Leveraging routine service statistics for impact evaluation:

lessons from a multi-country interrupted time series (ITS) study on contraceptive uptake

Authors: David Jodrell, Impact Evaluation Lead and Borja Marti, Quantitative Evaluator.

BACKGROUND

The Challenge Initiative (TCI) supports 214 local governments (LGs) across 13 countries to scale up family planning (FP) interventions, aiming for greater self-reliance in implementing high-impact practices and increasing modern contraception use among the urban poor.

Given its scale, since 2015, TCI has monitored progress using routine Health Management Information Systems (HMIS) FP service statistics - a low-cost alternative to primary data collection. In 2021, TCI engaged Itad to evaluate TCI, including its impact on contraceptive uptake.

Within budget constraints, automated impact evaluation using HMIS data is an attractive option for cost-effective evaluation supporting its integration into monitoring. However, these methods must be benchmarked against more rigorous analysis to understand trade-offs between cost and credibility. We compare:

- automated pipeline – minimal cost, highly scalable
- curated pipeline – labor-intensive, context-specific

RESEARCH QUESTION

Under what conditions can routine HMIS FP data produce credible impact estimates of contraceptive uptake?

METHODS OF ANALYSIS

Contraceptive uptake was proxied using TCI's monitoring algorithm: Net Accumulated Clients (NAC) quantifies service statistics into contraceptive uptake, accounting for method duration (permanent, long-acting or short-acting), measured as net change in clients per 1,000 women of reproductive age (WRA).

Impact evaluation used ITS analysis with autoregressive integrated moving average (ARIMA) models: ITS leverages long pre-TCI time series without requiring controls; ARIMA adjusts for autocorrelation and non-stationarity common in HMIS data. Results are adjusted for service delivery point (SDP) reporting growth and other confounders.

Figure 1 shows our treatment effect. Net contraceptive uptake (NCU) represents the average monthly NAC difference per 1,000 WRA, when comparing observed trends during TCI with the counterfactual trend (dashed line) predicted from pre-TCI data.

