Provider level fraud detection for federal healthcare programmes

OPEN METHODOLOGY · FREE TO RUN, VERIFY, AND ADOPT

VALIDATION PROTOCOL · VERSION [VERSION]

How to find out whether this works on *your* claims data

A staged self test an institution runs on its own data, inside its own environment, without contacting the author and without sending anything anywhere. It is written to be capable of returning the answer no, which is the only kind of validation worth running.

STAGE MAP — 12 STAGES · 8 GATES · ABOUT 3 TO 5 MONTHS END TO END

STAGE	WHAT IT ESTABLISHES	WHO IS NEEDED	ELAPSED	GATE
PHASE 1 · CAN THIS BE RUN AT ALL				
S0	Authority, governance, and the rules the outputs will be held to	Privacy, legal, investigations lead	1 to 2 weeks	GATE
S1	An extract that contains the required fields over a usable window	Data engineering	3 to 5 days	NONE
S2	That the extract is intact, joinable, and correctly coded	Analyst	1 day	GATE
PHASE 2 · DOES THE METHOD REPRODUCE				
S3	That the feature construction reproduces on your data	Analyst	1 day	GATE
S4	Baselines, including your existing rule set, measured first	Analyst, investigations lead	2 days	GATE
S5	A trained model with a recorded, reproducible configuration	Analyst	Hours	NONE
S6	Measured performance, reported against the baselines	Analyst	Hours	NONE
PHASE 3 · IS IT WORTH AN INVESTIGATOR'S TIME				
S7	An operating point derived from investigative capacity	Investigations lead	1 day	NONE
S8	Precision and lift at the top of the queue you would actually work	Analyst	Hours	GATE
S9	That the scoring does not concentrate on a group without cause	Analyst, compliance	1 to 2 days	GATE

STAGE	WHAT IT ESTABLISHES	WHO IS NEEDED	ELAPSED	GATE
PHASE 4 · DOES IT HOLD UP IN THE FIELD				
S10	Blind investigator review of flagged entities against controls	Investigators	6 to 12 weeks	GATE
S11	A recorded adoption decision and the numbers behind it	Programme leadership	1 day	GATE
S12	Monitoring and re validation cadence, if adopted	Analyst	Ongoing	NONE

Stage 10 dominates the schedule, and that is the honest shape of this work. Stages 1 to 9 can be completed in about two weeks by one analyst. At the end of them you will know whether the method is worth a pilot. You will not know whether it works, because nothing before Stage 10 involves an investigator opening a case.

FOR

Programme integrity units

State Medicaid, MAC, UPIC, OIG analytic contractors

RUN BY

One analyst plus an investigations lead

No machine learning specialist required

DATA MOVEMENT

None

Nothing leaves your environment at any stage

AUTHOR CONTACT

Not required

Optional, and only to report a defect

COMPANION

Feature dictionary

[LINK]. Read it first.

VERSION

[VERSION]

Published [DATE]

LICENSE

CC BY 4.0

Adapt it to your own governance

ARCHIVED AT

[DOI]

Permanent identifier

0.1 Why an institution should not adopt this on published numbers

Including the published numbers in the companion documents, which were produced on a synthetic dataset and establish less than they appear to.

A reported result tells you that a method separated fraud from legitimate records in the data it was built on. It does not tell you how the method behaves on production claims, which carry missingness patterns, coding drift between years, and provider behaviour that a generated dataset does not contain. It tells you nothing at all about the burden your investigators would carry working the flags it produces, and that burden is the entire operational question.

This protocol exists so that the answer comes from your data rather than from a claim made about somebody else's. It is designed around three commitments. Nothing leaves your environment.

Every stage has a written stop condition, so that abandoning the exercise is a documented outcome rather than a failure. And the acceptance criteria are set by you, before you look at the results, because a threshold chosen after seeing the numbers is not a threshold.

