

Provider level fraud detection for federal healthcare programmes

OPEN METHODOLOGY · FREE TO RUN, VERIFY, AND ADOPT

FEATURE DICTIONARY · VERSION [VERSION]

Every measurement, how it is computed, and why it indicates fraud risk

The complete specification of the feature space used by the healthcare claims anomaly detection work. Written so that an analyst at a state programme integrity unit can reproduce every measurement on their own claims data, and so that a reviewer can check that no measurement was asserted without a definition.

FEATURE INDEX		29 ENTRIES · 19 IMPLEMENTED · 10 SPECIFIED		
ID	FEATURE	TYPE	ROLE	STATUS
A · IDENTIFIERS AND DEMOGRAPHICS				
A1	BeneID	String key	Join key, not a model input	IMPLEMENTED
A2	DOB	Date	Source for Age	IMPLEMENTED
A3	DOD	Date	Source for Alive and Age, then dropped	DROPPED
A4	Gender	Binary	Model input	IMPLEMENTED
A5	Race	Categorical	Retained with restrictions	RESTRICTED
A6	State	Categorical	Model input	IMPLEMENTED
A7	County	Categorical	Model input	IMPLEMENTED
B · COVERAGE AND ENROLMENT				
B1	NoOfMonths_PartACov	Integer	Model input	IMPLEMENTED
B2	NoOfMonths_PartBCov	Integer	Model input	IMPLEMENTED
C · CLINICAL INDICATORS				
C1	ChronicCond_* (10)	Coded flag	Model input	IMPLEMENTED
C2	RenalDiseaseIndicator	Binary	Model input	IMPLEMENTED
D · FINANCIAL				
D1	IPAnnualReimbursementAmt	Currency	Model input	IMPLEMENTED
D2	IPAnnualDeductibleAmt	Currency	Model input	IMPLEMENTED
D3	OPAnnualReimbursementAmt	Currency	Model input	IMPLEMENTED

ID	FEATURE	TYPE	ROLE	STATUS
D4	OPAnnualDeductibleAmt	Currency	Model input	IMPLEMENTED
E · ENGINEERED				
E1	Alive	Binary	Model input	IMPLEMENTED
E2	Age	Integer	Model input	IMPLEMENTED
E3	ChronicDiseaseIndex	Integer 0 to 10	Model input	IMPLEMENTED
E4	BirthYear	Integer	Interim only	DROPPED
F · PROVIDER LEVEL AGGREGATION LAYER				
F1	ProviderID	String key	Unit of analysis	SPECIFIED
F2	prov_claim_count	Integer	Volume base	SPECIFIED
F3	prov_distinct_benes	Integer	Breadth	SPECIFIED
F4	prov_claims_per_bene	Ratio	Intensity	SPECIFIED
F5	prov_mean_reimb	Currency	Cost profile	SPECIFIED
F6	prov_reimb_peer_z	Float	Peer deviation	SPECIFIED
F7	prov_posthumous_rate	Rate 0 to 1	Hard implausibility	SPECIFIED
F8	prov_short_tenure_rate	Rate 0 to 1	Front loading	SPECIFIED
F9	prov_cond_mismatch_rate	Rate 0 to 1	Clinical implausibility	SPECIFIED
F10	prov_bene_overlap	Float	Cross provider structure	SPECIFIED

APPLIES TO Healthcare claims anomaly detection Reference implementation [LINK]	UNIT OF RECORD Beneficiary Provider level layer specified in Section F	SOURCE TABLE Beneficiary master 23 raw columns before engineering
READER Programme integrity analyst No machine learning background assumed	ENVIRONMENT Python 3.10, pandas, scikit-learn CPU only, no accelerator required	VERSION [VERSION] Published [DATE]
LICENSE CC BY 4.0 Code under [CODE LICENSE]	ARCHIVED AT [DOI] Permanent identifier	

0.1

What this document is for

 A detection method that cannot be reproduced on another institution's data is a result, not a capability. This document is the part that makes the difference.

Every entry below defines one measurement: what it is, exactly how it is computed, over what window, what happens when the underlying value is missing, why it carries fraud signal, and what it should not be trusted to do. An analyst at a state Medicaid programme integrity unit should be able to compute each one against their own historical claims without contacting the author, and obtain a measurement that is comparable to the one computed anywhere else.

Cross institutional comparability is the point. Billing fraud is routinely operated across several institutions in several states at once, precisely because no single institution sees the whole pattern. A measurement defined loosely produces numbers that cannot be compared between institutions, and a distributed detection capability built on loosely defined measurements is not a capability at all.

