U.S. DEPARTMENT OF AGRICULTURE

Stopping Payment Errors Before They Happen

An Introduction to Risk Assessment With Examples From the Supplemental Nutrition Assistance Program

December 2025

Nondiscrimination Statement

In accordance with federal civil rights law and USDA civil rights regulations and policies, the USDA, its agencies, offices, employees, and institutions participating in or administering USDA programs are prohibited from discriminating based on race, color, national origin, religion, sex, disability, age, marital status, family/parental status, income derived from a public assistance program, political beliefs, or reprisal or retaliation for prior civil rights activity, in any program or activity conducted or funded by USDA (not all bases apply to all programs). Remedies and complaint filing deadlines vary by program or incident.

Persons with disabilities who require alternative means of communication for program information (e.g., Braille, large print, audiotape, American Sign Language, etc.) should contact the state or local agency that administers the program or contact USDA through the Telecommunications Relay Service at 711 (voice and TTY). Additionally, program information may be made available in languages other than English.

To file a program discrimination complaint, complete the USDA Program Discrimination Complaint Form, AD-3027, found online at How to File a Program Discrimination Complaint and at any USDA office or write a letter addressed to USDA and provide in the letter all of the information requested in the form. To request a copy of the complaint form, call (866) 632-9992. Submit your completed form or letter to USDA by:

1. **Mail:** USDA Food and Nutrition Service, 1320 Braddock Place, Room 334, Alexandria, VA 22314; or
2. **Email:** FNSCIVILRIGHTSCOMPLAINTS@usda.gov.

USDA is an equal opportunity provider, employer, and lender.

U.S. DEPARTMENT OF AGRICULTURE

Stopping Payment Errors Before They Happen

An Introduction to Risk Assessment With Examples From the Supplemental Nutrition Assistance Program

December 2025

Contract: GS10F0136X **Order:** 140D04221007

Authors

Jacob Beckerman-Hsu, Kevin Baier, and Betsy Thorn

Submitted to

USDA Food and Nutrition Service
Office of Evaluation, Analysis, and Regulatory Affairs
1320 Braddock Place
Alexandria, VA 22314

Project Officer: Kameron Burt

Submitted by

Westat Insight
1310 North Courthouse Road
Suite 880
Arlington, VA 22201

Project Director: Betsy Thorn

This study was conducted by Westat Insight under Contract No. GS10F0136X/140D04221007 with the U.S. Department of Agriculture's (USDA) Food and Nutrition Service. The findings and conclusions in this report are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Suggested Citation

Beckerman-Hsu, J, Baier, K., & Thorn, B. (2025). *Stopping payment errors before they happen: An introduction to risk assessment with examples from the Supplemental Nutrition Assistance Program*. Westat Insight. U.S. Department of Agriculture, Food and Nutrition Service.

Contents

Tables

What Are Payment Errors?

An improper payment, or payment error, is "any payment that should not have been made or that was made in an incorrect amount under statutory, contractual, administrative, or other legally applicable requirement" (PaymentAccuracy.gov, n.d.). Payment errors can result in a financial loss to the Federal government when a payment is made to the wrong recipient and/or when the payment is higher than it should be (overpayment). Payment errors also happen when payments are lower than they should be (underpayment), which can harm payment recipients. To ensure sound stewardship of taxpayer dollars, the Federal government proactively works to minimize payment errors and recover funds paid out in error. Agencies across the Federal government, including those in the following list, provide updated statistics on payment errors and agency plans to prevent payment errors, which can be found at https://www.paymentaccuracy.gov.[1]

- Department of Homeland Security
- Department of Labor
- Department of Transportation
- Department of Education
- Department of Health and Human Services
- Office of Personnel Management
- Small Business Administration
- Social Security Administration
- Department of Treasury
- Department of Agriculture
- Veterans Administration

What Is a Risk Assessment (RA) Tool, and How Can It Prevent Payment Errors?

