International Journal of Computer Applications Technology and Research
Volume 14–Issue 203 – 211, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1019

# An Intelligent Process Optimization Framework Integrating Predictive Analytics and Workflow Modeling to Improve U.S. Tax Administration Efficiency

[1]Juliet Nankunda, [2]David Kaberia, [1]Ismail Katamba
[1]Maharishi International University
[2]Albers School of Business and Economics at Seattle University

## Abstract

Federal tax administration combines large transaction volumes, statutory obligations, and interdependent review workflows. This article develops an Intelligent Process Optimization Framework (IPOF) that integrates workload forecasting, process-state modeling, capacity-constrained case routing, human oversight, and performance monitoring. Following design-science principles, the study conducts a thematic synthesis of scholarly research and a descriptive secondary analysis of U.S. administrative records published by December 2024. The official evidence provides an operational baseline rather than a test of IPOF. The IRS reported 141.317 million individual returns processed by May 10, 2024. The Treasury Inspector General for Tax Administration reported 455,463 amended returns pending in Submission Processing on March 2, 2024, of which 73.1% had exceeded that function's 30-day aging threshold. Existing FixERS automation resolved 321,357 of 493,072 returns placed into its production process as of February 29, 2024, while 171,715 were returned to manual processing. These findings identify workload scale, age-sensitive queues, and automated-to-human handoffs as measurable design requirements. IPOF is specified through a five-layer architecture, predictive expressions, explicit assignment constraints, and a proposed evaluation protocol. No case-level IRS data, deployed system, or causal efficiency improvement is claimed. The contribution is an auditable, testable integration design grounded in verifiable administrative measurements and clearly delimited future validation requirements (Internal Revenue Service [IRS], 2024a; Treasury Inspector General for Tax Administration [TIGTA], 2024).

**Keywords:** tax administration; predictive analytics; workflow modeling; process mining; constrained optimization; government AI; taxpayer service; operational efficiency

## 1. INTRODUCTION

### 1.1 Background of the Study

Federal tax administration combines high-volume transaction processing with legally consequential decisions about filing, payments, correspondence, examinations, collections, and appeals. The IRS reported processing more than 271.4 million returns and other forms and collecting approximately $4.7 trillion in fiscal year 2023 (IRS, 2024b). Separately, its filing-season statistics reported 141,317,000 individual income tax returns processed by May 10, 2024, compared with 140,224,000 at the May 12, 2023 reporting point (IRS, 2024a). These two indicators cover different reporting periods and product populations and should not be combined into a single annual throughput series. They establish scale, not the efficiency of any proposed technology.

Service and exception handling present more specific operational challenges. In a January 2024 review, the U.S. Government Accountability Office (GAO, 2024) reported that 61% of taxpayer mail inquiries were classified as late at the end of the 2023 filing season despite improved telephone-service performance. TIGTA (2024, p. 5) recorded 455,463 amended returns awaiting work in Submission Processing on March 2, 2024, of which 73.1% were beyond that function's 30-day aging threshold. These findings support examination of case inventory, waiting time, resource constraints, and cross-stage handoffs, rather than predictive accuracy alone.

The IRS also projected a $696 billion gross tax gap for tax year 2022, including $539 billion attributable to underreporting, in a report issued in October 2024 (IRS, 2024c). The tax gap measures projected unpaid tax liability and is not an administrative-processing cost, an observed 2024 revenue shortfall, or an amount that IPOF can automatically recover. It motivates responsible compliance infrastructure but is not used as an estimated project benefit. Internationally, the Organisation for Economic Co-operation and Development (OECD, 2024) documents varied digital administrative practices across jurisdictions, underscoring the need for institution-specific evaluation.

These observations motivate IPOF as a proposed, auditable decision-support framework. Its purpose is not to automate a taxpayer liability judgment, but to connect authorized administrative work with predictive insight, documented workflow states, staff capacity, and rights-based safeguards. Operational efficiency is defined through timely resolution, pending inventory, rework, staff requirements, service access, and accuracy.

### 1.2 Problem Statement

The central problem is a **decision-to-execution gap**. A statistical model may indicate that an administrative case is complex or time-sensitive without identifying the authorized unit, available staff, required documentation, and time limits needed to resolve it. Conversely, a static workflow can document handoffs and delays without forecasting incoming demand or allocating work under changing capacity. The practical consequence may be unresolved exceptions and aged inventories even where some tasks are automated.

TIGTA (2024) shows why the issue is not a simple absence of technology. The IRS's FixERS tool had already handled 493,072 returns placed into its production process by February 29, 2024; 321,357 returns were resolved automatically and 171,715 were sent back for manual processing. The policy and research question is therefore how existing automated and manual work can be modeled, monitored, and coordinated, not how to introduce automation into an entirely manual agency. The proposed contribution is an integrated and testable architecture, not evidence that an IRS implementation has occurred.

### 1.3 Objectives of the Study

The objectives are to (a) critically synthesize scholarly research and official U.S. administrative evidence published through December 2024; (b) extract reproducible historical indicators of filing workload, aged queues, correspondence timeliness, and automated-to-manual handoffs; (c) specify IPOF's data, process, predictive, optimization, and oversight layers; (d) formulate measurable allocation constraints and an empirical validation plan; and (e) distinguish verifiable baseline observations from untested efficiency hypotheses. The article does not equate faster processing with recovered tax revenue.

### 1.4 Research Questions

RQ1. What do publicly reported pre-2025 U.S. tax-administration statistics establish about workload, aged processing queues, and automation handoffs? RQ2. How can predictive analytics and event-log workflow modeling be integrated into authorized, capacity-aware case routing? RQ3. Which objective functions, constraints, KPIs, and safeguards are needed to test whether the proposed framework improves timeliness without creating unfair or unlawful burdens? RQ4. What additional case-level data and prospective comparison would be necessary to estimate IPOF's effectiveness?

