CIPHER

Automated PPD Anonymization Tool for Clinical Documents Submission under EMA Policy 0070 and Health Canada PRCI

Shubham Patil
Clinical Trial Results and Business Automation Manager II

Divya Singh
Clinical Trial Disclosure Manager II

Bristol-Myers Squibb

PHUSE 2026

ROADMAP

How this talk flows

From the problem, to the principle, to the proof.

1

The challenge

Why redaction is hard — and uneven

2

The principle

The datasets as the source of truth

3

The tool

Four modules, one interface

4

Salient features

Custom approaches & Refine

5

Results

Metrics and the pilot

THE WHY

The redaction challenge

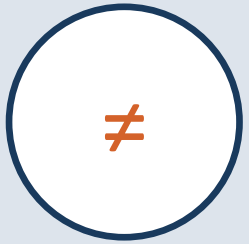
This is not a story about software. It is a story about protecting patients.

Volume

A single submission runs to thousands of pages across dozens of documents.

Hidden in plain sight

One patient may appear as a full ID, a short ID, or a date — written many ways.



It differs by reviewer

For medical history and medications, what to redact depends on the reviewer's understanding and threshold — so first-hand results vary from person to person.

THE TREE = THE DATASETS

ADaM / SDTM — the single source of every identifier that can appear anywhere

THE PRINCIPLE

“An apple never falls far from the tree.”

Every document is grown from the same datasets. So an identifier on a page **must trace back** to a value the datasets already hold.

It may be shortened, reformatted, split across lines, or printed inside a picture — but it cannot be something the datasets have never seen.

THE HOW

Reading the tree first

Ingestion happens before a single page is opened.

What we take from the datasets

- **Participant identifiers** full and short subject IDs
- **Every date** screening, dosing, visit, outcome
- **Context that identifies** country, age, demographics, medical history, concomitant medications, events

One value becomes many

A single date is expanded into every written form it could take on the page:

15APR2024 **15 Apr 2024**

2024-04-15 **Apr 2024**

The same is done for identifiers — this is how we cover the distance an apple can fall.

THE WHAT

CIPHER at a glance

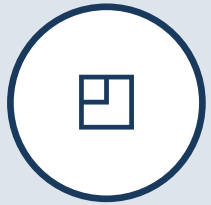
Four modules, one principle — the datasets decide what to protect.

Risk Assessment

Reads the subject-level dataset and fills the risk form automatically.

Mark Documents

Marks every identifier in the documents, anchored to the datasets.



Threshold Assessment

Shows how many participants sit behind each term before rules are set.

