

Data De-Identification: Balancing Patient Privacy, Data Utility and Good Programming Practices (PP30)

Dave Handelsman, Senior Director of Industry Strategy
(dave.handelsman@d-wise.com)

Introduction

In the past two years, the energy and focus on clinical trial data transparency have demonstrated the need to expedite the process of patient-level data de-identification. The historic tools to de-identify patient-level data don't behave consistently across industry, and aren't sufficiently scalable to meet the growing need for de-identified data.

To be successful in 2015 and going forward, any de-identification solution must:

Protect patient confidentiality

AND

Comply with national regulations

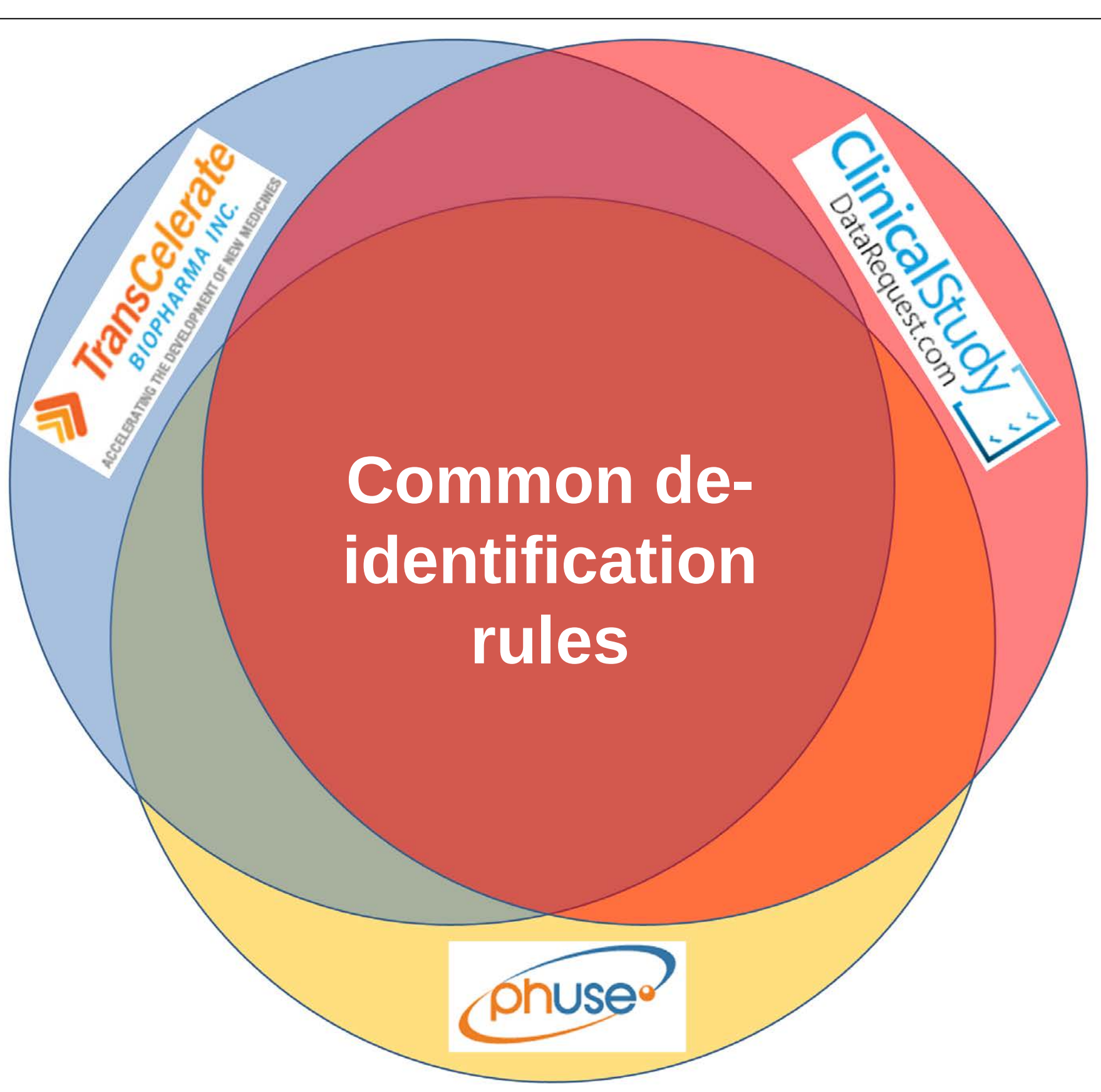
AND

Maintain data utility and potential analysis value