WHAT THE OUTPUT IS, AND WHAT IT IS NOT

The model produces a risk score and a peer relative rank for a billing entity, together with an itemised account of which measurements drove that score. It is a prioritisation instrument for human investigators. **It does not determine that fraud occurred, and no entry in this dictionary should be described to anyone as evidence of fraud.** A high value on any feature here means the entity's billing pattern differs from its peers in a way that has historically warranted a look. That is all it means.

Status labels used throughout

IMPLEMENTED	Computed in the published reference implementation. The code shown in the entry is the code that runs.
SPECIFIED	Defined here for the provider level layer but not yet implemented . The definition is fixed so that an institution can compute it, and so that the eventual implementation can be checked against it.
DROPPED	Present in the source data, used to derive another feature, then removed before training. Dropped for leakage or redundancy control, with the reason recorded.
RESTRICTED	Available in the data but subject to a documented restriction on use. See Section G.

Conventions

- Code is Python with pandas. `df` denotes the working beneficiary frame; the raw frame is never mutated.
- Currency values are United States dollars, unrounded, as stored by the programme.

- "Window" states the period a measurement summarises. Where the source field is an annual total, the window is the programme year and cannot be narrowed without a claim level rebuild.
 - Feature identifiers (A1, B2, F7 and so on) are stable across versions. If a measurement's definition changes, it receives a new identifier rather than a redefinition, so that results computed under an older version remain interpretable.
 - Anything marked [LIKE THIS] must be filled in before this document is relied upon.
-

0.2 Order of operations, which is not optional

Three of the engineered features depend on being computed before a value is overwritten. Executing these steps out of order produces a frame that looks correct and is wrong.

```
# 1. Normalise the renal indicator before any numeric operation
df['RenalDiseaseIndicator'] = df['RenalDiseaseIndicator'].replace({'Y': 1, '0': 0}).astype(int)

# 2. Parse dates
df['DOB'] = pd.to_datetime(df['DOB'])
df['DOD'] = pd.to_datetime(df['DOD'])

# 3. Alive MUST be derived here, while DOD nulls still mean "living"
df['Alive'] = df['DOD'].isna().astype(int)

# 4. Only now fill DOD, so that Age is computable for every row
df['DOD'] = df['DOD'].fillna(df['DOD'].max())

# 5. Age is derived from the filled DOD
df['Age'] = ((df['DOD'] - df['DOB']).dt.days / 365.25).round().astype(int)

# 6. Chronic condition index
df['ChronicDiseaseIndex'] = df[CHRONIC_COLS].isin(PRESENT_VALUES).sum(axis=1)

# 7. Drop the sources that would leak or duplicate
df = df.drop(columns=['DOD', 'BirthYear'])
```

STEP 3 BEFORE STEP 4

If DOD is filled before Alive is derived, every beneficiary is marked deceased and the single strongest hard implausibility signal in the feature set is silently destroyed. Nothing downstream fails, no error is raised, and model performance degrades in a way that is easy to attribute to the data rather than to the pipeline. Any institution reimplementing this should assert on the proportion of living beneficiaries immediately after step 3 and compare it against their own enrolment records.

A

Identifiers and demographics

Seven source fields. Two are keys or date sources rather than model inputs, one carries a use restriction, and the rest enter the model directly.

A1 BeneID		IMPLEMENTED	
Unique identifier for a beneficiary. Groups the multiple claims belonging to one person.			
TYPE		SOURCE	WINDOW
String		Raw	Static
MODEL INPUT			
No			
COMPUTATION	None. Used as a join key between the beneficiary master and the inpatient and outpatient claim tables, and as the grouping key for any beneficiary level aggregate.		
MISSING	A row with no BeneID cannot be joined and must be quarantined rather than imputed. Count and report such rows; do not drop them silently.		
RATIONALE	The beneficiary identifier is the required input on every claim. No claim exists without one, which is why it is the single item a billing scheme has to obtain, and why the identifier itself is the object of the theft rather than the money.		
CAUTION	Never a model input. Passing an identifier to a tree based model lets it memorise individuals, which inflates validation performance and produces a model that fails on any beneficiary it has not seen. In production this field is direct identifying information under HIPAA. It should be pseudonymised before analysis and the mapping held separately under the institution's own controls.		