Government agencies employ many strategies to prevent payment errors. As advanced analytic platforms and methods become more accessible, more agencies are adding RA tools to their payment error prevention approach. In the context of payment accuracy, RA tools use information about an individual, household, or other payment recipient to identify those most likely to receive payments in an improper amount. RA tools flag those potential over- and underpayments so the agency can dedicate resources to ensuring the payments are correct, ideally fixing any errors before making improper payments.

RA tools can take many forms. An example of a simple RA tool is a checklist of characteristics that put a case at risk for a payment error. To provide examples from the Supplemental Nutrition Assistance Program (SNAP), a checklist might include large household size (e.g., four or more people) and multiple sources of income, which could be household characteristics associated with payment error (i.e., SNAP benefits higher or lower than they should be). Program staff using this RA checklist would flag households with those characteristics as high risk. A more complex RA tool could be programmed within a SNAP State agency's eligibility system and employ

[1] Under the Payment Integrity Information Act of 2019 (Public Law 116–117) and Appendix C to Office of Management and Budget Circular No. A-123, Requirements for Payment Integrity Improvement, Federal agencies must report on improper payments.

machine learning algorithms to automatically assess a case's risk for a payment error and instruct eligibility workers on the correct course of action: immediate approval for benefit issuance or submission for a secondary review.

Creating and implementing successful RA tools requires collaboration between many people, including, as relevant, program administrators, policy experts, IT staff, external contractors, frontline staff, supervisors, and statisticians. When everyone involved has a foundational understanding of RA tools, they can ask the right questions and identify how their expertise can best contribute to a successful RA tool.

This guide is designed to equip local, State, and Federal staff with an understanding of key RA tool concepts that can help them design, use, and oversee the use of RA tools to improve payment accuracy. Although this guide uses examples from SNAP, the principles described here apply to other human service agency settings.

Understanding Risk Assessment in SNAP Payment Accuracy

This guide was developed as a part of a larger study on the use and effectiveness of RA tools in SNAP administration. For more information on the project, see:

Thorn, B., Baier, K., Beckerman-Hsu, J., Giesen, L., Calvin, K., McCall, J., Esposito, J., Campbell, N., & Chance, S. (2025). *Understanding risk assessment in Supplemental Nutrition Assistance Program payment accuracy*. Westat Insight. U.S. Department of Agriculture, Food and Nutrition Service.

Key Questions to Ask When Developing an RA Tool

To develop a successful RA tool, agency staff need to ask the following questions:

- Which cases should the tool flag?
  - Should it flag all cases with a payment error?
  - Should it flag only cases with a payment error above a certain dollar amount?
  - Should it flag some other set of cases?
- Which case characteristics best differentiate between the cases that should and should not be flagged?
- What is the most accurate source of data to use for the RA tool?
- What performance metrics can tell you how well the RA tool is working?
- Once an initial tool has been developed, how well is it working?
- During development, how can you test whether the RA tool works better for some protected classes than others?
- Should local agencies be allowed to customize the tool?

A. Which cases should the tool flag?

The question of which cases an RA tool should flag may have more than one "right" answer; it may vary by agency. Administrators, policy experts, and others should collaborate to determine which cases the tool should flag. It may seem natural for RA tools to provide a yes/no flag indicating whether a case is likely to have a payment error. However, there are other options for RA tool output. For instance, an RA tool could be designed to output the probability that a given case has a payment error. If the agency has the capacity to review all the cases with at least a 50-percent probability of having a payment error, it can do so. If that would be too many cases to review, the agency could instead review only the cases with a higher probability of having an error (e.g., 70 percent). Yet another output to consider is the estimated dollar amount of the payment error on a case; that approach could help agencies prioritize reviewing the cases predicted to have the largest payment errors, thereby having the greatest effect on the payment error rate (PER).[2]

When agencies are first starting to explore their data to find an effective algorithm for flagging cases, it may be worthwhile to try algorithms that flag different types of cases and then compare (1) how well those algorithms work overall (refer to the "Once an initial tool has been developed, how well is it working?" section for more information), (2) how well they work for each protected class (refer to the "During development, how can you test whether the RA tool works better for some protected classes than others?" section for more information), and (3) how challenging it will be to implement a full RA tool using each algorithm (e.g., will it be easier to train staff to use one tool than another?).

