Risk Management Software and Customs Enforcement in Pakistan: *Analysis Plan*

Michael Carlos Best (Columbia) Faraz Hayat (IGC) Tim Dobermann (IGC)

January 30, 2026

1 Context and study design

1.1 Operational context (summary)

A Goods Declaration form (GD) is filed by traders when expecting a shipment of imports. Under the status quo, the Risk Management System (RMS1) assigns an initial channel (Green/Yellow/Red) and sends GDs to a scheduler that assigns them to assessment officers. Red GDs receive document assessment plus physical examination; Yellow GDs receive document assessment only; Green GDs are cleared with no assessment.

The anonymous third party software (henceforth referred to as RMS2) is a decision-support system available at the assessment stage (document inspection). RMS2 produces a GD-level channel (Green/Yellow) and flags three risks: (i) valuation or unit price mis-declaration (tax recovery risk), (ii) HS code mis-declaration, and (iii) country-of-origin mis-declaration. RMS2 channel is Yellow if any of the three risks is flagged; otherwise Green. RMS2 also produces predicted/counterfactual values (e.g., predicted unit prices and predicted tax loss). A GD is classified as a tax recovery risk if the predicted positive tax collected is more than 2% of the declared taxes and higher than 100,000 PKR.

1.2 Experimental design

The study spans approximately 2.5 months and has two orthogonal design elements.

Officer-level randomized access. Assessment officers are randomly assigned to a treated group (access to RMS2 UI) or a control group (no RMS2 UI). Let $T_o \in \{0, 1\}$ indicate whether officer o is treated.

Forced-red (ground truth) sample. A stratified random sample of GDs that were initially classified Green or Yellow by RMS1 is selected into a "forced-red" sample. For these GDs, the initial RMS1 channel is hidden and the GD's RMS1 channel is forced to Red so it receives maximal scrutiny. These forced-red GDs then enter the scheduler and can be assigned to treated or control officers.

2 Data, units of observation, and construction

2.1 Expected raw files / tables

The analysis uses the following sources:

- **GD-level file**: GD identifiers, timestamps, initial/final RMS1 channel, scores, declared/assessed values and taxes, trader metadata, and item-level fields.
- **Officer file**: officer characteristics (join year, specialization, placement, etc.).
- **Collectorate file**: collectorate descriptors.
- **Officer–GD assignment and event log**: assignment, completion times, assessment IDs, and indicators/timestamps for RMS2 UI access.
- **RMS2 outputs**: RMS2 channel, risk flags, predicted values, predicted tax loss, predicted HS/origin, and a per-GD indicator of UI access.

3 Variable definitions

3.1 Treatment and exposure

- Officer treatment: $T_o = 1$ if officer o is treated.
- Assigned officer for GD g: $o(g)$. Define $T_g \equiv T_{o(g)}$.[1]
- UI access: $Access_g = 1$ if the assigned officer accessed RMS2 UI for GD g; $Access_g = 0$ otherwise.

3.2 Forced-red indicator

Define $FR_g = 1$ for forced-red GDs. Operationally $FR_g = 1$ if (i) RMS1 channel is Green or Yellow and GD is randomly selected to be forced red or (ii) RMS1 channel is Red.

3.3 Valuation/tax outcomes (truth)

Let[2]

$$\begin{aligned} Tax_g^{decl} &= \text{Declared Total Duty and Taxes}, \\ Tax_g^{assess} &= \text{Assessed Total Duty and Taxes}, \\ Recov_g &= \max\{Tax_g^{assess} - Tax_g^{decl},\ 0\}, \\ RecovPct_g &= \frac{Recov_g}{Tax_g^{decl}} \end{aligned}$$

Define "true high-risk" valuation loss using the rule decided by policymaker:

$$High_g = \mathbf{1}\{RecovPct_g > 0.02 \text{ and } Recov_g > 100{,}000 \text{ PKR}\}. \tag{1}$$

Let $Low_g = 1 - High_g$. And we can do write similar definitions for HS Code and Country of Origin.[3]

3.4 RMS2 predictions and RMS benchmark

- $High_g^{R2}$: RMS2 predicted high tax recovery risk under the (2%, 100k) rule.
- $R2Yellow_g$: RMS2 predicted yellow channel
- $High_g^{R1}$: RMS1 predicted high risk. These are GDs with RMS1 channel assignment of either Red or Yellow.
- $NoRisk_g$: No true risk under any set of rules - defined over forced red sample

[1] One concern is that multiple officers can look at the same GD. We can maybe define $T_g = 1$ if $T_o = 1$ for at least one $o \in o(g)$

[2] We can define recovery without the max too to account for reductions

[3] We can define other "true high risk" rules for tax recovery

3.5 Completion time variables

$$\begin{aligned} ClearTime_g &= \text{Fully cleared timestamp} - \text{GD file timestamp}, \\ AssessTime_g &= \text{Assessment completion time} - \text{Assignment time}, \\ QueueTime_g &= \text{Assignment time} - \text{Scheduler entry time}. \end{aligned}$$

3.6 Forced-red sampling weights

Forced-red sampling is stratified by RMS1 channel. Let stratum selection probabilities be p_s and define weights $w_g = 1/p_s$. These can be used to construct weighted means. In all of the analysis below, we will estimate both unweighted effects and weighted effects using these weights to estimate population-level effects.

