Data Recipient Report for the Janssen Clinical Trial Data Set TMC114FD2HTX3001

“A Study to Evaluate Efficacy and Safety of Darunavir/Cobicistat/Emtricitabine/Tenofovir Alafenamide (D/C/F/TAF) Fixed Dose Combination (FDC) Versus a Regimen Consisting of Darunavir/Cobicistat FDC With Emtricitabine/Tenofovir Disoproxil Fumarate FDC in Treatment-naive HIV Type 1 Infected Subjects”

Product Name	Symtuza
Active Substance	Darunavir/Cobicistat/Emtricitabine/Tenofovir Alafenamide
Dataset Type	SDTM
Study Code	TMC114FD2HTX3001
NCT Number	NCT02431247
Reporting Effort	Week 48
Version	1.0
Date	June 27, 2024

Contents

1 Introduction

The purpose of this project was to perform anonymization of the Janssen TMC114FD2HTX3001 clinical trial data set.

The anonymization of this data set was performed to allow the data to be shared with external research teams. Access to clinical trial data provides opportunities to conduct further research that can help advance medical science and improve patient care. This helps ensure the data provided by study participants are used to maximum effect in the creation of knowledge and improving patient care. The data release is subject to certain criteria being met, including a requirement to effectively anonymize the data.

Statistical anonymization was used to preserve the utility required by recipients, while accounting for the context of the data sharing scenario.[2] Unlike a rules-based framework that removes dates (except years) and aggregates all ages over 89 as 90 or older, such as HIPAA Safe Harbor, this approach is adaptive to population distributions, sample size, and the desired utility of the anonymized data.

The data sharing environment and contracts in place with the data recipient are assumed to be at a level which would result in a Privacy and Security Context Assessment score of High and a Recipient Trust Context Assessment score of Medium.

This report describes the anonymization approach used for the study TMC114FD2HTX3001, based on the re-identification risk determination that was performed on the data.

1.1 Data Set Model

The data set described in this report for study TMC114FD2HTX3001 was received in the Study Data Tabulation Model (SDTM) standard. For more information on this standard see https://www.cdisc.org/standards/foundational/sdtm

1.2 Definitions

Definitions of key terms (such as the different types of identifiers) and acronyms are provided in Section A *Definitions*. Additional terms and definitions are provided elsewhere [1].

2 Anonymization Process

2.1 Use of Software

The analysis described in this report was performed using a re-identification risk measurement software application.

2.2 Supporting Documentation

The following documents were provided to assist with the analysis:

- TMC114FD2HTX3001 Transformation Summary
- Annotated CRF
- define.pdf
- define.xml
- define1-0-0.xsd
- study-data-reviewers-guide.pdf

2.3 Transformations

In order to bring the risk of re-identification below the determined threshold, some transformations were required on the dataset. The transformations are described based on the fields used in the risk measurement. In all cases, modifications to these fields are applied to all other linked fields, e.g. where concomitant medication start date is suppressed, concomitant medication end date will also be suppressed as it is a linked field.

The anonymization strategy required the following modifications to the original datasets:

Identifier	**Transformation**
Subject IDs (USUBJID)	Masked
Site IDs (SITEID)	Suppressed
Free-text	Suppressed
Patient dates	PHUSE shifted
Age	Generalized to 2-year intervals
Country	Suppressed
Ethnicity	Suppressed
Concomitant medication dates	Suppressed
Medical history dates	Suppressed

2.4 Implemented Transformation Types

The following data transformations have been applied in this dataset:

Masking Masking of the unique subject ID was performed using Format-Preserving Encryption (FPE). This type of encryption creates an encrypted value that has the same length as the original ID.

Generalization Reduce the precision of a field.

For this specific project, the age of subjects was generalized to 2-year intervals. Table 1 summarizes the mapping of age to generalized age for age greater or equal to 0 years.