B. Which case characteristics best differentiate between the cases that should and should not be flagged?

As part of their current efforts to minimize payment errors, many agencies already identify the types of cases that most often have payment errors. For example, in SNAP, quality control (QC) staff review representative samples of cases to identify those with payment errors and highlight which aspect(s) of the original case files were incorrect. SNAP State agencies can use this information to reduce their payment errors by implementing solutions like improved staff training on specific types of cases.

[2] In SNAP, larger households have higher maximum benefit levels. Focusing on cases with larger predicted payment errors would likely result in reviewing more cases from larger households, which are the households more likely to have children. Under the U.S. Department of Agriculture (USDA) non-discrimination statement (USDA, n.d.), USDA programs are prohibited from discriminating on the basis of family/parental status; agencies should be cautious when considering this approach. For more detail on this topic, refer to the "During development, how can you test whether the RA tool works better for some protected classes than others?" section.

The type of data analysis required to develop an RA tool is similar to work that agencies already do, but it does have an important difference. In addition to focusing on the types of cases most likely to have payment errors, agencies also need to identify the types of cases least likely to have payment errors when developing an RA tool. In other words, effective RA tools need to be able to differentiate between cases with and without payment errors; it is not sufficient for them to identify all the cases with payment errors without regard for accurately determining which cases do not have payment errors.

To provide a concrete example of why it is important to examine cases without payment errors, consider a SNAP State agency whose QC data indicate that cases with payment errors tended to be in larger households. The issue is that among all large-household cases, there are likely many without payment errors, so the State agency would waste time reviewing all cases for large households. Household size alone, in this example, is not sufficient to accurately differentiate between cases with and without payment errors. If the State agency could identify the characteristics of cases that rarely have payment errors, the State agency could add that information to its RA tool to improve it. For example, the State agency's QC data might reveal that cases with only unearned income rarely have payment errors. As such, the State agency could modify the tool to flag large households with earned income and skip those with only unearned income. Making sure that an RA tool can (1) accurately flag cases with payment errors and (2) accurately skip those without payment errors ensures staff focus only on the cases most likely to need correction, thereby making efficient use of time.

Not everyone needs to know how to do the data analysis, but everyone needs to understand the results

A statistician or other data expert will be able to conduct specialized analyses that can improve the final RA tool, but developing a good RA tool is not simply a math problem to be solved. Good RA tools also require policy knowledge and frontline program experience. It is crucial for everyone to understand the results of the data analysis and provide feedback based on their expertise. For example, State agency staff might be able to quickly spot results that conflict with their knowledge of the types of cases most and least likely to have payment errors. Raising these concerns can help ensure accurate data analysis and final RA tools that reflect a proper understanding of SNAP policy and practice.

1. Consider Using Multivariable Approaches

In the previous example, the State agency is conducting bivariate analyses; it is looking at one case characteristic at a time (variable 1) and assessing how strongly each is associated with payment error (variable 2). This is an important step regardless of the final methods used. However, agencies could consider additional analyses to develop their final RA tools. Statistical and machine learning models can use many variables simultaneously to identify the types of cases most likely to have payment errors. This approach can yield better performance than could be achieved by building an RA tool based on bivariate analyses alone.

Although multivariable methods hold great promise for developing accurate RA tools, this suggestion should not be misunderstood as a recommendation to use the most advanced and complicated methods available. In fact, many would recommend against that approach because those methods do not necessarily yield better results and can be so complicated that the results are difficult to fully understand.[3] Using methods that can readily be interpreted and understood, such as logistic regression models or decision trees (Rudin et al., 2022), can help RA tool developers refine the tools and make it easier to communicate how the final tool works, which is important for transparency (Thorn et al., 2025).