### 1.5 Scope and Delimitations

The design's conceptual unit is an **administrative case**, such as a processing exception, correspondence item, or amended-return review requiring staff attention. The empirical analysis uses only published aggregate statistics for defined reporting periods, not individual taxpayer records, administrative event logs, or experiments on an IRS system. The framework does not determine liability, autonomously select examinations, or replace statutory notices and appeals. Figures from different reports are used to characterize distinct processes and are not pooled into a common denominator. Every cited publication and government record used in this article was released no later than December 31, 2024.

## 2. LITERATURE REVIEW

### 2.1 Theme I: Tax Digitalization, Administrative Capacity, and Service Outcomes

Hesami et al. (2024) systematically reviewed research on electronic invoicing and prefilled tax returns using structured evidence synthesis. Their work identifies possible efficiency and compliance benefits associated with better information capture, but does not establish how downstream correspondence, specialist review, or exception queues should be scheduled. Digital intake may shift a bottleneck without resolving it. The OECD (2024) examines digital channels and administrative practices across 58 jurisdictions, but its cross-sectional comparisons cannot demonstrate the causal impact of a specific U.S. routing algorithm.

In the U.S. setting, GAO (2024) reviewed filing-season data, operating controls, and staff interviews. It found that 61% of taxpayer-mail inquiries were late at the end of the 2023 filing season, despite progress in processing returns and answering calls. In a separate 2024 operational review, TIGTA (2024, p. 5) reported materially different pending inventories and aging thresholds across return-processing programs. For example, 73.1% of 455,463 pending amended returns in Submission Processing exceeded 30 days; 63.1% of 502,368 pending amended returns in Accounts Management exceeded 44 days. These figures should not be compared as a single standardized lateness rate because the case populations and criteria are different. They do, however, support a disaggregated approach to workflow-state definitions, throughput, and queue-aging metrics.

Darmayasa and Hardika (2024) examined discourse surrounding Indonesia's core tax administration system through qualitative content analysis, associating system integration with administrative authority and taxpayer trust. That work does not measure IRS operational gains and should not be transferred directly to U.S. federal processes. Taken together, the studies support digital coordination as a plausible design objective while emphasizing the need for process-specific measurement.

### 2.2 Theme II: Risk Prediction, Administrative Decisions, and Optimization Objectives

Prediction and allocation solve different problems. A predictive model estimates an outcome conditional on available information, while an allocation rule chooses an action under resource, timing, and legal constraints. Keen and Slemrod (2017) develop an economic framework for optimal tax administration, emphasizing that enforcement responses and administrative costs must be evaluated together. Their formal analysis clarifies why maximizing a single detection statistic may differ from maximizing overall administrative or social value. It does not, however, specify a case-level workflow engine linking event logs to a daily staffing schedule.

Elzayn et al. (2023) offer particularly relevant U.S. evidence. Using IRS audit data, partial-identification methods, and counterfactual selection models, they estimate sizable disparities between audit rates for Black and non-Black taxpayers despite race-blind selection rules. Their analysis demonstrates that the **objective function itself** can affect who bears an administrative burden. Because their outcome concerns audits rather than return-processing queues, the paper does not validate IPOF's scheduling algorithm. It nevertheless provides an empirical rationale for measuring the distribution of case interventions and delays rather than treating a risk score as an ethically neutral priority.

The literature therefore supports separate evaluation of (a) model calibration and forecast error, (b) resource-allocation outcomes, and (c) taxpayer-facing consequences. IPOF proposes case-duration forecasts and deadline constraints but does not claim to have trained or validated a fraud prediction model. Prediction is only useful when it changes an authorized decision and the effects of that change are observed.

## 2.3 Theme III: Workflow Modeling, Process Mining, and Queue Dynamics

The Object Management Group (2014) formalizes Business Process Model and Notation (BPMN) as a way to represent activities, gateways, and organizational interactions. BPMN offers an interpretable target process but does not by itself quantify how work actually moves across offices. Van der Aalst (2016) develops process-mining methods that use timestamped event traces to discover process variants, test conformance with expected paths, and identify performance bottlenecks. For tax operations, this distinction is central: a documented target process may conceal repeated requests for evidence or long intervals between successive case events.

Event-log analysis can estimate elapsed time, repeated activity, queue arrival rates, and handoff counts. Incomplete timestamps, non-work status changes, and censoring of open cases can distort process mining and understate waiting time. Because discovered delays do not establish policy effects, process discovery must be paired with controlled evaluation.

A queueing perspective links the components: arrivals introduce workload, processing capacity constrains service, and prioritization redistributes waiting time. Consequently, the value of a routing policy should be tested under matched workload rather than inferred from its risk-ranking performance. IPOF treats queue behavior as an explicit outcome that must be reported alongside any predictive diagnostic.

## 2.4 Theme IV: Trust, Algorithmic Accountability, and Taxpayer Rights

Kirchler et al. (2008) conceptualize tax compliance through interacting dimensions of enforcement power and trust in authorities. Darmayasa and Hardika's (2024) tax-system discourse analysis reflects this dual concern, although it does not demonstrate causal effects from digital services to voluntary compliance. For IPOF, the implication is narrower and operational: unexplained delays, inconsistent treatment, or opaque interventions may impose burdens even when processing throughput increases.