A2 DOB		IMPLEMENTED
Date of birth. Retained as the anchor for Age.		
TYPE	SOURCE	WINDOW
datetime64	Raw, parsed	Static
MODEL INPUT		
Via E2		
COMPUTATION	<pre>df['DOB'] = pd.to_datetime(df['DOB'])</pre>	
MISSING	None observed in the source table. If encountered in production, Age cannot be computed for that row; flag it rather than imputing a birth date, because an imputed age propagates into every age consistency check.	
RATIONALE	Age enables detection of services inconsistent with the beneficiary's life stage, such as paediatric procedure codes billed against an elderly enrollee.	
CAUTION	Parse with an explicit format string in production. Silent misparsing of ambiguous day and month ordering shifts ages by up to eleven months and is invisible in summary statistics.	

Date of death. Used to derive Alive and Age, then removed before training.

TYPE		SOURCE	WINDOW
datetime64		Raw, parsed	Static
MODEL INPUT			
No			
COMPUTATION	<pre>df['DOD'] = pd.to_datetime(df['DOD']) df['Alive'] = df['DOD'].isna().astype(int) # before the fill df['DOD'] = df['DOD'].fillna(df['DOD'].max()) # after Alive exists df = df.drop(columns=['DOD']) # after Age exists</pre>		
MISSING	A null DOD denotes a living beneficiary and is meaningful rather than absent. It is captured in Alive first, then filled with the latest observed date of death so that a lifespan is computable for every row.		
RATIONALE	A service date after a recorded date of death cannot correspond to care delivered. This is the clearest hard implausibility available in claims data, and it is the basis for feature F7.		
CAUTION	Dropped to prevent leakage. Retaining the raw date alongside a target related to mortality or posthumous billing lets the model observe the outcome it is meant to predict, which produces validation performance that will not reproduce in deployment. Filling with the maximum observed date is a modelling convenience, not a claim about the person. Any downstream analysis that treats the filled value as a real date of death is misusing the field, which is a second reason the raw column does not survive into training.		

A4 Gender

IMPLEMENTED

Beneficiary sex as recorded by the programme, coded as an integer.

TYPE

Binary integer

SOURCE

Raw

WINDOW

Static

VALUES

0, 1

COMPUTATION

Used as stored. Coded 0 and 1 in the source table. [VERIFY which value denotes which sex in your own source system before relying on any sex specific check.]

MISSING

None observed. In production, treat as its own category rather than imputing the majority value.

RATIONALE

Supports detection of procedures billed against a beneficiary for whom they are anatomically impossible, which is a long standing and still productive rule based check.

CAUTION

A sex inconsistent procedure code is frequently a coding error rather than fraud. This feature should raise a question for a human reviewer, never an inference on its own.

Race code as recorded by the programme.

TYPE	Categorical	SOURCE	Raw	WINDOW	Static
MODEL INPUT	Restricted				
COMPUTATION	Label encoded for tree based models. Present in the source table as a small set of integer codes.				
MISSING	Treat as an explicit "not recorded" category. Race is unevenly recorded across programmes and imputing it manufactures a distribution.				
RATIONALE	Included in the source data for population level analysis. It carries no mechanism by which it would indicate that a billing entity is behaving fraudulently.				
RESTRICTION	This feature should be excluded from any deployed scoring model. A protected characteristic with no causal relationship to billing behaviour can still correlate with a fraud label if enforcement history was itself unevenly distributed, in which case the model learns the enforcement pattern and reproduces it as a prediction. The consequence lands on beneficiaries and on the providers who serve them. The recommended use is diagnostic only: compute score distributions across race groups as a fairness audit of the model, and keep the field out of the feature matrix. Section G sets out the audit procedure.				

Numeric geographic codes for the beneficiary's state and county of residence.

TYPE	Categorical integer	SOURCE	Raw	WINDOW	Static
CARDINALITY	State low, County high				
COMPUTATION	<pre>for col in ['Gender', 'Race', 'RenalDiseaseIndicator', 'State', 'County']: df[col] = LabelEncoder().fit_transform(df[col]) # tree based models</pre> One hot encoding is used instead for models that benefit from binary splits.				
MISSING	Treat as an explicit category. A missing geography is itself informative about record quality.				
RATIONALE	Fraud operations concentrate geographically, both because schemes are marketed locally and because a billing entity's catchment is bounded. Geography supports hotspot detection and gives peer comparison a defensible reference group.				
CAUTION	These are nominal codes. Label encoding imposes an ordering that does not exist, which tree ensembles tolerate and linear models do not. Use one hot encoding for any linear or distance based model, and note that county cardinality makes that expensive. Fit the encoder on training data only and apply transform to held out data. Fitting independently on each split produces inconsistent code assignments and a silently corrupted evaluation.				