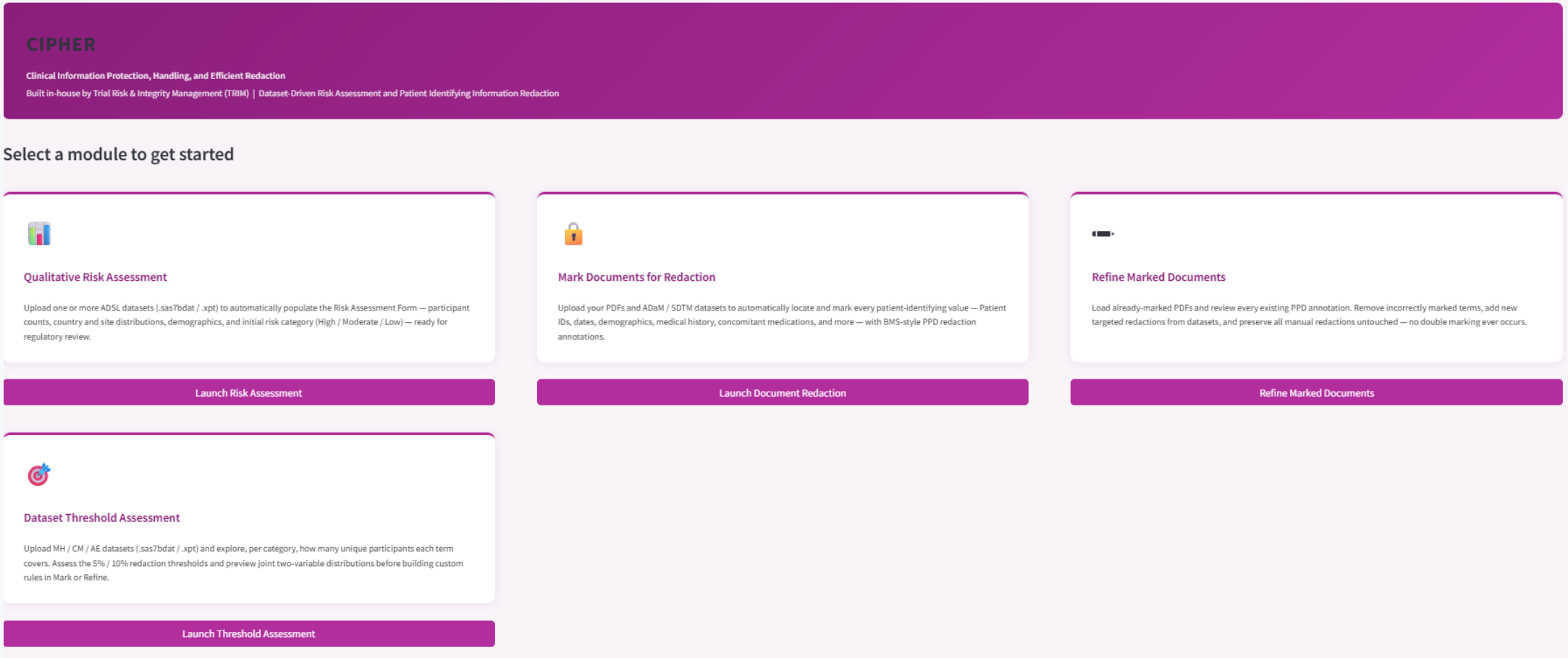
Refine

Review what was marked, remove the unwanted, add what is missing.

