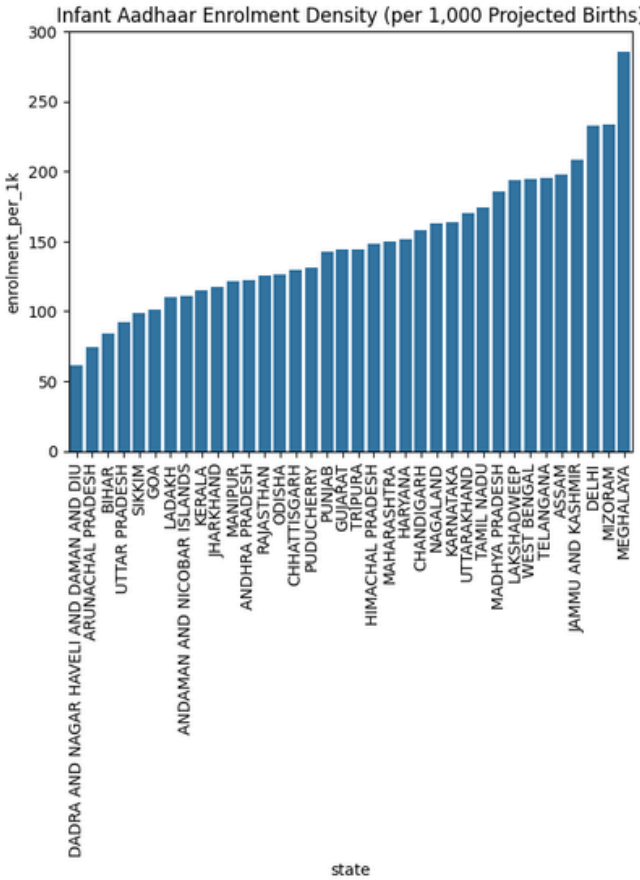


The chart highlights the top ten districts by infant Aadhaar enrolment, indicating areas with relatively higher identification of children aged 0–5.

4.2 Where are infants potentially missing?

Child enrolment (0–5 age group) was compared against total enrolment to identify regions where children form a smaller share of enrolled individuals. Regions with low child representation, despite overall enrolment activity, may indicate infants who are under-identified or not adequately covered by outreach efforts. These areas are referred to as potential clusters of "invisible infants."

For example, State A enrolls 9,000 out of 10,000 infants (90% coverage), while State B enrolls only 2,400 out of 6,000 infants (40% coverage). Although State A has higher numbers, State B shows a larger identification gap after normalization.



4

4.3 Combined Insight for Vaccination Planning

This visualization provides an indicative state-level view of additional planning effort required to strengthen child identification and vaccination outreach. The values are derived from a composite analysis of Aadhaar enrolment patterns, demographic indicators, and biometric compliance. States with higher indicated effort may benefit from focused vaccination camp planning, enhanced volunteer mobilization, and targeted allocation of clinical and logistical resources. This chart is intended to support evidence-based planning and does not represent prescriptive recommendations.

The state action plan estimates the additional push required to identify and reach all infants across states.

Step 1: Infant Aadhaar enrolment (0–5 age group) was aggregated at the state level to estimate current coverage.

