PharmaSUG 2017 - Paper PO05

Data Transparency, de-identification strategies, and platforms available for sharing data

Arun Raj Vidhyadharan and Sunil Mohan Jairath, inVentiv Health

ABSTRACT

The idea of data transparency became a reality year 2013 onwards. Pharmaceutical companies were started to share their own company trial data, and enables that data to be combined with data from other pharmaceutical companies. Many Regulatory authorities now require clinical trial sponsors to make clinical trial data available. Data sharing from clinical trials can lead to better understand and faster development and approval of medicines for rare diseases.

One of such initiatives led to launch of platform to share clinical trial data related to cancer studies and the platform is called as PDS (Project Data Sphere). Other platforms available are Yale Open Data Access (YODA) and ClinicalStudyDataRequest.com (CDSR) .

Data transparency in Pharmaceutical industry requires data de-identification to protect patient confidentiality and comply with legal governing laws. In order to put data for research and other purpose we need to first de-identify the clinical data.

INTRODUCTION

Data sharing can lead to discovery of new trends and associations that generate new insights or hypotheses for further research. But for all the benefits, there are some disadvantages as well. If we do not carefully de-identify clinical trial data it can compromise patient privacy, enable faulty science, and be a resource-intensive burden for trial sponsors. Some obvious questions that may arise are:

(a). What information should be shared, with whom and for what purposes?

(b). How do you ensure patient privacy without unduly limiting the research value of the data?

(c). How should this access and use be managed? Shall we allow data downloading or not?

(d). How far back in time should study data be made available? Etc.

At a basic level, de-identification is a process in which we remove some identifiers, dates, site id etc. but still our data makes sense when an analysis is run against the data.

This paper will talk about concept and techniques followed for de-identifying data and discuss data transparency and few platforms for uploading clinical trial data and future strategies of de-identification.

CHALLENGES IN DATA SHARING

Looking at pharmaceutical companies, there are four main issues that they will need to face.

a) Need to address variability in source data
b) Reduce and minimize effort, time, resources and cost
c) Minimize risks to the privacy and confidentiality of research participants
d) Ensure compliance with data privacy legal requirements

Some of these, such as the data privacy legal requirements, are very complex and are not yet fully defined. Furthermore, data privacy laws differ greatly between countries, so a company needs to ensure that releasing data in one country does not contradict with the laws in another where patients may have been based.