THE TOOL

Inside the tool

One interface — upload, assess, run, and review

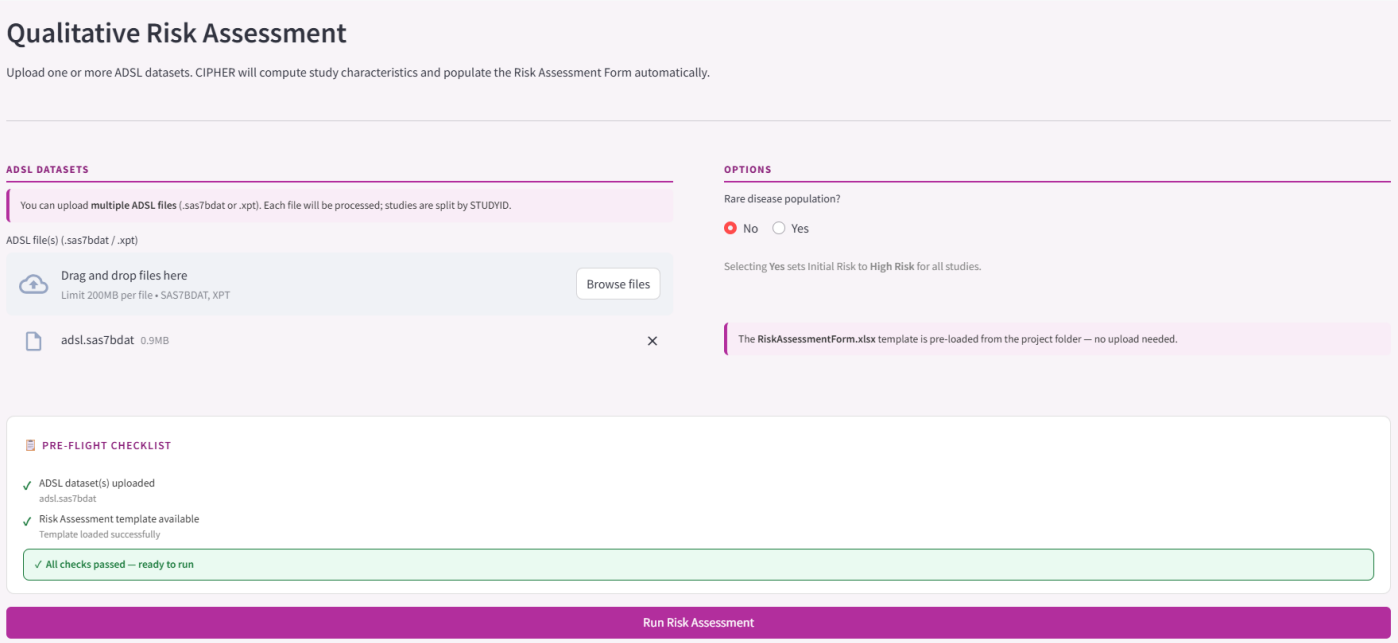


THE TOOL

Risk Assessment

Start with the study — the dataset fills the form.

- **Upload the subject dataset**
  One or more ADSL files — several studies can run together.
- **Counts done for you**
  Participants, countries, sites and demographics are tallied automatically.
- **An initial risk category**
  The tool proposes High, Moderate or Low for the reviewer to confirm.
- **A form, not a blank page**
  What used to be a manual tally arrives ready to check.

