compliance assist

Step 5: Analyse the Results

Purpose and Importance

With the test execution complete, the next step is to analyse the outcomes to evaluate the system's performance and identify any issues. This is arguably the most important part of the process. You interpret the raw data to determine whether your screening system is working effectively and where it needs improvement. Regulators expect firms not just to run tests, but to "identify issues" and perform a root cause analysis of any deficiencies. In other words, find out why something went wrong and how severe it is. During this analysis, you will categorise results into true positives, false positives, true negatives, and false negatives and see how they align with expectations. For any mismatches (especially misses of sanctioned names or heavy false positives), you'll investigate why the system behaved that way. The findings here will directly inform your tuning in the next step. The analysis will produce metrics and documentation that demonstrate to management and regulators your system's effectiveness (or gaps). For example, if you discover the system missed 2 out of 50 critical names, that's a problem to fix and evidence of why a change is needed. Conversely, if you find false positives are 80% of alerts, that quantifies an efficiency issue. Thorough analysis transforms test data into actionable insights and justifications for any system adjustments.

Key Activities and Decisions

In analysing test results:

- **Classify each test outcome vs expectation**: Go through each test case one by one using the expected outcome list and actual results from previous steps:
    - If a test name was expected to alert and did alert on the correct target, mark that as a true positive.
    - If a test name was expected to alert but did not alert, that's a false negative. List all of these prominently as they are high priority.
    - If a test name was expected not to alert and indeed did not, that's a true negative.
    - If a test name was expected not to alert but did trigger an alert, that's a false positive. Create groups or a table: e.g. "False Negatives: [list names]", "False Positives: [list names + what they matched erroneously]". Count them.
- **Identify patterns and root causes**: For each group of interest:
    - **Analyse False Negatives (misses)**: Are there commonalities? (e.g. all involve a certain type of spelling variation or all have a particular special character). Investigate each case. Why might the system have missed it? Check logs or alert scores if available. Perhaps the match score was just below threshold or maybe the variant wasn't on the list (could indicate a list coverage issue, like missing alias). Maybe the name was not split as expected and the system didn't catch it. Write down a hypothesis for each miss. Example: "Missed 'Mohamad Al Aziz', likely because list has 'Mohammed Al-Aziz' and one 'm' difference dropped score below threshold." or "Missed 'ABC Ltd', system possibly didn't match abbreviation vs full name." These hypotheses guide what to adjust.
    - **Analyse False Positives**: Look at what each false positive matched to. Often, you'll see the system matched a clean name to something on a list that isn't actually the same person/entity (because of

similarity). E.g. "John Major (clean) matched to John Major (sanctioned PEP)". Why didn't the system distinguish them? Possibly because no secondary identifier (DOB) was considered. Or maybe "XYZ Trading" matched "XYZ Trading Company" which is banned, maybe too partial matching. Identify patterns, are a lot of false positives due to common names? Or certain titles (like "General" flagging things)? Also, note if any false positive is a really insignificant issue versus a big one (some false positives might be very easy to clear, others might be truly burdensome). Document reasons where you can: "Many False Positives caused by single-name matches without context, system flags 'Iran' in company name even if not related to country, etc."

  - **For True Positives**: Check if any inefficiency exists, did any real match generate multiple alerts? Sometimes one input can match multiple listings (e.g. "Ali Mohammad" might match 5 different people on the list). If that happened, note it, might be something to refine (like finding ways to reduce duplicate alerts for same name). Generally, true positives are good, they show system capability.
  - **For True Negatives**: Not much to do here except maybe spot-check if any test might have slipped through incorrectly (unlikely if you properly labelled expected outcomes).

- **Quantify overall performance**: Calculate key metrics:
  - **False Negative rate**: e.g. "System missed 2 out of 50 expected hits = 4% false negative rate." Ideally this is 0, if not, that's your critical gap.
  - **False Positive rate**: e.g. "Out of 50 clean names, 10 triggered alerts = 20% false positive rate" (or you could express as precision: "Out of 70 alerts (including expected ones), X were false = Y% false positive among alerts"). Regulators look for an understanding of how much "noise" the system produces.
  - **True Positive detection rate**: e.g. "96% of test sanctions were detected" and Precision (if useful): "Out of all alerts our test produced, 85% were correct matches."
  - These numbers give a baseline to improve upon and may be used in reporting to show progress after tuning.
- **Assess compliance with objectives**: Refer back to the objectives defined previously. How do results stack up?
  - If an objective was "no misses", any false negative means that objective isn't met yet (and becomes a must-fix).
  - If an objective was "reduce false positives by 20%" and this is a baseline test, note the current false positive rate. Improvement comes after tuning.
  - If an objective was list coverage, ensure all lists had representation and see if any list's cases failed disproportionately.
  - Basically, identify which objectives are not yet satisfied and will need addressing in the next step.
- **Prioritise issues**: Not all issues are equal. A false negative on a critical sanctions name is a critical severity. Regulators consider missing a hit a serious compliance failure. Those go to the top of the fix list. False positives are usually moderate, they're efficiency issues, sometimes high if they're overwhelming. Within false positives, some might be tolerable (e.g. if it's a rare name and easy to clear) versus others that might

generate daily work. Rank issues (Critical, High, Medium, Low) based on impact and frequency. This priority will help you decide what changes to make first or how to trade-off if needed.