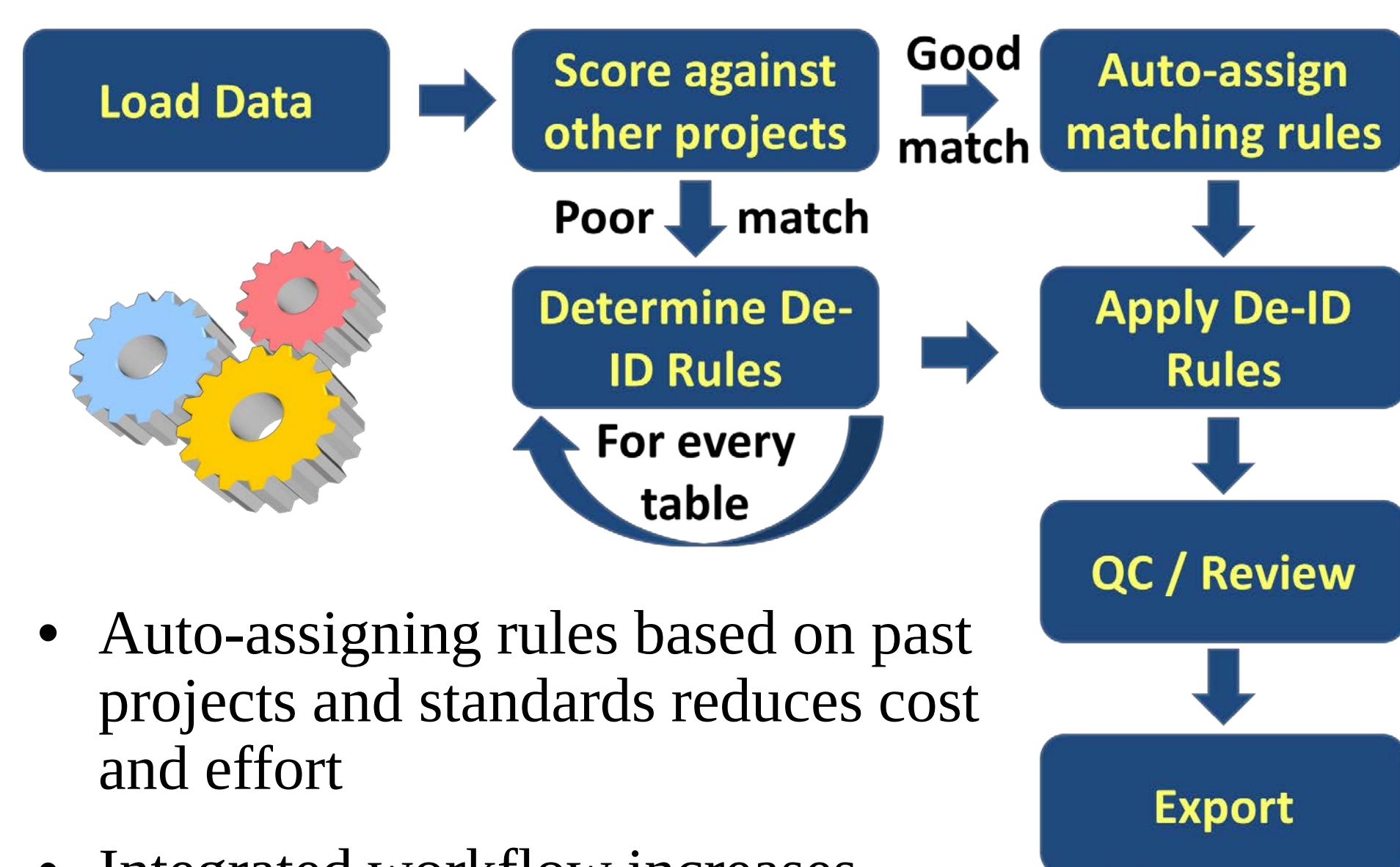
This poster will review the current state of the industry as it relates to patient-level data de-identification, assess the strategies being used and provide recommendations regarding how to bring scale, quality and efficiency to the de-identification process

De-Identification Rule Sets

There are currently three well-documented efforts regarding the development of de-identification rules for clinical trials. These rule sets all build on existing strategies related to HIPAA, Safe Harbor and Expert Determination. While there is certainly overlap across these rule sets, it's also clear that there is no current consensus regarding a canonical set of de-identification rules that are applicable across the biopharmaceutical industry.



