PharmaSUG2021 - Paper SA-145

Preserving the Privacy of Participant's Data by Data Anonymization

Chaithanya Velupam and Kaushik Sundaram, Covance by Labcorp

ABSTRACT

While the global landscape of clinical data sharing has become prominent day by day, it is challenging to store, mine, and analyze heterogeneous data across multiple data sources and multiple regions. Proactive sharing of clinical trial data has been a key strategic aim since last few years.

Data must be shared in such a way as to ensure the protection of participant's privacy. Whilst this is the foremost priority in any data sharing exercise, the changing technology is increasingly disabling the ability to share data in enough depth and detail. We explore how anonymization of internationally sourced clinical trial data may be achieved while maintaining the scientific utility of the data. We focus on anonymization, which plays an important role for the re-use of clinical data and for the sharing of research data.

We present a flexible solution for anonymizing distributed data in the semi-honest model by incorporating specific anonymization methods, so that the important insights of research data still prevail. Based on this case study, we provide useful recommendations that address some of the central questions of anonymization and consider the strengths and weaknesses of the anonymization process.

INTRODUCTION

The landscape of clinical trial data sharing has changed significantly, a number of sponsors and journals now require data sharing statements which requires data sharing plans at the time of registration of a clinical trial. As data typically flows through several sources, some of which are open to the public, cross referencing these data sources could possibly expose personal information of subjects. The intent of this paper is to provide information on techniques that could be applied programmatically in anonymizationof clinical data, which can be a source for tracing out the identity of subjectsparticipating in clinical trial.

Data Anonymization can be defined as removing identifiable and traceable links of an individual, the links from the original to the anonymized dataset are destroyed and it is no longer possible to trace back the subject in original dataset. Data anonymization is crucial for clinical data transparency since any subject-level data that is shared must to be anonymized to protect subject's identity and privacy.

Anonymization is relevant when clinical data is used for secondary purposes such as research, public health, certification or accreditation, and marketing.

Most privacy laws around the world are consent-based, if patients give their consent or authorization, the data can then be used for the purposes they authorize. If the data is anonymized, however, then no consent is required. In general, anonymized data is no longer considered personal health information and it falls outside of privacy laws. The legal framework in the European Union (EU) is provided by the General Data Protection Regulation (GDPR). While GDPR does not change the core data protection principles of the previous directives, it introduces new safeguards in the light of technological advances which must be considered for clinical data sharing.

Application of anonymization techniques may not necessarily ensure that the data does not pose any serious risk of re-identification, it can be said that the risk of re-identification is minimum. Also, anonymization inevitably causes information loss, and thus, various methods have been proposed to reduce information loss. However, existing data anonymization methods cannot avoid excessive information loss and preserve data utility. So, future innovation, research initiatives, and an ever evolving scientific landscape demand continuous and proactive exploration of new approaches to anonymization.

GENERAL PROCESS FLOW OF DATA ANONYMIZATION:

➢ IDENTIFYING SENSITIVE DATA:

Figure 1. Types of Clinical Data

The foremost part of the process is to identify the type of sensitive source data that is fetched from the available repositories of a particular therapeutic area or study.

Clinical data could be of a broad range of data types in the practice of medicine and the allied health sciences. They range from narrative, textual data to numerical measurements, genetic information, recorded signals, drawings, and even photographs or other images

Generally, we anonymize the source that either follow the CDISC standards (Clinical Data Interchange Standards Consortium) or the non-CDISC standards.

In CDISC datasets as there is a standard in place, domain and variables containing the sensitive information can easily be identified to be worked upon. On the contrary, for non-CDISC data it is tedious and cumbersome to identify data.

➢ CATEGORIZATION OF DATA:

Categories of data can be multiple types depending on the study and therapeutic area we are working on. Categorization of data is required to apply appropriate anonymization technique at later stage.