The NIST AI Risk Management Framework sets out govern, map, measure, and manage functions for controlling AI-related risks (Tabassi, 2023). Separately, GAO (2023) found incomplete or inaccurate AI-use inventory data across multiple U.S. federal agencies, reinforcing the need for documented model ownership, intended benefit, and monitoring. These are cross-government governance findings, not a measurement of IRS-specific IPOF readiness. IRS guidance in 2024 reiterates taxpayers' rights to quality service and to challenge agency positions (IRS, 2024d); Section 6103 restricts unauthorized disclosure of return information (26 U.S.C. § 6103, 2024). Nyombi et al. (2024) examined cybersecurity in tax accounting practices and highlighted inconsistent safeguards and the need for standardized protection measures. Their findings motivate security requirements for IPOF, but do not establish the effectiveness of a tax-case routing algorithm.

Elzayn et al. (2023) provide empirical evidence that a formally race-neutral selection process can generate unequal administrative effects. Accordingly, IPOF proposes audit trails, human review, consent-independent legal access controls, and disparity assessment at both the prediction and service-allocation stages. These controls are requirements for proposed deployment; the present aggregate statistics cannot establish algorithmic fairness.

## 2.5 Theoretical and Conceptual Framework

The article combines **design science**, **operations research**, **process mining**, and **taxpayer-rights governance**. Hevner et al. (2004) characterize design science as developing and evaluating artifacts that respond to practical organizational problems. Peffers et al. (2007) provide a six-stage design-science research methodology, including problem identification, artifact construction, demonstration, evaluation, and communication. These works justify treating IPOF as an artifact whose structure, assumptions, and testability must be documented, rather than as a claim of empirical policy success.

The conceptual chain is: **authorized event data -> validated process states -> forecasts of case complexity and arrivals -> legally feasible routing recommendations -> human execution -> observed service and fairness metrics -> recalibration**. Each link is a potential failure point. Incorrect time stamps distort process discovery; labels observed after resolution create leakage; an infeasible schedule ignores staffing limits; and unmonitored prioritization may shift burden to lower-priority cases. The framework therefore requires joint evaluation of predictive quality, queue performance, administrative legitimacy, and taxpayer-data security. Nagalila et al. (2024) discuss the importance of coordinating financial controls with cybersecurity, although their study does not test U.S. tax-processing optimization.

## 2.6 Empirical Synthesis and Specific Research Gap

The reviewed sources differ in methods and inferential strength. The tax-accounting security review by Nyombi et al. (2024) informs confidentiality and incident-response design, not queue performance. Hesami et al. (2024) synthesize digital-tax research but do not measure U.S. case-routing choices. Darmayasa and Hardika (2024) provide qualitative institutional analysis rather than post-deployment queue estimates. Elzayn et al. (2023) analyze U.S. audit-selection outcomes and disparities, but do not evaluate how tax-processing correspondence or amended-return cases should be assigned across staff pools. GAO (2024) and TIGTA (2024) provide measured administrative baselines, but neither evaluates an integrated predictive workflow algorithm. Van der Aalst (2016), Hevner et al. (2004), and Peffers et al. (2007) contribute transferable process-mining and design-science methods rather than evidence of an IPOF deployment.

The specific **researchable gap** concerns how predictions about future workload and case effort can be converted into feasible, auditable routing recommendations under legal and staffing constraints, and then evaluated against independently defined service KPIs. The present paper answers this at the design level and documents selected historical baselines. It does not claim to have estimated a causal benefit, to have identified every existing IRS internal system, or to have established the novelty of each technical component in isolation.

International Journal of Computer Applications Technology and Research
Volume 14–Issue 203 – 211, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1019

# 3. METHODOLOGY

## 3.1 Research Design and Approach

### *3.1.1 Design-science strategy and evidence boundaries*

The study follows design-science principles from Hevner et al. (2004) and Peffers et al. (2007). Its sequence is to identify a documented administrative problem, translate official findings into measurable requirements, develop the IPOF decision-support artifact, define the mathematical assignment problem, and specify how a real implementation should be evaluated. The empirical element is a descriptive secondary analysis of already published administrative statistics. No simulation of IRS outcomes, trained predictive classifier, prospective pilot, or demonstrated efficiency gain is included. The distinction prevents a conceptual framework from being confused with a tested intervention.

### *3.1.2 Documentary synthesis and source selection*

The literature review is a structured thematic synthesis, not a registered systematic review. Included sources are scholarly publications, technical specifications, and primary government records released by December 31, 2024, selected for direct relevance to administrative workload, process mining, optimization, fairness, and governance. The quantitative extraction prioritizes four official records: IRS filing-season statistics for May 10, 2024 (IRS, 2024a); GAO's January 2024 review of the 2023 filing season (GAO, 2024); TIGTA's April 2024 filing-season audit (TIGTA, 2024); and the IRS's October 2024 tax-gap projections (IRS, 2024c). Each numeric finding is reported with its source, observational cutoff, unit, denominator where applicable, and interpretive limits. An exhaustive database search, meta-analysis, and unseen confidential administrative records are not claimed.

### *3.1.3 Units, assumptions, and reproducibility*

Three observable aggregate units are used: individual tax returns processed at a named filing-season reporting date; pending cases in a named IRS program at a snapshot date; and return errors entering the FixERS automated process by February 29, 2024. GAO's late-correspondence percentage is separately retained at the level of the reported inquiry inventory. The conceptual IPOF system instead operates on individual case histories, which are not available in these public summaries. The procedure is reproducible at the level of source-to-figure transcription and arithmetic checks, not at the level of fitting predictions or identifying causal treatment effects.

## 3.2 Analytical Model and System Architecture

### *3.2.1 Five-layer architecture*

The proposed pipeline comprises (i) protected intake and provenance checks; (ii) event-log process discovery and state reconstruction; (iii) forecasting of arrivals, duration, and narrowly defined case outcomes; (iv) constraint-aware task allocation; and (v) human approval, audit logging, and KPI monitoring. Figure 1 shows the data flow. Failure or uncertainty at any layer reverts the case to a documented conventional route. This modular structure distinguishes the proposed operational system from the aggregate historical data available for the present analysis. The intake layer includes access controls, encryption, and security-event logging informed by the tax-accounting threats examined by Nyombi et al. (2024).