- **Document findings thoroughly**: As you analyse, document everything in an analysis report or matrix. For each issue (especially each false negative and significant false positive pattern), write a brief description of the suspected cause and potential solution ideas if apparent. For example:

Issue	Example Case(s)	Suspected Cause	Severity
Missed hyphenated name	"Abdul Rahman" (no match to "Abdul-Rahman" on list)	Hyphen handling – system treats "Abdul Rahman" as two names and maybe needed exact match.	High (miss)
High FP on common surnames	"Lee Teng" (matched sanctioned "Lee Teng-hui")	Threshold too low causing partial match on common surname "Lee".	Medium (noise)
…	…	…	…

- This will serve as a blueprint for changes in the next step. Also, compile overall stats (perhaps in a summary section). This analysis document, with all identified weaknesses and their root causes, will be part of your documentation package to demonstrate due diligence.
- **Consider external validation if needed**: If some results are perplexing (e.g. you expected a hit but got none and you can't tell why), you might contact the vendor or consult with a sanctions tech expert. Sometimes there are hidden logic features (like "soundex" or alias tables) that you might not be aware of. Getting clarification can help pinpoint the exact cause. Use this sparingly as needed.
- **Continuous improvement mindset**: Think beyond just pass/fail of this test. What do results say about the program? For instance, if there were misses because of missing aliases, maybe your list feed isn't including them, a process issue to fix beyond just the tool. Or if many false positives are due to data quality (like names input in all caps with no spaces), maybe an upstream data cleanup is needed. Note any such broader insights.

Practical Tips and Examples:

- **Focus on the "why"**: For every false negative, ask "Why did the system not alert?" If you have access to a matching score or log, use it. If "Mohamad Al Aziz" scored 78 and threshold is 80, that's a clear clue (score slightly under threshold). If there's no trace, you might re-run that one name in a sandbox with verbose logging or incrementally adjust it to see what the system catches (e.g. does adding a missing letter make it catch? That implies spelling sensitivity issue).
- **Leverage system's explanation features**: Some systems provide an explanation for matches or the ability to test scenarios in a diagnostic mode. If available, use that for tricky cases. E.g. input two names in a testing console to see the similarity score. That can confirm, "Ah, it saw these as 70% similar which is below 80% cutoff."

- **Example root cause identification**: Suppose 3 sanctioned entity names with "Ltd" at the end were missed (like "ABC Ltd"). If you find in the configuration that "Ltd" is on a stop-word list (words to ignore), the system might have ignored "Ltd" and only matched "ABC" which was too generic to flag. Root cause: stop-word removal weakening match for short names. Solution might be to adjust that logic.
- **Document success too**: Note the things the system did well, maybe it caught every alias of a certain notorious individual, showing strong fuzzy logic in that area. This helps you not "over-fix" something that isn't broken. It also is positive evidence. E.g. "System successfully matched all 10 variations of 'Alexander' -> indicates robust phonetic matching for Cyrillic names."
- **Communicate interim findings**: If analysis reveals any immediate critical issue (like "we missed a major OFAC name completely"), escalate that to compliance management right away, even before full tuning. They might need to enact a quick manual control (e.g. temporarily add that name as a manual watchlist or increase alert scrutiny on it) until a system fix is applied. Early heads-up on serious problems shows proactiveness.
- **Peer review**: Have at least one other team member review your analysis or replicate parts. They might catch something you missed (like "Oh, I see all these false positives are actually matching this minor list we might not need, maybe turn that list off?"). Collaboration can enrich the root cause analysis.
- **Check against objectives again**: For example, if an objective was "screen all fields properly," and one test was to see if an alias in the "aka" field is caught (and it wasn't), mark that objective as not met and note specifically "system not screening alias field – configuration gap." This keeps a clear line of sight from objectives -> findings -> remediation.
- **Data visualisation**: Sometimes making a simple chart can reveal patterns. For instance, a bar chart of "match score for each variant tested" might show all misses clustered just below the threshold. Or a pie chart of alert outcomes (how many correct vs incorrect) could be telling. This is optional but can be useful if you have a lot of data.
- **Remember external expectations**: For example, EBA expects "transparency around system performance". Ensure your analysis quantifies performance. FCA expects senior management MI on sanctions, the metrics and understanding you develop here can feed into that MI.