B

Coverage and enrolment

Two fields carrying beneficiary tenure. Their value comes from being read against cost rather than on their own.

Number of months the beneficiary has been enrolled under Medicare Part A (hospital insurance) and Part B (medical insurance).

TYPE Integer RANGE 0 to 12 per programme year	SOURCE Raw	WINDOW Cumulative to period end
COMPUTATION	Used as stored. During exploration, records were grouped by coverage duration to identify common durations and outliers, and beneficiaries were segmented into short tenure and long tenure groups to support downstream classification. <pre>df.groupby('NoOfMonths_PartACov').size() df['ShortTenure'] = (df['NoOfMonths_PartACov'] <= SHORT_TENURE_MONTHS).astype(int)</pre> [SET SHORT_TENURE_MONTHS from your own enrolment distribution. Do not import a threshold from another programme.]	
MISSING	None observed. A null here indicates an enrolment record that did not join and should be quarantined.	
RATIONALE	Short coverage duration paired with high reimbursement is the signature of front loaded billing: a scheme extracts as much as possible early in a coverage period, before any pattern can accumulate against the entity. New enrollees are also targeted deliberately in the approach phase, because a person who has never seen a Medicare Summary Notice does not know what a normal one looks like.	
CAUTION	Tenure alone carries almost no signal and will generate a large false positive burden if thresholded on its own. It is only meaningful in interaction with the financial features in Section D, which is one of the reasons tree ensembles outperform linear models on this feature set.	

c

Clinical indicators

Eleven flags describing the beneficiary's documented disease burden. These are the reference against which claimed services are judged plausible.

C1 ChronicCond_* (ten columns)			IMPLEMENTED
One coded flag per chronic condition, indicating whether the beneficiary has a documented diagnosis.			
TYPE	SOURCE		WINDOW
Coded flag	Raw		Cumulative diagnosis history
COUNT	10 columns		
COLUMNS	ChronicCond_Alzheimer	Alzheimer's disease	ChronicCond_Diabetes
	ChronicCond_Heartfailure	Congestive heart failure	ChronicCond_IschemicHeart
	ChronicCond_KidneyDisease	Chronic kidney disease	ChronicCond_Osteoporosis
	ChronicCond_Cancer	Any form of cancer	ChronicCond_rheumatoidArthritis
	ChronicCond_ObstrPulmonary	Chronic obstructive pulmonary disease	ChronicCond_Depression
	Column spellings above are reproduced exactly as they appear in the source table. Do not silently correct them in code, or the selector will fail to match.		
MISSING	None observed. A missing condition flag should be treated as "not documented" rather than "absent", because the two are different claims about the world and only the first is supported by the data.		

RATIONALE	Cross referencing documented conditions against claimed services exposes care billed for a condition the record does not support. This is the measurement behind the participant facing guidance that a screening or item offered to everyone of a given age, rather than for a reason a clinician found, is the one to be careful about.
UNRESOLVED	The encoding of these columns must be confirmed against your own extract before use. Two codings appear in the source documentation for this work: • 0 = No, 1 = Yes, in which case presence is <code>value == 1</code>. • 1 = Yes, 2 = No, which is the convention used in several public Medicare beneficiary extracts, in which case presence is <code>value == 1</code> and a test of <code>value > 1</code> counts absences. The consequence is not cosmetic. Under the second coding, a <code>> 1</code> test inverts <code>ChronicDiseaseIndex (E3)</code> so that it measures health rather than disease burden, and the model still trains, still converges, and still reports plausible performance. [VERIFY the coding in your extract and set <code>PRESENT_VALUES</code> accordingly. Record the result here before publication.]

C2 RenalDiseaseIndicator

IMPLEMENTED

Flag for end stage renal disease.

TYPE

Binary integer

SOURCE

Raw, normalised

WINDOW

Cumulative

VALUES

0, 1

COMPUTATION

```
df['RenalDiseaseIndicator'] = (df['RenalDiseaseIndicator']  
                               .replace({'Y': 1, '0': 0})  
                               .astype(int))
```

The source column mixes the string 'Y' with the string '0', which blocks any numeric operation until normalised. This is the first cleaning step in the pipeline for that reason.

MISSING

None observed. Any value outside {'Y', '0'} should raise rather than coerce, because a silent coercion to zero understates disease burden and makes legitimate high cost care look anomalous.

RATIONALE

End stage renal disease carries a distinct and legitimately expensive care pattern, including regular dialysis. Without this flag, the highest cost genuinely sick beneficiaries are indistinguishable from over billed ones, and the model spends its capacity flagging real nephrology.