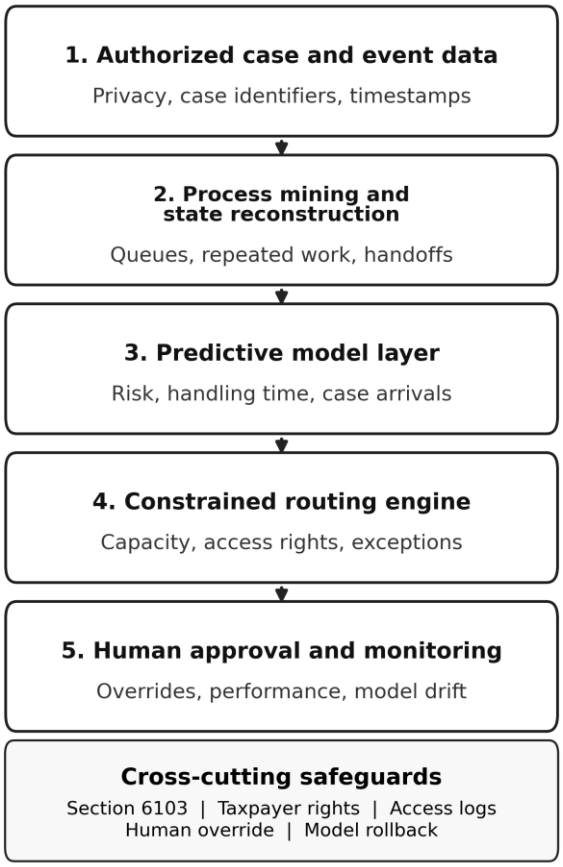


***Figure 1. Proposed five-layer IPOF architecture. Source: Author-developed framework.***

### *3.2.2 Forecasting model and calibration requirements*

For case i with features $x_i$ available at decision time t and state $s_i$, a future application could estimate a narrowly defined binary outcome as $p_i = P(Y_i = 1 \mid x_i, s_i, t)$, and remaining work duration as $\hat{d}_i = E(D_i \mid x_i, s_i, t)$. For processing unit k, the forecast workload in period t is $\hat{W}_{kt} = \Sigma_{i \in Q_{kt}} \hat{d}_i + \hat{\lambda}_{kt} \times E(D_{kt})$, where $Q_{kt}$ is pending eligible work, $\hat{\lambda}$ is arrivals forecast for that period, and E(D) is mean new-case effort. This expression is a planning approximation, not an empirically calibrated IRS forecast. Prediction labels must be observed only after the decision and must have lawful, operational meaning. Training data must exclude information that would not have been available at the routing time.

### *3.2.3 Feasible assignment and deferral*

Let $z_{ik} \in \{0,1\}$ indicate assigning case i to unit k for the day and $u_i \in \{0,1\}$ indicate deferring the case. The identity $\Sigma_k z_{ik} + u_i = 1$ prevents the assignment model from becoming infeasible when total demand exceeds capacity. A candidate objective is $J = \Sigma_i \Sigma_k z_{ik}(\alpha \hat{T}_{ik} + \beta \hat{C}_{ik} + \delta \hat{B}_{ik} - \gamma \hat{U}_{ik}) + \Sigma_i u_i A_i$, where $\hat{T}$ is expected delay, $\hat{C}$ staff effort, $\hat{B}$ expected additional taxpayer burden, $\hat{U}$ the lawful value of timely resolution, and $A_i$ an aging or deadline-related

deferral penalty. Weights α, β, γ, and δ are policy choices requiring transparent approval, not universal statistical constants.

Capacity constraints $\Sigma_i z_{ik}\hat{d}_{ik} \leq H_{kt}$ limit assignments to unit k's available hours $H_{kt}$. Eligibility constraints $z_{ik} \leq e_{ik}$ prevent work from being assigned to an unauthorized or unqualified unit. Cases with statutory deadlines, security restrictions, unresolved evidentiary requirements, or mandatory human review are constrained to legally permitted paths. Sensitivity analyses should vary the objective weights because the chosen priority definition can alter burdens (Elzayn et al., 2023). **This study does not estimate or solve the assignment formulation using IRS case-level records.** The equations state what a later authorized implementation would need to estimate and optimize.

## 3.3 Algorithm Flowchart and Process Design

### *3.3.1 State transitions and decision gates*

A BPMN representation should distinguish case received, eligibility checked, evidence requested, review pending, staff assigned, decision documented, and case closed states (Object Management Group, 2014). Case IDs and timestamps permit process mining for transitions, repeated steps, and waiting intervals (van der Aalst, 2016). Each model recommendation passes authorization, data-quality, confidence, available-capacity, deadline, and human-review gates before being eligible for operational action. A missing timestamp or unsupported prediction triggers fallback rather than an automated adverse determination.

### *3.3.2 Algorithmic procedure*

```
FOR each operating period t:
    LOAD authorized case events; check access rights
and completeness
    RECONSTRUCT case states and timestamped waiting
histories
    FORECAST arrivals and unresolved-case work; flag
uncertain outputs
    DIVERT restricted, overdue, or low-confidence
cases to officer review
    FORM feasible assignments and explicit deferrals
under capacity limits
    PRESENT routing rationale and permitted
alternatives to staff
    RECORD staff approval or override, action time,
and reason code
    UPDATE delay, open inventory, rework, access,
and disparity measures
    REVERT to the standard route if model or service
controls fail
END FOR
```

### *3.3.3 Proposed algorithm versus documented operations*

The algorithm and assignment equations are specifications for future development, not software shown to be operating within the IRS. TIGTA's evidence that 321,357 FixERS cases were automatically resolved and 171,715 were returned for manual processing (TIGTA, 2024, p. 7) establishes a real workflow handoff that an authorized future implementation could represent as separate states. It does not supply the time stamps, handling durations, or eligibility conditions needed to calculate an optimal schedule. Consequently, no estimated change in cycle time, staff utilization, or audit yield is attributed to IPOF.