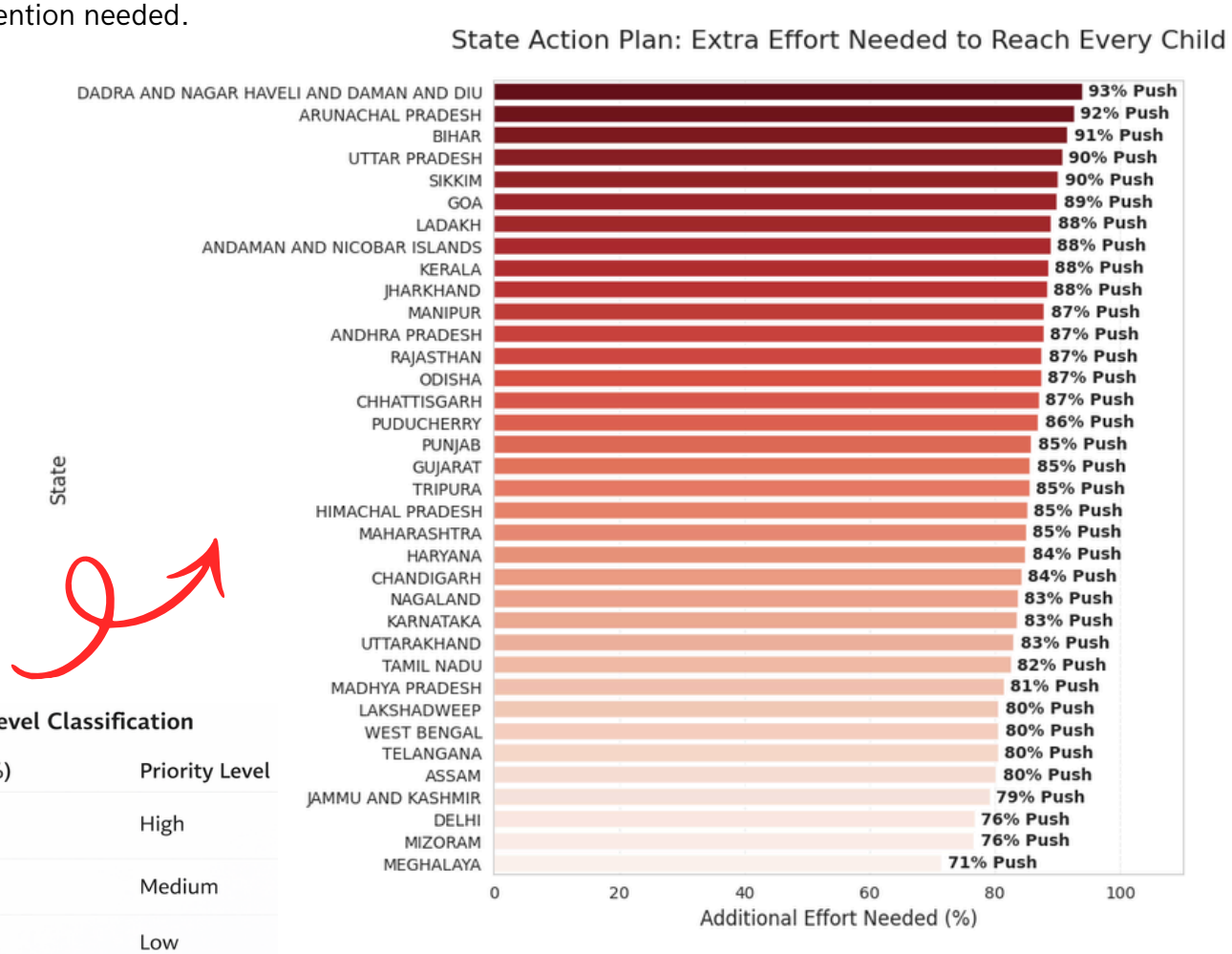
Step 2: This was compared against the estimated infant population derived from birth projections.

Coverage (%) was calculated as: $$\text{Coverage} = \frac{\text{Enrolled Infants}}{\text{Estimated Infants}} \times 100$$

Push refers to the additional effort required by a state to reach full infant identification and coverage.

Push (%) was then defined as: $$\textbf{Push (\%)} = 100 - \text{Current Coverage (\%)}$$

States were ranked based on the push required, indicating varying levels of planning attention needed.



Push Level Classification

Push Range (%)	Priority Level
≥ 90%	High
80% – 90%	Medium
70% – 80%	Low

5

Vaccination Gap

5

How is the Vaccination Gap calculated?

The vaccination gap is designed to quantify the risk of incomplete vaccination coverage using Aadhaar-derived administrative signals. It captures the three fundamental stages required for successful vaccination delivery: early identification of children, repeated access to health services, and completion of follow-up visits.