> **WHAT A PASSING RESULT WOULD AND WOULD NOT MEAN**
>
> A model that clears every gate here has demonstrated that it ranks billing entities in a way that puts more genuine problems near the top of an investigator's queue than your current method does. **That is the whole claim.** It is not a finding that any flagged entity committed fraud, it is not admissible as evidence of anything, and it does not license contacting an entity on the strength of a score. The output is a prioritisation instrument and should be described that way in every document your institution produces about it.

Before you start, read the feature dictionary

Stage 3 assumes you have it open. It defines every measurement, its exact computation, its missing value handling, and the two order dependent steps that fail silently. One item in it is unresolved and must be settled against your own extract at Stage 2: the coding of the chronic condition columns. `[LINK to feature dictionary]`

Phase 1

Can this be run at all

Three stages that end either with a usable extract and written authority, or with a documented reason to stop before any analytical work is done.

S0 Authority and governance

GATE

Establish that this test can be run lawfully under your existing authority, and fix in writing what the outputs may and may not be used for.

WHO

Privacy or compliance officer, investigations lead, legal counsel where enforcement could follow

ELAPSED

1 to 2 weeks

OUTPUT

Signed scope note

ACTIONS

1. Confirm that programme integrity analytics falls within your existing authority to use claims data. This protocol requires no new collection and no new data sharing.
2. Apply the minimum necessary standard to the extract specification at S1. Fields not listed in the feature dictionary should not be pulled.
3. Record that all processing occurs inside your environment. **No stage of this protocol requires data to leave, and any instruction that appears to require it is a defect worth reporting.**
4. Pseudonymise beneficiary and billing entity identifiers before analysis, holding the mapping separately under your own controls.
5. Agree in writing that during validation **no entity will be contacted, audited, or referred on the basis of a score**. Stage 10 works cases through your normal process, blind to the score.
6. Name the person accountable for the exercise and the person who will sign the Stage 11 decision.

PROCEED IF

- Written scope note exists and is signed.
- Data stays inside the environment for every stage.
- The no contact commitment for Stages 1 to 10 is recorded.

STOP IF

- Any step would require an external transfer of claims data.
- You cannot secure a commitment that scores will not drive contact during validation.
- No named owner. An unowned analytics pilot produces results nobody acts on and nobody retires.

S1 Assemble the extract

NO GATE

Produce two extracts covering the required fields: a development window and a later, untouched holdout window.

WHO

Data engineering, with the analyst specifying

ELAPSED

3 to 5 days

OUTPUT

Two dated extracts plus a specification note

REQUIRED

Beneficiary level fields, per the feature dictionary: beneficiary key, date of birth, date of death, sex, geography, Part A and Part B coverage months, the ten chronic condition flags, the renal disease indicator, and the four annual reimbursement and deductible amounts.

Claim level fields, required for the provider layer and for any measurement narrower than a year: billing entity key, beneficiary key, service date, and reimbursed amount.

Minimum scale. At least 24 months of history, and enough billing entities above your volume floor that peer comparison is meaningful. `[SET the minimum entity count from your own programme size. Below roughly a few hundred entities above the floor, peer standard scores are unstable and this exercise will mislead you.]`

ACTIONS

1. Split by time, not at random: an earlier development window and a later holdout window with no overlap. A random split lets the model learn from the future, and production never offers that.
2. Record the extract date and the reference date used for age. The feature dictionary notes that age for living beneficiaries is measured against a moving reference, so it is not comparable across extracts unless the date travels with it.
3. Document what your fraud label actually records. Closed investigations, recoveries, referrals, and suspensions are different things with different lags, and the difference determines what the model can learn.
4. Write down the label lag. Recent periods look clean because investigations have not concluded, not because behaviour improved.

CAUTION

The single most common way this exercise fails is a label whose meaning nobody wrote down. If two people in your unit would describe the fraud flag differently, settle that before S2 rather than after S10.