CAUTION

This feature protects a specific patient population from being scored as anomalous. If it is dropped for convenience, the false positive burden concentrates on providers serving dialysis patients, which is both operationally wasteful and a defensible complaint from those providers.

D

Financial

Four annual currency totals. They carry most of the model's discriminative weight and most of its potential to mislead.

IMPLEMENTED

Total amount reimbursed by the programme for inpatient and for outpatient services, per beneficiary, per year.

TYPE

Currency, float

SOURCE

Raw

WINDOW

Programme year

RANGE

0 to unbounded

COMPUTATION

Used as stored. Not scaled for tree ensembles. Standardised for distance based and margin based models, which are sensitive to feature magnitude.

```
SCALE_MODELS = ('KNN', 'SVM') # standardise for these only
```

MISSING

A zero and a null mean different things: no reimbursed care versus no record. Do not fill nulls with zero. Quarantine and count them.

RATIONALE

Reimbursement out of line with the beneficiary's documented condition profile is the primary quantitative fraud indicator in this feature set, and the interaction between these fields and Sections B and C is where the signal actually lives.

The beneficiary facing form of the same measurement is the instruction to look at the amount and not only the service name, because a large number attached to a small thing is the discrepancy.

CAUTION

These are annual totals. Any measurement requiring a narrower window, including the short interval repeat detection in F9, cannot be computed from this table and needs a claim level rebuild.

Both fields are heavily right skewed. Reporting a mean without a median and a high percentile will misdescribe the distribution to anybody reading the output.

Annual deductible amounts borne by the beneficiary for inpatient and outpatient care.

TYPE	SOURCE	WINDOW
Currency, float	Raw	Programme year
RANGE		
0 to unbounded		
COMPUTATION	Used as stored. The ratio of deductible to reimbursement is a candidate derived measurement and is not currently computed. <pre>df['IPDeductibleShare'] = (df['IPAnnualDeductibleAmt'] / df['IPAnnualReimbursementAmt'].replace(0, np.nan))</pre>	
MISSING	As D1 and D3. Zero is a legitimate value and must be distinguished from absence.	
RATIONALE	Deductible amounts constrain what a beneficiary could plausibly have consumed, and a deductible pattern inconsistent with the reimbursement pattern indicates that the billing does not describe care the beneficiary participated in.	
CAUTION	Deductible structures are set by programme rules that change between years. A model trained across a rule change learns the rule change as though it were behaviour. Segment by programme year or confirm that no structural change falls inside the training window.	

E Engineered features

Four derived measurements. Three enter the model; the fourth exists only long enough to produce another and is then removed.

E1 Alive		IMPLEMENTED
Binary survival status derived from the presence or absence of a recorded date of death.		
TYPE	SOURCE	WINDOW
Binary integer	Derived from A3	As at extract date
VALUES	1 living, 0 deceased	
COMPUTATION	<pre>df['Alive'] = df['DOD'].isna().astype(int) # step 3, before any fill</pre>	
MISSING	Not applicable by construction. Every row receives a value.	
RATIONALE	Encodes the mortality signal as a clean binary so that models need not reason about nulls in a date column, and so that the posthumous billing check in F7 has a stable field to key on. This is the measurement whose participant facing form is the instruction to keep reading a late relative's notices for at least a year. The scheme depends on nobody opening the mail; the model depends on somebody having recorded the date.	
CAUTION	Order dependent. See Section 0.2. Derived after the fill, this feature is constant and useless, and nothing in the pipeline will say so. Date of death arrives in programme data with a lag. A beneficiary who has died recently may still show <code>Alive = 1</code>, so a low posthumous rate is partly a statement about reporting latency and not only about the entity.	

E2 Age

IMPLEMENTED

Age in whole years, at death for deceased beneficiaries and at the extract date for living ones.

TYPE

Integer, years

SOURCE

Derived from A2, A3

WINDOW

As at extract date

RANGE

0 to about 110

COMPUTATION

```
df['Age'] = ((df['DOD'] - df['DOB']).dt.days / 365.25).round().astype(int)
```

Computed after DOD is filled with the latest observed date of death, so that living beneficiaries receive an age relative to that reference date rather than a null.

MISSING

Not applicable by construction, provided DOB is present.

RATIONALE

Supports detection of age inconsistent procedures and gives peer comparison a demographic reference, since expected cost varies strongly with age and a comparison that ignores it will flag geriatric practices as anomalous.

CAUTION

For living beneficiaries this is **age as at the latest date of death in the extract**, not age today. The reference date moves whenever the extract is refreshed, so ages are not comparable across extracts without recording the reference date alongside them.