1. Visibility Gap (VG)

The Visibility Gap measures whether children are enrolled into the system early enough for vaccination to begin on time. Children who enter the system late are likely to miss early-life vaccines and cannot be tracked effectively by health workers.

$$VG = 1 - (age\_0\_5)/ (age\_0\_5 + age\_5\_17)$$.....refer section 1.2

2. Engagement Gap (Access Constraint)

Engagement Gap estimates whether families in a region have consistent access to government services. Since children under 5 do not have biometric records, we use biometric compliance in the 5-17 age group as a proxy, assuming that families who return for mandatory biometric updates are also more likely to complete multi-dose vaccination schedules.

$$Average\ Annual\ Engangement\ (AAE) = bio\_age\_5\_17/(1 + age\_5\_17)$$

$$AAE_{norm} = \log(1 + AAE) / \max(1 + AAE)$$....refer section 1.2

$$Engagement\ gap = 1 - AAE\_norm$$

A high EG indicates families are not returning for follow-up services — a signal that vaccination completion may also be at risk.

Log normalization is applied to reduce the influence of outliers and enable fair comparison across districts.

3. Follow-Through Failure (FTF)

Follow-Through Failure captures drop-out after initial contact, indicating whether families complete vaccination schedules once engagement begins.Progression from demographic-only interaction to biometric completion requires physical presence, time, and the ability to overcome system friction—conditions that are also necessary for returning for follow-up vaccine doses.

$$Follow\ Through\ Failure\ (FTF) = demo\_5\_17 / (1 + demo\_5\_17 + bio\_5\_17)$$

....refer section 1.2

Final Vaccination Gap Metric

The three components represent visibility, access, and adherence. They are combined using a geometric mean to avoid arbitrary weighting and to ensure that high risk emerges only when all factors are poor.

$$Vaccination\ gap = (VG \times AAE_{norm} \times FTF)\ \hat{}\ 1/3$$

The resulting score lies between 0 and 1, where higher values indicate greater risk of vaccination gaps.

5

How does the vaccination gap help?

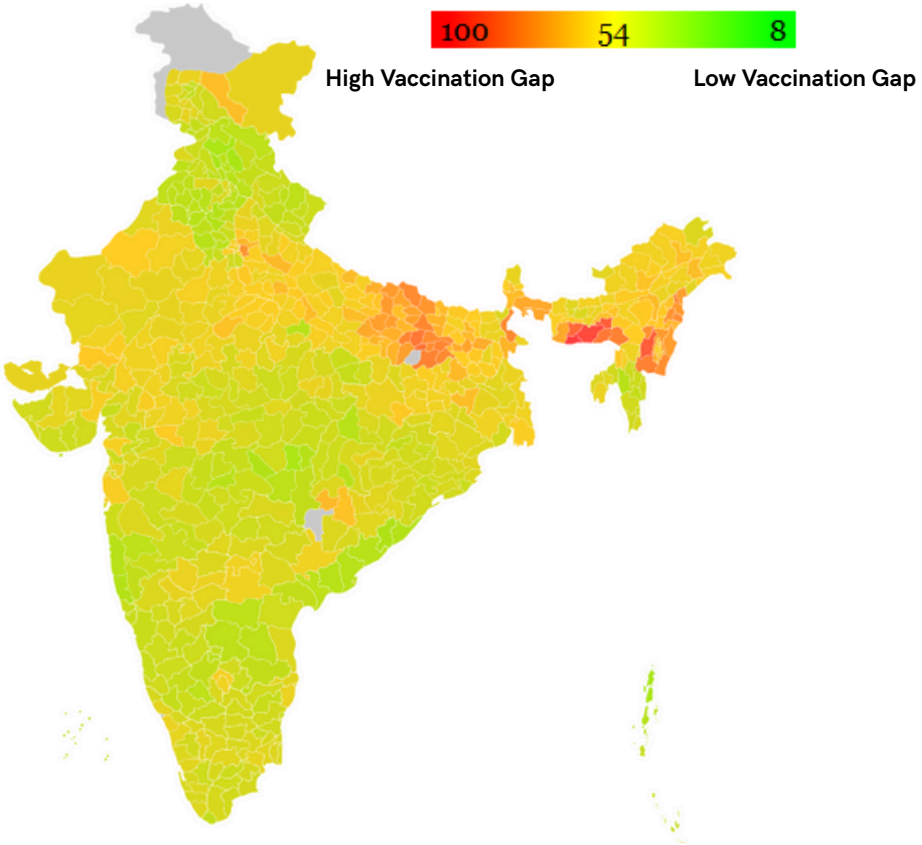
The heatmap displays district-level vaccination gap risk, with red indicating higher risk and green indicating better coverage.