S2 Integrity checks

GATE

Establish that the extract is complete, joinable, and coded the way you think it is, before any feature is computed on it.

WHO	ELAPSED
Analyst	1 day

OUTPUT

Integrity report, one page

ACTIONS

1. Row counts against the source system. A silent truncation at a round number is the classic extract failure.
2. Uniqueness of the beneficiary key. Duplicates inflate every aggregate downstream.
3. Join yield from claims to beneficiaries. Record the percentage that joins and inspect a sample of those that do not.
4. Missingness by column, reported as a table rather than a summary sentence.
5. Label prevalence. Report it as a percentage and compare it against what your unit believes the rate to be.
6. **Verify the chronic condition coding.** This is the unresolved item in the feature dictionary and it must be settled here.
7. Date ranges for birth and death, checked for impossible values and for placeholder dates such as a repeated default.
8. Currency sanity: negatives, zeros, and the top percentile of each of the four amount fields.

VERIFICATION

```
assert df['BeneID'].is_unique, "duplicate beneficiary keys"

# chronic condition coding: settle this before Stage 3
print(df[CHRONIC_COLS].stack().value_counts())
# two values expected. Identify which denotes presence in YOUR extract,
# then set PRESENT_VALUES in the feature pipeline accordingly.

join_yield = claims['BeneID'].isin(df['BeneID']).mean()
print(f"join yield {join_yield:.3%}")

print(df['PotentialFraud'].value_counts(normalize=True))   # label
prevalence
```

PROCEED IF

- Join yield at or above `[SET, suggested 95%]` and the shortfall is explained.
- Chronic condition coding is identified and recorded.
- Label prevalence is within a range your unit finds credible.

STOP IF

- The chronic condition coding cannot be determined. Proceeding risks an inverted health burden index that will not announce itself.
- Label prevalence is below about half a percent. At that rate this design will not produce a usable estimate and needs a different approach.
- Join yield is low and unexplained. Every provider level measurement inherits the gap.

Phase 2

Does the method reproduce

Four stages ending with a measured result that can be read against something. The baseline stage is the one people skip and the one that determines whether any number afterwards means anything.

S3 Reproduce the feature construction

GATE

Compute every feature in the dictionary on your extract, in the documented order, and confirm each one behaves as described.

WHO

Analyst

ELAPSED

1 day

OUTPUT

Feature frame plus assertion log

ACTIONS

1. Run the pipeline in the order given in the feature dictionary. Two steps are order dependent and fail silently if reversed.
2. After deriving the survival flag, compare the proportion of living beneficiaries against your own enrolment records. A value near zero means the flag was derived after the date of death column was filled.
3. Compute the correlation between the chronic disease index and the two reimbursement fields. A negative correlation indicates the index is inverted and sends you back to Stage 2.
4. Inspect the age distribution. Confirm the shape matches the population you know you serve.
5. Assert that no identifier column survives into the feature matrix.
6. Fit encoders on the development window only, and apply them to the holdout with transform.

VERIFICATION

```
assert 0.5 < df['Alive'].mean() < 1.0, "Alive derived after the DOD fill"

r =
df[['ChronicDiseaseIndex','IPAnnualReimbursementAmt']].corr().iloc[0,1]
assert r > 0, f"chronic index appears inverted (r={r:.3f}); revisit S2"

assert not {'BeneID','ProviderID'} & set(X.columns), "identifier leaked
into X"
assert X.isna().sum().sum() == 0, "unhandled nulls entering the model"
```

PROCEED IF	STOP IF
• Every assertion passes. • The living proportion and the age distribution match your own records. • The chronic index correlates positively with reimbursement.	• Any assertion fails. Fix the cause rather than relaxing the assertion. • You cannot reconcile the living proportion against enrolment. The strongest hard implausibility signal is unreliable until you can.

S4 Establish the baselines, before training anything

GATE

Produce the three numbers against which every later result will be read. Compute them now, while you still have no stake in the outcome.