## 3.4 Data Sources, Processing, and Validation

### *3.4.1 Required production data and lawful use*

A future restricted-data evaluation would need authorized case identifiers, case type, intake channel, time-stamped process events, assigned unit, closure status, recorded rework, and permissible case characteristics. The technical preprocessing plan comprises duplicate-event removal, timestamp and time-zone normalization, transition validation, data-leakage checks, missingness reporting, and case-level chronology reconstruction. Unresolved cases must remain in outcome analyses to avoid falsely optimistic completion estimates. U.S. return-information access is limited by 26 U.S.C. § 6103 (2024), and pseudonymization alone is not legal authorization for additional uses. Following the protection priorities discussed by Nyombi et al. (2024), the proposed data pipeline also requires logged access, verification of authorized transfers, and an incident-escalation procedure; these remain untested design requirements.

### *3.4.2 Model assessment protocol*

A future binary-outcome predictor would be evaluated with temporally ordered training and test cohorts, calibration, Brier score, precision at a fixed review budget, and subgroup error checks. Duration forecasts would require mean absolute error, interval coverage, and late-case tail error. For workflow evaluation, the primary outcomes would be elapsed case cycle time, proportion past an agency-defined deadline, 90th-percentile delay, work-in-progress inventory, manual rework, staff hours, and taxpayer contacts per resolved case. The present public evidence establishes only selected baselines for queue sizes, category-specific aging, and automated-versus-manual disposition. It is insufficient to compute individual duration prediction error or any before-after IPOF effect.

### *3.4.3 Governance and operational resilience*

Before a real pilot, an administrative owner would approve the purpose and permitted features; security personnel would enforce access and retention controls; caseworkers would control decisions; and independent reviewers would validate and monitor the artifact. Versioned models, override logs, monitoring of drift, and a manual fallback are required. This follows risk-governance principles set out by Tabassi (2023) and reflects broader federal AI inventory and accountability weaknesses described by GAO (2023). Neither source demonstrates that these safeguards have been installed in IPOF.

## 3.5 Secondary Data Extraction and Descriptive Analysis

### *3.5.1 Source-to-indicator mapping*

The extracted indicators are reported in Table 1. For the May 10, 2024 filing-season comparison, the measure is cumulative individual returns processed (141,317,000), not cases closed in a correspondence queue (IRS, 2024a). For the March 2, 2024 TIGTA inventory, the measure is outstanding work by distinct program, not arrivals or annual throughput. Two amended-return categories have separate aging rules: more than 30 days for Submission Processing and more than 44 days for Accounts Management (TIGTA, 2024, p. 5). For the FixERS tool, the denominator is 493,072 returns placed into its production process as of February 29, 2024 (TIGTA,

2024, p. 7). The GAO correspondence measure refers to the end of the 2023 filing season and cannot be interpreted as the fraction of all IRS filings that were late (GAO, 2024).

### *3.5.2 Descriptive computations and reporting rules*

The FixERS automated disposition rate is defined as 321,357 / 493,072 = 65.2% before rounding, and the fraction returned for manual processing is 171,715 / 493,072 = 34.8%; TIGTA reports the corresponding rounded values as 65% and 35%. The two groups sum exactly to the stated FixERS population. Snapshot case inventories are presented as counts and category-specific percentages in separate tables or graphics, without adding categories where workflow overlap may exist. Where an indicator includes an aging threshold, the threshold is displayed adjacent to the percentage. No unobserved case arrival process, processing duration, labor hours, or taxpayer characteristics are imputed.

### *3.5.3 Threats to inference and planned empirical test*

The observations describe different systems, source windows, and denominators. An increased inventory may reflect greater incoming demand, slower service, different eligibility rules, or changes in reporting; it does not identify a causal mechanism without event-level measurements. Similarly, the FixERS disposition split does not establish the cost or delay saved by the tool. Future research must obtain an authorized event log, validate the case-state map, estimate arrival and service-time distributions, and compare baseline routing with IPOF using a prospective or appropriately matched evaluation. Legal permissions and independent review would precede any use of taxpayer information.

## 4. RESULTS AND DISCUSSION

## 4.1 Presentation of Key Findings

### *4.1.1 Actual workload scale and service delays*

The IRS reported 141,317,000 individual tax returns processed by May 10, 2024, compared with 140,224,000 at its May 12, 2023 reporting point (IRS, 2024a). This approximately 0.8% reported increase is useful as evidence of processing scale, not as a controlled annual productivity comparison. In a different operating process and reporting period, GAO (2024) found that 61% of taxpayer-mail inquiries were late at the end of the 2023 filing season. The figures make different points: large-scale completed return processing can coexist with substantial unresolved service demand.

### *4.1.2 Program-specific aged inventories*

TIGTA (2024, p. 5) recorded 455,463 pending amended returns in Submission Processing and 502,368 in Accounts Management as of March 2, 2024. The respective proportions beyond their program aging thresholds were 73.1% (over 30 days) and 63.1% (over 44 days). The same audit reported 432,672 unpostable returns, of which 43.8% had exceeded age criteria that varied from more than three to more than fifteen weeks. Figure 2 displays the two amended-return categories because these provide a relatively coherent example of specialist queues; they remain distinct functions with different lateness definitions. A larger pending count does not necessarily imply a slower staff member or a single shared bottleneck. The observations support disaggregated queue and exception-state modeling.