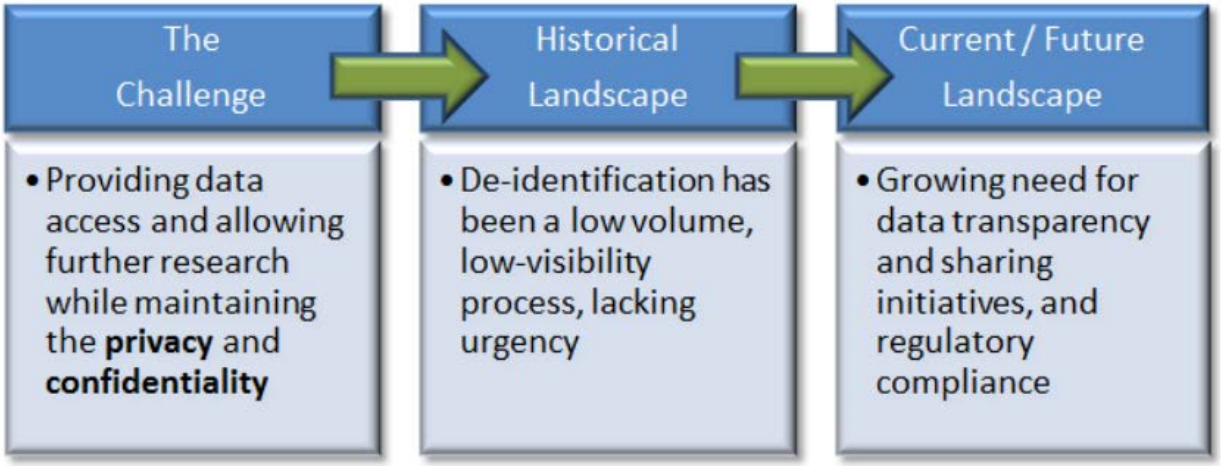

Figure1 : Biggest challenge with Data Transparency :Privacy and Confidentiality

Clinical trial patient data falls under Protected Health Information (PHI) requirements, and the de-identification of patient data covers not just the more obvious identifiers of a patient, such as subject's race, ethnicity, age, date of birth, but also disease population, geographical location, etc., that singly or in combination could allow a subject to be re-identified.

HIPAA, SAFE HARBOR AND EXPERT DETERMINATION

In the United States, one of the primary standards used to provide guidance for de-identifying personally identifiable information (PII) and personal health information (PHI) is the HIPAA Privacy Rule (45 CFR 164.514) from the US Department of Health and Human Services

These mechanisms center on two HIPAA de-identification standards: HIPAA Safe Harbor and the Statistical or Expert Determination methods.

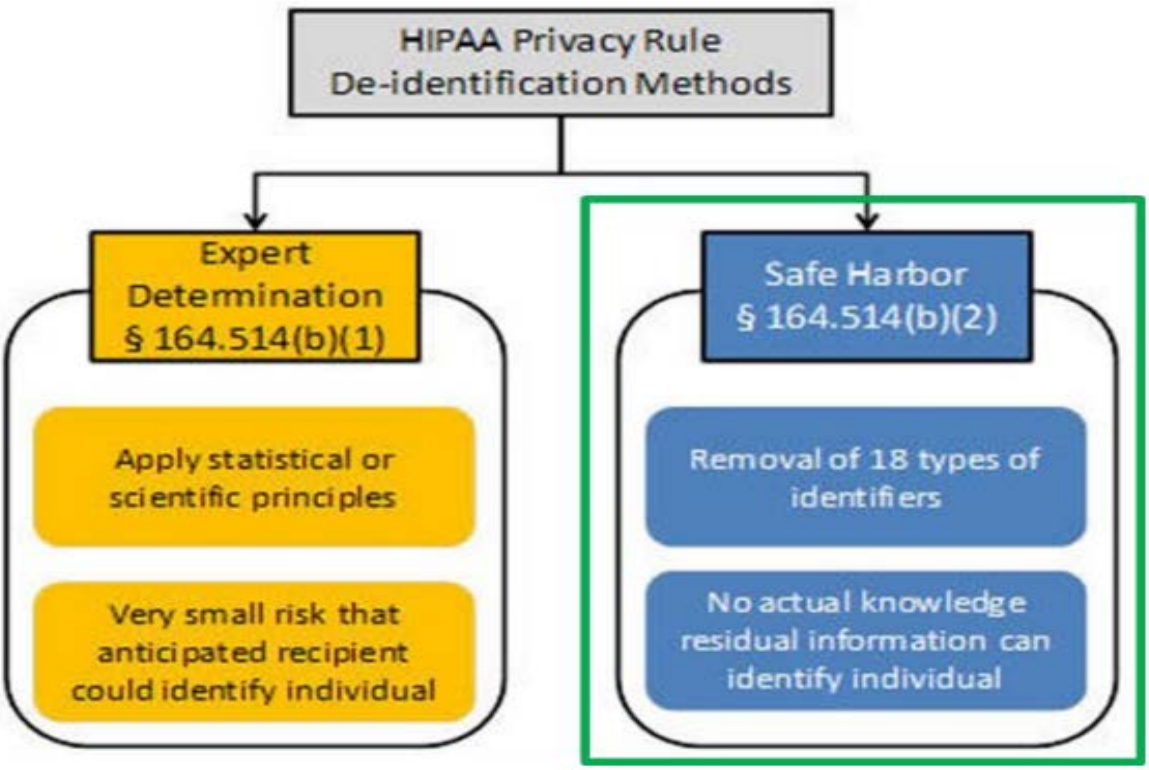

Figure 2 - HIPPA Privacy Rule: De-identification Methods

EXPERT DETERMINATION

In the case of Expert Determination or Statistical Method, this first requires finding a person (or persons) who has appropriate experience with the rules governing identifiable information. They are fluent with the statistical methods and scientific principles for adjudicating the risk of data in terms of individual identification potential.