C. What is the most accurate source of data to use for the RA tool?

Agencies gather information from a variety of sources to make payment decisions. It is critical to ensure that all data analyzed to develop an RA tool (and the data the RA tool will use going forward to flag cases at risk for a payment error) are accurate. For example, in SNAP, State agencies collect written applications, interview households, and conduct data matches. If a State agency knows a particular piece of information collected from written applications is frequently corrected during interviews, it would be advisable not to use that information from the written application for the RA tool. Similarly, agencies should carefully review any data collected by third parties to ensure their accuracy before using them in RA tools.[4]

In addition to ensuring the data used are accurate, it is best practice to design RA tools with feedback channels that allow people to correct any inaccurate information that could lead to potentially harmful results (Schwartz et al., 2022). These channels already exist in fields such as consumer finance to ensure individuals can identify and correct mistakes in their financial records, but they may be less feasible in human services. In SNAP, State agencies must disclose to households the information used to determine their eligibility and the amount of their benefits, and households can contact their local office to correct any errors or omissions they see on their case files. For other programs, if possible, providing individuals with a straightforward way to check and correct data gives them more agency over the use of their information and ensures an RA tool accesses the most current and accurate information.

[3] See Rudin (2019), for example.

[4] Data from data brokers are often contested as inaccurate. For example, in a study of Facebook users, Venkatadri et al. (2019) found that Facebook users described more than 40 percent of data points about them from data brokers as "not at all accurate."

D. What performance metrics can tell you how well the RA tool is working?

The first step to understanding how well an RA tool works is to identify what positive and negative effects it could have for the agency and its beneficiaries. Then, you can decide on the best way to measure those effects. Choosing appropriate tool performance metrics is critical because (1) tool developers will seek to optimize those performance metrics when creating the tool and (2) those metrics will be used to monitor the tool after it is implemented to ensure it works as well as anticipated.

Example RA Tool Performance Metrics

Accuracy: the number of cases classified correctly (i.e., cases with payment errors flagged as high risk plus cases without payment errors not flagged as high risk) divided by the total number of cases

Sensitivity: the proportion of cases with a payment error that are correctly flagged as high risk

Specificity: the proportion of cases without a payment error that are correctly not flagged as high risk

There is no universal single metric or set of metrics best able to measure the performance of all RA tools. Some general metrics like accuracy (see the performance metrics textbox on the right) can be useful, but other specialized performance metrics may provide more meaningful insight into specific positive and negative effects of an RA tool.[5] To find the best way to measure tool performance, the study team recommends everyone involved in developing an RA tool first work together to identify the potential effects of the tool. Then, the RA tool developers can identify metrics that best capture those effects.

To provide an example of how to first identify the potential effects of a tool and then choose performance metrics that capture those effects, consider a SNAP RA tool designed to flag cases likely to have payment errors. The SNAP State agency would collect a written application and conduct an interview with households applying to the program, and a caseworker would make an initial eligibility and benefit level determination according to standard procedure. Then, a caseworker could apply the RA tool to the case to see whether it flags the case as high risk. If so, a supervisor could review all the case information to ensure the eligibility and benefit level determination is correct before benefits are issued.

To determine the best ways to measure the performance of this SNAP RA tool, the first step is to think through the potential effects of the tool. The RA tool can have four results, each with its own unique potential effects on households and the agency: true positives, false positives, false negatives, and true negatives. These RA tool results are typically laid out in a table called a confusion matrix[6] (table 1). Not only are confusion matrices helpful for thinking through the

[5] See Hand (2012) and Canbek et al. (2022) for examples.

[6] The term "confusion matrix" refers to how this table shows where the RA tool is incorrectly classifying, or "confusing," some cases (Yang & Berdine, 2024).

potential effects of all results of an RA tool, but they also serve as a foundation for measuring tool performance.