Figure 1: Illustrative ITS analysis

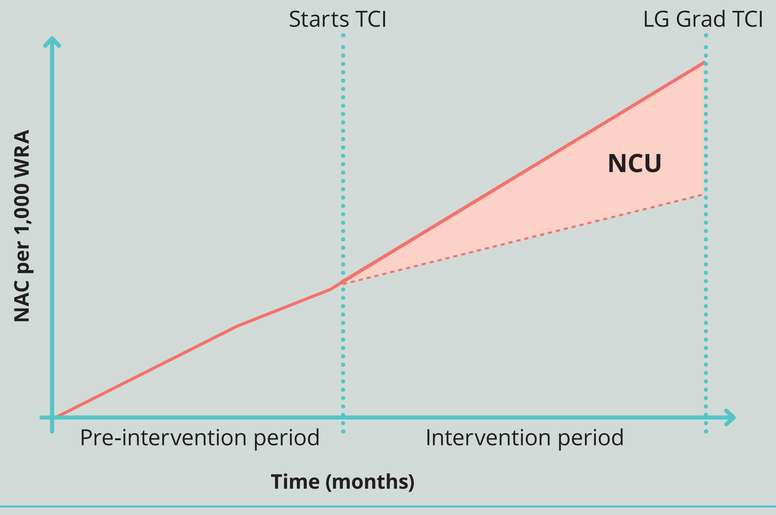
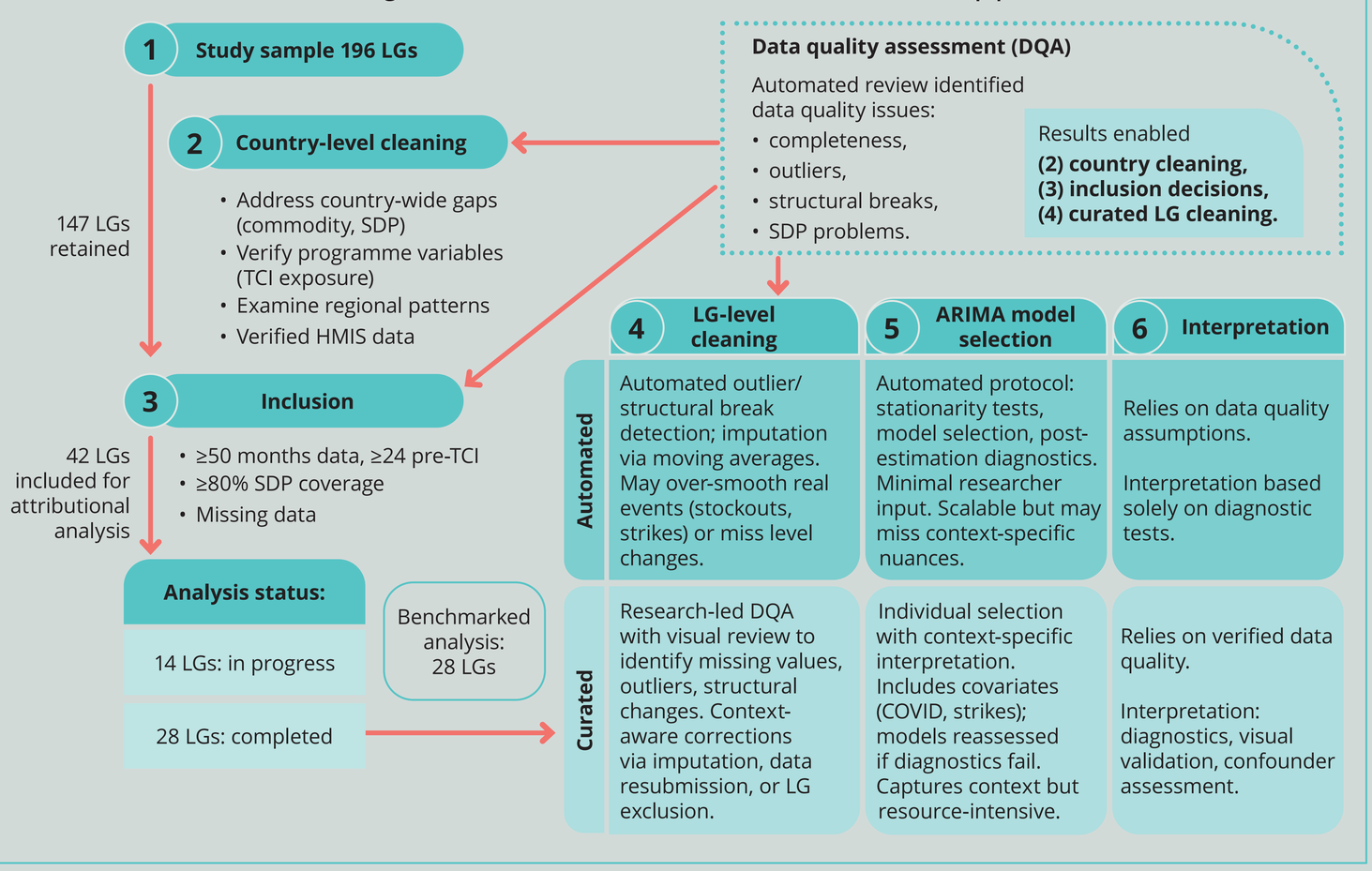


Figure 2 shows study inclusion and automated vs curated pipelines: both used identical stopping rules, but curated analysis used researcher-led specification, improving diagnostic performance.

Figure 2: Inclusion flow and automated vs curated ITS pipelines



RESULTS

Figure 3 benchmarks automated against curated pipelines using two measures: (1) percentage change in NCU (n=24; e.g., +100% = doubling of NCU), and (2) mean standardized NCU difference (n=28; direction-removed, as presented as standard deviation (SD),[1] distinguishing proportional from absolute shifts in NCU.

Results are classified by the available automated diagnostic results:

Core diagnostic failed[2]

Autocorrelation unresolved (n=13).

Supplementary diagnostics failed[3]

Conditional volatility or structural breaks (n=6).

Adequate

All diagnostics passed (n=9).

The % change in NCU visual only includes 24 observations, owing to directional shifts between automated and curated pipelines. These directional changes were not predicted by diagnostics (adequate n=1, supplementary n=1, core n=2); three of the four directional shifts were from negative to positive NCU between pipelines.

Core failed

Clustering around zero with several widely scattered estimates.