Record that reference date in the run metadata. An age distribution that shifts between runs for this reason has been mistaken for population drift before.

Count of documented chronic conditions per beneficiary. A single health complexity score replacing ten separate flags.

TYPE

Integer

SOURCE

Derived from C1

WINDOW

Cumulative

RANGE

0 to 10

COMPUTATION

```
CHRONIC_COLS = [c for c in df.columns if c.startswith('ChronicCond_')]
assert len(CHRONIC_COLS) == 10, CHRONIC_COLS
```

```
PRESENT_VALUES = {1} # see C1: confirm against your extract
df['ChronicDiseaseIndex'] =
df[CHRONIC_COLS].isin(PRESENT_VALUES).sum(axis=1)
```

The assertion is deliberate. A prefix selector that silently matches nine or eleven columns produces an index on a different scale, and the model will fit it without complaint.

MISSING

Not applicable by construction. A beneficiary with no documented conditions receives 0, which is a meaningful value rather than an absence.

RATIONALE

Health complexity is the denominator for cost. High reimbursement against a high index is expected; high reimbursement against an index of zero or one is the discrepancy worth investigating. Collapsing ten binary flags into one ordinal score makes that interaction available to the model in a compact form.

CAUTION

Direction depends on the C1 coding and must be verified. If the source codes conditions as 1 for yes and 2 for no, a presence test of > 1 produces an index of absences, inverting the feature's meaning while leaving every downstream diagnostic looking normal. This is the single most consequential ambiguity in this dictionary.

After computing, check the correlation between ChronicDiseaseIndex and the reimbursement fields. A negative correlation is a strong indication that the index is inverted.

E4 BirthYear		DROPPED	
Year extracted from date of birth, used for cohort exploration and then removed.			
TYPE		SOURCE	WINDOW
Integer		Derived from A2	Static
MODEL INPUT			
No			
COMPUTATION	<pre>df['BirthYear'] = df['DOB'].dt.year # exploration only df = df.drop(columns=['BirthYear']) # before training</pre>		
MISSING	Not applicable.		
RATIONALE	Useful for demographic cohort description during exploration. Carries no information beyond Age once Age exists.		
DROPPED	Redundant with E2 and more likely to be misread. Retaining both invites a model to split on birth year as a proxy for extract vintage, which is an artefact of when the data was pulled rather than anything about the beneficiary.		

F The provider level aggregation layer

Specified here, not yet implemented. Every entry in this section is a definition an institution can compute, and a commitment the eventual implementation can be measured against.

Sections A to E describe the beneficiary level feature space of the completed work. The design objective of the endeavour is different in one specific way: **the unit of analysis is the billing entity, not the claim and not the beneficiary.**

That distinction is the whole argument. A rule evaluating one claim at a time cannot observe that an entity's overall pattern is implausible when every individual claim is defensible in isolation, and that is exactly the evasion strategy sophisticated billing fraud uses. Aggregating to the entity makes the aggregate visible.

The features below are stated in full so that the specification is fixed before the code exists rather than reverse engineered from it afterwards. Where a measurement cannot be computed from the annual beneficiary table and requires a claim level rebuild, that is said in the entry.

HONEST STATUS

None of F1 to F10 is implemented in the published reference implementation as at version [VERSION]. They are specified. Any statement elsewhere that this detection work computes provider level aggregates today would be inaccurate, and this dictionary is the document that says so.

ID	FEATURE	DEFINITION AND COMPUTATION	FRAUD RISK RATIONALE
F1	ProviderID	The billing entity identifier. Grouping key for every measurement in this section. Never a model input, for the same memorisation reason given at A1.	Establishes the unit of analysis. Everything in this section is a property of an entity rather than of a claim.
F2	prov_claim_count	<code>claims.groupby('ProviderID').size()</code> over a fixed window. Requires the claim level tables.	Volume base. Carries little signal alone and is the denominator for most of the rates below, so it must be computed first and reported alongside them.
F3	prov_distinct_benes	<code>claims.groupby('ProviderID')['BeneID'].nunique()</code> .	Breadth of reach. A scheme run through a health fair or a call centre touches many beneficiaries briefly; a genuine small practice touches fewer, repeatedly.