Table 1. Sample confusion matrix with definitions for SNAP RA tools

RA tool result	Actual payment error status	
	Payment error **Actual positives: cases with payment errors**	**No payment error** **Actual negatives: cases without payment errors**
High risk **Predicted positives:** cases predicted to have payment errors	**1. True positives:** cases with payment errors classified as high risk	**2. False positives:** cases without payment errors classified as high risk
Low risk **Predicted negatives:** cases predicted not to have payment errors	**3. False negatives:** cases with payment errors classified as low risk	**4. True negatives:** cases without payment errors classified as low risk

Note: RA = risk assessment

Table 2 describes the potential effects of the example SNAP RA tool on households and the State agency for the four outcomes. Note that for cases with payment errors (i.e., true positives and false negatives), the effects differ between overpayments and underpayments. The effects described in table 2 are specific to the example RA tool described previously. Effects for RA tools implemented in a different fashion or implemented in other programs with different policies may have different effects.

Table 2. Effects of an example SNAP RA tool on households and the SNAP State agency

Outcome	Effect on household	Effect on State agency
1. True positive (underpayment)	Household benefits increase relative to the initial underpayment.	The State agency catches and corrects what would have been a payment error.
1. True positive (overpayment)	Household benefits decrease relative to the initial overpayment, but the household will not be subject to claims, as it would have been had the case not been flagged and the agency identified the overpayment later (e.g., through a QC review).	
2. False positive	No effect on the household. The household's benefits will not change, and the household may not even be aware the State agency flagged and reviewed its case.	The State agency will invest staff time in reviewing the case, only to find there was nothing wrong with it.
3. False negative (underpayment)	The household receives lower benefits than it should. If the State agency reviews the case later (e.g., as part of SNAP QC), the agency will restore benefits for up to 12 months (7 CFR 273.17(b)). If the State agency does not review the case and identify the underpayment, the household will never receive the full benefits for which it qualified.	The case has a payment error. If it is sampled for QC, it will contribute to the State agency's PER.
3. False negative (overpayment)	If the State agency never identifies the overpayment, the household will receive higher benefits. If the State agency identifies the overpayment later (e.g., through a QC review), the agency will establish a claim on the household (7 CFR 273.18(a)(2)). Paying claims may impose hardship on SNAP participants because they have low incomes and frequently report food insecurity despite receiving SNAP benefits (Brady et al., 2023).	
4. True negative	No effect on the household or State agency. The State agency takes no further action on this case. If the State agency had taken further action on this case, nothing would have changed about the household's benefits.	

Note: PER = payment error rate; QC = quality control

In this example, from the household's perspective, the RA tool has no effect when there is no payment error (i.e., false positives and true negatives). However, there are meaningful effects for households that would have had payment errors had the RA tool not flagged their cases (i.e., true positives). There are also meaningful effects for any households not flagged that had payment errors (i.e., false negatives); they will be issued benefits in the incorrect amount. In this example, the most relevant metrics of RA tool performance for households should be based on the number of true positives and false negatives. A common measure that meets these criteria is ***sensitivity***, which is the proportion of cases with a payment error the RA tool flags as high risk. The higher the sensitivity, the more households will have the correct benefit amounts.

From the State agency's perspective, true positives are the value of having an RA tool: The greater the number of true positives, the more payment errors the State agency can catch and correct. False negatives are also important to State agencies because they represent missed opportunities. Had the RA tool been able to flag them, the State agency would have caught more payment errors. So, just as sensitivity is a valuable metric of RA tool performance for households, it also captures vital information about the RA tool from the State agency perspective.

One more RA tool outcome is critical for State agencies: false positives. False positives are a cost of having an RA tool because the greater the number of false positives, the more time the State agency spends reviewing cases that already had the correct benefit amount. As such, State agencies also have an interest in metrics of RA tool performance that incorporate information about the number of false positives. A common measure that serves this purpose is ***specificity***, which is the proportion of cases without a payment error that the RA tool classifies as low risk. The higher the specificity, the less time the State agency will spend reviewing cases with no payment errors.