- **Patient Identifiers:**Patient Identifiers can also be known as Personally Identifiable Information (PII). It is any data that can be used to identify a specific individual. Name, social security numbers, email address and phone numbers have most commonly been considered PII, but technology has expanded the scope of PII considerably. It can include an IP address, login IDs, social media posts, or digital images. Geolocation, biometric, and behavioral data can also be classified as PII.

  For clinical data, in addition to the above one can consider medical identification numbers, health insurance status, Blood test results, X-rays, Admission and discharge dates, medical device identifiers and mental health records etc.

- **Staff Personnel Information:** It is mostly related to Medical investigators, staff personnel's information or study site related information. It is data that specifies about the location of the site, educational qualifications of the investigators/staff or maybe the College/University from the where staff has completed his/her degree.

- **Free Text:**It comprises of long strings of natural language that may include personal data.
- **Date Variables:**It includes datevalues like treatment start date, AE onset date etc.
- **Proprietary Information:**Proprietary information is any information that deals with the activities, business or products of a company. More specifically, some things that commonly fall under this umbrella include trade secrets, financial data, product research and development, computer software, business processes and marketing strategies.
  In Clinical world, it can be medical dictionaries like WHODD (World Health Organization Drug Dictionary) and MedDRA (Medical Dictionary for Regulatory Activities).

Since these medical dictionaries require license if needed to be accessed by third party, it is required to anonymize this information while sharing it to unauthorized users.

For example, in clinical datasets, they can be stored in the form of ATC codes.

CMTRT	ATC4_S	ATC4_L	ATC3_S	ATC3_L	ATC2_S	ATC2_L	ATC1_S	ATC1_L	ATC_S	ATC_L
IBUPROFEN	M01AE	PROPIONI	M01A	ANTIINFLA	M01	ANTIINFLA	M	MUSCULO-S	M01AE	PROPIONIC
TYLENOL	N02BE	ANILIDES	N02B	OTHER AN	N02	ANALGESI	N	NERVOUS S	N02BE	ANILIDES,
ASPIRIN	N02BA	SALICYLIC	N02B	OTHER AN	N02	ANALGESI	N	NERVOUS S	N02BA	SALICYLIC
SUDAFED	R01BA	SYMPATHO	R01B	NASAL DEC	R01	NASAL PR	R	RESPIRATO	R01BA	SYMPATHOI
IBUPROFEN	M01AE	PROPIONI	M01A	ANTIINFLA	M01	ANTIINFLA	M	MUSCULO-S	M01AE	PROPIONIC
BACTRIM D	J01EE	COMBINAT	J01E	SULFONAN	J01	ANTIBACT	J	ANTIINFEC	J01EE	COMBINATI

Figure 2. Concomitant Medication(CM) dataset consisting of Proprietary Information.

- **Questionnaires with Personal Information:**They can be subject's personal data recorded in the form of Questionnaires. So these must also be considered.

- **Comments:** These are mostly the free text data, where the information might be about Investigator, Study, Site, Vendor or any other information that can be unique for a subject.

➢ DE-IDENTIFICATION STANDARDS/RULES

De-identification process can be accomplished using different methods based on the type of data we have identified to anonymize.

Below are the different methods generally applied on various categories of data.

- **Masking:**It is hiding data with altered values that replaces private identifiers with fake identifiers or pseudonyms.It is generally applied for identifying variables, the most common programmatic approach followed for masking is by generating a random number which is unique to each subject.

STUDYID	SITEID	DESITEID	SUBJID	DESUBJID	USUBJID	DEUSUBJID	INVID	DEINVID
ABC	102	SAL	01345	00789	ABC-304-01345	ABC-SAL-01345	1456	3456
ABC	102	PPT	01359	00889	ABC-304-01359	ABC-PPT-01359	1676	7867
ABC	102	GSG	01367	00791	ABC-304-01367	ABC-GSG-01367	1456	4532
ABC	103	KER	01412	00118	ABC-304-01412	ABC-KER-01412	1472	6747
ABC	104	PAR	01349	00224	ABC-304-01349	ABC-PAR-01349	1472	5565
ABC	106	TSG	01350	00794	ABC-304-01350	ABC-TSG-01350	1472	4351

Figure 3. Dataset showing Identifier variables along with their De-identified versions

- **Offset Date:**It is a process of providing modified date by adding or removing (unique number per subject) random days for the subject consistently across all domains for the subject. Partial dates are handled by imputing and just offsetting the imputed dates. By adding unique random value for a particular subject, paired dates are handled properly so that the study days are not impacted.