Common Pitfalls:

- **Missing the forest for the trees**: Don't get so bogged down in one weird case that you lose sight of bigger trends. If one test name failed due to an extremely odd reason that wouldn't occur elsewhere, note it but focus on broader patterns that affect many cases.
- **Jumping to conclusions without evidence**: Ensure your root cause hypotheses are backed by evidence or logical reasoning. If uncertain, mark it as something to confirm (maybe test a mini scenario to verify). For instance, don't assume "the list must be missing alias" unless you verify the alias isn't in the data feed.
- **Overlooking data errors**: If a supposed miss or false positive seems inexplicable, double-check that the test data and expected outcome were correct. Perhaps the "miss" was because the test name itself was wrong (typo in your expected list), meaning the system might be fine. Always rule out test setup error.

- **Tunnel vision on false positives only**: Sometimes teams focus heavily on reducing false positives (because it's a pain point) and might rationalise away a false negative ("oh, that variant was too unlikely"). Remember that from a compliance perspective, false negatives are much more critical. Treat each miss as a serious issue to address unless you can firmly justify its outside any reasonable scenario (rarely the case if you included it in testing).
- **Not documenting rationale**: Months later, you'll want to remember why you decided a certain fix was needed. If you just write "Adjusted fuzzy threshold from 80 to 70" without context, an auditor might ask "why?". Document in analysis: "We lowered threshold to 70 because analysis showed multiple misses in 75-79 score range, meaning real sanctioned names were being lost." This links evidence to action.
- **Ignoring minor improvements**: If everything mostly works, don't just celebrate and stop. There's almost always some tweak to optimise. Conversely, if you got perfect results (0 misses, negligible false positives), reflect on if your test was rigorous enough, perfect may raise scepticism unless double-confirmed.
- **Blaming the tool for data issues and vice versa**: Distinguish system config issues versus data issues versus list issues. For example, if the system missed because the alias isn't on the list, the shortcoming is in sanctions data management (maybe need a better list source or manual addition), not the matching algorithm. Fixing that might involve a different approach than tuning. Your analysis should separate these: e.g. "Issue: Missing alias – this is a data coverage problem, not algorithm; solution might be enhancing data feed."
- **Neglecting to consider frequency/impact**: Some false positives you found might theoretically occur, but if in reality, they'd be extremely rare (like a very unusual name similarity), they might be lower priority to solve. Conversely, a false positive on "Common Name" will happen daily. Analysis should weigh how often an issue would crop up in real operations, which might require input from knowledge of your customer base.
- **Not involving compliance officers in interpretation**: If you're an analyst or IT person, ensure a compliance officer or sanctions SME looks at the false negatives especially. They can gauge the regulatory severity. They might say "Missing that name would be a breach of OFAC requirements – unacceptable," reinforcing that fix's priority.

compliance assist

Checklist – Analyse the Results:

- [ ] All test cases evaluated: Every test case's actual outcome has been compared against the expected outcome and classified (TP, TN, FP, FN). No case is left unexamined.
- [ ] False negatives (misses) identified: Any instance where a sanctioned name did not trigger an alert is listed, with details of the name and which sanction it corresponded to.
- [ ] False positives identified: Any instance where a non-sanctioned name triggered an alert is listed, along with what it matched to (which list entry or why it was flagged).
- [ ] Patterns analysed: Common causes or themes for misses and false alerts are determined (e.g. specific name patterns, certain list, or field issues).
- [ ] Root causes hypothesised: For each distinct issue or pattern, a likely root cause is documented (e.g. "threshold too high for minor spelling differences" or "list does not include alias X").
- [ ] Performance metrics calculated: Overall detection rate, false positive rate and any other relevant performance metrics are computed to quantify system effectiveness.
- [ ] Objectives check: The outcomes are measured against the objectives set previously, noting which objectives are met and which are not (with evidence).
- [ ] Severity prioritised: Each identified issue is assigned a severity/priority level based on risk impact (e.g. misses are critical/high priority; false positives moderate unless very numerous).
- [ ] Draft remediation ideas: Initial thoughts on how to address each issue are noted (to be refined in next step) – e.g. "consider lowering threshold" or "need to add alias to list", linking issue to potential fix.
- [ ] Analysis documented: A report or matrix capturing all the above (findings, causes, metrics) is prepared and saved. This will be used for stakeholder communication and forms part of the audit trail.
- [ ] Key stakeholders informed: Any urgent or significant findings are communicated to appropriate parties immediately (especially any critical false negatives), so interim risk mitigations can be done if needed ahead of final fixes.
- [ ] Ready for tuning phase: The team has a clear understanding of what issues need to be fixed or tuned in the system, setting the stage for the next step.