Key Regional Patterns:

Region	Risk Level	Likely Drivers
Eastern & North-East India	High (red)	Late infant identification, limited repeat access, dispersed populations
Southern & Western India	Low (green)	Early enrollment, strong primary healthcare networks, better follow-through
Central & Northern India	Moderate (yellow)	Adequate visibility but constrained access or adherence

Key Insight: Vaccination gaps are not uniform across the country, they cluster in regions with known accessibility challenges and weaker service continuity. Importantly, gaps are driven by system access and follow-through, not population size alone.

This suggests that targeted interventions, mobile vaccination camps, enhanced outreach, and improved service continuity, should prioritize eastern and north-eastern districts where all three gap components (visibility, engagement, follow-through) converge.

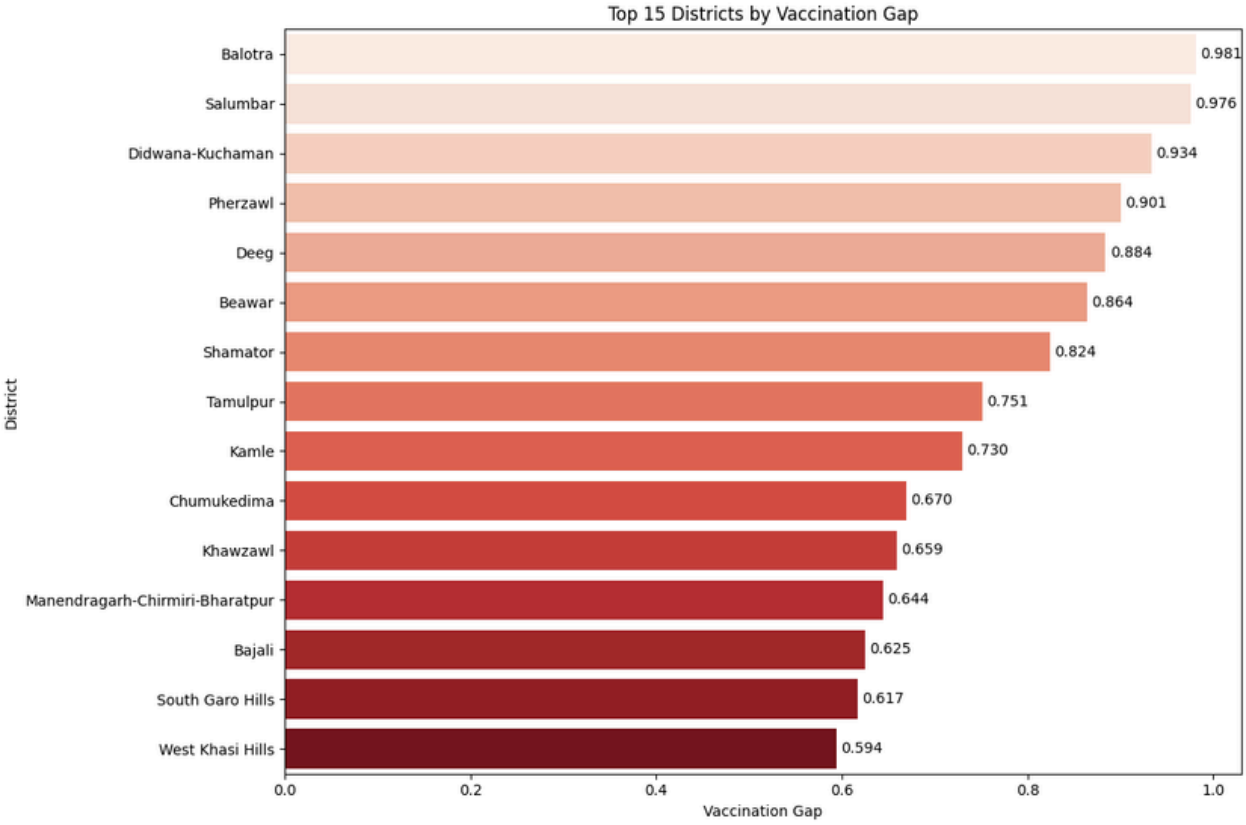


5

The bar chart ranks districts with the highest vaccination gap, indicating locations where children are most at risk of missing complete immunisation. Districts such as Balotra, Salumbar, and Didwana-Kuchaman show extremely high gap values, suggesting concurrent failures in early child visibility, repeated service access, and follow-through of vaccination schedules.