DOM	VAR	USUBJID	DEUSUBJID	OFFSET	DATE	DE_DATE	DAY
DM	RFSTDTC	ABC-300-235	ABC-PVR-776	20	2020-10-25	2020-11-14	1
DM	RFENDTC	ABC-300-235	ABC-PVR-776	20	2021-03-15	2021-04-04	142
CM	CMSTDTC	ABC-300-235	ABC-PVR-776	20	2021-01-03	2021-01-23	71
LB	LBDTC	ABC-300-235	ABC-PVR-776	20	2020-11-18	2020-12-08	25
AE	AESTDTC	ABC-300-235	ABC-PVR-776	20	2021-02-23	2021-03-15	122

Figure 4. Dataset with OFFSETvalues and DE_DATE (i.e. Date+ Offset no. of days)

- **Redaction:**The simplest method of anonymizing data is to simply delete personal data so that it can be shared, distributed, or posted without revealing the subject's identity. This is particularly useful for personal data that can directly identify an individual person, such as like ethnicity, race, country, region, birthdate and free text.

STUDYID	USUBJID	INVNAM	BRTHDTC	RACE	ETHNIC	COUNTRY
ABC	ABC-0344-1001	ULRIKE	1946-03-17	Black	NOT HISPANIC OR LATINO	USA
ABC	ABC-0344-1002	LORCH	1933-07-18	Asian	NOT HISPANIC OR LATINO	USA
ABC	ABC-0344-1003	DAYAL	1956-09-29	White	NOT HISPANIC OR LATINO	USA
ABC	ABC-0344-1004	COLLIN	1961-05-20	Asian	NOT HISPANIC OR LATINO	USA
ABC	ABC-0344-1005	PARKER	1988-02-21	Asian	NOT HISPANIC OR LATINO	USA
ABC	ABC-0344-1006	MEHTA	1971-12-22	Black	HISPANIC OR LATINO	USA

Figure 5. Example of variables that can be redacted (Invnam, Birthdate, Race,Ethnicity,Country).

- **Partial reduction:**It is mainly applied on the variables like concomitant medication indication, proprietary information and also in the other variables wherever applicable.

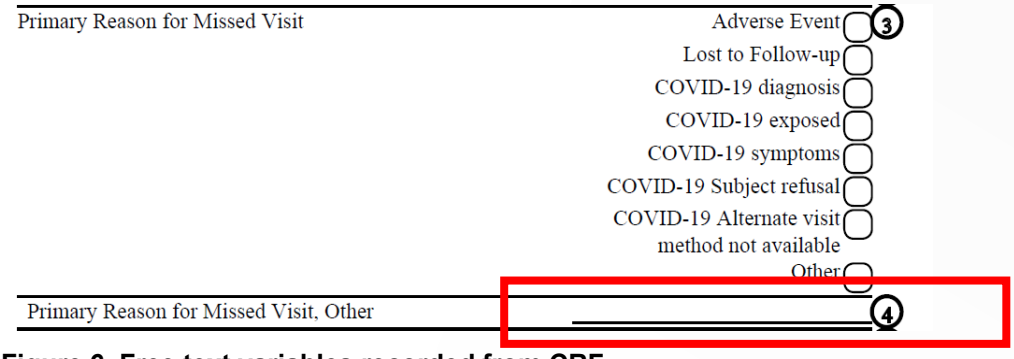

Figure 6. Free text variables recorded from CRF.

