Aligning Historical Data: Harmonizing Under-5 Mortality Records in Türkiye across the 20th Century

I. Introduction: The Critical Need for Spatially and Temporally Consistent U5M Data

A. Background and Significance of Under-Five Mortality (U5M)

Under-five mortality (U5M) is a globally recognized indicator of public health, socio-economic development, and human capital. Longitudinal analysis of U5M trends is essential for evaluating national policy effectiveness and aligning with global health goals such as the Sustainable Development Goals (SDGs). Since its founding in 1923, the Republic of Türkiye has prioritized demographic data collection, recognizing its importance for resource allocation, policy design, and national development.

B. The Challenge of Historical Data Heterogeneity in Türkiye (1923–2008)

Türkiye's historical demographic data stem from three main sources: population censuses, sample surveys (e.g., TDHS), and decentralized registration systems. These sources vary in methodology, coverage, and frequency, creating inconsistencies that hinder longitudinal analysis—especially for sensitive indicators like U5M. Manual data collection, regional disparities, and evolving registration systems have led to fragmented records. Without harmonization, these inconsistencies risk misinterpretation and limit the utility of historical data for policy and research.

C. Study Objective and Contribution

This study aims to harmonize Türkiye's historical U5M data (1923–2008) through digitalization, reclassification, and statistical modeling of both death counts and zero-age population figures (P_0). The focus is on establishing spatial and temporal continuity at the provincial level. The contribution is twofold: (1) an empirically robust dataset for calculating age-specific mortality rates (M_x), and (2) a replicable methodological framework for integrating fragmented historical data into modern demographic analysis.

II. Methodological Framework: Data Acquisition and Standardization

A. Comprehensive Data Sourcing and Digitalization

Mortality data were sourced from TURKSTAT's electronic and physical archives, covering deaths by age, sex, and region. Population data were drawn from TURKSTAT census reports. Historical documents were digitized using a hybrid approach: Optical Character Recognition (OCR) for efficiency, supplemented by manual transcription to correct errors from faded prints and inconsistent formatting. Microsoft Excel served as the coordination platform, enabling structured data entry and verification.

Quality control measures included manual screening, duplicate checks, and interoperator validation. This ensured that digitized data retained the integrity of original records while meeting the analytical standards required for longitudinal modeling.

B. Standardization of Mortality Data for Model Integration

Historical mortality records used non-uniform age groupings—daily, monthly, and annual formats—resulting in up to 44 sub-age categories. To enable consistent analysis, these were reclassified into a standardized 22-category structure: five early-life intervals (0, 7, 14, 21, 28 days), eleven monthly intervals (2–12 months), and six broader categories (15, 18, 21 months; 2–5 years).

This reaggregation was achieved through coding schemes and validated via double-entry checks. The standardized format allows for precise differentiation between neonatal, post-neonatal, and child mortality, aligning with international demographic modeling standards.

III. Zero-Age Population Data Acquisition and Estimation

A. National Population Denominators

Zero-age population figures (P_0) are essential for calculating mortality rates (M_x). National data were obtained from TURKSTAT census records for benchmark years. To fill gaps in non-census years (1950–2008), estimates from the UN World Population Prospects 2024 were integrated. This dual-source approach ensures a continuous, gender-disaggregated national P_0 series.

Demographic shifts are evident in the harmonized data: P_0 peaked around 1980 (1.43 million) and declined to 1.27 million by 2008, reflecting Türkiye's fertility transition.

B. Provincial-Level Estimation

Provincial census data were sparse, with some provinces having only eight data points across 72 years. To construct continuous provincial P_0 series, curve estimation was performed using SPSS. Ten regression models were tested per province and sex. Model selection prioritized:

1) Demographic plausibility (no negative estimates),
2) Compatibility with census benchmarks,
3) High R^2 values (as supportive, not decisive).

Segmented modeling was applied in metropolitan provinces to account for demographic shifts due to urbanization and migration. This ensured stable, non-negative estimates across all provinces and years.

IV. Conclusion: Empirical Contributions and Policy Relevance

This study successfully harmonizes Türkiye's fragmented under-five mortality records into a standardized, longitudinal dataset spanning 1923–2008. By reclassifying mortality data and statistically estimating missing population denominators, it enables reliable calculation of age-specific mortality rates (M_x) at both national and provincial levels.

The dataset provides a foundation for analyzing long-term regional disparities in child survival and supports evidence-based public health planning. It also offers a replicable methodological framework for researchers working with historical demographic data in other contexts. Future research can build on this work to explore causal links between mortality trends and regional investments in health, education, and infrastructure.