4 Outcome: System Accuracy or Risk detection rate

The main objective here is to assess the predictive performance of the two RMS systems in identifying high-risk GDs. We focus on the forced-red GDs assigned to control officers to construct a ground-truth sample where the channel is held fixed at Red.

Primary sample Define the ground-truth sample as forced-red GDs assigned to control officers:

$$\mathcal{S}_{GT} = \{g : FR_g = 1 \text{ and } T_g = 0\}. \tag{2}$$

4.1 Binary Outcomes

To study binary outcomes we will focus on the classification of GDs into high-risk versus low-risk using the (2%, 100k) tax recovery rule in equation (1).

We will evaluate the performance of each algorithm using standard tools: The Receiver Operating Curve (ROC) and the confusion matrices. From these we can construct scalar measures of precision such as the Area Under the Curve (AUC), precision, and recall of the algorithms.

To compare the two algorithms directly we will compute 2 tests. First, we will use the McNemar (1947) test to compare the accuracy of the two algorithms in the forced red sample we defined above. Second, we will use the Alpaydm (1999) to estimate the difference in the out-of-sample performance of the two algorithms.

4.2 Continuous Outcomes

We will also evaluate the performance of the two algorithms in predicting continuous outcomes such as the actual tax recovery amount $Recov_g$ and the recovery percentage $RecovPct_g$. For each GD in the ground-truth sample $\mathcal{S}_{GT}$ we will compare the predictive performance of the two algorithms by computing the mean squared error (MSE) between the actual amounts and the predicted amounts. We will also look for complementatities or substitutabilities between the two algorithms by estimating the following regression:

$$Recov_g = \alpha + \beta_1 PredRecov_g^{R1} + \beta_2 PredRecov_g^{R2} + \beta_3 PredRecov_g^{R1} \times PredRecov_g^{R2} + \varepsilon_g, \tag{3}$$

where $PredRecov_g^{R1}$ and $PredRecov_g^{R2}$ are the predicted tax recovery amounts from RMS1 and RMS2 respectively. The sign and significance of β_3 will indicate whether the two algorithms are complements or substitutes in predicting tax recovery amounts.

4.3 Heterogeneity

We will also explore heterogeneity in the performance of the two algorithms across different dimensions such as HS categories, trader types, and ports of entry. We will estimate the performance metrics (AUC, MSE) separately for each subgroup and compare the results.

5 Outcome: Customs clearance time reduction

Here we quantify how often RMS2 would classify a GD as Green (low risk) when RMS would not, and where forced-red indicates low risk. We consider three notions of time savings (A) potential routing improvements and (B) implied time savings using baseline clearance-time differences; and additionally (C) any observed processing-time changes in the pilot.

5.1 Potential rerouting shares

Sample. Use $\mathcal{S}_{GT}$.

Safe downgrades. Define and compute the share of forced-red GDs that RMS2 classifies Green, RMS classifies Yellow/Red, and truth is low risk:

$$SD = \Pr(R2Yellow_g = 0 \text{ and } R1High_g = 1 \text{ and } NoRisk_g = 0). \tag{4}$$

Unnecessary upgrades. Define and compute the share of GDs RMS2 classifies Yellow, RMS classifies Green, and truth is low risk:

$$UU = \Pr(R2Yellow_g = 1 \text{ and } R1High_g = 0 \text{ and } NoRisk_g = 0). \tag{5}$$

5.2 Implied clearance-time savings

Step 1: Estimate baseline clearance times by channel. Using non-forced GDs ($FR_g = 0$) assigned to control officers, estimate the mean (and median) clearance time by RMS channel:

$$\bar{T}_c = \mathbb{E}[ClearTime_g \mid Channel_g = c,\ FR_g = 0], \quad c \in \{Green, Yellow, Red\}. \tag{6}$$

Step 2: Translate rerouting shares into time changes. Under a policy that uses RMS2 to downgrade SD fraction of cases from Yellow/Red to Green, implied time saved per 1,000 GDs is:

$$TS_{1000} \approx 1000 \cdot SD \cdot (\bar{T}_{Yellow} - \bar{T}_{Green}). \tag{7}$$

Similarly, implied added time from unnecessary upgrades is:

$$TA_{1000} \approx 1000 \cdot UU \cdot (\bar{T}_{Yellow} - \bar{T}_{Green}). \tag{8}$$

Report net implied change $TS_{1000} - TC_{1000}$.[4]

5.3 Observed time impacts of RMS2 UI

Even without rerouting at entry, RMS2 may affect processing times through UI usage.