De-Identification Workflow



- Auto-assigning rules based on past projects and standards reduces cost and effort
- Integrated workflow increases quality and reduces risk
- Flexible approach for evolving key mapping management needs is critical

Analysis – Similarities

Data Type	De-Identification Action
Personally identifiable information (PII)	Remove
Unique subject ID	Re-code
Subject ID	Re-code
Initials	Remove
Investigator names and locations	Remove
Age > 89	Remove age and categorize as Age >=90
Dictionary coded terms	Retain

Analysis – Differences

Data Type	De-Identification Approaches
Date of Birth	• Delete month and day while keeping the year • Remove the entire date value • Replace the value with a calculated age
Date fields, except for Date of Birth	• Calculate study day from designated anchor point for each patient • Apply a random offset, per patient, to each patient's date value
Free-text values	• Review manually and redact selectively • Always remove
Patients per Site, Sites per Study, Sites per Country	• Aggregate sites with few[†] patients • Aggregate countries with few[†] sites and/or patients • Avoid de-identifying single-site trials [†] "Few" patients is frequently defined as <10; "Few" sites are not defined
Variable names	• Review for location information • Ignore

Conclusions

Significant Variation in Industry

Clinical trial data transparency and sharing, and its companion process of patient-level de-identification, have been addressed with historic speed in the biopharmaceutical industry. Not surprisingly, comprehensive and consistent industry-driven guidelines, rules and best practices still need to be established.

Ongoing Collaborations

Formal industry collaborations regarding de-identification continue to progress. The PhUSE De-Identification Working Group published their de-identification guidelines for CDISC SDTM 3.2 data in Spring 2015. The details regarding this effort can be found at www.phuse.eu/Data_Transparency_download.aspx. In the same timeframe, the TransCelerate de-identification project team published their de-identification guidelines. Their details are located at www.transceleratebiopharmainc.com/assets/clinical-data-transparency/.

These newly published guidelines, like those associated with the eight biopharmaceutical companies here, are harmonized in some respects but differ in others.

Recommendations

As patient-level de-identification matures, a well-defined set of de-identification rules directly lends itself to a more streamlined approach than the time-consuming and resource-intensive SAS programming process in use today.

A robust software solution that can apply industry rules and provide flexibility to handle both historic and modern clinical trial data will be of tremendous benefit in addressing this emerging problem space. Such a solution will need to be designed to evolve with the business and regulatory needs of the industry, while never losing sight of the vital need to protect patient privacy.

References

- www.clinicalstudydatarequest.com/Documents/Astellas-Data-Anonymization-Standards.pdf
- www.clinicalstudydatarequest.com/Documents/Bayer-Data-Anonymization-Standards-Apr-14-2014.pdf
- www.clinicalstudydatarequest.com/Documents/Anonymisation of Clinical Trial Datasets-BL.pdf
- www.clinicalstudydatarequest.com/Documents/Anonymisation of clinical trial datasets Eisai.pdf
- www.clinicalstudydatarequest.com/Documents/Anonymisation of Clinical Trial Datasets.pdf (GSK)
- www.clinicalstudydatarequest.com/Documents/Anonymisation_clinicaltrialdata_Lilly_update.pdf

- www.clinicalstudydatarequest.com/Documents/Novartis-Global-Data-Anonymization-Standards.pdf
- www.clinicalstudydatarequest.com/Documents/Anonymisation of Clinical Trial Datasets.pdf (Roche)
- www.clinicalstudydatarequest.com/Documents/Sanofi-Deidentification-Guide.pdf
- www.clinicalstudydatarequest.com/Documents/Takeda-Anonymization-Standard-10Oct-2014-Final.pdf
- www.clinicalstudydatarequest.com/Documents/UCB-Anonymization-Clinical-Trial-Datasets.pdf
- www.clinicalstudydatarequest.com/Documents/Anonymisation of Clinical Trial Datasets.pdf (ViiV)

d-Wise
www.d-wise.com/blur