WHO

Analyst with the investigations lead

ELAPSED

2 days

OUTPUT

Baseline table, written down and dated

THE THREE

1. **Majority class.** Predict legitimate for everything. Record the accuracy this achieves. At a nine percent fraud rate it scores about ninety one percent while detecting nothing, and having that number on the page is the cheapest protection against a misread result.
2. **Random ranking.** Rank entities at random and record precision at your intended queue depth. This is the floor any ranking must clear.
3. **Your existing rule set.** Score it on the same holdout window, with the same metrics, at the same queue depth. **This is the comparator that actually matters**, because adopting a new method means replacing or supplementing this one.

WHY NOW

Baselines computed after a model result are chosen, however honestly, in the light of that result. Compute them first, write them down, date the file, and do not revise them.

PROCEED IF

- All three baselines are recorded and dated.
- The investigations lead agrees the rule set was scored fairly, at the depth they actually work.

STOP IF

- Your existing rule set cannot be scored on the same data. Without it, an improvement cannot be interpreted and the exercise cannot support a decision.
- Nobody can state the queue depth your team actually works. Stage 7 depends on it.

S5 Split and train

NO GATE

Fit the model with a configuration you can reproduce exactly, and record it before looking at any result.

WHO
Analyst

ELAPSED
Hours

OUTPUT
Trained model plus configuration record

ACTIONS

1. Use the temporal split from S1. Train on the earlier window, evaluate on the later one. A random split is acceptable only for a first pass and must be labelled as such.
2. Handle class imbalance by weighting rather than resampling, so that the data distribution is unchanged and the result stays interpretable.
3. Run the depth progression rather than jumping to the deepest setting. The sensitivity curve tells you whether performance is being bought by capacity, and where it stops paying.
4. Record the random seed, library versions, hyperparameters, feature list, and both window dates in one file committed alongside the results.

CONFIG

```
RandomForestClassifier(
    n_estimators=1000,
    max_depth=25,
    class_weight='balanced',
    random_state=42,
)
# depth progression to run: 4, 10, 15, 25
# reference environment: Python 3.10, pandas 1.5.3, scikit-learn 1.2.1, CPU only
```

This configuration was best on the study data. Treat it as a starting point rather than a recommendation for your data, and record whatever you actually ran.

S6 Measure, and report against the baselines

NO GATE

Produce the full metric set on the holdout window, with the S4 baselines on the same page.

WHO
Analyst

ELAPSED
Hours

OUTPUT
Results table and curves

REPORT

- Confusion matrix at the operating threshold, in counts rather than proportions.
- Precision, recall, and F1 for the fraud class specifically, not the weighted average.
- Area under the ROC curve, as a ranking quality measure.
- **Area under the precision recall curve.** Under heavy imbalance this describes performance more faithfully than ROC and should lead your report.
- A calibration check, if any downstream process will treat the score as a probability.
- The three S4 baselines alongside, on the same page.

REPORT Y ALONE

Accuracy at a nine percent base rate is dominated by the majority class and will overstate capability to anybody reading quickly, including your own leadership. Where it is reported, put the majority class baseline immediately beside it.

For the same reason, note that a high accuracy and a high area under the ROC curve are not competing claims about the same thing. Accuracy is measured at one threshold; the ranking measure describes performance across all of them. For a prioritisation instrument the ranking measure is the relevant one, because the output is a queue rather than a verdict.

REFERENCE

PUBLISHED RESULTS ON THE SYNTHETIC STUDY DATA, FOR ORIENTATION ONLY

MODEL	DEPTH	TREES	ACCURACY	PRECISION	RECALL	F1	AUC
RF 1	4	500	64.70%	0.558	0.354	0.433	0.66
RF 3	15	500	78.13%	0.735	0.666	0.699	0.85
RF 5	**25**	**1000**	**86.88%**	**0.865**	**0.777**	**0.819**	**0.94**
SVM	n/a	n/a	62.98%	0.577	0.108	0.181	n/a
KNN	n/a	n/a	56.40%	0.413	0.341	0.373	n/a

Do not treat these as targets. They were produced on generated data with a clean label. Falling short of them on production claims is expected and is not by itself a reason to stop. What matters is Stage 8, not this table.