COVANCE by labcorp

USUBJID	QNAM	QLABEL	QVAL
ABC-0076-110	SVREASND	Primary Reason for Missed Visit	Other: Subject suffering from Dementia
ABC-0076-118	SVREASND	Primary Reason for Missed Visit	Other: Abnormal cognitive
ABC-0129-104	SVREASND	Primary Reason for Missed Visit	Other: Delusions and Hallucinations
ABC-0214-103	SVREASND	Primary Reason for Missed Visit	Other: Social Factors

Figure 7. Supplementary Dataset containing highlighted data that have to be partially redacted

- **Proprietary information:**The Proprietary information can be shared to the parties who are authorized or who have license to access this information. Separate dataset with redacted proprietary information need to be created for unauthorized parties.

USUBJID	CMDECOD	CMDECOD_REDACTED
ABC-ACY-00906	BENADRYL /01563701/	BENADRYL
ABC-ACY-00906	BENADRYL /01563701/	BENADRYL
ABC-ACY-00906	CALCIUM D3 /01483701/	CALCIUM D3
ABC-ACY-00906	LORATADINE	LORATADINE
ABC-ACY-00906	EPINEPHRINE	EPINEPHRINE
ABC-ACY-00906	FOLIC ACID	FOLIC ACID
ABC-ACY-00906	INFLIXIMAB	INFLIXIMAB

Figure 8. Dataset containing proprietary Information whose values are procured from WHODD.

- **Grouping:** Grouping can be used across gender, regions, treatment details, stratification factors etc. A General method that can be used for grouping of variable values is by using K-anonymity. For *k*-anonymity to be achieved, there need to be **at least *k* individuals in the dataset who share the set of attributes that might become identifying for each individual**.

  For example, assume that we have four regions in our dataset. Some regions are very small and when combined with other key variables in the dataset, produce high re-identification risk for some individuals in those regions. One way to reduce risk would be to combine some of the regions by recoding them. We could, for example, make two groups out of the four, call them ‘urban’ and ‘rural’ and re-label the values accordingly.

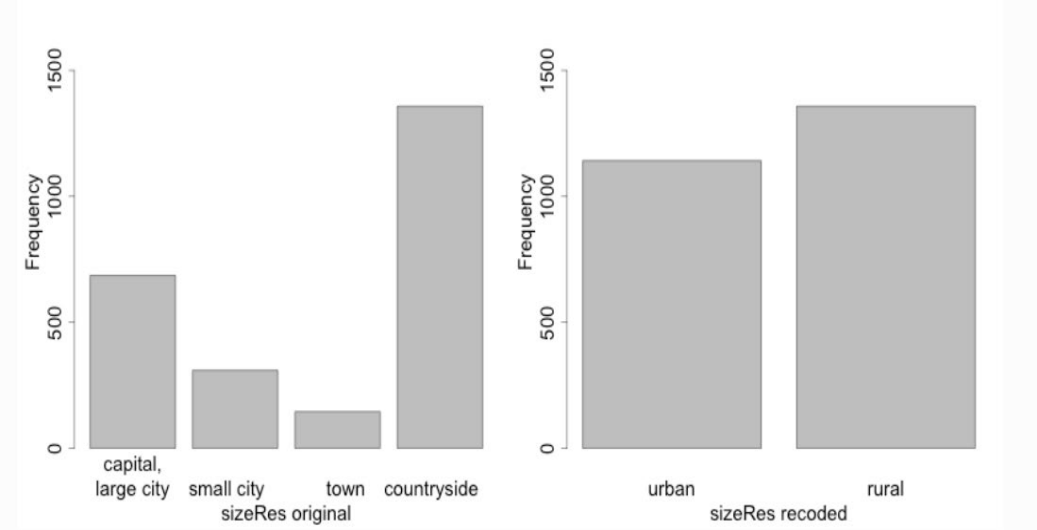

Figure 9.Illustrates the effect of recoding the variable "sizeRes" and show respectively the frequency counts before and after recoding. We see that the number of categories has reduced from 4 to 2 and the small categories ('small city' and 'town') have disappeared.