ID	FEATURE	DEFINITION AND COMPUTATION	FRAUD RISK RATIONALE
F4	prov_claims_per_bene	F2 / F3. Report with a minimum volume floor, since the ratio is unstable for entities below [MIN_CLAIMS].	Separates breadth from intensity. High breadth with low intensity is the health fair pattern; low breadth with very high intensity is the repeat billing pattern.
F5	prov_mean_reimb	Mean reimbursed amount per claim for the entity, reported with the median and the 90th percentile because the distribution is right skewed.	Cost profile. On its own it distinguishes specialties more than it distinguishes behaviour, which is why F6 rather than F5 is the scored measurement.
F6	prov_reimb_peer_z	Standard score of F5 within a peer group defined by specialty and geography: $(F5 - \text{peer_mean}) / \text{peer_sd}$. Peer group definition must be recorded with the output. [FIX the peer grouping rule and record it here.]	The core comparative measurement. An entity billing several standard deviations above genuinely comparable peers is the finding an

ID	FEATURE	DEFINITION AND COMPUTATION	FRAUD RISK RATIONALE
			investigator can act on, and the peer definition is what makes it defensible when challenged.
F7	prov_posthumous_rate	Share of the entity's claims whose service date falls after the beneficiary's recorded date of death. Joins claim service dates to A3 before the fill described in Section 0.2.	Hard implausibility rather than a statistical outlier. Care cannot be delivered to a person who has died. Subject to the death reporting lag noted at E1, a non trivial rate is difficult to explain innocently.
F8	prov_short_tenure_rate	Share of the entity's reimbursement attributable to beneficiaries below the short tenure threshold defined at B1, weighted by amount rather than by claim count.	Front loading. Billing concentrated into the first months of a beneficiary's enrolment is the pattern of a scheme extracting value before

ID	FEATURE	DEFINITION AND COMPUTATION	FRAUD RISK RATIONALE
			scrutiny accumulates.
F9	prov_cond_mismatch_rate	Share of the entity's claims where the service billed has no supporting documented condition for that beneficiary, using the C1 flags as the reference set. Requires a mapping from procedure code to supporting conditions. [The mapping is a separate deliverable and does not exist yet.]	Clinical implausibility. Services billed for conditions the record does not support are the machine readable form of the screening that was offered to everybody rather than indicated for anybody.
F10	prov_bene_overlap	Share of the entity's beneficiaries who also appear at a defined set of other entities within the window. Computable only where the institution holds claims across those entities.	Cross entity structure. Schemes cultivated across several billing entities are invisible to each entity individually, which is the reason the same detection capability distributed across many institutions degrades the

ID	FEATURE	DEFINITION AND COMPUTATION	FRAUD RISK RATIONALE
			scheme everywhere at once.

MINIMUM VOLUME FLOORS ARE NOT OPTIONAL

Every rate in this section is unstable at low denominators. An entity with four claims and one implausible one scores at 25 percent and will dominate any ranking, which wastes investigative capacity on the smallest entities in the programme and produces exactly the pattern of complaint that discredits an analytics programme. Set and publish a minimum volume floor, report entities below it separately rather than scoring them, and record the floor alongside every ranked output.

G Encoding, scaling, leakage, and fairness

Handling rules that apply across features rather than to any one of them.

Encoding by model family

MODEL FAMILY	CATEGORICAL HANDLING	SCALING	NOTE
Tree ensembles	LabelEncoder	None	Splits are invariant to monotone transforms, so imposed ordinal codes are tolerated.
Linear models	One hot	Standardise	An imposed ordering on nominal codes is a modelling error here, not a tolerance.
Distance based (KNN)	One hot	Standardise	Unscaled currency fields dominate every distance computation.
Margin based (SVM)	One hot	Standardise	As above. County cardinality makes one hot encoding expensive here.

Leakage controls in force

- **Raw DOD is dropped** after Alive and Age are derived, so that no model can observe a mortality outcome it is being asked to predict.
- **BirthYear is dropped** as redundant with Age and as a proxy for extract vintage.
- **Identifiers are never model inputs.** A1 and F1 are keys.
- **Encoders are fitted on training data only** and applied to held out data with `transform`. Fitting independently on each split produces inconsistent codes and an evaluation that overstates performance.
- **Working copies are used throughout.** All transformations run on copies of the source frames, so the raw extract is reproducible and a pipeline error can be diagnosed rather than merely repeated.

Fairness audit, required before any deployment

A fraud label in historical data is a record of what was detected and pursued, which is not the same thing as what occurred. Where past enforcement attention was unevenly distributed across geographies or populations, a model trained on those labels will learn the distribution of attention and return it as a prediction, with the appearance of objectivity.