SAFE HARBOR

Safe Harbor describes 18 elements of data that must be removed or generalized in a data set in order for it to be considered “de- identified.” The 18 data elements (aka direct identifiers) include the following:

1. Names
2. Zip codes (except first three)
3. All elements of dates (except year)
4. Telephone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers

10. Account numbers
11. Certificate or license Numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic or code

DATA DE-IDENTIFICATION STEPS

RECODING IDENTIFIERS

In order to protect participant privacy, the identifiers like Subject IDs, Reference IDs, Sponsor IDs, Investigator IDs and Site IDs must be recoded, and to anonymize the data, the key code that was used to generate these new random identifiers should be irreversibly destroyed. The investigator/site name and contact information, as well as that for any third party vendors (such as laboratories and providers of imaging and biomarker data) should be set to blank or removed. All other types of identifiers (eg, treatment kit numbers, device numbers, laboratory IDs) should be also re-coded using new randomly generated identifiers, or removed/set to blank.

HANDLING DATES

All time-related information is important in clinical research in particular dates, they present them self in two forms date and relative days. Some of the dates recorded in clinical studies include visit dates, dates of birth, date of randomization, dates of adverse events, etc. Removal of the following date elements from datasets is required in order to achieve the Safe Harbor method of de-identification. Also Birth must be derived into “Age at baseline” and patients over 89 years old must be aggregated into one category.

OFFSET DATE METHOD

All dates in a study are replaced with a new date generated using a random offset for each participant. By using one offset for all dates for a participant, the relative distance between a participant’s dates is maintained from their original dates to their de-identified dates. However, a disadvantage to this approach is that it may be perceived as not being as secure as having a different random offset for each participant. For this reason, an algorithm that assigns different random offsets to each participant in a study is considered a stronger approach when using this method.

RELATIVE STUDY DAY METHOD

If a variable containing Relative Study Day is not already present in the data provider’s datasets, it is calculated for each observation as days relative to a reference date, eg, date of study entry or date of randomization. The same algorithm is applied to all dates across the study in order to maintain the relationship between events for each participant (eg, their visit schedule). All date variables are then removed from or set to blank in the de-identified datasets.

HANDLING DATE OF BIRTH AND AGE

In order to adhere to the requirements of the HIPAA Privacy Rule using the Safe Harbor method, additional requirements are stipulated to protect the privacy of participants aged over 89 years by aggregating their ages into a single category rather than presenting their exact age. De-identified datasets must also not display any dates indicative of age >89 years, eg, year of disease diagnosis or year a prior medication was started.

MEDICAL DICTIONARIES AND CODING

The most common dictionaries currently used by data providers are MedDRA for adverse events and diseases, and WHO Drug for medications, though some data providers use their own in-house dictionaries. Dictionaries are upgraded at regular intervals, and datasets can be up-versioned as needed.

FREE-TEXT VERBATIM FIELDS

All free text verbatim terms and comments variables should be set to blank or removed if redaction is required for every record in a dataset. Certain free text fields (or parts thereof) may be considered for retention in their original form if removal of this information would impact the scientific value of the dataset, eg, a free text field in an oncology study where tumor site was recorded. Such fields should be reviewed carefully to ensure they do not contain personal information.