Similar approach can be followed for categories of age to be chosen so that they still allow data users to make calculations relevant for the subject being studied.

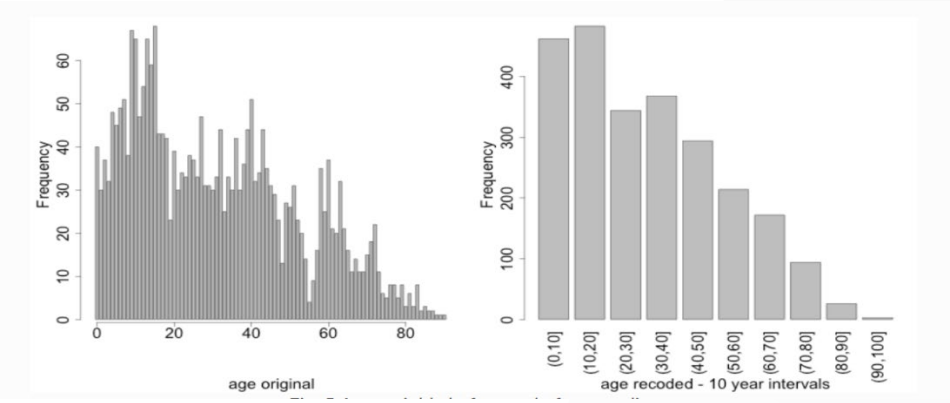

Figure 10.Shows the effect of grouping of the variable "age"

Outliers: In statistics, an outlier is a data point that differs significantly from other observations.

To determine the outlier values, generally we use Mean and Standard Deviation(σ). The upper and lower limits are calculated by the following:

Lower limit(.L) : Mean – 3*σ

Upper Limit (.H): Mean + 3*σ

The outlier values of the numeric variables are set to special missing.L/.H and their respective character variables values are made to redacted-low/redacted-high, also the dependent variables are redacted.

Example for dependent variables: If there is HEIGHT, WEIGHT & BMI recorded in a vital signs datasets and suppose weight value for a particular patient is an outlier and has .H or “redacted-high” assigned, then the BMI value also must be redacted and set to “redacted”.

This is done because as standard BMI is calculated byBMI = Weight (kg) / Height $(m)^2$,one can easily be able to re-calculate weight by having values of both BMI & Height.Hence we also redact the dependent variables so that it cannot be re-calculated.

STUDYID	DOMAIN	USUBJID	HEIGHT	DE_HEIGHT	WEIGHT	DE_WEIGHT	BMI	DE_BMI
ABC	VS	ABC-165-01758	132.1	redacted-low	90.3	90.3	51.7	redacted-high
ABC	VS	ABC-186-01693	136	136	107.3	redacted	58	redacted-high
ABC	VS	ABC-189-01431	168.1	168.1	60.3	60.3	21.3	21.3
ABC	VS	ABC-189-01482	170.2	170.2	61.2	61.2	21.1	21.1
ABC	VS	ABC-189-01574	154.9	154.9	53.9	53.9	22.5	22.5

Figure 11. Dataset after redacting outlier information.