Phase 3

Is it worth an investigator's time

The stages that convert a metric into an operational statement. The order here is deliberate and is the reverse of the usual one.

S7 Set the operating point from capacity, not from the model

NO GATE

Derive the score threshold from how many cases your unit can actually work, rather than choosing a threshold and discovering how many flags it produces.

WHO
Investigations lead, with the analyst

ELAPSED
1 day

OUTPUT
A queue depth and the threshold implied by it

ACTIONS

1. Ask the investigations lead one question: how many billing entity reviews can this unit open per month, sustainably, alongside everything else it does?
2. Convert that to a queue depth over the holdout window. Call it k.
3. Rank entities by score and take the top k. The threshold is whatever score sits at position k. You do not choose it.
4. Record the recall implied by that depth. This is the honest statement of what the method would catch at your real capacity, and it is usually lower than the headline recall.

WHY THIS ORDER

Choosing a threshold first produces a flag volume the unit cannot absorb, a backlog, and a quiet decision by investigators to ignore the queue. That failure is attributed to the model afterwards, but it was designed in at this step.

Starting from capacity also makes the trade off visible to the person who owns it. A recall figure is an analyst's number. A statement that the unit would work forty entities a month and catch roughly a third of what is there is a decision the programme can actually take.

S8 Precision and lift at the top of the queue

GATE

Measure the only thing that determines whether an investigator's day improves: what fraction of the cases they would actually open turn out to be worth opening.

WHO
Analyst

ELAPSED
Hours

OUTPUT
Precision at k, lift, and expected yield per investigator month

COMPUTE

- **Precision at k.** Of the top k ranked entities, the share carrying the label.
- **Lift.** Precision at k divided by the base rate. A lift of four means the queue is four times richer than working entities at random.
- **Expected yield.** Precision at k multiplied by the monthly review capacity, giving true positives per investigator month.
- The same three figures for your existing rule set at the same k, from S4.

PROCEED IF	STOP IF
• Lift at k materially exceeds the random baseline. • Precision at k exceeds your existing rule set at the same depth, by a margin you set at S4 and did not revise. • The investigations lead agrees the expected yield would change how they allocate the month.	• The rule set you already run performs as well or better at your real queue depth. There is no case for adoption and the correct outcome is a documented no. • Lift is near one. The ranking is not carrying information at the depth you would work.

S9 Fairness and disparate impact audit

GATE

Establish whether the scoring concentrates on any group in a way that case mix does not explain, before anyone is scored in production.

WHO	ELAPSED
Analyst with compliance	1 to 2 days

OUTPUT

Published audit, filed with the model

WHY THIS IS NOT OPTIONAL

A fraud label records what was detected and pursued, which is not the same as what occurred. Where past enforcement attention fell unevenly across geographies or populations, a model trained on those labels learns the distribution of attention and returns it as a prediction, wearing the appearance of objectivity. The consequence falls on beneficiaries and on the providers who serve them.

ACTIONS

1. Confirm the race field was excluded from the feature matrix, per the restriction recorded in the feature dictionary.
2. Compute score distributions and flag rates at your operating threshold, broken down by race, state, county, entity size, and specialty.
3. Where a group shows an elevated flag rate, test whether case mix explains it. Serving a dialysis population, or an older population, legitimately raises cost.
4. Record the result and publish it alongside the model, so that an adopting colleague does not have to ask whether it was done.
5. Schedule this audit to repeat at every re validation, not only at first adoption.