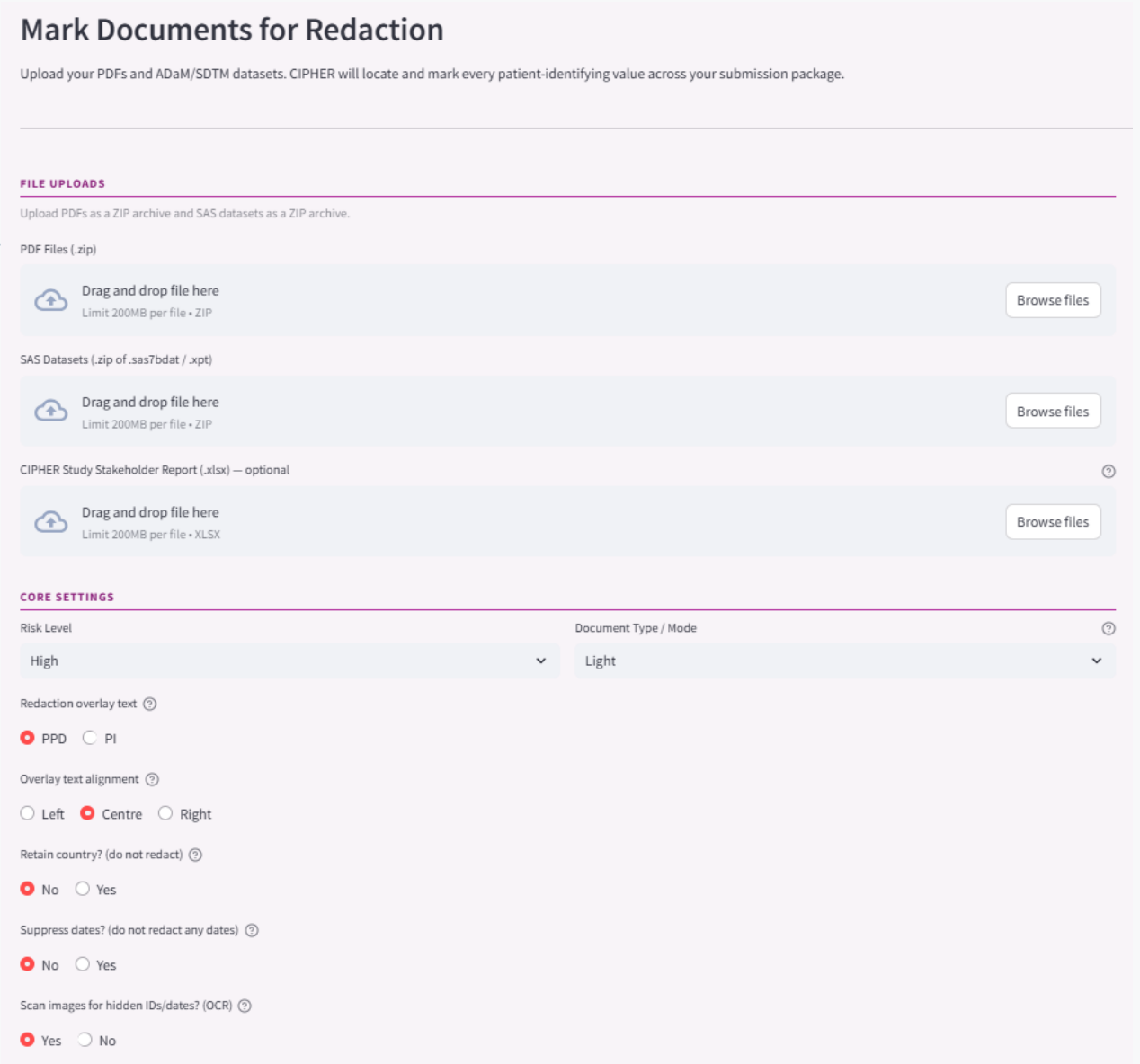


THE TOOL

Mark Documents

Choose exactly what to protect — and how it looks.

- **Pick the categories**
  Age, ethnicity, country, dates, medical history, medications and more — switched on as the document needs.
- **Set the marking properties**
  Choose the label and appearance, so every mark matches the house standard.
- **Add personnel**
  Point to the contact report to mark names, emails and phone numbers.
- **One run, a marked copy**
  The datasets drive it; every mark is listed for review.



SALIENT FEATURE

One tool, three depths

The same engine, scaled to what the document actually needs.

Light

Participant identifiers and dates only — where context is not a concern.

Heavy

Adds country, age, demographics, history, medications and events.

Personnel

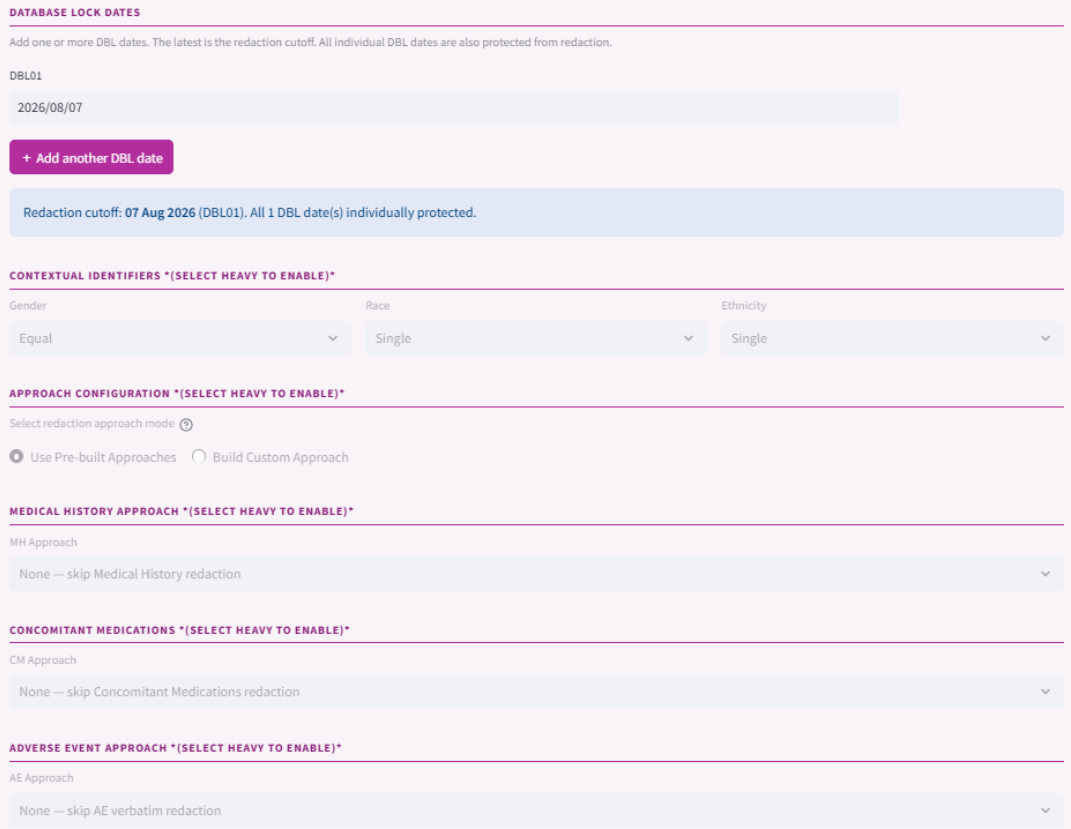
Study staff contact details — available alongside any mode.