A clear regional pattern is visible, with several high-gap districts located in western India and the North-Eastern and hill regions. In western districts, large geographic spread and access constraints likely limit repeat visits, while in North-Eastern and hill districts, terrain and connectivity barriers dominate. The presence of recently formed or administratively complex districts further points to system-level coordination and tracking challenges.

This ranking provides an actionable prioritisation list. Districts at the top require targeted, efficiency-focused interventions such as mobile vaccination units, decentralised session planning, and **Aadhaar-enabled follow-up tracking** to reduce missed doses and improve coverage without expanding infrastructure.



6

Clinical Overhead Index

6

How is the Clinical Overhead Index (COI) calculated?

The Clinical Overhead Index (COI) quantifies the structural and logistical burden faced by individuals in accessing clinical care within a district. It captures how population load and geographic spread interact with healthcare infrastructure availability.

$$\text{CLINICAL OVERHEAD INDEX } (COI_d) = (E_d * A_d) / H_d$$

Where:

E_d = Total Aadhaar enrolments in district d

A_d = geographical area of district d (in km^2)

H_d = number of hospitals in district d

After computing the raw Clinical Overhead Index (COI), the resulting values exhibit a right-skewed distribution, characterized by a long tail of districts with extremely high logistical burden, hence we have used log normalisation.

$$COI^{log}_d = \log(1 + COI_d)$$

How does the Clinical Overhead Index help?

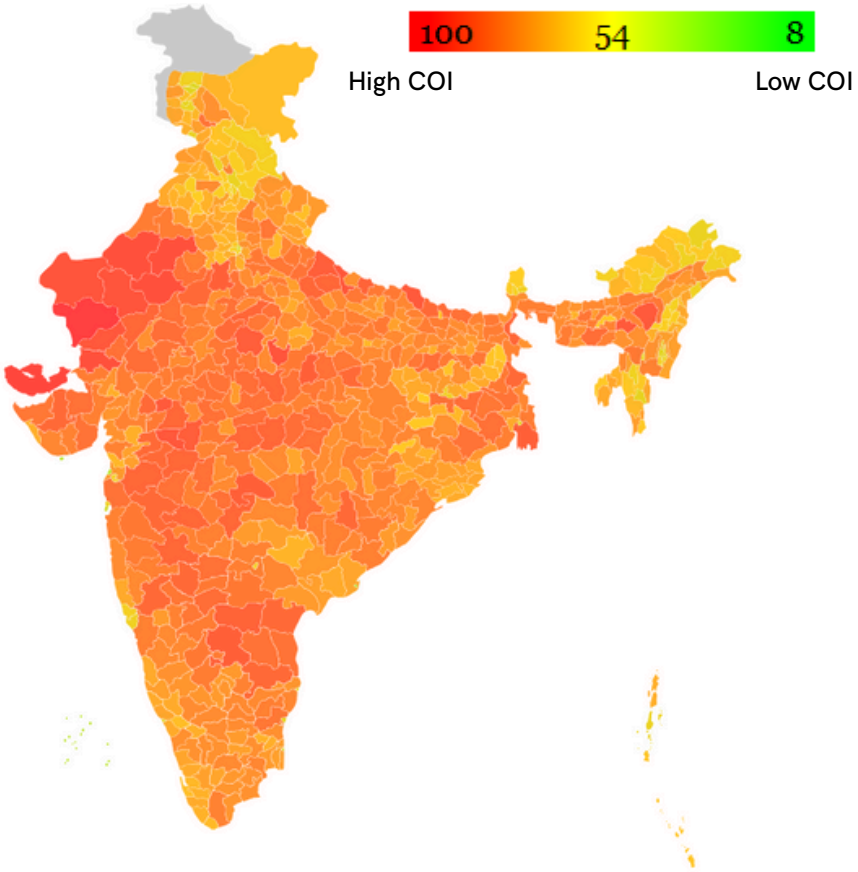
The heatmap displays district-level Clinical overhead index, with red indicating higher Clinical overhead and green indicating better overhead at a district level