SENSITIVE INFORMATION AND LOW FREQUENCY EVENTS

This refers to studies with rare diseases (eg, small denominators where the total eligible patient population is small), rare events (eg, small numerators), genetic information, extreme values (eg, height, weight, BMI), or sensitive data (eg, illicit drug use or “risky behavior”). Other data items may be of an increased sensitivity, and therefore, additional steps may be required to further protect participant data privacy such as setting variables to blank, or replacing the “sensitive” records (or parts thereof) with “--redacted--”. Alternative approaches include adding noise (eg, using an offset method for dates) or aggregating data (eg, defining age bands).

QUALITY CHECKS AND VALIDATION

Validation and QC checks are very important part of data de-identification process. It ensures that all necessary data have been de-identified appropriately and consistently between the datasets. The enhanced Safe-Harbor approach combines removal of the relevant 18 HIPAA identifiers with the removal of additional personal information that may be present in a study dataset.

AVAILABLE PLATFORMS FOR CLINICAL TRIAL DATA SHARING

CLINICALSTUDYDATAREQUEST.COM OR ‘CSDR’ (12)

The launch of the first data sharing platform (in 2014) marked the beginning of a new era of data transparency within the Pharmaceutical Industry. This is a multi-sponsor patient-level data request site.

The original concept was based on a GSK-only request site which was expanded to accommodate additional sponsors, in theory both commercial and non-commercial (academics, charities etc.). The website allows external researchers to browse studies listed as available to request, put together a research proposal and submit. It also allows researchers to enquire about studies not listed as to whether they could be available to request. The research proposal requires information such as lay summary of the research intent, study design, studies required and rationale, primary and secondary endpoints, statistical analysis and publication plans.

Thirteen study sponsors have now committed to the sharing of participant level data from their clinical trials via multi-sponsor platform. CSDR provides a structured format for requesting data, including a step-by-step diagram, user guide, supporting guidance videos and the opportunity to communicate with the sponsor throughout the process. Functionalities of the platform include the ability to:

- select available studies from a list provided by the sponsor
- submit a research proposal for the studies required for the research, including statistical analysis plan for
- review by an independent review panel
- signing data sharing agreements by the researcher and sponsor for approved proposals
- remote access to requested de-identified datasets and all related documentation is provided via a secured SAS analytic environment

YALE OPEN DATA ACCESS OR 'YODA'

the YODA PROJECT | Forging a unified scientific community

The YODA website states: "The YODA Project seeks mutually beneficial partnerships with Data Holders, promoting independence, responsible conduct of research, good stewardship of data, and the generation of knowledge in the best interest of society. To participate, each Data Holder must transfer full jurisdiction over data access to the YODA Project".

Johnson and Johnson (J&J) have been in partnership with YODA since October 2014. YODA performs the independent scientific reviews of research requests (requests for both PLD and reports) and as such makes all decisions on release of clinical trial data.

PROJECT DATA SPHERE® (PROJECTDATASPHERE.ORG)

Project Data Sphere®

The Project Data Sphere® platform was developed by SAS. The Project Data Sphere® repository is a universal platform to share oncology clinical trial knowledge and data to accelerate cancer research. Thus it will help cancer researchers, industry, academia, providers, and other organizations in a collaborative effort to transform "big data" into solutions for cancer patients.

At present, PDS contains data from more than 41,000 research participants from 72 oncology trials, covering multiple tumor types. The data have been donated by academic, government, and industry sponsors. These numbers are increasing quickly as use of the PDS accelerates. More than 1400 unique researchers have accessed the PDS database more than 6500 times. As one interesting example, a

challenge was issued in 2014 that asked respondents to use PDS to create a better prognostic model for advanced prostate cancer. A total of 549 registrants from 58 teams and 21 countries responded. Accessible data included control groups from prospective, randomized, industry-sponsored trials. Solvers had backgrounds in statistics, data modeling, data science, machine learning, bioinformatics, engineering, and other specialties. Unexpectedly, the winning entrant, a team of researchers from Finland, had never worked on prostate cancer in the past, and this team considerably outperformed the best existing model for predicting overall survival among men with advanced prostate cancer.7 Thus, the PDS Prostate Cancer DREAM Challenge confirmed that an open-access model empowers global communities of scientists from diverse backgrounds and promotes crowd-sourced solutions to important clinical problems. This level of engagement is not possible with gatekeeper models.

