Paper RE03

Privacy Considerations of Using RWD from Statistical Programming Perspective

Zhouzhou (Joe) Xi, Bristol Myers Squibb, New Jersey, USA
Lorraine Fang, Bristol Myers Squibb, New Jersey, USA

ABSTRACT

With the increasing use of Real-World Data (RWD) in clinical research, privacy compliance has become a critical focus for pharmaceutical companies. This paper presents practical approaches adopted by the statistical programming team at Bristol Myers Squibb (BMS) to ensure privacy compliance when handling RWD, including company-level policies, data storage and access controls, pseudonymization, and de-identification techniques. The discussion is grounded in recent regulatory guidance and industry best practices.

INTRODUCTION

While the use of Real-World Data (RWD) offers great opportunities for medical research, it also presents complex privacy and data protection challenges. Regulatory authorities such as the FDA along with industry organizations including PHUSE, have set high standards for protecting patient privacy and ensuring secure data handling. At BMS, we receive RWD from many data vendors, each with its own set of privacy and governance requirements. This paper examines how statistical programmers can manage these diverse rules, focusing on practical strategies to protect patient confidentiality while ensuring the analytical utility of the data.

COMPANY-LEVEL POLICY ON RWD PRIVACY

BMS enforces a comprehensive company-level policy on RWD privacy, with a strong emphasis on internal controls over data use, access/share, storage, and anonymization. These policies are designed to comply with legal and regulatory requirements as well as the specific privacy requirements set by our data providers, thereby safeguarding real-world patient's privacy and identity.

For statistics and programming team, this translates clearly defined daily workflows that ensure patient identities are protected at all stages of data handling. The following sections describe the practical approaches in our daily work.

DATA STORAGE AND ACCESS CONTROLS

In accordance with BMS RWD privacy policies, data received under vendor agreements needs to be stored on dedicated and validated servers, or on standard data servers with restricted folder access.

In most cases, RWD are stored on regional servers (such as Data Commons) aligned with vendor agreement terms. These servers have been validated and restricted to only authorized users. Physical servers are located in either US, EU or other regions complying with the vendor's requirement on physical location of data storage. In general, copying of patient level data outside of designated server is prohibited unless specifical exemptions (e.g. data submission to FDA for review) are granted.

In certain circumstances, RWD must be transferred to statistical programming friendly data server to be conveniently processed (e.g. to be accessible using SAS). When RWD are stored on standard servers, additional safeguards are applied, including layered access controls, double-restricted areas and firewalls protections to separate study folders and anonymized data repositories.

