Breaking the Bottleneck:
Agentic AI for Timely Government Statistics

The Committee on National Statistics (CNSTAT) Workshop, National Academies of Sciences, Engineering, and Medicine
KiDeuk Kim / Senior Fellow / The Urban Institute
April 30, 2026

Problem Statement

Federal statistical agencies produce official statistics that guide policy, advance research, and strengthen public understanding. In the criminal justice field, these statistics are especially vital for tracking crime trends, offender and victim populations, and the operations of criminal justice agencies. Yet producing these data products often depends on labor-intensive backend work: cleaning data submissions from state and local agencies, standardizing variables across jurisdictions, linking fragmented records, and resolving missing or conflicting information. These burdensome processes create significant bottlenecks and can delay the release of critical indicators.

In large, multi-jurisdictional data collections, these challenges are compounded by **asynchronous reporting cycles**: **some agencies report quickly, while others may lag by months or longer**. The result is slower production, repetitive and inefficient workflows, reduced comparability, and statistical information that often arrives too late to inform fast-moving policy and operational decisions. This is especially consequential in criminal justice, where agencies need timely evidence to respond to emerging crime patterns, shifts in justice-system populations, staffing pressures, and evolving public safety concerns.

Overview

This project addresses those barriers by integrating agentic AI for administrative data processing with estimation methods that account for non-response on a rolling basis.

The proposed system automates core production tasks—including cross-jurisdiction variable standardization, validation, record linking, and weighting—while generating transparent, machine-readable documentation for every key step. This creates a workflow that is faster, more consistent, and easier to implement than traditional approaches that rely on repeated processing and manual review.

Rather than treating timeliness and rigor as competing goals, **the current approach combines automation through an agentic AI with established statistical methods and human oversight**. The result is stronger comparability across jurisdictions, improved reproducibility, and a modernized production pipeline capable of delivering continuously refreshed, dissemination-ready statistics.

Central Dilemma for Statistical Agencies

The current approach replaces the traditional wait-until-complete approach with a rolling estimation framework in which outputs are updated as new submissions are received. Historically, this approach has seen limited use in criminal justice because repeated data processing, reweighting, and estimate updates would have imposed substantial operational costs. In many settings, the staff time and resources required to continuously refresh statistics outweighed the benefits of releasing provisional estimates before data collection and processing were fully complete.

That logic is understandable, but it creates a less visible and often overlooked tradeoff. **By the time statistics are finalized, the period in which they are most valuable for policy decisions, operational response, or public understanding may have narrowed—or passed entirely**. This is the central dilemma: producing earlier estimates was traditionally too costly, yet waiting for complete data can make the final statistics far less useful when decisions must be made in real time.

From Delayed Reporting to Rolling Estimates

With advances in AI and workflow automation, tasks that once required repeated manual effort—such as file intake, variable standardization, validation checks, anomaly detection, record linkage, and documentation—can now be executed automatically and consistently at scale. This substantially reduces the operational burden that historically made iterative production difficult to sustain. As data submissions arrive, data can be processed, harmonized, quality-checked, and integrated into the production system with minimal delay.

Once reporting reaches a predefined coverage threshold, provisional estimates are generated using established statistical adjustment methods designed to reduce bias from non-responding jurisdictions or delayed submissions.

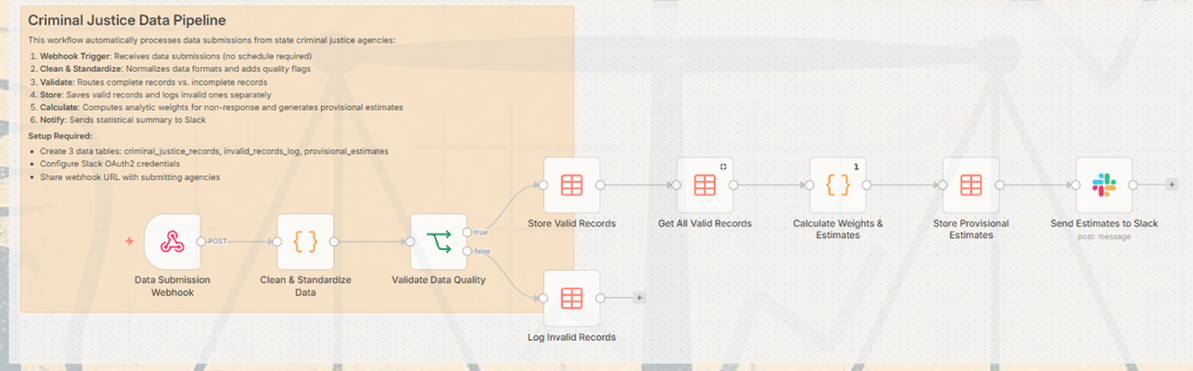
Reported jurisdictions are weighted to represent those still missing, and weights are updated as coverage improves:

$$\hat{Y} = \Sigma_{(i \in R)} w_i y_i$$

Where:
- R = responding jurisdictions
- y_i = reported value from jurisdiction i
- w_i = adjustment weight for jurisdiction i
- $\hat{Y}$ = provisional estimate for all jurisdictions

With each additional data submission, coverage expands and estimates are refreshed accordingly

Criminal Justice Data Pipeline



This workflow creates a faster and more reliable way to turn incoming agency data into usable statistics. Instead of relying on iterative manual processing, the system automatically checks, organizes, analyzes, and updates results whenever new data are submitted.

1. **Data Submission**
   State or local agencies send their data whenever they are ready. No need to wait for a fixed reporting deadline.
2. **Clean and Standardize**
   Incoming files are automatically transformed into a common structure so data from different agencies can be harmonized, integrated, and compared consistently.
3. **Data Quality Check**
   The system reviews submissions for missing fields, unusual values, and incomplete records. Certain issues can be corrected automatically, while records requiring additional attention are flagged for human review.
4. **Store Valid Records**
   Clean and validated records are added to the main production database for analysis and reporting. Records containing unreconcilable errors, missing information, or unresolved issues are stored separately in a review log, ensuring they are available for follow-up correction if necessary.
5. **Generate Interim Estimates**
   Once enough data have been received, the system automatically produces provisional statistics using established adjustment methods.
6. **Update and Share Results**
   The newest estimates are saved and summary updates can be sent automatically to dashboards, reports, or communication tools.

Rethinking Government Statistics

Official statistics are vital to public decision-making, but their **value depends not only on accuracy, but also on timeliness, transparency, and usability**. The current approach shows how **agentic AI and workflow automation can modernize statistical production by reducing bottlenecks, accelerating outputs**, and preserving rigor through human oversight and established methods. **The broader opportunity is not just faster reporting, but public data systems that adapt and inform decisions in real time**.

Acknowledgements

This work outlines a prototype framework for AI-enabled statistical production using n8n, large language models (e.g., ChatGPT, Claude), and Python for workflow automation and data processing. Development of this “Agent” for timely government statistics was informed in part by operational insights from the Bureau of Justice Statistics’ Criminal Cases in State Courts project (2018-85-CX-K029). Opinions, findings, conclusions, and recommendations expressed herein are those of the author and do not necessarily reflect those of the Urban Institute or its funders, including BJS.

For inquiries, please contact KiDeuk Kim, the Urban Institute (email: kkim@urban.org).