There are some differences between the 3 platforms. Both CSDR and YODA allow “enquiries” for datasets of studies which are not listed on the websites. In the spirit of transparency, full reasons are provided where enquires (and research proposals) result in a negative outcome

Project Data Sphere provides an alternative format to clinical trial data sharing. PDS allows researchers, whether independent or affiliated to industry, hospitals or academic institutions, access to historical, participant level, comparator arm phase III trial datasets and accompanying documentation with the aim of development and improvement of trial design and methodology, as well as acceleration of future research hypotheses. Unlike CSDR and YODA, access to de-identified datasets is granted to all approved researchers, without the need for a formal research proposal. While this format has the advantage of immediate access to datasets, PDS only has the control arm of trials rather than the full trial data set. It must also be noted that CSDR and YODA initiatives fall under controlled data releases while PDS falls under semipublic data release, which are 2 different contexts.

PHUSE SDTM 3.2 DE-IDENTIFICATION STANDARDS

The PhUSE de-identification standards aim at SDTM domains. It helps in:

1. Assigning rules for de-identification
2. Understanding rationale and address exceptions and special considerations
3. Facilitate the identification of direct and quasi identifiers.
4. Provide generalized SDTM rules to apply together with technical guidelines
5. Ensure consistency in data de-identification across various sponsors

DELIVERABLE STRUCTURE

The deliverable consists of an MS Excel spreadsheet with different tabs:

1. Cover tab: Document information.
2. Intro tab: Introduction including background, important considerations, out of scope, disclaimer and approach.
3. Definitions tab: List of important terms with their definitions and examples when applicable.
4. Decisions tab: Important areas with rationale for decisions.
5. Rules tab: The different rules to be applied together with technical guidance.
6. SDTMIG tab: The SDTMIG 3.2 variables (1300+) together with their assessment for direct/quasi identifiers,
7. Primary rules, Secondary rules and Comment for De-Identification. See Figure 1.
8. References tab: Sources used for the elaboration of the deliverable.

9. Appendices tab: Appendices for guidance on “Dates Offset” and “Low Frequency”
10. Change log tab: Different versions of the deliverable together with list of changes.

Observation_Class	Domain_Prefix	Variable_Name	Variable_Label	Length	Controlled_Terms_or_Format	Role	CDISC_Notes	Core	Direct_Quasi_Identifier (Direct/Quasi)	DI_Primary_Rule	DI_Alternative_Rules	DI_Comment
Interventions	SU	SUDOSFRM	Dose Form	20		Variable Qualifier	Dose form for SUTRT. Examples: INJECTABLE, LIQUID, or POWDER.	Perm		Review and only redact values with personal information		Extensible codelist attached. Generally the data is entered correctly.
Interventions	SU	SUDOSFRQ	Use Frequency Per Interval	20	(FREQ)	Variable Qualifier	Usually expressed as the number of repeated administrations of SUDOSE within a specific time period. Example: Q24H (every day)	Perm		Review and only redact values with personal information		Extensible codelist attached. Generally the data is entered correctly.
Interventions	SU	SUDOSTOT	Total Daily Consumption	8		Record Qualifier	Total daily use of SUTRT using the units in SUDOSU. Used when dosing is collected as Total Daily Dose. If sponsor needs to aggregate the data over a period other than daily, then the aggregated total could be	Perm				
Interventions	SU	SUROUTE	Route of Administration	20	(ROUTE)	Variable Qualifier	Route of administration for SUTRT. Examples: ORAL, INTRAVENOUS.	Perm		Review and only redact values with personal information		When free text value is entered in "other" value, it should be reviewed and recoded if required during the trial conduct. This is expected to take place during the data cleaning process. This variable is
Interventions	SU	SUSTDTC	Start Date/Time of Substance Use	20	ISO 8601	Timing		Perm	Quasi Level 2	Offset		
Interventions	SU	SUENDTC	End Date/Time of Substance Use	20	ISO 8601	Timing		Perm	Quasi Level 2	Offset		
Interventions	SU	SUSTDY	Study Day of Start of Substance Use	8		Timing	Study day of start of substance use relative to the sponsor-defined RFSTDTC.	Perm	Quasi Level 2	No further de-identification		
Interventions	SU	SUENDY	Study Day of End of Substance Use	8		Timing	Study day of end of substance use relative to the sponsor-defined RFSTDTC.	Perm	Quasi Level 2	No further de-identification		
Interventions	SU	SUDUR	Duration of Substance Use	20	ISO 8601	Timing	Collected duration of substance use in ISO 8601 format. Used only if collected on the CRF and not derived from start and end date/times.	Perm	Quasi Level 2	No further de-identification		Extreme values may be considered for low frequency.