PROCEED IF	STOP IF
• The audit is complete, written, and filed with the model. • Any elevated flag rate is explained by case mix, and the explanation is recorded.	• A material disparity is unexplained. Do not proceed to Stage 10 while it is open. • The audit cannot be run because the breakdown fields were not extracted. Return to S1.

Phase 4

Does it hold up in the field

The stage that costs the most, takes the longest, and is the only one that measures value rather than a proxy for it.

S10 Blind investigator review

GATE

Have investigators work a mixed sample of flagged and unflagged entities through the normal process, without knowing which is which, and compare what they find.

WHO

Investigators, coordinated by the investigations lead

ELAPSED

6 to 12 weeks

OUTPUT

Case outcome log and a yield comparison

DESIGN

1. Draw `[N]` entities from above the operating threshold and `[N]` matched controls from below it, matched on volume, specialty, and geography so that the comparison is not confounded by size.
2. Shuffle them into a single list with no score, no rank, and no indication of source. Investigators must not be able to infer group membership from the ordering or the file.
3. Work every case through your normal process, at normal depth, with the normal standard of evidence.
4. Record for each case: the outcome category, the hours spent, and whether the reviewer would have opened it under existing triage.
5. Unblind only after every case in the sample is closed. Partial unblinding contaminates the remainder.

CATEGORIES

No issue found	Billing consistent with services and documentation.
Documentation error	A discrepancy explained by coding or record keeping, with no overpayment.
Overpayment identified	Recoverable amount established, without a finding of intent.
Referred	Escalated for investigation under your normal referral criteria.

Report the first category as honestly as the last. A method that produces a high share of no issue findings is telling you its precision at your queue depth, and that is exactly what this stage exists to learn.

YOU UNBLIND

Decide now, in writing, how much higher the flagged yield must be than the control yield to justify adoption, and who will judge it. A margin chosen after seeing the split is not a margin, and everyone in the room will know it.

PROCEED IF

- Flagged yield exceeds control yield by the margin fixed in advance.
- Hours per productive case are acceptable to the investigations lead.
- Blinding held. Record any case where it did not.

STOP IF

- Flagged and control yields are comparable. The ranking is not improving how the unit spends its time, whatever the earlier metrics said.
- Blinding failed broadly enough that the comparison cannot be trusted. Rerun rather than discount it.

S11 Decision and record

GATE

Convert the result into a decision with a name against it, and record the numbers that produced it.

WHO	ELAPSED
Programme leadership	1 day

OUTPUT

Signed decision plus the completed record sheet

OPTIONS

- **Adopt** as a triage input alongside existing rules, with the operating threshold, the fairness audit, and the monitoring cadence recorded.
- **Pilot with limits** for a defined period and a capped queue depth, with a fixed date to revisit.
- **Defer pending data improvement**, naming the specific defect: label quality, claim level granularity, entity count, or missing procedure codes.
- **Reject**, with the reason recorded. **This is a legitimate and useful outcome**, and a protocol incapable of producing it would not be worth running.

RECORD

Complete the record sheet at the end of this document. Whatever the decision, the sheet is what lets a successor understand why, and what lets you compare against a later version of the method without repeating the whole exercise.

S12 Monitoring, if adopted

NO GATE

Keep the thing honest after it stops being a project.

WHO	ELAPSED
Analyst	Ongoing

OUTPUT

Standing monitoring report

ACTIONS

1. Re run Stages 6 through 9 on a fresh holdout window at a fixed cadence. `[SET, suggested every 6 months]`
2. Monitor feature distributions for drift. A shift in coding practice or a programme rule change moves the inputs without anybody announcing it.
3. Account for label lag in every refresh. Recent periods look clean because investigations have not concluded, and a model retrained naively on them will learn that recency means innocence.
4. Repeat the fairness audit at every refresh, not only at first adoption.
5. Track realised precision at k in production against the S8 estimate. Divergence is the earliest signal that something upstream has changed.

L

What this protocol cannot tell you

Stated here so that a passing result is not read as more than it is.