- Exclude A5 from the feature matrix of any deployed scoring model. See the restriction recorded at A5.
- After training, compute the distribution of scores and of flag rates across race, state, and county groups, and record the result alongside the model.
- Where a group shows a materially elevated flag rate, establish whether it is explained by legitimate case mix before deployment, not after a complaint.
- Publish the audit with the model. An institution adopting this method should not have to ask whether it was done.

H

Performance context

Reproduced here so that the feature set is read against what it actually achieved, rather than against a number recalled from elsewhere.

Class balance in the study data is approximately 9 percent fraudulent against 91 percent legitimate. At that ratio a classifier predicting the majority class for every record scores about 91 percent accuracy while detecting nothing, which is why accuracy alone is not reported and why the fraud class was weighted with `class_weight='balanced'` rather than resampled.

RANDOM FOREST PROGRESSION AND COMPARISON MODELS, HELD OUT TEST SET

MODEL	DEPTH	TREES	ACCURACY	PRECISION	RECALL	F1	AUC
RF 1	4	500	64.70%	0.558	0.354	0.433	0.66
RF 2	10	500	72.54%	0.663	0.569	0.612	0.77
RF 3	15	500	78.13%	0.735	0.666	0.699	0.85
RF 4	25	500	86.80%	0.864	0.776	0.818	0.94
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

ON THE TWO HEADLINE NUMBERS

Accuracy of 86.88 percent and AUC of 0.94 are not in tension and are not describing different models. They are two properties of the same configuration, RF 5. Accuracy is measured at one decision threshold; AUC measures ranking quality across all thresholds and says that a randomly chosen fraudulent record is ranked above a randomly chosen legitimate one 94 percent of the time. For a prioritisation instrument, the ranking property is the relevant one, because the output is a queue for investigators rather than a verdict.

WHAT THESE NUMBERS DO NOT ESTABLISH

This is a synthetic public dataset structured to mirror Medicare claims. Performance on it is evidence that the feature construction captures signal, and it is not evidence of how the method behaves on production claims, nor of the false positive burden a real investigative team would carry. Establishing that is the purpose of the validation protocol [\[LINK\]](#), which is a separate document, and no institution should adopt this method on the strength of the table above alone.

I

Known limitations of this feature set

Recorded here rather than discovered later by the institution that adopts it.

- **No procedure or diagnosis codes.** The beneficiary table carries no ICD or CPT level detail, so upcoding, unbundling, and service specific implausibility are

outside the reach of Sections A to E entirely. F9 depends on a procedure to condition mapping that does not yet exist.

- **Annual granularity.** Financial fields are yearly totals, so short interval repeat billing, the pattern most visible to a beneficiary reading a quarterly notice, cannot be computed without a claim level rebuild.
- **No temporal features.** Service date, day of week, and claim submission lag are absent. Timing is one of the more productive signal families in transaction fraud and it is unexploited here.
- **Beneficiary rather than provider unit.** The implemented model scores records, not entities. Section F closes this gap by specification only.
- **Synthetic source data.** Real claims carry missingness patterns, coding drift between years, and provider behaviour that a generated dataset does not reproduce. Expect performance to fall.
- **Labels are enforcement records.** The fraud label marks what was detected, not what occurred. Undetected fraud sits in the negative class, which bounds achievable recall and biases the model toward the schemes that were already being caught.
- **The C1 coding ambiguity is unresolved.** Until verified against a specific extract, E3 has a documented risk of pointing in the wrong direction. It is stated at C1, at E3, and here, because it is the item most likely to be skipped.

Use it, check it, and tell me where it is wrong

This dictionary is published so that an institution can reproduce every measurement without contacting the author, and so that a reviewer can verify that no measurement was claimed without a definition. If a definition here is ambiguous enough that two analysts would compute it differently, that is a defect in this document and worth reporting.

COMPANION DOCUMENTS

- Reference implementation [\[LINK\]](#)
- Validation protocol [\[LINK\]](#)
- Implementation guide [\[LINK\]](#)
- Methodology paper [\[LINK\]](#)

VERSIONING RULE

Feature identifiers are stable. A changed definition receives a new identifier rather than a redefinition, so that outputs computed under an earlier version stay interpretable. Changelog: [\[LINK\]](#)

CORRECTIONS

Report an ambiguous definition, a coding discrepancy, or a limitation not recorded here: [\[CONTACT\]](#)

Ayeni, A. [YEAR]. Feature dictionary: provider level fraud detection for federal healthcare programmes, version [VERSION]. [DOI] · Licensed CC BY 4.0.

This document describes a prioritisation instrument for human investigators. It does not determine that fraud has occurred, and no measurement defined here constitutes evidence of fraud by any billing entity or beneficiary.