***Table 1. Selected observed IRS operational indicators with source-specific definitions***

| Operating measure | Published observation<br>Source and observation cutoff |
|---|---|
| Individual returns processed | 141,317,000, cumulative<br>IRS (2024a), May 10, 2024 |
| Taxpayer mail inquiries classified late | 61% of inquiries, not all IRS returns<br>GAO (2024), end of 2023 filing season |
| Amended returns, Submission Processing | 455,463 pending; 73.1% > 30 days<br>TIGTA (2024, p. 5), Mar. 2, 2024 |
| Amended returns, Accounts Management | 502,368 pending; 63.1% > 44 days<br>TIGTA (2024, p. 5), Mar. 2, 2024 |
| FixERS disposition, eligible production cases | 321,357 automated; 171,715 sent for manual work (N = 493,072)<br>TIGTA (2024, p. 7), Feb. 29, 2024 |

***Note. These observations are reported for different dates and administrative processes. Case inventories, filing counts, and automation dispositions are not interchangeable or additive.***

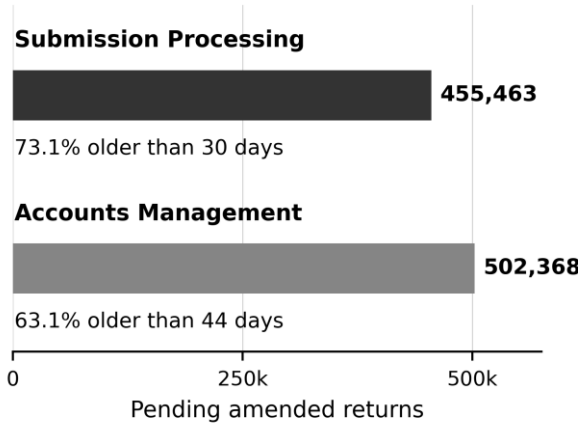


***Figure 2. Actual amended-return processing inventories on March 2, 2024. Source: TIGTA (2024, p. 5, Figure 4).***

### *4.1.3 Existing automation and remaining manual handoffs*

TIGTA (2024, p. 7) reported that FixERS had received 493,072 returns into production as of February 29, 2024. It automatically resolved 321,357 cases, while 171,715 were resuspended for manual processing. Figure 3 reports these two mutually exclusive outcomes using the audit's denominator. The 65% automated resolution share establishes that tax processing already includes functioning automation. The 35% manual portion highlights a workflow boundary where specialized review, routing, and transparent disposition tracking may matter. Neither figure measures the time or cost impact of IPOF.

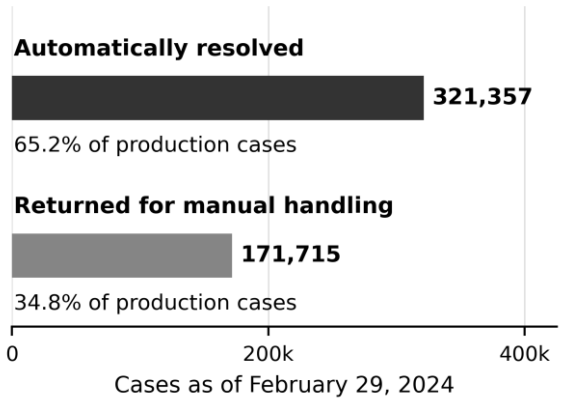


*Figure 3. Actual FixERS automatic versus manual dispositions as of February 29, 2024. Source: TIGTA (2024, p. 7).*

## 4.2 Interpretation and Analysis

### *4.2.1 What the administrative observations establish*

The evidence reveals an **integration problem**, not a demonstrated failure of every existing tool. IRS 2024 filing statistics document high-volume intake and processing; GAO (2024) documents unresolved taxpayer correspondence; and TIGTA (2024) documents aged specialist inventories alongside successful automation of a defined error subset. A predictive model concerned only with incoming returns would not by itself remove a constraint in amended-return review. Similarly, increasing automated error resolution could transfer work to a manual queue unless end-to-end demand, case pathways, and capacity are measured. IPOF's value proposition should therefore be evaluated by comparing full case-resolution paths rather than a standalone classification score.

### *4.2.2 Boundaries of the available evidence*

The observed counts and percentages are sourced, but no matched treatment and control groups, individual process traces, or staff utilization records were available. These measures cannot identify the causes of delay or calculate realized benefits from the proposed framework. In particular, the 73.1% and 63.1% amended-return aging rates use different thresholds and should not be ranked as if both measured a common 30-day service standard. An implementation study must use harmonized, pre-registered KPIs and assess whether any shorter waits for prioritized cases come at the expense of legally time-sensitive or vulnerable taxpayers (Elzayn et al., 2023; Tabassi, 2023).

## 4.3 Policy and Implementation Implications

### *4.3.1 Due process and confidentiality*

An IPOF deployment should be restricted to decision support and administrative routing until statutory authority, auditability, and evidence standards have been established for the specific use. Human officials should remain responsible for determinations affecting liabilities, enforcement, or appeal rights. Any use of taxpayer records requires an authorized purpose, access-control logging, retention limits, and protections consistent with Section 6103 (26 U.S.C. § 6103, 2024). Model explanations must support internal accountability without exposing sensitive examination-selection methods. Transparency can therefore combine meaningful taxpayer notices and oversight access with lawful protection of enforcement-sensitive information.