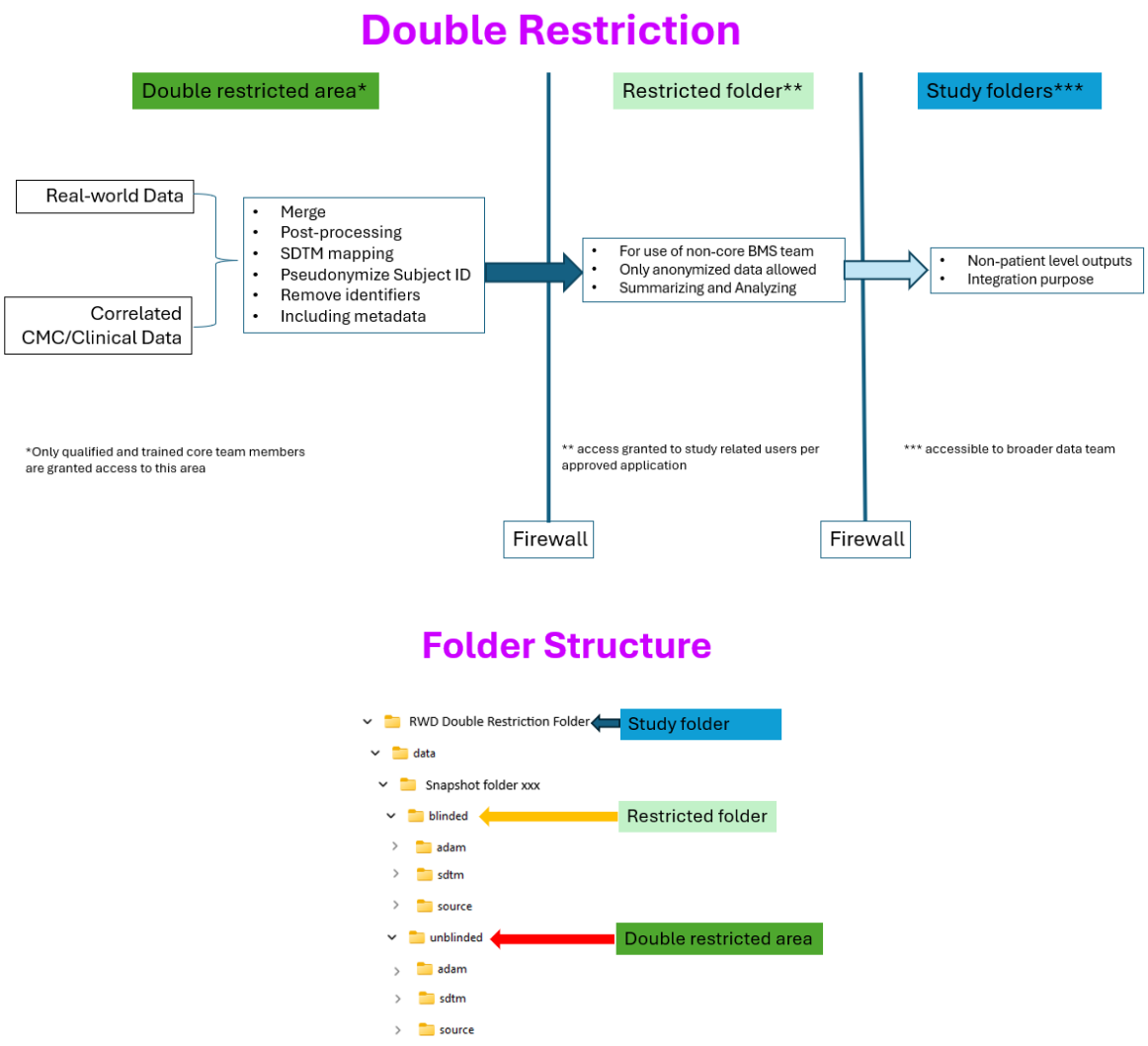

Figure 1: Double Restriction Flow Chart

Under the double-restriction model described above, a core team maintains raw data in a highly restricted 'inner' folder. Before other users or third parties can access the data, the core team must de-identify the data and transfer it to a separate, yet still secure, 'outer' folder. This approach preserves the integrity and confidentiality of the original data while enabling its appropriate use for research and analysis.

DE-IDENTIFICATION OF PATIENT PERSONAL INFORMATION

The core objective of these policies and rules is to protect patient privacy and prevent re-identification is central to regulatory compliance. Under HIPAA 18 specific identifiers must be removed or masked, including names, detailed geography (smaller than a state), and specific dates (except years), etc. Similarly, GDPR requires anonymization or pseudonymization of patient data that can directly or indirectly identify an individual, though GDPR does not provide a fixed list of identifiers to mask. To ensure compliance with these regulations, as well as BMS internal policies, the following de-identification approaches have been applied to data prior to sharing out of core team:

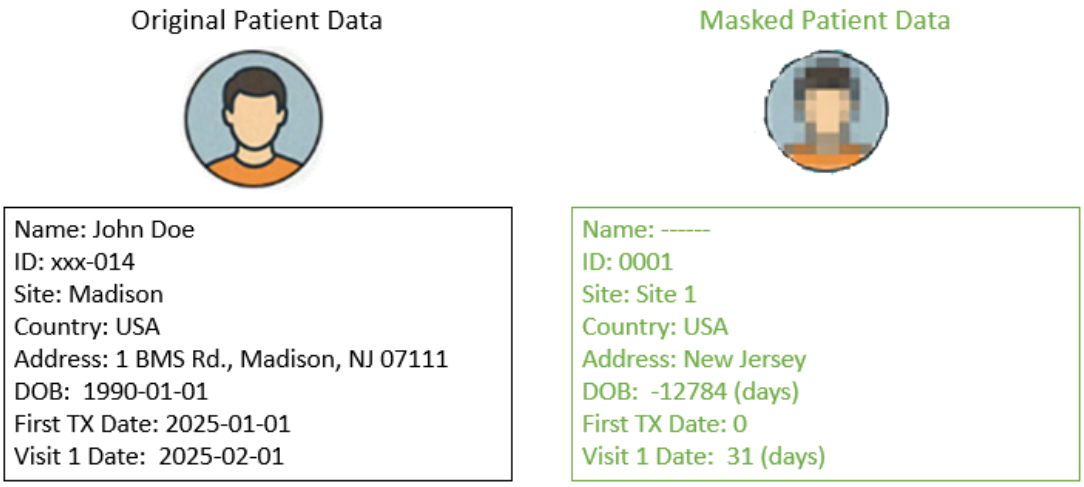

Figure 2: Sample of De-identification

Patient ID Pseudonymization: although RWD vendors typically provide masked patient identifiers, BMS implements a secondary layer of pseudonymization to bolster security for external data sharing. Prior to sharing data outside of the core team, subject IDs undergo a robust de-identification process:

- Numeric IDs are transformed via a secure mathematical algorithm.
- Alphanumeric identifiers are mapped using a one-to-one cryptographic crosswalk table.
- Algorithms or crosswalk tables stored in restricted areas.