In summary, the study team recommends a two-step process to determine the best way to measure how well an RA tool is working. First, agencies need to rely on their program and policy expertise to make their version of table 2 specific to the RA tool they are considering. Note that table 2 is merely an example; if a SNAP State agency implemented an RA tool in a way that resulted in different application, recertification, and/or reporting requirements for households flagged as high risk (e.g., a longer interview, more required verifications), the effect of the tool on the households would be different from what the table shows. Other differences between the example RA tool discussed here and a particular RA tool an agency is considering could also change table 2.

Second, based on table 2, State agencies can choose which RA tool performance metrics best measure how well the RA tool is working. Note that the study team encourages anyone involved in developing and overseeing the use of an RA tool to use the approach described here—not the conclusions drawn here. Sensitivity and specificity may not always be the best performance metrics for all RA tools. Weighing the pros and cons of different performance metrics requires policy and programmatic expertise as well as expertise in statistics, machine learning, or related fields. In practice, this step will require collaboration among many people.

E. Once an initial tool has been developed, how well is it working?

Before implementing any RA tool, it is best practice to test how well it will perform (The White House, 2022). Once an agency has developed an initial RA tool algorithm to flag cases, it should apply the initial tool to a set of cases to develop confusion matrices (see tables 3 and 4 for examples) and calculate the performance metrics selected to measure how well the tool works (see previous section for a discussion of selecting performance metrics). The ideal source of data to generate a confusion matrix is one that (1) includes a sample of cases representative of the population the RA tool will be used on and (2) has been reviewed by experts to determine which cases had payment errors. For most SNAP State agencies, QC data are likely the best option.

Assessing performance is critical for all SNAP State agencies, not only those with RA tools

All SNAP State agencies do at least some additional review of some of their cases (e.g., quality assurance reviews). If SNAP State agencies assess the performance of their method for selecting cases, they can refine their approach to maximize the benefits of these reviews.

When reviewing RA tool performance, the agency should check the cases the RA tool classified incorrectly (i.e., false positives and false negatives). If the agency can identify why these errors occurred, it can revise the RA tool algorithm to improve performance. Agencies should expect to iterate through many versions of an RA tool during the development stage to optimize the performance of the final tool selected for implementation.

Table 3. Example confusion matrix for a proposed SNAP RA tool

RA tool result	SNAP QC result	
	Payment error (N) Actual positives: cases with payment errors	No payment error (N) Actual negatives: cases without payment errors
High risk **Predicted positives:** cases predicted to have payment errors	8	4
Low risk **Predicted negatives:** cases predicted not to have payment errors	42	46

Note: RA = risk assessment; QC = quality control

F. During development, how can you test whether the RA tool works better for some protected classes than others?

As agencies develop RA tools, they need to be aware of the legal implications of their tools. Tools that were not intended or designed to adversely affect any group of people may still do so, which could be considered disparate impact (Polek & Sandy, 2023). While the legal standards in disparate impact lawsuits continue to evolve (Schwartz et al., 2022),[7] core to disparate impact claims is demonstrating that individuals in a protected class have been disproportionately harmed. It is therefore important for agencies to proactively test how well their tools work for different protected classes, which can give them the opportunity to adjust their tools as needed to avoid disparate impact. For instance, family/parental status is a protected class listed in the USDA non-discrimination statement (USDA, n.d.). To test for the potential for disparate impact by family/parental status, the State agency from the previous example (see tables 2 and 3) would want to calculate separate confusion matrices for households with and without children (table 4).