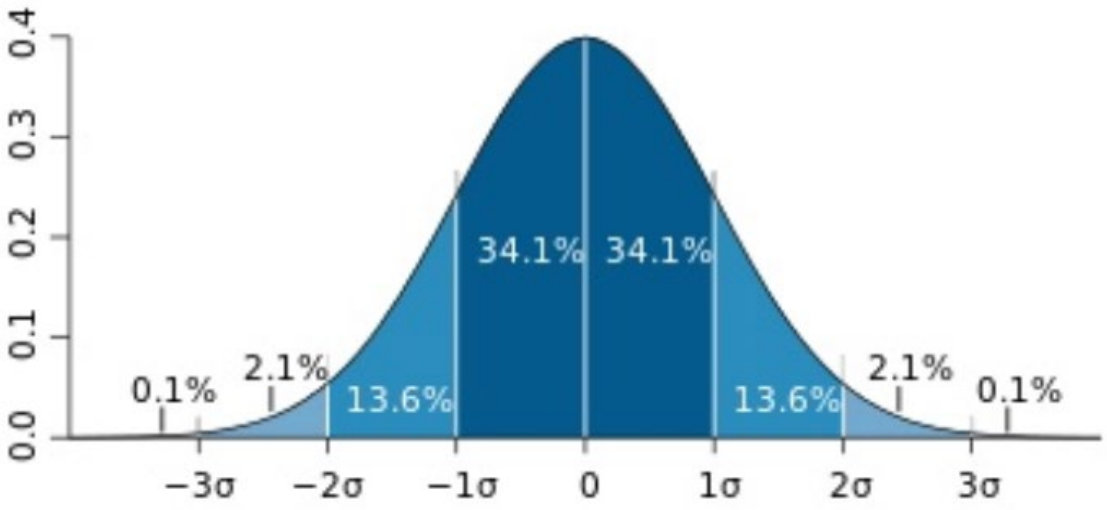

Figure 12.Consider the population graph with majority lying nearing the mean represented as 0. Here outliers are considered as the 0.1% on either ends i.e, Mean-3* σ and Mean+3* σ.

- **Post Anonymization Process:**It is process of breaking the link between the anonymized dataset and source dataset after performing anonymization. Since a unique random number is created for each subject to mask the identifying variables (ieusubjid, subjid etc), the master dataset acts as a link between the source datasets and anonymized datasets. So after we complete actual anonymization process we are required to break the link so that the original patient identity cannot be traced back. This process is called post anonymization, the most common programmatic approach involves making the seed ids to ‘0’ in ranuni and byte function and refreshing the datasets which gets sorted with new values in the identifying variables. Finally deleting the master dataset.

STUDYID	SITEID	DESITEID	SUBJID	DESUBJID	USUBJID	DEUSUBJID	INVID	DEINVID
ABC	101	SAL	01233	00789	ABC-101-01233	ABC-SAL-00789	3575	3468
ABC	101	PPT	01241	00889	ABC-101-01241	ABC-PPT-00889	3575	4537
ABC	101	GSG	01259	00872	ABC-101-01259	ABC-GSG-00872	3575	7864
ABC	102	XLX	01264	00248	ABC-102-01264	ABC-XLX-00248	1472	5012
ABC	103	TQG	01278	00118	ABC-102-01278	ABC-TQG-00118	1472	2398
ABC	104	HZF	01312	00615	ABC-102-01312	ABC-HZF-00615	1472	8876

Figure 13. Example of Master Dataset.

It is essential to document the seed ids before we rerun the programs as part of post anonymization process. We also document age bands and stat values before we proceed for post anonymization.