Figure 3: De-Identification Standards for CDISC SDTM 3.2

CURRENT AND POSSIBLE FUTURE DE-IDENTIFICA TION SCENARIOS

Current process to produce de –identified data is to take SDTM and ADAM data and apply the above mentioned de-identification rules and optionally produce anonymized TLFs too. Since data transparency is a relatively new concept, the suggested way to share data for older projects doesn’t need to be followed for newer studies. So figure 4 shows the possible future scenario where we can create anonymized SDTM directly from SDTM data and used the anonymized SDTM to generate anonymized ADAM and anonymized TLF.

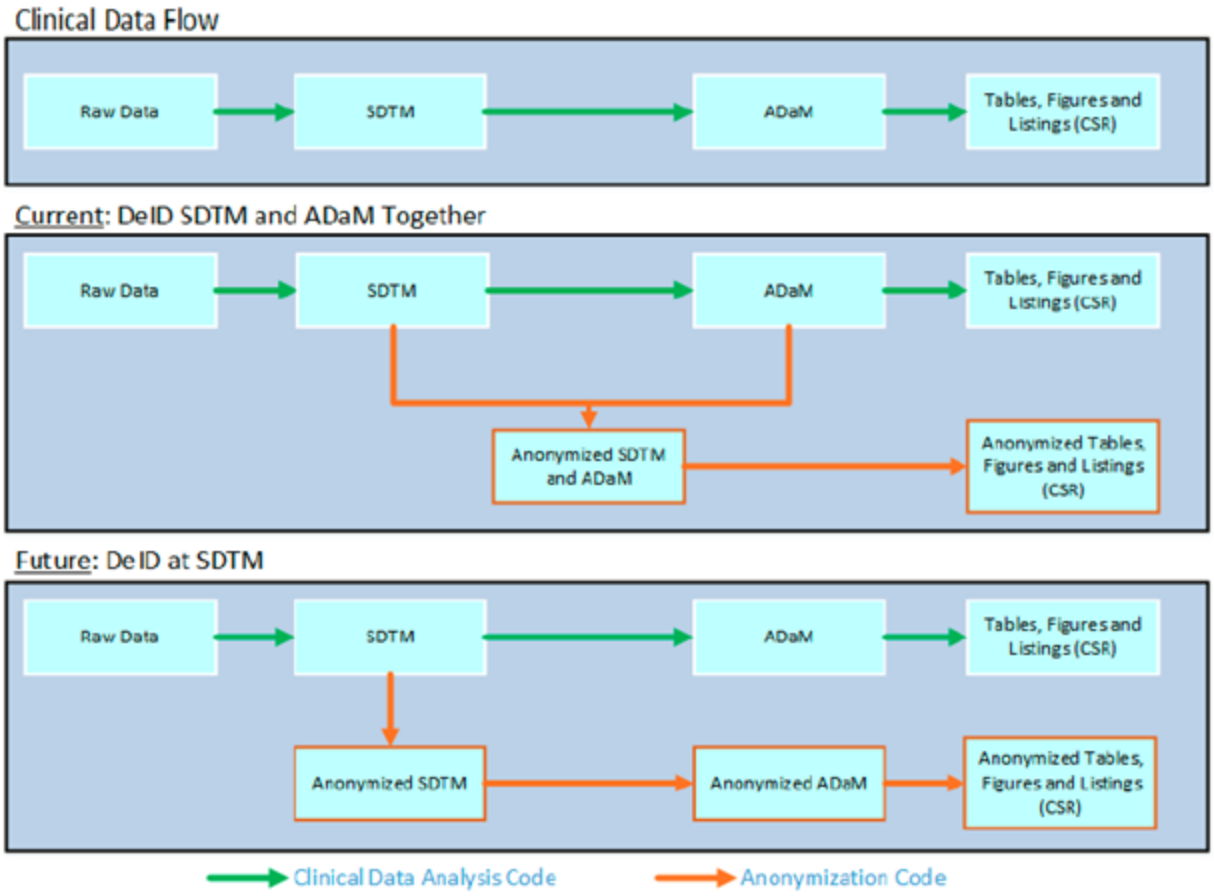

Figure 4: Current and poosible future scenarios in De-identification

CONCLUSION

Data transparency is a big initiative and a very useful one. However since it's still a relatively new concept, there is no industry standard to de-identify data. We are slowly getting there with availability of SDTM 3.2 De-identification standards and rules provided by TransCelerate BioPharma Inc. Increasingly, data providers are defining algorithms for de-identification and anonymization of data More Initiatives will continue to make data more accessible to researchers, and ensure that all sponsors of clinical trials are able to share data in a cost effective manner with the right incentives. As we will progress we can hope for more standardized approach across Pharma companies and hence we can continue to support this beautiful initiative with greater efficiency and more meaningful data.

REFERENCES

