

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

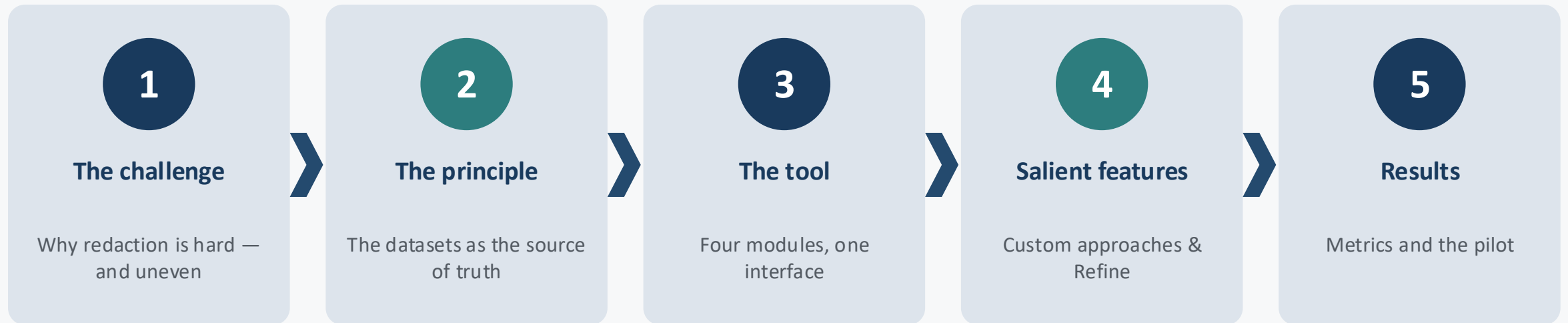
Clinical Trial Disclosure Manager II

Bristol-Myers Squibb

PHUSE 2026

How this talk flows

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



The redaction challenge

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



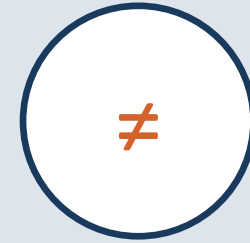
Volume

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.

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.

CIPHER at a glance

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



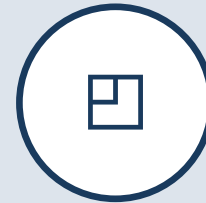
Risk Assessment

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



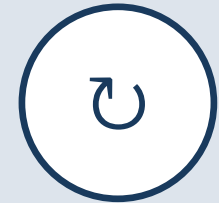
Mark Documents

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.

Inside the tool

One interface — upload, assess, run, and review

CIPHER

Clinical Information Protection, Handling, and Efficient Redaction
Built in-house by Trial Risk & Integrity Management (TRIM) | Dataset-Driven Risk Assessment and Patient Identifying Information Redaction

Select a module to get started



Qualitative Risk Assessment

Upload one or more ADaM datasets (.sas7bdat / .xpt) to automatically populate the Risk Assessment Form — participant counts, country and site distributions, demographics, and initial risk category (High / Moderate / Low) — ready for regulatory review.

Launch Risk Assessment



Mark Documents for Redaction

Upload your PDFs and ADaM / SDTM datasets to automatically locate and mark every patient-identifying value — Patient IDs, dates, demographics, medical history, concomitant medications, and more — with BMS-style PPD redaction annotations.

Launch Document Redaction



Refine Marked Documents

Load already-marked PDFs and review every existing PPD annotation. Remove incorrectly marked terms, add new targeted redactions from datasets, and preserve all manual redactions untouched — no double marking ever occurs.

Refine Marked Documents



Dataset Threshold Assessment

Upload MH / CM / AE datasets (.sas7bdat / .xpt) and explore, per category, how many unique participants each term covers. Assess the 5% / 10% redaction thresholds and preview joint two-variable distributions before building custom rules in Mark or Refine.

Launch Threshold Assessment

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.

Qualitative Risk Assessment

Upload one or more ADSL datasets. CIPHER will compute study characteristics and populate the Risk Assessment Form automatically.

ADSL DATASETS

You can upload multiple ADSL files (.sas7bdat or .xpt). Each file will be processed; studies are split by STUDYID.

ADSL file(s) (.sas7bdat / .xpt)

Drag and drop files here
Limit 200MB per file • SAS7BDAT, XPT

Browse files

adsl.sas7bdat 0.9MB

OPTIONS

Rare disease population?