- Average % change = 35%
- Standardised mean diff = 0.14 SD

Supplementary failed

Few cases, large deviations, consistently unreliable.

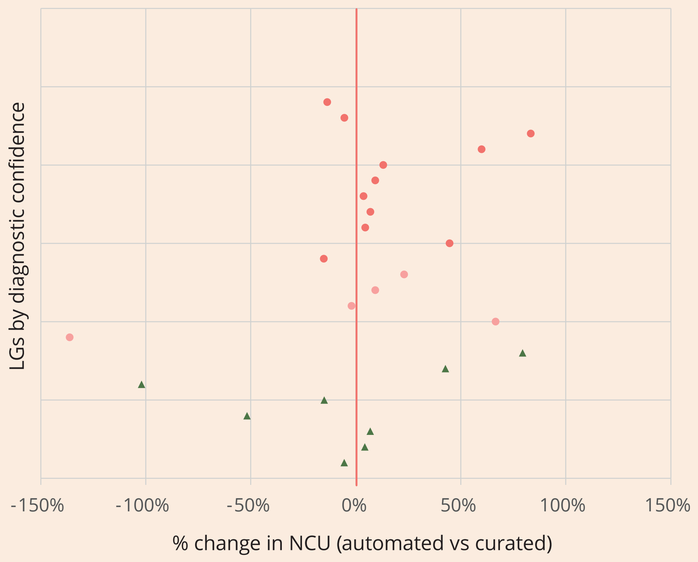
- Average % change = 47%
- Standardised mean diff = 0.20 SD

Adequate

Some clustering around zero but still variable.

- Average % change = 33%
- Standardised mean diff = 0.14 SD

Figure 3: % NCU change by automated diagnostics



Automated diagnostics poorly predicted divergence: all categories showed substantial relative differences (33% – 47%), including four unanticipated directional shifts (three from negative to positive). Moreover, moving from automated to the curated pipeline markedly improved diagnostics - autocorrelation control (56% → 100%), conditional variance control (40% → 98%), and addressing unexplained structural breaks (20% → 0%).

Box 1: LG-level cleaning, Pakistan

Pakistan illustrates how LG-level cleaning shaped ARIMA–ITS implementation and ensured rigor in the curated results.

- SDP covariate removed: unassociated with NAC, owing to compelled reporting; brief variability addressed with structural covariates.
- Directional changes explained: Korangi and Rawalpindi shifts reflected differing HMIS and Contraceptive Logistics Management Information System (cLMIS) reporting patterns pre-TCI.
- Review led to LG exclusion: Karachi Central excluded despite strong automated results, owing to uncorrectable reporting spike at TCI start.

CONCLUSIONS

Benchmarking automated against curated pipelines demonstrates that HMIS data can generate robust contraceptive uptake impact estimates using ARIMA ITS, but automated approaches have significant limitations.

Automated diagnostics poorly predicted divergence from curated analysis. All categories showed substantial average relative changes (33%–47%). Although standardized mean differences were modest (0.14–0.20 SD), they masked considerable variation between analysis pipelines for individual LGs (range = –0.36 to 0.68 SD) and large relative and directional shifts identified.

Despite TCI's data quality investments[4] and stringent inclusion criteria, curated analysis required substantial resources for verification, cleaning, and context-specific modeling, essential for reliable estimates and identifying changes missed by automation (Box 1).

We conclude that HMIS data can yield credible impact estimates of contraceptive uptake via ITS, but only under conditions that require substantial researcher input, which cannot be automated.

Footnote

[1] Calculated as: (Curated - Automated) / Automated × 100. Standardized mean difference, calculated as: |Automated NCU – Curated NCU | / SD(Curated NCU), averaged across all observations.

[2] Core diagnostic failures: Portmanteau test for white noise (minimum validity threshold). Failure indicates autocorrelated residuals and unreliable impact estimates.

[3] Supplementary diagnostic failures: Conditional heteroscedasticity (ARCH test for autocorrelated error volatility) or researcher-identified challenges (severe seasonality, unexplained variance shifts, or step changes).

[4] TCI addresses data quality through: (1) routine HMIS validation and correction; (2) quarterly data quality audits; (3) review meetings and (4) data quality tools and training