Table 4. Example confusion matrix for a proposed RA tool stratified by the presence of children in the household

RA tool result	SNAP QC result			
	Households with children		Households without children	
	Payment error (N) Actual positives: cases with payment errors	No payment error (N) Actual negatives: cases without payment errors	Payment error (N) Actual positives: cases with payment errors	No payment error (N) Actual negatives: cases without payment errors
High risk Predicted positives: cases predicted to have payment errors	8	1	0	3
Low risk Predicted negatives: cases predicted not to have payment errors	30	23	12	23

Note: RA = risk assessment; QC = quality control

The example in table 4 illustrates two important points. First, it shows how a tool might perform better for some groups than others; all eight cases with payment errors correctly flagged by the tool (i.e., true positives) were among households with children. Second, it shows how the number of cases in each cell of the confusion matrix can become small when dividing the sample by

[7] Similar to the work of Schwartz et al. (2022), this document is not meant to serve as legal guidance, but merely a reminder that RA tools have important legal implications. Agencies should work closely with legal experts to understand the current law in this area.

protected class status. These small samples can make it hard to detect differences in tool performance across protected classes. It is critical to note that just because a statistical test shows no significant difference across protected classes does not mean there is truly no difference; the sample may just be too small to detect a difference.

The study team recommends consulting legal experts and a statistician to help conduct and interpret tests for potential disparate impact.

G. Should local agencies be allowed to customize the tool?

Currently, most SNAP State agencies with an RA tool allow local agencies to customize the tool. When a local agency changes anything about the algorithm used to flag cases, a case flagged by one local agency may not necessarily be flagged by another. Potential benefits to this approach include local agencies having tools that better meet their specific situations, possibly yielding better tool performance and leading to greater payment accuracy. Potential drawbacks of this approach include the increased effort required to make data-driven customizations to a tool at the local agency level and to test and monitor multiple versions of an RA tool rather than a single version used by the entire State agency. Local agency customization may also create unintended consequences if some customized tools have higher performance than others, yielding differing payment accuracy for SNAP populations in different parts of the State. State agencies should weigh these possible benefits against the possible costs of local agency customization.

Key Considerations for Monitoring and Evaluating an RA Tool

Ongoing monitoring and evaluation are critical for agencies using RA tools. Key questions include the following:

- Once implemented, how can you tell if RA tools are performing as well as anticipated?
- How can you assess the overall effects of RA tools?

A. Once implemented, how can you tell if RA tools are performing as well as anticipated?

To get real-time feedback on how well an RA tool is working, agencies can track the number of cases the RA tool flagged and how many of those flagged cases initially had payment errors that agency staff corrected upon review. To use the example from table 3, eight of the 12 flagged cases had payment errors. The agency could therefore expect approximately two-thirds of its flagged cases to have payment errors its staff can correct. When the agency sees a decrease in the proportion of flagged cases that require correction, it is likely time to revisit the tool's

algorithm and make updates as needed. Tracking which flagged cases have errors corrected during a review serves a second benefit: helping agencies identify what types of errors are happening in real time so the agency can provide training, make policy updates, or take other actions to improve accuracy.

In addition to real-time tracking, it is best practice to periodically generate updated confusion matrices to provide a full assessment of RA tool performance and update the tool as needed to ensure continued high performance (The White House, 2022). For SNAP State agencies, just as was true at the tool development stage, QC data will most likely be the best data source. However, SNAP State agencies need to add at least two data points generated during the implementation of the RA tool:[8]

1. The case ID or another identifier for all cases flagged as high risk
2. The original benefit amount determined for all flagged cases (i.e., before any review resulting from the case getting flagged)

The addition of these two data points allows data users to distinguish between (1) cases that initially had an incorrect eligibility or benefit determination that was corrected after the case was flagged and (2) cases that always had a correct eligibility and benefit determination.[9] Once these data points are merged with the QC data, the State agency can calculate the difference between the original benefit amount and the correct benefit amount as determined during the QC review to determine whether the case would have had a payment error had the original benefit amount been issued. The agency can then compare which cases the RA tool flagged as high risk against which cases would have had a payment error in the absence of the RA tool, generating a confusion matrix that accurately shows when the RA tool made correct and incorrect classifications.

Agencies can use the updated confusion matrices to check the performance of their tools and test how changes to their tools would affect performance. Any change in the number or type of errors—because of changes in the population of program beneficiaries, policy changes, system changes, staff turnover, or other reasons—can affect RA tool performance. As a result, when enough time has passed for a large portion of the caseload to have churned or after an important change in policy or practice, it is advisable to generate new confusion matrices and conduct a full assessment of RA tool performance.

