Methodology Report: Data Collection for the “Most Common Newborn Names” Study (2024 Edition)

1. Overview

This report outlines the procedures, data sources, validation steps, and quality controls used to collect and prepare the dataset for the 2024 analysis of the most frequently used names for newborns. The goal of the study was to generate a nationwide ranking of the top names, identify demographic and regional trends, and ensure consistency across reporting authorities.

2. Data Sources

2.1 Civil Registration Authorities

The primary data was obtained from:

- **State and regional civil registration offices**
- **National Population Registry (NPR)**
- **Municipal birth certificate departments**

Each authority provided anonymized datasets containing:

- First name(s) registered at birth
- Baby’s sex
- Birth date (month-level granularity for privacy)
- Registration region code

These datasets represented **92%** of official births during the study period.

3. Supplemental Sources

3.1 Hospital Birth Logs

To fill gaps caused by delayed civil registrations, we requested aggregated counts from:

- Public hospitals (41 institutions)
- Private clinics (18 institutions)

Hospitals provided:

- Counts of baby names assigned at discharge
- Non-identifiable statistics only
- No personally identifiable information (PII)

This supplemental data accounted for **7%** of total birth records.

3.2 Community Birth Centers

Just **1%** of births occurred in non-hospital settings. Midwife associations provided:

- Name usage frequencies
- Postpartum registration disclosures

4. Data Collection Period

The study analyzed births occurring **January 1, 2024 – December 31, 2024**.
Data submissions were collected between **January 10–February 15, 2025** to allow for late registrations.

5. Data Cleaning and Normalization

5.1 Standardizing Name Variants

To ensure fair ranking, we normalized names using the following rules:

- Removal of hyphens (e.g., “Ana-Maria” → “Ana Maria”)
- Lowercase transformation for comparison
- Ignoring diacritics during grouping but preserving them in presentation
- Consolidating recognized spelling variants (e.g., “Mohammed,” “Muhammad,” “Mohamad” grouped under a canonical form)

5.2 Duplicate Detection

Duplicate entries were flagged through:

- Shared registration numbers
- Identical name + sex + region + birth month coincidences
- Hospital vs. civil registry reconciliation

Approx. **0.8%** of records were removed as duplicates.

6. Quality Assurance Procedures

6.1 Cross-Source Verification

For each region:

- Total birth counts from civil registries were compared with hospital discharge statistics.
- Variances >3% were escalated for manual review.
- Missing or inconsistent data triggered secondary verification requests.

6.2 Outlier Detection

Using automated checks:

- Names appearing fewer than 3 times per region were verified to avoid transcription artifacts.
- Unrealistic entries (e.g., numeric strings, emojis, or accidental punctuation) were removed.

6.3 Manual Audit

A random **5% sample** of regional datasets was manually inspected for:

- Correct name normalization
- Proper sex tagging
- Accurate frequency aggregation

7. Data Aggregation and Final Dataset Construction

After validation, we aggregated:

- Total count of each unique normalized name
- Per-sex statistics
- Per-region rankings

The final dataset included:

- **11,482 unique name variants**
- **486,029 birth records**
- **99.2% coverage of all registered births**

8. Limitations

- Late registrations submitted after February 15, 2025 are excluded.
- Name normalization may group variants that parents considered distinct.
- Rare names occurring only once may still reflect reporting inconsistencies.

9. Ethical Considerations

- All data was anonymized at the source.
- The study excluded any personal identifiers.
- Regional data was aggregated to avoid identifying individuals with rare names.

10. Summary

This methodology ensured that the final list of the most common newborn names for 2024 is accurate, representative, and derived from reliable and ethically sourced records. A combination of official registries, hospital reports, and community birth center statistics enabled near-total birth coverage and robust validation.

11. Results

The following table shows the top names given to new borns in 2025:

Name	Number	Rate per 1000
Sofia	432	16.50
Julia	396	15.12
Martí	352	12.75
Martina	349	13.33
Biel	346	12.53
Marc	344	12.46

The following infographic shows the number of names starting by letter A: