

Presentation SA08

Harnessing AI/ML Technologies to Expedite Anonymization of Unstructured Data: Implications for EU CTR

Niamh McGuinness, PhD

Sr. Advisor, Clinical Trial Transparency and Privacy

Muhammad Oneeb Rehman Mian , PhD

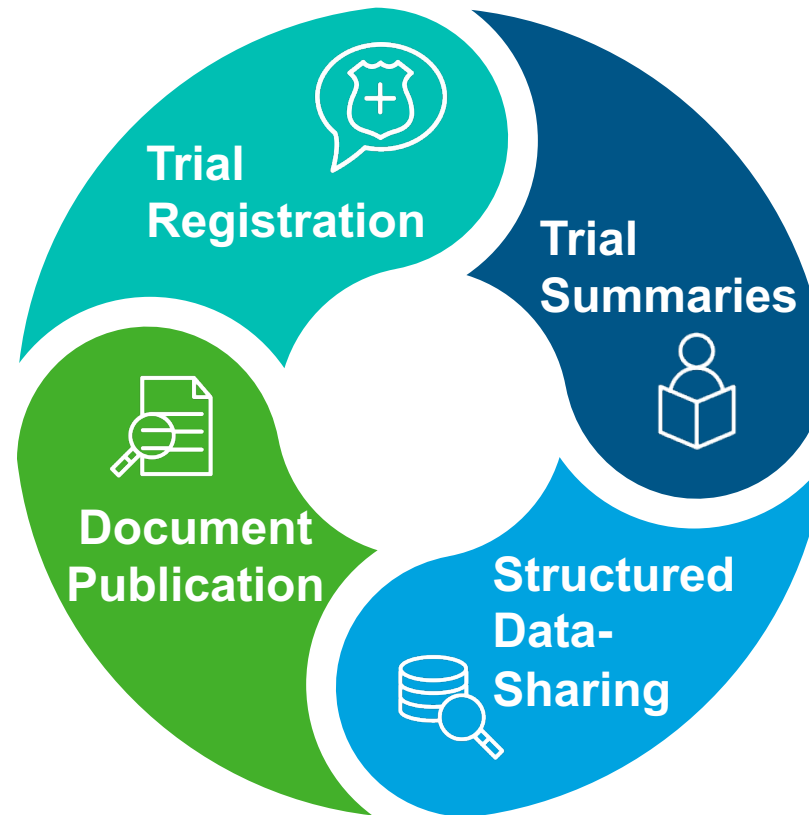
Sr. Manager, Data Enablement and Privacy

With thanks to David Di Valentino, PhD, AI/ML Solutions Lead

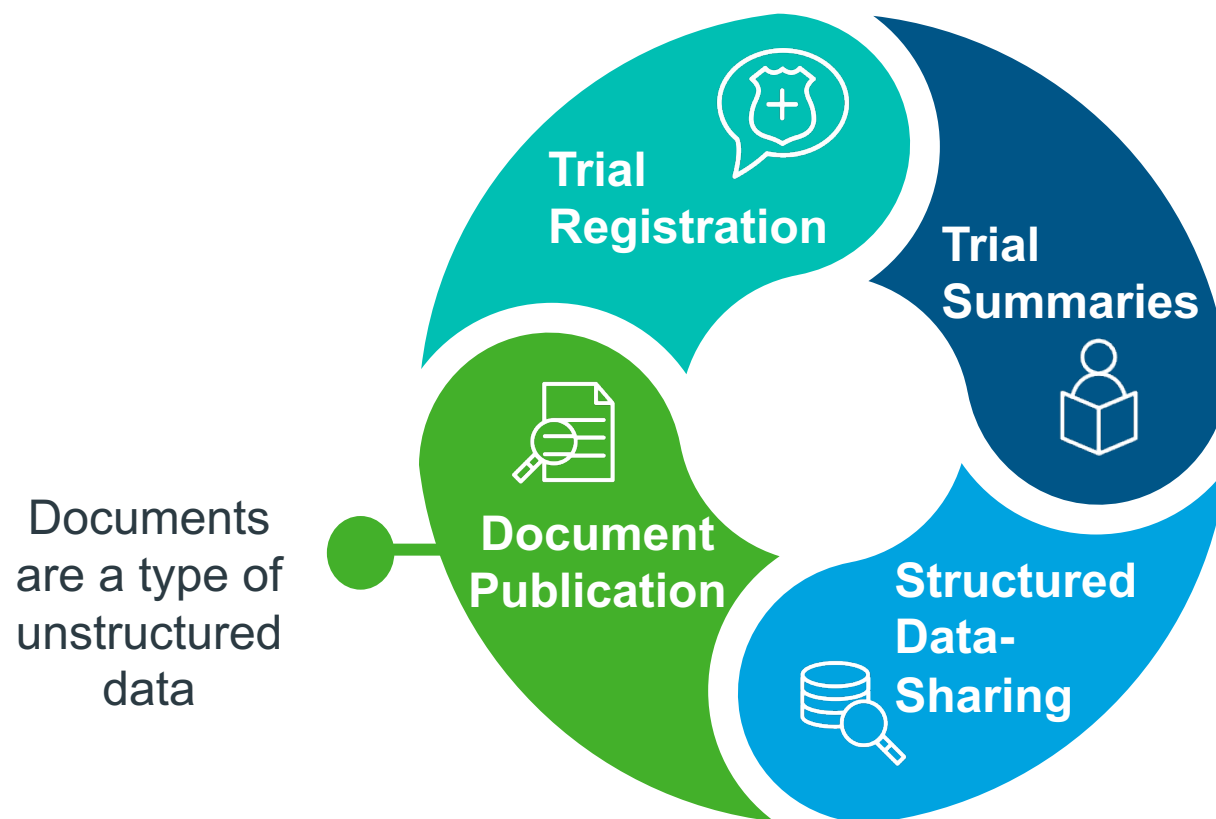


Introduction to Clinical Trial Transparency

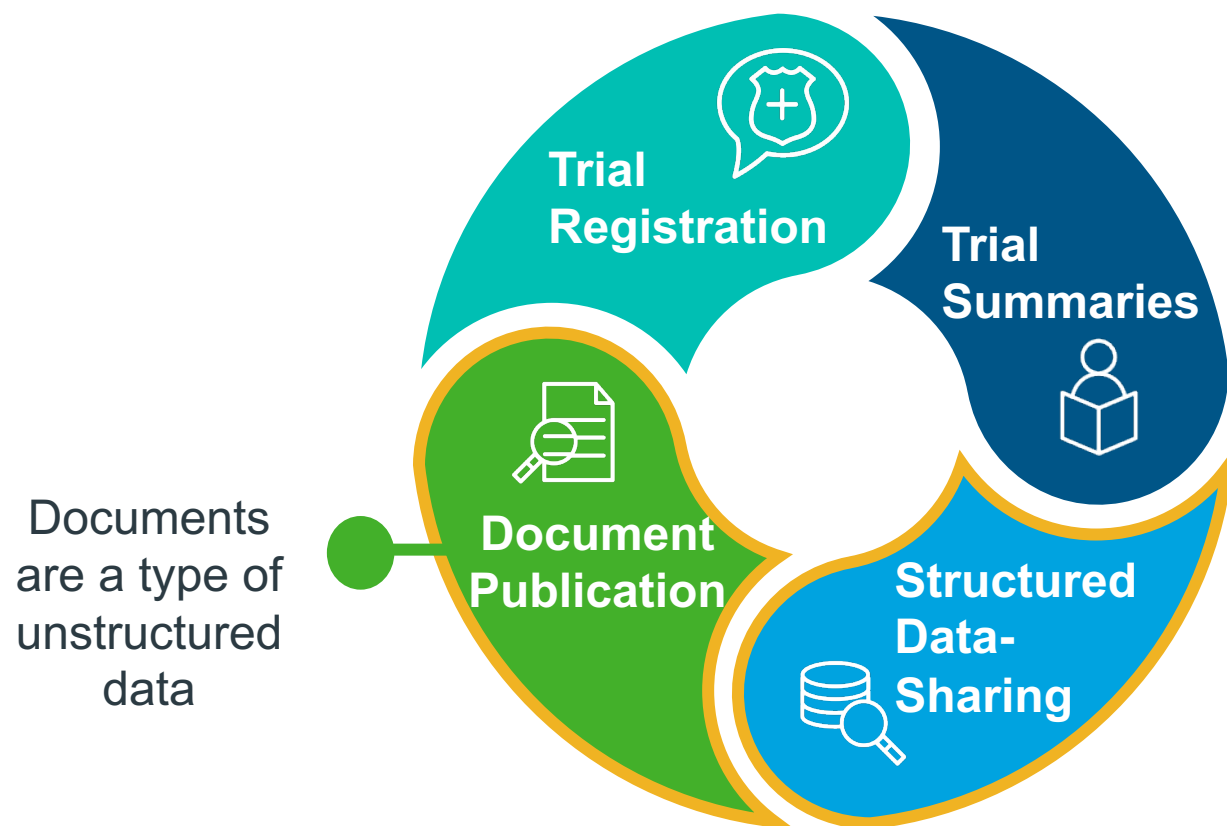
The Context: Clinical Trial Transparency and Disclosure



The Context: Clinical Trial Transparency and Disclosure



The Context: Clinical Trial Transparency and Disclosure



Both require protection of personal and/or commercially-confidential information



How EU CTR Fits into the Clinical Trial Transparency Landscape

Requirements for document publication



EMA Clinical Data Publication

- Proactive publication of clinical docs from central drug MAs from 2015
- Personal and commercially-confidential info must be redacted
- *On pause for non-COVID-related dossiers*

Requirements for document publication



EMA Clinical Data Publication

- Proactive publication of clinical docs from central drug MAs from 2015
- Personal and commercially-confidential info must be redacted
- *On pause for non-COVID-related dossiers.*



Health Canada PRCI

- Proactive publication of docs from drug or med. dev. submissions from 2019, and on-request for pre-2019
- Confidential business info (CBI) must be redacted
- Personal info must be redacted or statistically anonymized

Requirements for document publication



EMA Clinical Data Publication

- Proactive publication of clinical docs from central drug MAs from 2015
- Personal and commercially-confidential info must be redacted
- *On pause for non-COVID-related dossiers*



Health Canada PRCI

- Proactive publication of docs from drug or med. dev. submissions from 2019, and on-request for pre-2019
- Confidential business info (CBI) must be redacted
- Personal info must be redacted or statistically anonymized



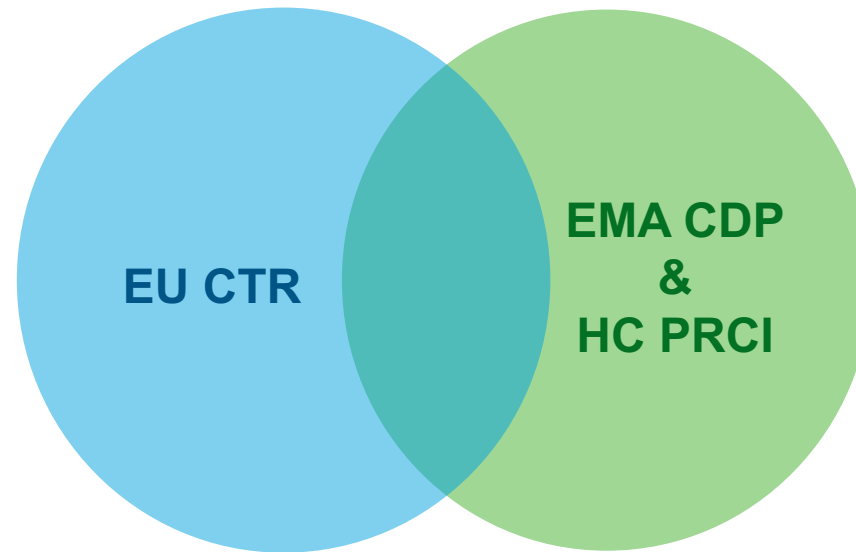
EU CTR (CTIS)

- Most clinical trial information generated during the lifecycle of a clinical trial from 31st January 2022
- CCI must be redacted
- Personal info must be redacted or statistically anonymized

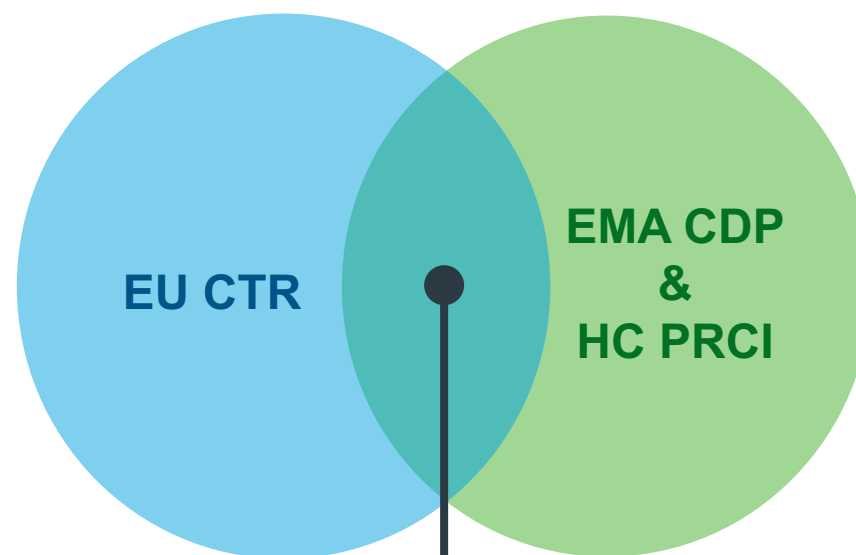
Mandatory Compliance for New CTAs – January 31st, 2023



Document Types Eligible for Publication

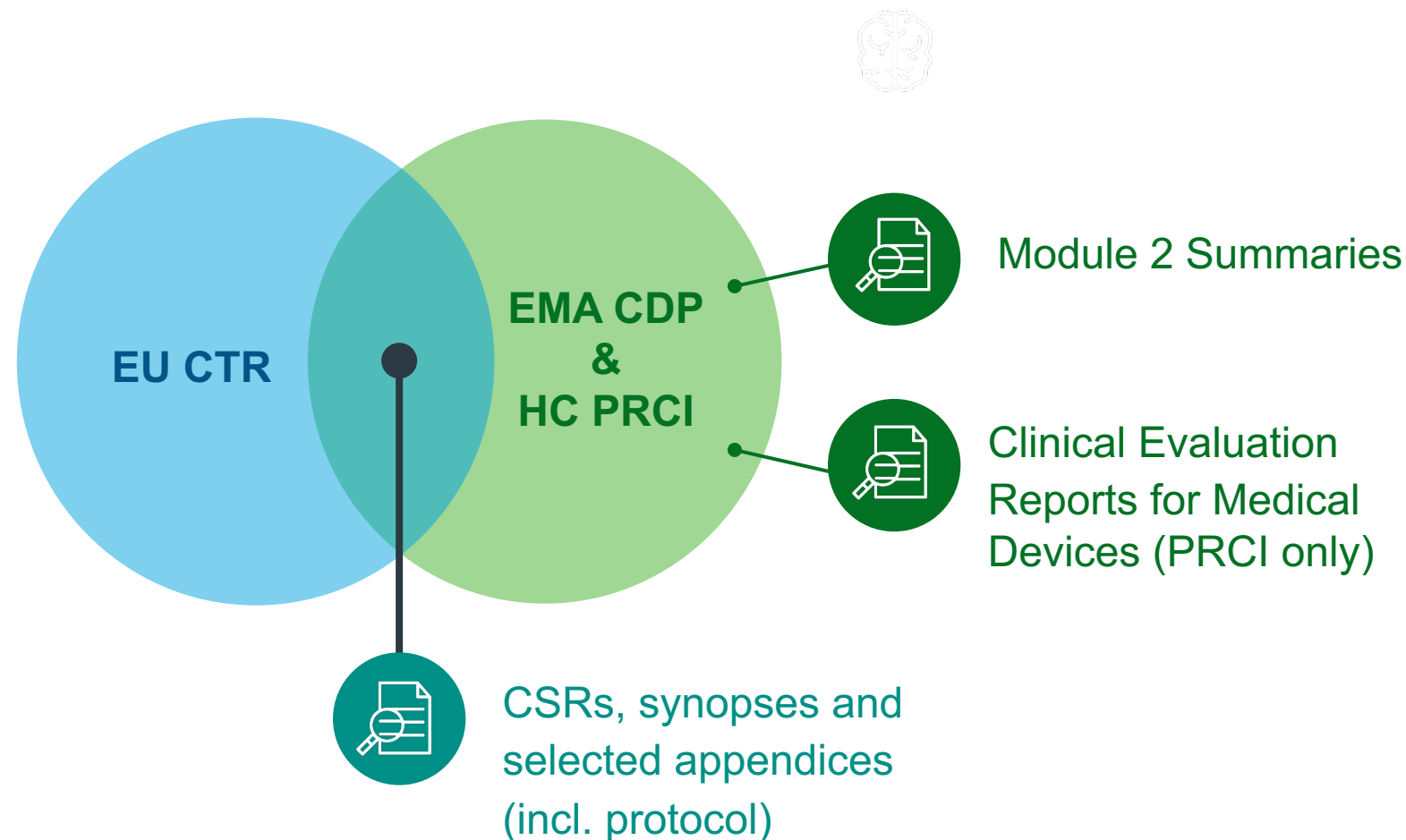


Document Types Eligible for Publication

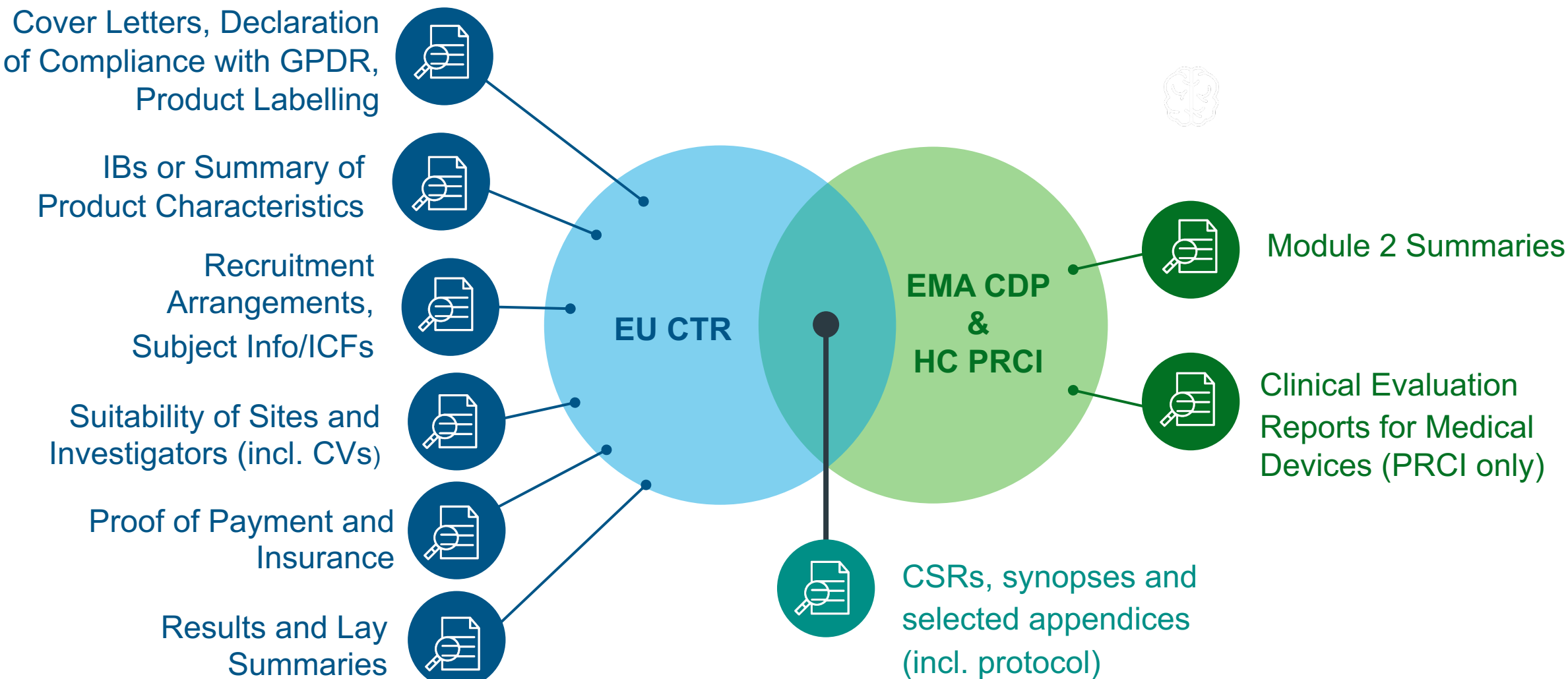


CSRs, synopses and
selected appendices
(incl. protocol)

Document Types Eligible for Publication



Document Types Eligible for Publication

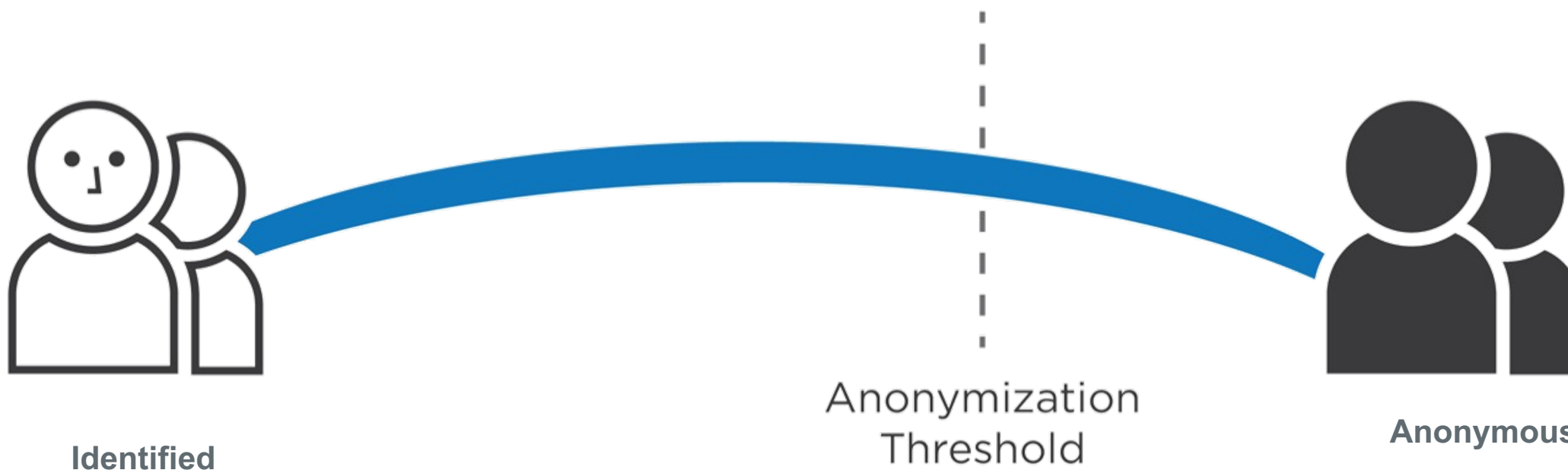




Introduction to Anonymization

Anonymization

Specific guidance **varies between regulators**, but broadly speaking, we can consider **anonymization** to mean rendering data subjects **non-identifiable/indistinguishable** through the **transformation or removal of identifying elements of personal information**.



Anonymization: Where to Start



What to
include



Anonymization: Where to Start



What to
include



How to
assess

Anonymization: Where to Start



What to
include

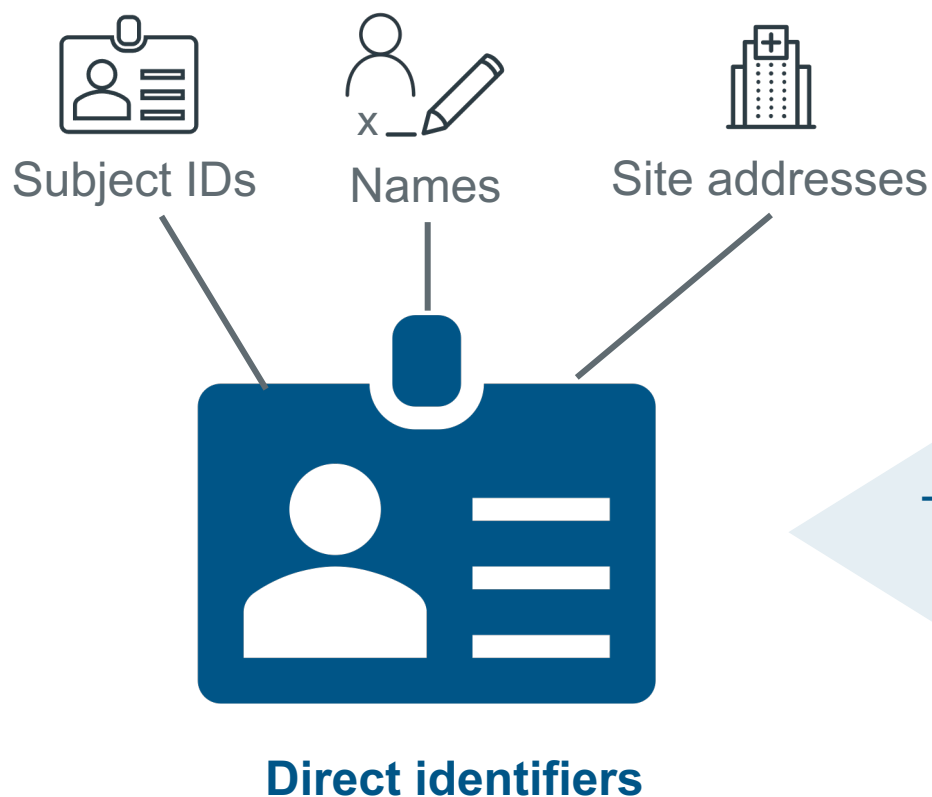


How to
assess



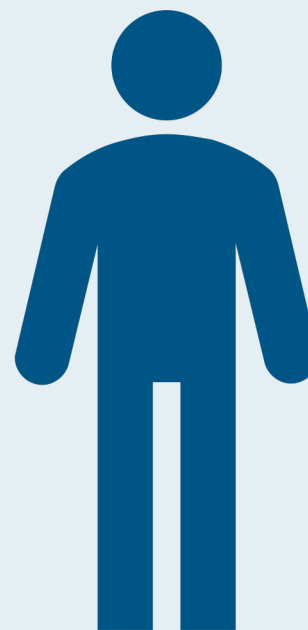
How to
mitigate

What to include – participants



Typical indirect identifiers include ...

Indirect identifiers



Age (or date of birth)



Date of death *



Vital signs data (e.g., height, weight)



Medical details



Notable calendar dates



Adverse event information

* If applicable

What to include – sponsor/site personnel (more limited scope)



How to assess

Qualitative

Rules-Based, Non-Analytical



Usually driven by **observations**, **assessments** by responsible parties

Near **complete removal** of PI; may preserve some PI if deemed acceptable

Not typically considered in assessment of data subject privacy



Privacy protection



Data transformations



Data utility

Quantitative

Statistically-derived, Analytical



Driven by mathematical model of **identifiability or re-identification risk**

Adjustable based on re-identification risk; varies based on **population** under study

Balances necessary **privacy protection** with usefulness of data

How to mitigate: redaction versus statistical anonymization

Redaction

Subject 899-456 was a 34-year-old female with a medical history of coeliac disease.



Subject PPD was a PPD-year-old PPD with a medical history of PPD.

Statistical Anonymization

Subject 899-456 was a 34-year-old female with a medical history of coeliac disease.



Subject 454-994 was a 30-35-year-old female with a medical history of autoimmune disorders.



AI/ML and Anonymization

A Focus on c-NLP

Challenges of c-NLP-driven statistical anonymization

To build the correct data model:



PI must be detected with sufficiently **high accuracy** to be modelled correctly and to ensure adequate protection in the published document.

Subject 899-456 was a 34-year-old female with a medical history of coeliac disease.

Challenges of c-NLP-driven statistical anonymization

To build the correct data model:



PI must be detected with sufficiently **high accuracy** to be modelled correctly and to ensure adequate protection in the published document.

Subject 899-456 was a 34-year-old female with a medical history of coeliac disease.

Challenges of c-NLP-driven statistical anonymization

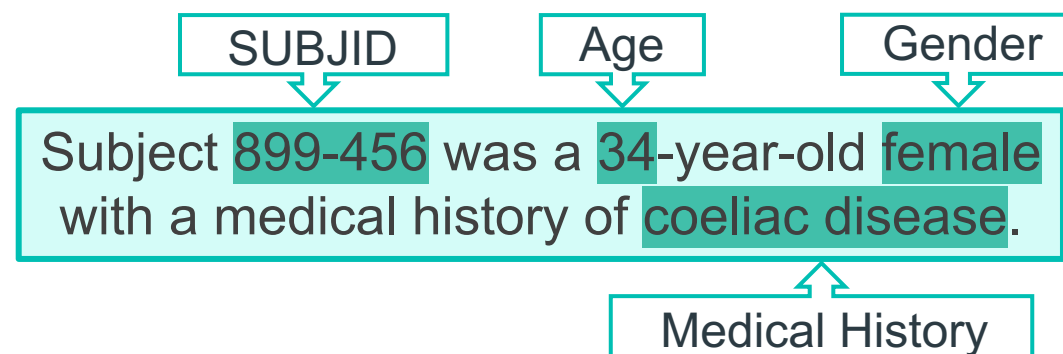
To build the correct data model:



PI must be detected with sufficiently **high accuracy** to be modelled correctly and to ensure adequate protection in the published document.



PI must be detected as the **correct type** (e.g., age, date, medical history, etc.) to be **modelled properly**



Challenges of c-NLP-driven statistical anonymization

To build the correct data model:



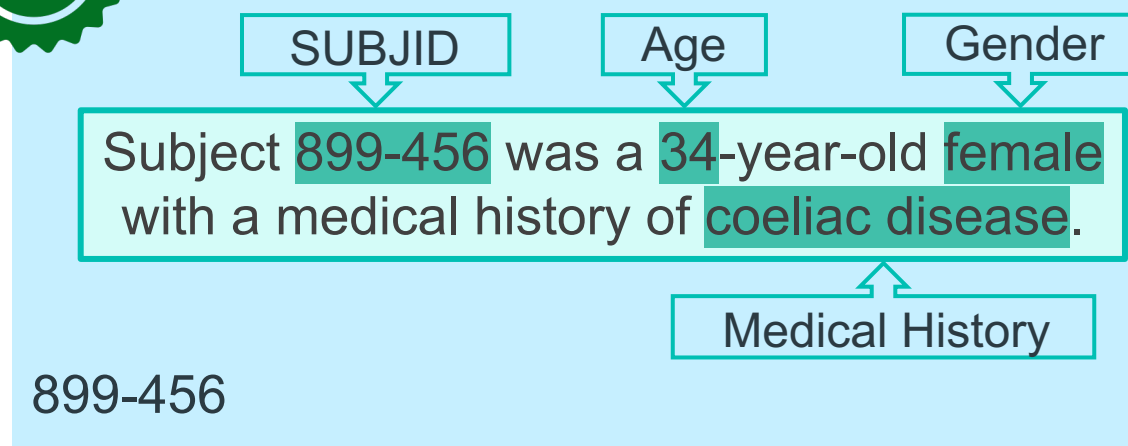
PI must be detected with sufficiently **high accuracy** to be modelled correctly and to ensure adequate protection in the published document.



PI must be detected as the **correct type** (e.g., age, date, medical history, etc.) to be **modelled properly**



Detected PI should be **correctly assigned** to the appropriate **data subject**



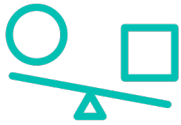
Redaction requirements in consideration of EU CTR

The use of redaction to transform documents **may substantially simplify** the path to compliance for EU CTR, as PI must only be **captured in absolute terms**.



So, we primarily require that the PI detection process is **accurate enough** to justify a determination that data subjects are **rendered anonymous**.

Redaction vs. statistical anonymization for EU CTR



The determination of whether statistical anonymization or redaction is most appropriate for transparency is ultimately within the **purview** of the **trial sponsor**.



Especially for Part 1 and Part 2 documents, which must be uploaded quickly, vary in size and format, and in which CCI will constitute a greater concern for sponsors, **redaction** is likely the best choice.



Clinical reports, which may also be eligible for publication under other transparency initiatives could be subject to **statistical anonymization** but will likely be **redacted** instead.



We anticipate that **c-NLP-driven redaction approaches** can be a more **efficient and viable option** for transparency purposes.

Redaction vs. statistical anonymization for EU CTR



The determination of whether statistical anonymization or redaction is most appropriate for transparency is ultimately within the **purview** of the **trial sponsor**.



Especially for Part 1 and Part 2 documents, which must be uploaded quickly, vary in size and format, and in which CCI will constitute a greater concern for sponsors, **redaction** is likely the best choice.



Clinical reports, which may also be eligible for publication under other transparency initiatives could be subject to **statistical anonymization** but will likely be **redacted** instead.



We anticipate that **c-NLP-driven redaction approaches** can be a more **efficient and viable option** for transparency purposes.

Redaction vs. statistical anonymization for EU CTR



The determination of whether statistical anonymization or redaction is most appropriate for transparency is ultimately within the **purview** of the **trial sponsor**.



Especially for Part 1 and Part 2 documents, which must be uploaded quickly, vary in size and format, and in which CCI will constitute a greater concern for sponsors, **redaction** is likely the best choice.



Clinical reports, which may also be eligible for publication under other transparency initiatives could be subject to **statistical anonymization** but will likely be **redacted** instead.



We anticipate that **c-NLP-driven redaction approaches** can be a more **efficient and viable option** for transparency purposes.

Redaction vs. statistical anonymization for EU CTR



The determination of whether statistical anonymization or redaction is most appropriate for transparency is ultimately within the **purview** of the **trial sponsor**.



Especially for Part 1 and Part 2 documents, which must be uploaded quickly, vary in size and format, and in which CCI will constitute a greater concern for sponsors, **redaction** is likely the best choice.



Clinical reports, which may also be eligible for publication under other transparency initiatives could be subject to **statistical anonymization** but will likely be **redacted** instead.



We anticipate that **c-NLP-driven redaction approaches** can be a more **efficient and viable option** for transparency purposes.

Considerations for different AI/ML technologies

We've found that a **hybrid model** tends to be most effective for detecting PI in clinical text:



Leverage **large, pre-trained language models** for contextual understanding of text



Fine-tune on **real-world data** from completed anonymization projects

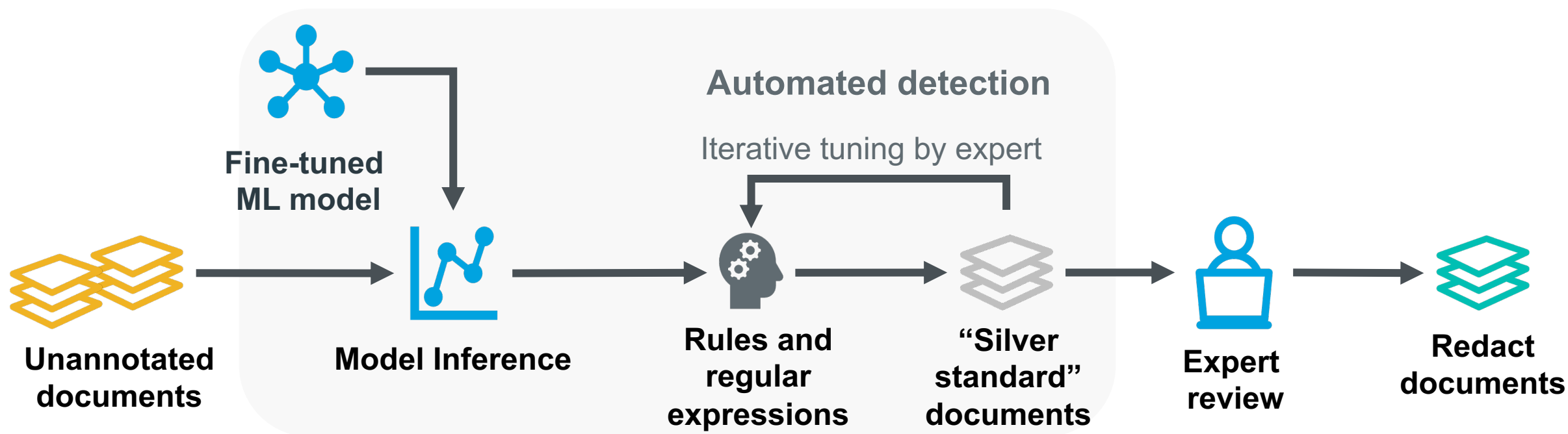


Apply **regular expressions** to detect attributes with known formatting (dates, ages, etc.)



Iteratively improve rules to target efficiency gains with each submission

C-NLP-driven PI redaction workflow for EU CTR



Redaction involves the **detection and masking** of nearly all **personal information** pertaining to participants in a clinical study.

Typically, in order to achieve this goal, the PI detection task requires some **expert input** to iteratively tune rules / regular expressions and **redact any remaining identifiers** missed by automated detection.

Testing our redaction approach on EU CTR documents

Documents tested:



Investigator Brochure

- Summarizes the **body of information** about an investigational product obtained during a **trial**



Protocol

- Outlines the **rationale, objective(s), methodology used, and analysis plan** for a clinical study



Investigator CVs

- Presents investigator eligibility and **experience.**



Clinical Study Reports

- Provides detail (including individual **participant data**) about the **safety and efficacy results** of a clinical trial.

Testing our redaction approach on EU CTR documents

Observations

 Investigator Brochure • ~100-300 pages • Some contain small participant narratives:  Participant/case ID  Age  Country  Medical history	 Protocol • ~100-300 pages • Usually contain some personnel info:  Name  Contact details  Address	 Investigator CVs • ~2-5 pages • Usually contain some personnel info:  Name  Contact details  Address	 Clinical Study Reports • ~100-100,000 pages • Rich participant info across document:  Subject ID  Age, Key Dates  Demography & VS  Medical history  Adverse events
---	--	--	---

Observations on EU CTR document testing



Sections of text with **sparse PI** tend to be best handled through **manual annotation**.



ML model predictions tend to be most effective when PI has **contextual dependence**, such as in clinical narratives (IBs and CSRs)



Writing **custom regular expressions** tended to be most efficient for capturing PI in **tabular data**.



Time spent **manually correcting** documents tended to vary more strongly with **document length** than **density of PI**

Results



In a particular CSR narrative format, the c-NLP model achieved **100% detection accuracy**.



Use of custom **regular expressions** for documents with predictably formatted personnel-related information reduced manual effort by **up to 50%**.



Using a c-NLP-driven PI detection process showed a **20% improvement in accuracy** relative to a fully manual approach on documents like **Investigator Brochures**.

Conclusions



EU CTR will **substantially grow** the scope of **public clinical document releases** required of trial sponsors.



Analysts found that an **AI/ML-driven approach** allowed for **successful** redaction of documents in scope of EU CTR.



Given the increased throughput requirements, **AI/ML-driven redaction** becomes more viable to **satisfy regulatory criteria**.

Thank you!



nmcguinness@privacy-analytics.com, mrehman@privacy-analytics.com



<https://privacy-analytics.com/>



[@privacyanalytic](https://twitter.com/privacyanalytic)



<https://www.youtube.com/user/PrivacyAnalytics>



<https://www.linkedin.com/company/privacy-analytics-inc-/>