[8] Other data points can also be valuable for RA tool monitoring, but the utility of other data may depend on the specifics of the State agency and its RA tool.

[9] Data sources used to monitor RA tools in other programs may require similar additions to be useful for monitoring.

B. How can you assess the overall effects of RA tools?

In addition to monitoring tool performance, it may be valuable to assess changes in the PER attributable to RA tool implementation. Agencies should begin by reviewing PERs before and after they introduced their tools. To estimate the impact, agencies can consider using quasi-experimental methods, such as the difference-in-differences model used by Thorn et al. (2025).[10] Accurately estimating the impact of an RA tool would require staff time, technical expertise, and rigorous data collection on a variety of factors aside from RA tool implementation that could affect PERs (e.g., changes in the population of program beneficiaries, program policy changes).

Agencies should also document (or precisely estimate) all costs related to developing, testing, implementing, monitoring, and evaluating their RA tools. They should then compare these costs against the dollar amounts of payment errors corrected and any possible liabilities they might otherwise have needed to pay for high PERs. This type of evaluation can help agencies determine whether they are generating a positive return on investment from their RA tool. Agencies can consider tool performance, changes in the PER, and tool cost-effectiveness to create a holistic understanding of the tool's strengths and areas for improvement, informing next steps to improve payment accuracy most effectively.

[10] Agencies could also use these methods for other outcomes of interest, such as the proportion of cases that have payment errors.

References

Brady, P. J., Harnack, L., Widome, R., Berry, K. M., & Valluri, S. (2023). Food security among SNAP participants 2019 to 2021: A cross-sectional analysis of current population survey food security supplement data. *Journal of Nutritional Science, 12,* Article e45. https://doi.org/10.1017/jns.2023.32

Canbek, G., Taskaya Temizel, T., & Sagiroglu, S. (2022). PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. *SN Computer Science*, *4*, 13. https://doi.org/10.1007/s42979-022-01409-1

Hand, D. J. (2012). Assessing the performance of classification methods. *International Statistical Review*, *80*(3), 400–414. https://doi.org/10.1111/j.1751-5823.2012.00183.x

PaymentAccuracy.gov. (n.d.). *About Paymentaccuracy.gov*. https://www.paymentaccuracy.gov/about-payment-accuracy/

Polek, C. & Sandy, S. (2023). The disparate impact of artificial intelligence and machine learning. *Colorado Technology Law Journal, 21*(1), 85–108. https://ctlj.colorado.edu/?p=958

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, *16*, 1–85. https://doi.org/10.48550/arXiv.2103.11251

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (NIST Special Publication 1270). U.S. Department of Commerce, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

Thorn, B., Baier, K., Beckerman-Hsu, J., Giesen, L., Calvin, K., McCall, J., Esposito, J., Campbell, N., & Chance, S. (2025). *Understanding risk assessment in Supplemental Nutrition Assistance Program payment accuracy*. Westat Insight. U.S. Department of Agriculture, Food and Nutrition Service.

U.S. Department of Agriculture. (n.d.). *Non-discrimination statement*. https://www.usda.gov/non-discrimination-statement

Venkatadri, G., Sapiezynski, P., Redmiles, E. M., Mislove, A., Goga, O., Mazurek, M., & Gummadi, K. P. (2019, May 13–17). *Auditing offline data brokers via Facebook's advertising platform* [Conference presentation]. Association for Computing Machinery WWW '19: The World Wide Web Conference, New York, NY, United States. https://doi.org/10.1145/3308558.3313666

The White House. (2022). *Blueprint for an AI Bill of Rights*. https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/

Yang, S., & Berdine, G. (2024). Confusion matrix. *The Southwest Respiratory and Critical Care Chronicles, 12*(53), 75–79. https://doi.org/10.12746/swrccc.v12i53.1391