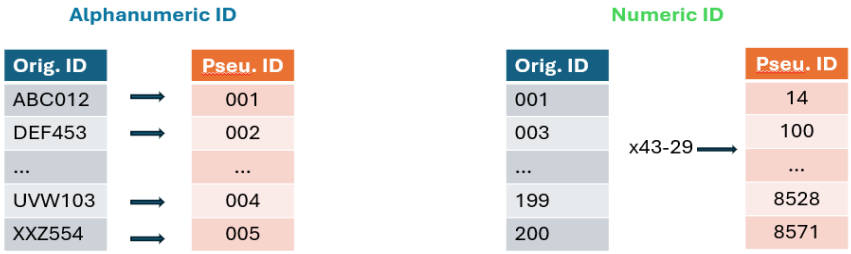

Figure 3: Sample of ID Pseudonymization

This methodology ensures data integrity and longitudinal consistency across datasets while strictly restricting access to the re-identification key to authorized personnel, preventing end-users from reversing the process.

Date Handling: To mitigate the risk of re-identification associated with absolute dates, all longitudinal data are provided in relative day format. These values are derived from a sponsor-defined reference anchor, typically the date of first study treatment (RFXSTDTC) or another significant clinical milestone. For external data sharing, absolute ISO 8601 date/time variables (e.g., --DTC, --STDTC) are removed from the SDTM datasets, and only derived relative day variables, such as --DY or --STDY, are retained, preserving temporal relationships and clinical utility while ensuring patient privacy.

Geographic Information: As specified by HIPAA, geographic information such as patient address, hospital/site/center or country need to be pseudonymized or masked up to no finer than state level in US. Different approaches are used based on the nature of the data or analysis requirement.

For example, if site level analysis is needed, then all sites need to be pseudonymized using alias (e.g. site 1, site 2...) to protect disclosure of patient's personal geographic information. Otherwise, irrelevant geographic information is either masked or simply removed to minimize re-identification risk.

Other Information: Study team reviews and agrees on variables to be de-identified or removed in accordance with HIPAA and/or GDPR guidelines. Examples include but are not limited to contact information, account information,

technical identifier, biometric data, images, job title etc. In general, such information shall not be shared with any third party. If use of these variables is necessary for analytical purposes appropriate, pseudonymization/anonymization measures need to be applied.

CONCLUSION

This paper describes BMS's approach to managing Real-World Data (RWD) privacy and how statistical programming teams operationalize these company-wide policies in daily practice. Our strategy is built on two foundational pillars:

- **Controlled Data Access and Secure Storage**: Ensuring that sensitive data are housed in a validated environments and only authorized users can access the sensitive data.
- **Rigorous De-identification Practice:** Using systematic and programmatic methods to protect patient identifiers and personal information.

Together, these measures enable BMS to remain aligned with evolving global regulations and industry's best practices. By implementing these controls, we preserve the analytical value of RWD while maintaining rigorous protection of patient privacy. Ultimately, this framework supports ethical, secure, and audit-ready evidence generation

REFERENCE

(1) GDPR https://gdpr-info.eu/ (accessed 12/02/2025).
(2) HIPAA: https://www.hhs.gov/hipaa/index.html (accessed 12/02/2025).
(3) Jaffe D.H., Malin B.A., Hendricks-Sturrup R.M. (2025). "*A real-world data challenge: guidance for aligning data privacy compliance and fit-for-purpose usability*". Health Affairs Scholar, 2025 Nov 12;3(11).
(4) Ferran J.M., Lanoue J. (2015). "*PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models*". PharmaSUG 2015 -Paper DS10.
(5) FDA Guidance (2021) "*Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry"*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Joe Xi
Company: Bristol Myers Squibb
Email: joe.xi@bms.com

Author Name: Lorraine Fang
Company: Bristol Myers Squibb
Email: lorraine.fang@bms.com