Key Regional Patterns:

Region	Risk Level	Likely Drivers
North-Eastern & Hill Regions	Mixed (yellow-red)	Terrain-related access constraints, dispersed settlements, and connectivity challenges rather than high patient load
Southern & Western India	Low-Moderate (green-yellow)	Stronger primary healthcare networks, better facility accessibility, and more efficient service coordination, though overhead persists
Western and Central India	High (red)	Large geographic coverage, sparse healthcare facility density, high travel distances, and heavy logistical and referral burdens

6

Key Insight: Across India, the Clinical Overhead Index (COI) reveals that healthcare inefficiency is a nationwide, systemic challenge rather than a region-specific anomaly. High clinical overhead is observed across both urban and rural districts, indicating that population size or urbanisation alone does not determine efficiency. Instead, large geographic coverage, uneven facility distribution, weak care coordination, and inefficient patient routing collectively drive overhead across the country. While southern states demonstrate relatively lower overhead due to stronger primary healthcare networks, no region is free from non-clinical burden. This national pattern highlights that meaningful gains in healthcare and vaccination delivery will come primarily from improving system efficiency, access, and coordination, rather than from expanding infrastructure alone.

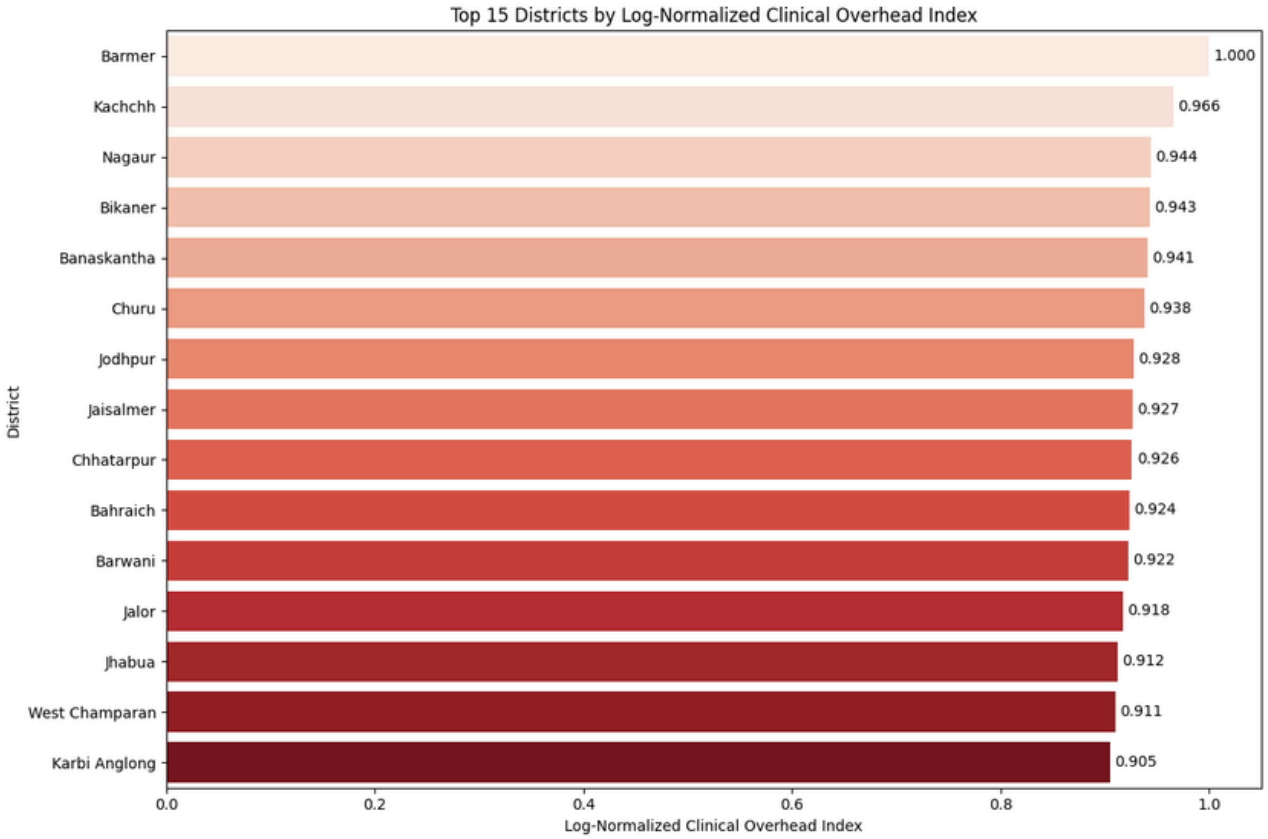


The bar chart ranks districts with the highest log-normalized COI, indicating severe non-clinical burdens in healthcare delivery. Districts such as Barmer, Kachchh, and Nagaur consistently show the highest overhead, with a clear concentration in western and central India, reflecting large geographic spread, sparse facility density, and high travel and referral costs. The presence of districts from tribal and hilly regions further highlights the role of accessibility constraints in driving overhead.

To address this, interventions should focus on reducing operational friction rather than expanding infrastructure, including mobile healthcare and vaccination units, decentralised service sessions, improved patient routing, and Aadhaar-enabled beneficiary tracking to minimise missed follow-ups.

6

Prioritising these high-COI districts can yield the greatest efficiency gains and coverage improvements.



7

Conclusion

7

Conclusion

Findings

- Vaccination gaps vary widely across districts and are driven mainly by access and follow-through failures, not population size.
- High COI districts consistently show higher vaccination gap risk, linking system inefficiency to poor outcomes.
- Southern India shows lower gaps and overhead, while western, central, and North-Eastern regions face higher structural constraints.