☒ No ☐ Yes

Selecting Yes sets Initial Risk to High Risk for all studies.

The RiskAssessmentForm.xlsx template is pre-loaded from the project folder — no upload needed.

PRE-FLIGHT CHECKLIST

- ✓ ADSL dataset(s) uploaded
adsl.sas7bdat
- ✓ Risk Assessment template available
Template loaded successfully

✓ All checks passed — ready to run

Run Risk Assessment

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.

Mark Documents for Redaction

Upload your PDFs and ADaM/SDTM datasets. CIPHER will locate and mark every patient-identifying value across your submission package.

FILE UPLOADS

Upload PDFs as a ZIP archive and SAS datasets as a ZIP archive.

PDF Files (.zip)

Drag and drop file here
Limit 200MB per file • ZIP

Browse files

SAS Datasets (.zip of .sas7bdat / .xpt)

Drag and drop file here
Limit 200MB per file • ZIP

Browse files

CIPHER Study Stakeholder Report (.xlsx) — optional

Drag and drop file here
Limit 200MB per file • XLSX

Browse files

CORE SETTINGS

Risk Level

Document Type / Mode

High

Light

Redaction overlay text ⓘ

☒ PPD ☐ PI

Overlay text alignment ⓘ

☐ Left ☒ Centre ☐ Right

Retain country? (do not redact) ⓘ

☒ No ☐ Yes

Suppress dates? (do not redact any dates) ⓘ

☒ No ☐ Yes

Scan images for hidden IDs/dates? (OCR) ⓘ

☒ Yes ☐ No

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.

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.

DATABASE LOCK DATES

Add one or more DBL dates. The latest is the redaction cutoff. All individual DBL dates are also protected from redaction.

DBL01

2026/08/07

+ Add another DBL date

Redaction cutoff: 07 Aug 2026 (DBL01). All 1 DBL date(s) individually protected.

CONTEXTUAL IDENTIFIERS *(SELECT HEAVY TO ENABLE)*

Gender	Race	Ethnicity
Equal	Single	Single

APPROACH CONFIGURATION *(SELECT HEAVY TO ENABLE)*

Select redaction approach mode ⓘ

☒ Use Pre-built Approaches ☐ Build Custom Approach

MEDICAL HISTORY APPROACH *(SELECT HEAVY TO ENABLE)*

MH Approach

None — skip Medical History redaction

CONCOMITANT MEDICATIONS *(SELECT HEAVY TO ENABLE)*

CM Approach

None — skip Concomitant Medications redaction

ADVERSE EVENT APPROACH *(SELECT HEAVY TO ENABLE)*

AE Approach

None — skip AE verbatim redaction

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.

MEDICAL HISTORY — CUSTOM RULES

Build rules to determine which Medical History terms get redacted. Standard rules process each filter value independently. Joint rules combine all filter values before applying the participant threshold.

Standard Rule 1

Dataset file

admh.xpt

Filter variable

MHBODSYS

Filter values (partial match — e.g. selecting 'PSYCHIATRIC' also matches 'PSYCHIATRIC DISORDERS')

Congenital, fami... x Injury, poisoning... x Nervous system ... x Psychiatric disor... x

Redact from variable(s)

MHTERM x MHDECOD x

Population flag (Y rows only)

SAFFL

Threshold

☒ Redact all matching terms ☐ Only if ≤ 10 unique participants ☐ Only if ≤ 5 unique participants ☐ Enter custom participant limit

+ Add Standard Rule (MH)

+ Add Joint Rule (MH)

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.

CONCOMITANT MEDICATIONS — CUSTOM RULES

Build rules to determine which Concomitant Medication terms get redacted. MDSNM/MDSNM1 terms should be added using an 'all' threshold rule with no filter variable if you want them redacted unconditionally.

Standard Rule 1

✕

④ Dataset file: Filter variable:

④ Filter values (partial match — e.g. selecting 'PSYCHIATRIC' also matches 'PSYCHIATRIC DISORDERS'):

✕ ✕ ✕

④ Redact from variable(s): ✕ Population flag (Y rows only):

④ Threshold:

☐ Redact all matching terms ☒ Only if ≤ 10 unique participants ☐ Only if ≤ 5 unique participants ☐ Enter custom participant limit

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.