[4] We may add actual distributions of Red, and Yellow to improve this

Regression specifications. Estimate:

$$\log(AssessTime_g) = \alpha + \tau T_g + X_g'\Gamma + \varepsilon_g, \tag{9}$$

and analogously for $QueueTime_g$ and $ClearTime_g$. Cluster at officer (and possibly add week or other time fixed effects). τ measures the difference in completion times for GDs assigned to treated versus control officers, capturing the impact on clearance speed from RMS2 UI usage.

6 Outcome: Information effect of RMS2 UI

We exploit the randomized assignment of RMS2 UI access to officers to estimate the information effect of RMS2 on assessed outcomes. We can do this both overall, and exclusively in our forced-red sample.

Primary outcomes. Our main primary outcomes are at the GD level:

- Tax recovery (levels): $Recov_g$.
- Any recovery: $\mathbf{1}\{Recov_g > 0\}$.
- High-risk detected: $High_g$.

6.1 ITT regression

We can estimate the treatment effects in a difference in differences design by estimating:

$$Y_{got} = \alpha + \beta(T_{o(g)} \times Post_{t(g)}) + \lambda_{o(g)} + \delta_{t(g)} + X_g'\Gamma + \varepsilon_{got}, \tag{10}$$

where:

- Y_got is one of the outcomes listed above.
- X_g are pre-determined GD controls (initial RMS channel $RMS0_g$, GD score, category/type, port/collectorate, trader rank, declared taxes/values).
- $\lambda_{o(g)}$ are officer fixed effects.
- $\delta_{t(g)}$ are week-of-filing or week-of-assignment fixed effects.

β measures the effect of RMS2 UI usage on the outcomes holding the channel fixed at Red. We cluster standard errors at the officer level (treatment assignment level).

6.1.1 Secondary: TOT using actual UI access

Because UI access may be incomplete among treated officers, estimate a complier effect using IV:

$$Access_{got} = \pi_0 + \pi_1(T_{o(g)} \times Post_{t(g)}) + \eta_{o(g)} + \rho_{t(g)} + X_g'\Pi + u_{got}, \tag{11}$$

$$Y_{got} = \alpha + \rho\,\widehat{Access}_{got} + X_g'\Gamma + \lambda_{o(g)} + \delta_{t(g)} + \varepsilon_{got}. \tag{12}$$

ρ is the effect of RMS2 UI use induced by an officer's treatment assignment. Here $Access_{got}$ is defined as the GD's information being accessed by the officer to whom the GD was assigned.

The first stage coefficients are also of interest here to track takeup over time. We can extend the specification to use an event study design to track the dynamics of takeup and treatment effects over time.

7 Officer-level Outcomes

Some of our outcomes of interest are best measured at the officer level over time.

7.1 Officer-week panel construction

For each officer o and week t, using non-forced GDs ($FR_g = 0$) assigned to o in week t, construct:[5]

$$\begin{aligned} N_{ot} &= \#\{g : o(g) = o,\ week(g) = t,\ FR_g = 0\}, \\ Cases_{ot} &= \#\{g : o(g) = o,\ week(g) = t,\ FR_g = 0,\ Recov_g > 0\}, \\ Amt_{ot} &= \sum_{g: o(g)=o,\ week(g)=t,\ FR_g=0} Recov_g. \end{aligned}$$

Also define workload-normalized outcomes:

$$CasesRate_{ot} = Cases_{ot}/N_{ot}, \quad AmtPerGD_{ot} = Amt_{ot}/N_{ot}. \tag{13}$$

7.2 Difference-in-differences

Estimate:

$$Y_{ot} = \alpha + \beta(T_o \times Post_t) + \lambda_o + \delta_t + \varepsilon_{ot}, \tag{14}$$

where Y_{ot} is one of $Cases_{ot}$, Amt_{ot}, $CasesRate_{ot}$, or $AmtPerGD_{ot}$; λ_o are officer fixed effects; δ_t are week fixed effects. We cluster standard errors at officer level. We wil also estimate officer-level TOT analogously to the GD-level outcomes by instrumenting for UI access with random assignment.

References

Alpaydm, Ethem. 1999. Combined 5 x 2 cv F Test for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, **11**(8), 1885–1892.

McNemar, Quinn. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, **12**(2), 153–157.

[5] We may relax analysis to the full sample as opposed to just GDs where $FR_g = 0$