THE TOOL

Mark Documents

Choose exactly what to protect — and how it looks.

- **Pick the categories**
  Age, ethnicity, country, dates, medical history, medications and more — switched on as the document needs.
- **Set the marking properties**
  Choose the label and appearance, so every mark matches the house standard.
- **Add personnel**
  Point to the contact report to mark names, emails and phone numbers.
- **One run, a marked copy**
  The datasets drive it; every mark is listed for review.

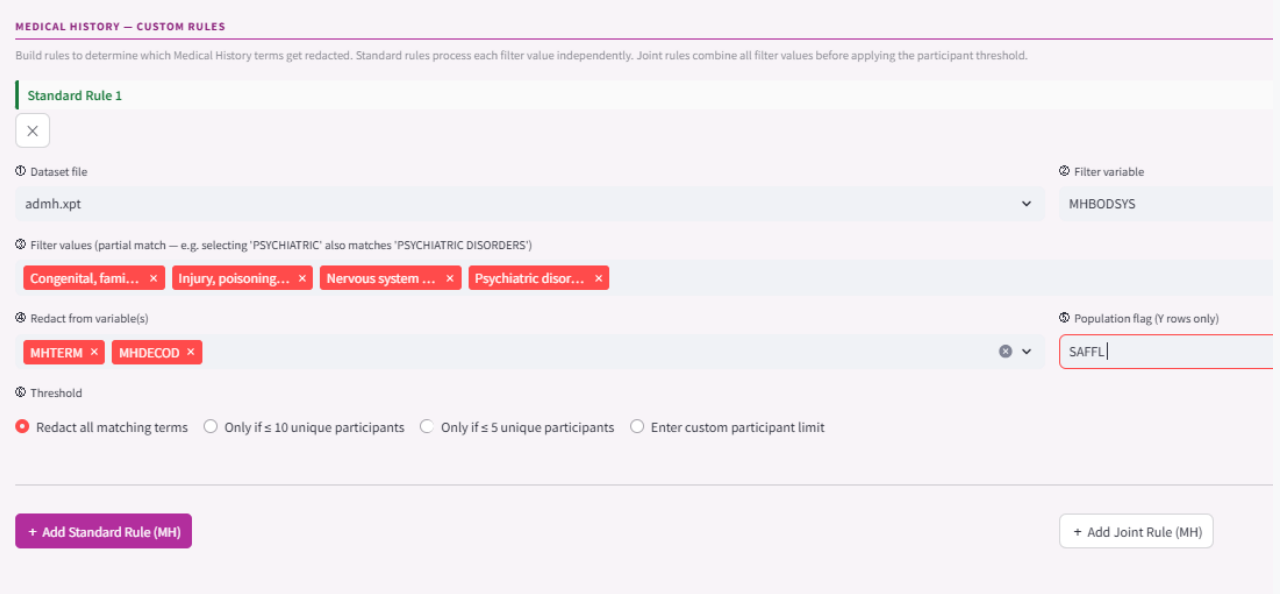


SALIENT FEATURE

Tailoring Medical History

Ready-made approaches, or terms you choose for the study.

- **Coded by body system**
  CIPHER groups medical history the way the data is organised — by body system and sub-category.
- **Choose the sensitivity**
  Pick a preset approach that includes or leaves out sensitive groupings, so redaction fits the document.
- **Add your own terms**
  Bring in study-specific conditions that should always be protected.
- **Only what identifies**
  Threshold-driven — conditions common enough not to single anyone out are left readable.