- **It cannot establish that any flagged entity committed fraud.** Every measurement is a statement about billing patterns relative to peers. Intent is established by investigators through evidence, and nothing in this exercise contributes to that.
- **It cannot measure undetected fraud.** Schemes your unit never caught sit in the negative class. That bounds achievable recall and biases the model toward the schemes already being found, which is the opposite of what you most want.
- **It cannot correct a bad label.** If the fraud flag records referral rather than finding, the model learns what gets referred. Stage 1 asks you to write down which it is for exactly this reason.
- **It cannot see what is not in the data.** Without procedure and diagnosis codes, upcoding and unbundling are out of reach entirely, as recorded in the feature dictionary limitations.
- **It cannot tell you the method will keep working.** Fraud adapts to detection. Stage 12 exists because a validated model is a validated model on a date.
- **It cannot substitute for your own governance.** Where anything in this protocol conflicts with your institution's rules, your rules govern.

R

Validation record

One page. Complete it at Stage 11 whatever the decision. Nothing on it identifies a beneficiary, a provider, or a case.

Validation record

Provider level fraud detection · Protocol version `[VERSION]` · No identifying information

INSTITUTION
Type is enough if you prefer

DATE COMPLETED

ACCOUNTABLE OWNER

DEVELOPMENT WINDOW

HOLDOUT WINDOW

SPLIT TYPE
Temporal or random

ENTITIES ABOVE VOLUME FLOOR

VOLUME FLOOR USED

LABEL PREVALENCE

LABEL DEFINITION
Referral, finding, recovery

LABEL LAG

CHRONIC CODING FOUND
Which value denotes presence

STAGES COMPLETED, AND ANY STAGE WHERE YOU STOPPED

BASELINE: MAJORITY CLASS

BASELINE: RANDOM AT K

BASELINE: EXISTING RULE SET AT K

QUEUE DEPTH K
Reviews per month

PRECISION AT K

LIFT AT K

RECALL AT K

AREA UNDER PR CURVE

AREA UNDER ROC CURVE

S10 SAMPLE SIZE
Flagged and control

S10 FLAGGED YIELD

S10 CONTROL YIELD

MARGIN FIXED IN ADVANCE

HOURS PER PRODUCTIVE CASE

FAIRNESS AUDIT OUTCOME

DEFECTS FOUND IN THE PUBLISHED METHOD OR DOCUMENTATION
The most useful line on this sheet

DECISION, AND THE REASONING BEHIND IT

SIGNED

RE VALIDATION DUE

SENDING IT BACK IS OPTIONAL AND USEFUL

There is no requirement to report anything to anybody. If you are willing, the defects line and the decision line are what improve the published method for the next institution, and a recorded no is more useful than a recorded yes. Send to [CONTACT], with nothing that identifies a beneficiary, a provider, or a case.

Run it, and tell me where it fails

This protocol is published so that an institution can decide about the method without trusting the person who wrote it. If a stage is ambiguous enough that two analysts would execute it differently, or if a gate turned out to be the wrong gate, that is a defect in this document and worth reporting.

COMPANION DOCUMENTS

- Feature dictionary [LINK]
- Reference implementation [LINK]
- Implementation guide [LINK]
- Methodology paper [LINK]

VERSIONING RULE

Stage identifiers are stable. A changed stage receives a new identifier rather than a redefinition, so that a record sheet completed under an earlier version stays interpretable. Changelog: [LINK]

CORRECTIONS

Report an ambiguous stage, a gate that produced the wrong answer, or a limitation not recorded here: [CONTACT]

Ayeni, A. [YEAR]. Validation protocol: institutional self test for provider level fraud detection in federal healthcare programmes, version [VERSION]. [DOI]

This protocol validates a prioritisation instrument for human investigators. Completing it does not establish that any billing entity committed fraud, and no result produced by it constitutes evidence against any entity or beneficiary.