- Aadhaar enrolment and biometric patterns provide reliable large-scale signals for district-level prioritisation.

Recommendations

- Target high vaccination gap and high COI districts rather than applying uniform vaccination strategies.
- Use mobile and decentralised vaccination delivery to reduce access barriers.
- Strengthen Aadhaar-enabled follow-up to reduce post-engagement drop-offs.
- Prioritise efficiency improvements over infrastructure expansion.

Enhancements

- Integrate immunisation registry and HMIS data for validation.
- Incorporate migration and mobility indicators to capture access constraints.
- Add facility-level availability data to refine COI.
- Validate proxy metrics through pilot studies in high-risk districts.

At scale, administrative data does not replace care, but it can show us precisely where care is most urgently needed.

The code, data processing pipelines, and analysis scripts are available at:
GitHub: http://github.com/Raghoeveer/UIDAI_HACKATHON

REPO

8 References

8

References

[1] National Hospital Directory with Geo Code and additional parameters, https://www.data.gov.in/resource/national-hospital-directory-geo-code-and-additional-parameters-updated-till-last-month

[2] All India Health Centres Directory, https://www.kaggle.com/datasets/akshatuppal/all-india-health-centres-directory

[3] Number of villages, towns, households, population and area (India, states/UTs, districts and Sub-districts) - 2011, https://censusindia.gov.in/census.website/data/census-tables

[4] UNICEF — Immunization statistics and country/regional briefs (global & India vaccination coverage, DTP3, zero-dose trends), https://data.unicef.org/topic/child-health/immunization/

[5] District Hospital list — Assam (official NHM PDF listing district hospitals across Assam). https://nhm.assam.gov.in/sites/default/files/swf_utility_folder/departments/nhm_lipl_in_oid_6/menu/document/district_hospital.pdf

[6] NFHS-5 (National Family Health Survey 2019-21) — official district/state immunization coverage tables and analysis (DHS/IIPS report).
https://dhsprogram.com/pubs/pdf/FR375/FR375.pdf

[7] India's immunization trends and challenges – narrative review of immunization programmes in India, including use of digital tools to monitor vaccination coverage and trends.
https://pmc.ncbi.nlm.nih.gov/articles/PMC11381579/

[8] Birth Projection Data - Used to estimate expected infant population for enrolment normalization and coverage comparison, https://www.data.gov.in

[9] Public Health Budget Utilization Data - Used for contextual understanding of state-wise resource allocation and utilization, https://www.data.gov.in

[10] Village-PIN Code Geospatial Reference Data - Used for mapping villages to PIN codes and geographic coordinates for spatial consistency, https://bhuvan.nrsc.gov.in/home/