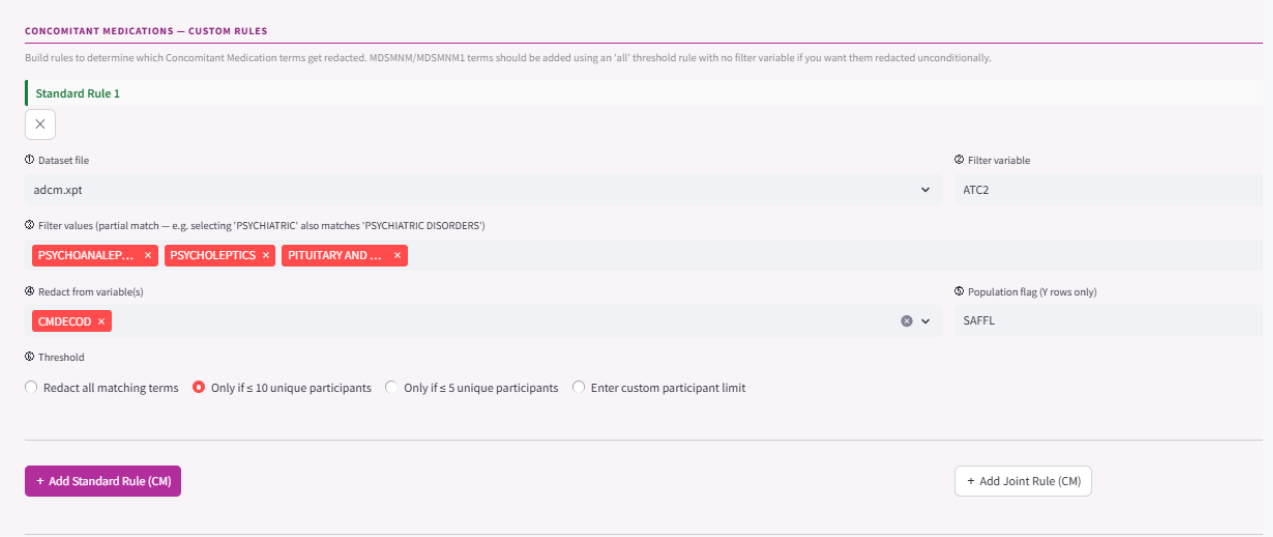


SALIENT FEATURE

Tailoring Concomitant Medications

The same evidence-led control, applied to medications.

- **Grouped by drug class**
  Medications are handled by their drug classification, not one name at a time.
- **Preset rules**
  Choose a ready-made approach; certain medication-name fields are always protected.
- **Add your own names**
  Include study-specific medications that should always be marked.
- **Only what identifies**
  Threshold-driven — widely used medications stay readable; rare ones are protected.