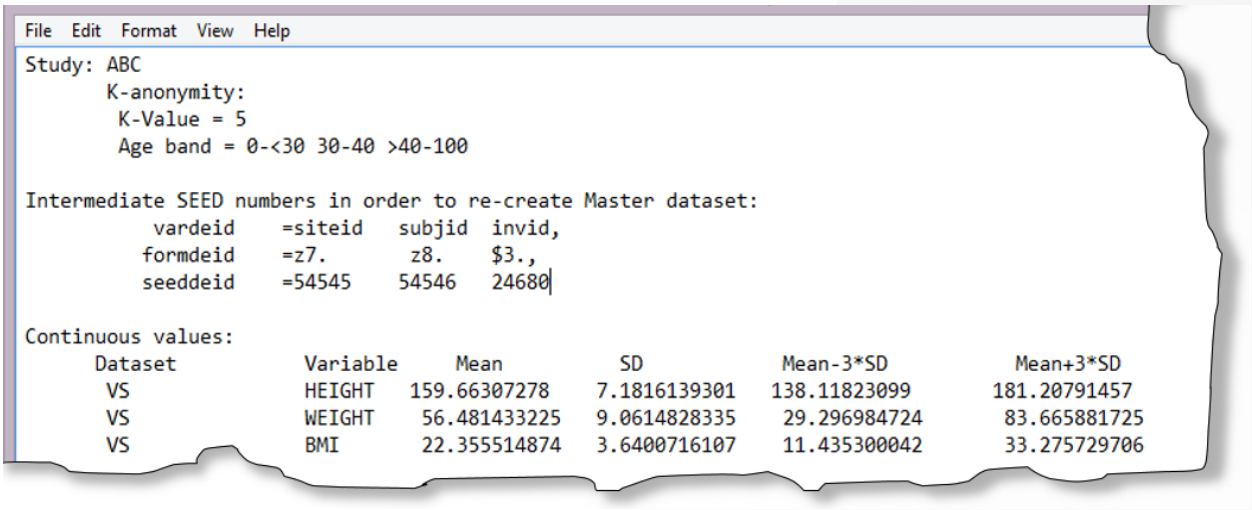

Figure 14. Example of Documentation

QUALITY CHECKS:

Quality of the anonymization can be ensured with double programming, also by manual recheck of de-identified data to back trace the subject.

The consistency of identifier variables such as usubjid and subjid in anonymized datasets need to be ensured across various domains in a particular study.

Based on the therapeutic area and SAP specific stratification factors, we also double check on indications of concomitant medications, questionnaire related data and any variable that can have sensitive information.Any contact information of investigators and third party vendors need to be cross checked for redaction.

FUTURE CHALLENGES:

- **NOT FULL PROOF:** Data Anonymization is done for minimizing the risk and possible harm of any data leaks. Since there is no single Implementation Guide document which is followed by all the companies for handling Clinical data.

- **COMPLEXITY:**The more dimensions there are to a dataset, that is more than dependencies of the dataset, more information has to be redacted to protect the patient privacy. Hence it makes it more difficult for the programmer to find multiple links to be personal identifiers across all the dependent domains.

- **DEPRIVES OF VALUABLE INSIGHTS:**Even though the approach looks robust and safe, it may deprive us of any detailed information or insight of the study that might be critical for the researchers, as subject privacy is considered paramount.

- **UNABLE TO MAINTAIN PATIENT DATA PRIVACY:** There are some scenarios where few of the clinical subjects were able to get traced back from the anonymized data during the secondary usage or from the data being available in the public domains. This must be taken strictly so that any future occurrences of these types of events must not take place.

CONCLUSION:

Since most of the clinical data describes or allows you to identify a person, be smart about what you do with it when sharing it for secondary research. Privacy is not just an on/off switch, but rather represents a continuum of options available with different tradeoffs between user privacy and data utility. It is still possible to derive amazing results out of research even by using anonymized data.

REFERENCES:

1. European Medicines Agency. “Data anonymization - a key enabler for clinical data sharing”. 01December2017. https://www.ema.europa.eu/en/documents/report/report-data-anonymisation-key-enabler-clinical-data-sharing_en.pdf

2. “Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule”. http://privacyruleandresearch.nih.gov/pdf

3. Official journal of the European communities. “General Data Protection Regulation”. 01December2001.http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:008:0001:0022:en:PDF

4. “Anonymization Methods”.https://sdcpractice.readthedocs.io/en/latest/anon_methods.html

5. “GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES”. Personal Data protection commission. 25January2018. https://iapp.org/media/pdf/resource_center/Guide_to_Anonymisation.pdf

COVANCE by labcorp

CONTACT INFORMATION:

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

Chaithanya Velupam
Covance by Labcorp
chaithanya.velupam@covance.com

Kaushik Sundaram
Covance by Labcorp
kaushik.sundaram@covance.com