### *4.3.2 Procurement, staffing, and governance*

GAO (2023) reported federal AI-use inventory and management shortcomings, while GAO (2024) identified IRS processing and service measurement issues. Together these findings suggest that better analytics alone are insufficient. Implementation would require data stewards, tax-process specialists, security staff, quantitative model validators, frontline caseworkers, and clear executive ownership of outcome measures. Pilot procurement should specify access to audit logs, data portability, model replacement, independent evaluation, and fallback operation. A phased design is preferable to wholesale replacement of legacy systems because it limits operational disruption and permits time-bounded learning from each process stage (GAO, 2024). The integrated financial-control and cybersecurity perspective discussed by Nagalila et al. (2024) further supports joint review of process controls and information security, without demonstrating IRS-specific cost savings.

## 4.4 Innovation and Research Contributions

### *4.4.1 Technical contribution*

The principal contribution is an operationally specified chain from event evidence to prediction and from prediction to a **constrained, reviewable administrative decision**. Unlike an isolated risk-scoring model, IPOF incorporates work estimates, available staff capacity, allowable case paths, aging safeguards, override records, and post-deployment monitoring. Unlike a static BPMN workflow, the proposed artifact predicts demand and supports scenario comparisons. Its mathematical objective can be tailored to multiple policy priorities while separating mandatory legal constraints from adjustable managerial preferences.

### *4.4.2 Evidentiary contribution and boundaries*

The empirical contribution is a traceable mapping of selected government-reported operational figures into requirements for future prediction, queue modeling, and workflow evaluation. Rather than inventing a performance comparison, the article distinguishes documented service conditions from a technical architecture that has not yet been deployed or validated. Earlier work covers digital transformation, optimal tax administration, process mining, and responsible automated decisions as distinct topics (Hesami et al., 2024; Keen & Slemrod, 2017; van der Aalst, 2016). The research contribution is the explicit, falsifiable integration design and associated source-grounded test plan, not a new algorithm or demonstrated tax-revenue effect.

## 4.5 Scalability and Broader Impact

### *4.5.1 Modular scaling across tax workflows*

Technical scaling would start with a bounded, low-risk workflow such as routine correspondence triage, proceed through shadow-mode scoring, and expand only after prospective controls are met. A common event taxonomy could support comparisons across queues while retaining local policy constraints. Reusable interfaces might reduce future integration costs, but heterogeneous source systems, staffing, and tax rules mean that 'scalable' should be tested

rather than assumed. Changes in federal tax law should trigger review of relevant features, routing rules, and model calibration.

#### *4.5.2 Fiscal, taxpayer, and social implications*

Potential benefits include shorter document-resolution cycles, better allocation of scarce specialist time, lower repeated-contact demand, and fewer unnecessary interventions. The possible fiscal effect should be measured as actual budgetary cost avoided or verified lawful revenue recovered, net of deployment and oversight costs. The IRS projected a $696 billion gross tax gap for tax year 2022 in its October 2024 publication (IRS, 2024c), but applying a hypothetical recovery percentage to that figure would not be a defensible estimate of IPOF savings. More timely case handling could improve the ability of households and small firms to obtain clarifications and resolve refund questions, but such outcomes require direct measurement of taxpayer burden and differential impact rather than inference from queue throughput alone.

# 5. CONCLUSION

## 5.1 Summary of Findings

### *5.1.1 Scholarly and institutional findings*

The evidence reviewed supports the importance of modernizing tax administration through data integration and better workflows, while also documenting limitations involving staffing, correspondence, legacy technology, and AI governance. A meaningful research contribution must not equate the existence of predictive analytics with a demonstrated improvement in government performance. The proposed IPOF makes the critical decision pathway explicit and specifies how to observe its effects.

### *5.1.2 Findings from the administrative data*

The official baseline reveals the coexistence of 141.317 million processed individual returns by May 10, 2024, aged amended-return inventories in March 2024, and an established automated error-resolution tool that nevertheless returned 171,715 cases for manual processing as of February 29, 2024. These observations motivate integrated workflow planning, but do not establish that IPOF reduces processing time or administrative cost. The interpretation is deliberately bounded by source-specific reporting populations, dates, and age definitions.

## 5.2 Recommendations

### *5.2.1 Technical validation*

An initial pilot should use an authorized historical event-log extract to characterize baselines, estimate handling-time distributions, and conduct off-policy or shadow-mode evaluation. Independent reviewers should pre-register service, accuracy, cost, and fairness metrics, including acceptable changes in the longest-delay tail. Randomized or staggered prospective evaluation would be preferable where lawful and operationally feasible. The pilot should publish non-sensitive methodology and aggregated findings to enable external scrutiny.

### *5.2.2 Administrative safeguards*

The agency should establish decision rights, model documentation, reproducibility checks, incident escalation, data-access restrictions, and a fully functional manual fallback. Authorized officials should review model recommendations and record overrides. Governance should examine whether queue gains introduce new taxpayer burden or disparate treatment, with reapproval following material policy, data, or algorithm changes.

## 5.3 Limitations

### *5.3.1 Data limitations*

The evidence consists of published, aggregate administrative observations from different dates and workflow categories. No confidential taxpayer records, case-level event log, assigned staff time, agency pilot, or linked outcomes were available. Hence, the study cannot estimate case-level handling times, forecast calibration, taxpayer-level disparities, or financial savings from IPOF. Category counts are not combined into an agency-wide case inventory.

### *5.3.2 Methodological limitations*

The review is a thematic synthesis rather than an exhaustive systematic review, and the descriptive analysis does not establish a causal mechanism. Historical inventories may reflect demand variation, staffing, policy constraints, or recording practices. The proposed optimization model remains unsolved with production inputs; it also excludes real-world details such as specialist availability, multilingual service, evidence dependencies, and appeal-related constraints. No causal before-and-after efficiency effect is inferred.