SALIENT FEATURE

Rules built by the Dataset Threshold Reviewer

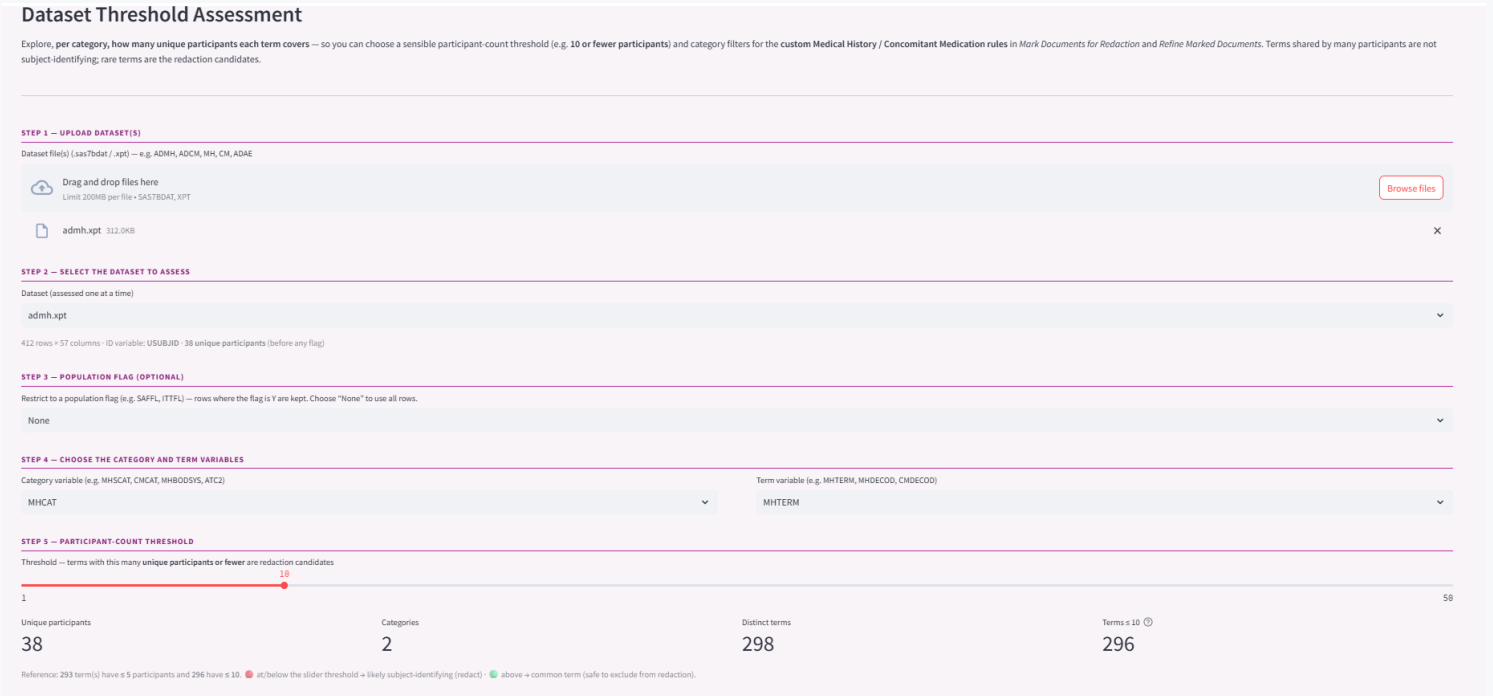
No two studies are alike — so the reviewer sets the line, on evidence.

1. **Look first**
   See how many participants sit behind each term.

2. **Decide the line**
   Judge which terms are rare enough to identify.

3. **Build the rule**
   Choose categories & terms — preset or custom.

4. **Apply & confirm**
   Run across documents; every mark is listed.



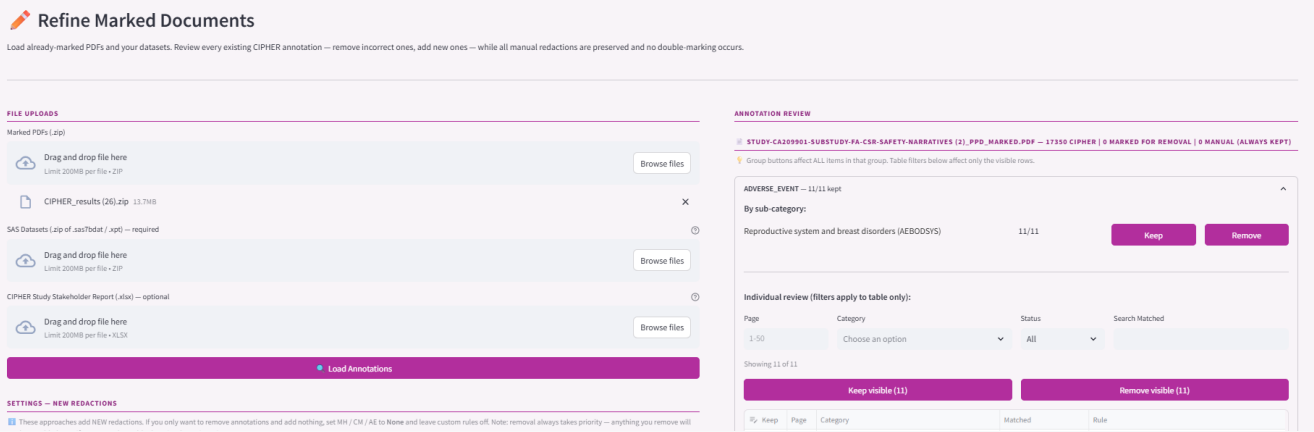
The reviewer keeps the judgement. CIPHER takes over the reading, counting and marking.