Dataset Threshold Assessment

Explore, per category, how many unique participants each term covers — so you can choose a sensible participant-count threshold (e.g. 10 or fewer participants) and category filters for the custom Medical History / Concomitant Medication rules in Mark Documents for Redaction and Refine Marked Documents. Terms shared by many participants are not subject-identifying; rare terms are the redaction candidates.

STEP 1 — UPLOAD DATASET(S)

Dataset file(s) (.sas7bdat / .xpt) — e.g. ADMH, ADCM, MH, CM, ADAE

Drag and drop files here

Limit 200MB per file • SAS7BODAT.XPT

Browse files

admh.xpt 312.0KB

STEP 2 — SELECT THE DATASET TO ASSESS

Dataset (assessed one at a time)

admh.xpt

412 rows × 57 columns • ID variable: USUBJID • 38 unique participants (before any flag)

STEP 3 — POPULATION FLAG (OPTIONAL)

Restrict to a population flag (e.g. SAFFL, ITTFL) — rows where the flag is Y are kept. Choose "None" to use all rows.

None

STEP 4 — CHOOSE THE CATEGORY AND TERM VARIABLES

Category variable (e.g. MHSCAT, CMCAT, MHBODSYS, ATC2)

MHSCAT

Term variable (e.g. MHTERM, MHDECOO, CHDECOO)

MHTERM

STEP 5 — PARTICIPANT-COUNT THRESHOLD

Threshold — terms with this many unique participants or fewer are redaction candidates

18

1

58

Unique participants

38

Categories

2

Distinct terms

298

Terms ≤ 10 ⓘ

296

Reference: 293 term(s) have ≤ 5 participants and 296 have ≤ 10. ● at/below the slider threshold → likely subject-identifying (redact) • ● above → common term (safe to exclude from redaction).

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

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.

✏ Refine Marked Documents

Load already-marked PDFs and your datasets. Review every existing CIPHER annotation — remove incorrect ones, add new ones — while all manual redactions are preserved and no double-marking occurs.

FILE UPLOADS

Marked PDFs (.zip)

Drag and drop file here
Limit: 200MB per file • ZIP

Browse files

CIPHER_results (26).zip 13.7MB

✕

SAS Datasets (.zip of sas7bdat / .xpt) — required

Drag and drop file here
Limit: 200MB per file • ZIP

Browse files

CIPHER Study Stakeholder Report (.xlsx) — optional

Drag and drop file here
Limit: 200MB per file • XLSX

Browse files

Load Annotations

SETTINGS — NEW REDACTIONS

These approaches add NEW redactions. If you only want to remove annotations and add nothing, set MH / CH / AE to None and leave custom rules off. Note: removal always takes priority — anything you remove will

ANNOTATION REVIEW

STUDY-CA209901-SUBSTUDY-FA-CSA-SAFETY-NARRATIVES (2), PPD_MARKED.PDF — 17350 CIPHER | 0 MARKED FOR REMOVAL | 0 MANUAL (ALWAYS KEPT)

Group buttons affect ALL items in that group. Table filters below affect only the visible rows.

ADVERSE_EVENT — 11/11 kept

By sub-category:

Reproductive system and breast disorders (AEBOOYS)

11/11

Keep

Remove

Individual review (filters apply to table only):

Page Category Status Search Matched
1/50 Choose an option All

Showing 11 of 11

Keep visible (11)

Remove visible (11)

Keep Page Category Matched Rule

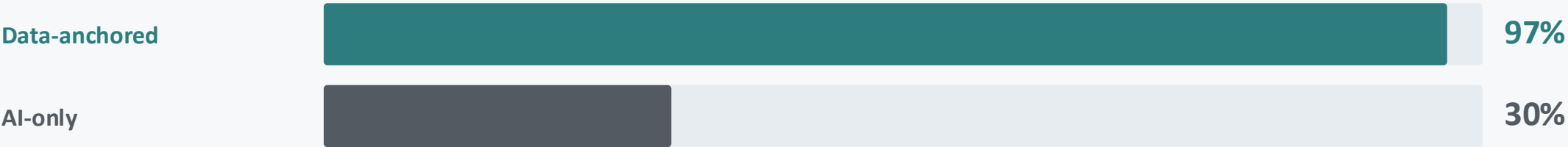
Results

Scale, accuracy, and what the team gets back

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

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.**”*

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.

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

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.

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.

Bristol-Myers Squibb. PHUSE 2026