1. https://www.projectdatasphere.org/projectdatasphere/html/resources/PDF/DEIDENTIFICATION
2. Clinical Study Data Request Site. https://clinicalstudydatarequest.com/
3. http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-FINAL.pdf
4. http://www.phuse.eu/Data_Transparency.aspx
5. Clinical Data Interchange Standards Consortium, "CDISC SDTM Implementation Guide (version 3.2)," 2015.

6. PhUSE De-Identification Working Group, “De-Identification Standards for CDISC SDTM 3.2,” 2015. http://www.phuse.eu/Data_Transparency_download.aspx
7. Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J. P. (2014).
8. Reanalyses of randomized clinical trial data. Jama,312 (10), 1024-1032.
9. YODA: the Yale School of Medicine’s Open Data Access (YODA) Project. https://yoda.yale.edu/
10. Project Data Sphere: https://www.projectdatasphere.org/ (accessed 24/08/2015).
11. TransCelerate Biopharma, “Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach,” 2015.
12. https://support.sas.com/resources/papers/proceedings15/1884-2015.pdf
13. http://www.lexjansen.com/phuse/2015/dh/DH09.pdf
14. PHUSE SDE Frenchtown, NJ

ACKNOWLEDGMENTS

The authors would like to thank John Durski, Associate Director, inVentiv Health, for all his support and motivation in writing this paper.

CONTACT INFORMATION

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

Name: Arun Raj Vidhyadharan
Enterprise: inVentiv Health
Work Phone: 908.421.4667
E-mail: arunraj.vidhyadharan@inventivhealth.com

Name: Sunil Mohan Jairath
Enterprise: inVentiv Health
Work Phone: 862.240.4182
E-mail: sunil.jairath@inventivhealth.com