SALIENT FEATURE

Refine: the second pass

Marking is rarely the last word — so review is built in.

- **See every mark**
  Each one shown with its page, its category and the dataset value behind it.
- **Remove the unwanted**
  Deselect a single mark, or a whole category, in one action.
- **Add what is missing**
  Bring in a newer dataset and mark only the new values.
- **Manual work is safe**
  Marks added by hand are preserved exactly — nothing is ever marked twice.



Results

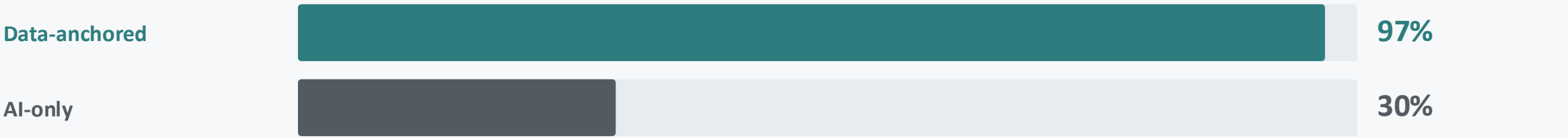
Scale, accuracy, and what the team gets back

THE PROOF

AI vs data-anchored — the metrics

Same documents, measured side by side.

Accuracy on complex patient narratives



Metric	Data-anchored	AI-only
Same result every run	Yes — reproducible	Can vary
Every mark traceable	Yes — to a dataset value	Hard to evidence
Missed identifiers	Study personnel not part of CTMS reports	Patient Ids, Sensitive MH, CM
False marks	MH Terms under abbreviations, Study Personnel like PI	Diagnosis, insensitive MH and CM terms

THE PROOF

Results from the pilot

99.6%

Overall accuracy

across all modes

21,782

Identifiers caught

hits across the pilot

78

Misses

of 21,860 total marks

WHAT WE SAW

- **Scale is real** 42 files and over 16000+ pages went through in the pilot.
- **Accuracy holds on hard content** Even dense narratives ran at 99.6%.
- **Consistent across reviewers** The same study gives the same marks, whoever runs it.

*"Reviewers checked a marked copy instead of **building the marks by hand**."*

THE PROOF

Accuracy by mode

Hits, misses and accuracy for each depth of redaction.

Mode	Files	Pages	Hits	Misses	Accuracy
Light	24	13,152	540	3	**99.4%**
Heavy (narratives)	3	1,100	21,189	65	**99.7%**
Study Personnel	15	1,932	53	10	**84.1%**
Overall	42	16,184	21,782	78	**99.6%**

Accuracy = hits ÷ (hits + misses). Study Personnel is the current focus for improvement.

THE PROOF

Why anchoring wins

The same documents, two approaches.

CIPHER — anchored

- Knows exactly what to look for (Build your own approach)
- Same inputs give the same marks, every run
- Every mark traces to a dataset value
- Defensible in an audit

AI-only — unaided

- Infers what looks like an identifier
- Results can vary between runs
- Hard to evidence why a mark was made
- Difficult to defend

THE VALUE

Why it matters

Beyond the mechanics, the point is trust.

1

Complete

Every identifier traces to the datasets — nothing that matters is missed.

2

Defensible

Every mark can be explained and reproduced. A clear audit story.

3

Efficient

Reviewers check a marked document instead of building the marks by hand.

Complete protection, proven and repeatable — that is what patients and regulators are owed.

KEY TAKEAWAYS

What we learned

Five lessons from building CIPHER.

1	**Anchor to the data**	The datasets are the only reliable source of truth.
2	**Cover the variations**	The value matters more than the format it appears in.
3	**Keep humans in charge**	The tool marks; the reviewer decides.
4	**Build for repeatability**	Same inputs, same output — every time.
5	**Design in-house**	Close to the problem, quick to improve.

Thank you

The datasets are the tree. Every apple traces home.