## 5.4 Future Research

### *5.4.1 Empirical and computational extensions*

Future work should evaluate event-log completeness, perform real-world process conformance analysis, compare predictive approaches using temporal holdouts, and stress-test the assignment problem under volatile demand. More robust optimization methods could account for uncertainty in durations and model errors, while discrete-event simulation could represent specialist pools and dependency-driven rework. Research should compare FIFO, human discretionary triage, predictive-only, optimization-only, and fully integrated workflows.

### *5.4.2 External validity and socioeconomic assessment*

Subsequent studies should examine outcomes across different case types, taxpayer service channels, and income or business categories where lawful and ethically justified. An economic evaluation would report fully burdened implementation costs, administrative labor effects, verified collection outcomes, and changes in taxpayer time costs. Independent replication, statutory review, and taxpayer-centered impact assessment would be prerequisites for credible broader-impact conclusions.

## 5.5 Concluding Synthesis

### *5.5.1 Principal proposition*

The contribution of intelligent tax-process optimization lies not in attaching machine learning to a tax database, but in linking evidence, process state, uncertainty, authorized decisions, and observable service consequences. IPOF provides a technically specified starting point for that integration while preserving explicit legal and human control over consequential action.

### *5.5.2 Final implication*

The next evidentiary step is **validation**, not large-scale claims of transformation. If carefully tested, an integrated predictive-and-workflow architecture could help government agencies pursue service timeliness and operational discipline without sacrificing accountability. Until such tests are completed, the value of this paper is its architecture, its falsifiable hypotheses, and its verifiable baseline-to-requirements analysis.

## References

26 U.S.C. § 6103 (2024). https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title26-section6103

Darmayasa, I. N., & Hardika, N. S. (2024). Core tax administration system: The power and trust dimensions of slippery slope framework tax compliance model. *Cogent Business & Management, 11*(1), Article 2337358. https://doi.org/10.1080/23311975.2024.2337358

Elzayn, H., Smith, E., Hertz, T., Ramesh, A., Fisher, R., Ho, D. E., & Goldin, J. (2023). *Measuring and mitigating racial disparities in tax audits* [Working paper]. Stanford Institute for Economic Policy Research. https://siepr.stanford.edu/publications/working-paper/measuring-and-mitigating-racial-disparities-tax-audits

Hesami, S., Jenkins, H., & Jenkins, G. P. (2024). Digital transformation of tax administration and compliance: A systematic literature review on e-invoicing and prefilled returns. *Digital Government: Research and Practice, 5*(3), Article 18. https://doi.org/10.1145/3643687

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105. https://doi.org/10.2307/25148625

Internal Revenue Service. (2024a). *Filing season statistics for week ending May 10, 2024*. https://www.irs.gov/newsroom/filing-season-statistics-for-week-ending-may-10-2024

Internal Revenue Service. (2024b, April 18). *IRS releases 2023 Data Book describing agency's transformation through statistics* (IR-2024-115). https://content.govdelivery.com/accounts/USIRS/bulletins/397422c

Internal Revenue Service. (2024c, October). *Tax gap projections for tax year 2022* (Publication 5869). https://www.irs.gov/pub/irs-pdf/p5869.pdf

Internal Revenue Service. (2024d, January 16). *The Taxpayer Bill of Rights protects all taxpayers year-round* (IRS Tax Tip 2024-02). https://www.irs.gov/newsroom/the-taxpayer-bill-of-rights-protects-all-taxpayers-year-round

Keen, M., & Slemrod, J. (2017). Optimal tax administration. *Journal of Public Economics, 152*, 133-142. https://doi.org/10.1016/j.jpubeco.2017.04.006

Kirchler, E., Hoelzl, E., & Wahl, I. (2008). Enforced versus voluntary tax compliance: The slippery slope framework. *Journal of Economic Psychology, 29*(2), 210-225. https://doi.org/10.1016/j.joep.2007.05.004

Nagalila, W., Nyombi, A., Sekinobe, M., Ampe, J., & Happy, B. (2024). Fortifying national security: The integration of advanced financial control and cybersecurity measures. *World Journal of Advanced Research and Reviews, 23*(2), 1095-1101. https://doi.org/10.30574/wjarr.2024.23.2.2444

Nyombi, A., Nagalila, W., Happy, B., Sekinobe, M., & Ampe, J. (2024). Enhancing cybersecurity protocols in tax accounting practices: Strategies for protecting taxpayer information. *World Journal of Advanced Research and Reviews, 23*(3), 1788-1798. https://doi.org/10.30574/wjarr.2024.23.3.2838

Object Management Group. (2014). *Business process model and notation (BPMN), version 2.0.2*. https://www.omg.org/spec/BPMN/2.0.2

Organisation for Economic Co-operation and Development. (2024). *Tax administration 2024: Comparative information on OECD and other advanced and emerging economies*. OECD Publishing. https://doi.org/10.1787/2d5fba9c-en

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77. https://doi.org/10.2753/MIS0742-1222240302

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Treasury Inspector General for Tax Administration. (2024, April 30). *Interim results of the 2024 filing season* (Report No. 2024-408-024). https://www.tigta.gov/sites/default/files/reports/2024-05/2024408024fr.pdf

United States Government Accountability Office. (2023). *Artificial intelligence: Agencies have begun implementation but need to complete key requirements* (GAO-24-105980). https://www.gao.gov/products/gao-24-105980

United States Government Accountability Office. (2024). *2023 tax filing: IRS improved customer service, but could further improve processing and evaluate expedited hiring* (GAO-24-106581). https://www.gao.gov/products/gao-24-106581

van der Aalst, W. M. P. (2016). *Process mining: Data science in action* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-49851-4