Age (Years)	**Generalized Age Value (Years) (AGE)**
$0 \leq$ Age < 2	0
$2 \leq$ Age < 4	2
$4 \leq$ Age < 6	4
$6 \leq$ Age < 8	6
...	...
$64 \leq$ Age < 66	64
$66 \leq$ Age < 68	66
$68 \leq$ Age < 70	68
$70 \leq$ Age < 72	70

Table 1: Age generalization

PHUSE date shifting Offset a date value according to the scheme defined in the Pharmaceutical Users Software Exchange (PHUSE) CDISC SDTM anonymization standard [3]. This scheme determines a delta for each patient based on a difference between a date in the trial available for all patients (in this case the first visit date) and an anchor date (in this case, 06 July 2015).

Suppression The original value is replaced with an empty cell. The following types of suppression were applied for this project:

global suppression (GS): Occurs when risk measurement determines that no suitable generalized value can be retained and all values in the column are therefore suppressed.

parameter-value suppression (PV): Occurs when values in a column are suppressed based on the values of a parameter-column in the same dataset. For example, a vital sign dataset may include a parameter-column specifying the type of measurement such as “systolic blood pressure”, “height”, “weight” and “temperature”, and one or more value-columns containing the values of the mesurements (for example, height measured in centimeters when the parameter is “height”).

Parameter-value suppression occurs when all values in the value-column associated with one or more identifiers in the parameter-column are suppressed as part of the anonymization strategy.

Please see the file “TMC114FD2HTX3001 Transformation Summary.csv” for a catalog of all transformations applied to the data set.

3 Conclusions

The re-identification risk of the Janssen TMC114FD2HTX3001 clinical trial database, after the anonymization as described in this report, is below the data risk threshold given the assumed level of mitigating controls and motives and capacity in the context of the data sharing environment.

References

[1] Khaled El Emam. *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[2] International Standards Organization. ISO/IEC 27559:2022: Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework. Technical report, ISO, 2022.

[3] PhUSE De-Identification Working Group. De-Identification Standards for CDISC SDTM 3.2. Technical report, 2015.

[4] Pierangela Samarati. Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[5] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

A Definitions

A.1 Acronyms

FPE Format-Preserving Encryption

PHUSE Pharmaceutical Users Software Exchange

SDTM Study Data Tabulation Model

A.2 Identifiers

It is useful to differentiate among the different types of variables in a disclosed data set or document. The way the variables are handled during the risk measurement and anonymization process will depend on how they are categorized.

A distinction is made among three types of variables [4, 5]:

Directly identifying variables. One or more direct identifiers can be used to uniquely identify an individual, either by themselves or in combination with other readily available information. In clinical trial data sets and documents, the only patient direct identifier will likely be the subject ID. There will be direct identifiers pertaining to staff and investigators; however, these are treated differently than patient information.

Indirectly identifying variables (quasi-identifiers). The quasi-identifiers are the background knowledge variables about individuals in the disclosed data set that an adversary can use, individually or in combination, to probabilistically re-identify a trial participant. If an adversary does not have background knowledge of a variable then it cannot be a quasi-identifier. The manner in which an adversary can obtain such background knowledge will determine which attacks on a data set are plausible. For example, the background knowledge may be available because the adversary knows a particular target individual in the disclosed data set, an individual in the data set has a visible characteristic that is also described in the data set, or the background knowledge exists in a public or semi-public registry.

Examples of quasi-identifiers include sex, date of birth or age, locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, aboriginal identity, total years of schooling, marital status, criminal history, total income, visible minority status, event dates (such as admission, discharge, procedure, death, specimen collection, visit/encounter), codes (such as diagnosis codes, procedure codes, and adverse event codes), country of birth, birth weight, and birth plurality.

Other variables. These are the variables that are not really useful for determining an individual's identity. They may be clinically relevant or not.

A.3 Glossary

data recipient The data recipient is the researcher who accesses the anonymized data to perform an analysis.

Privacy and Security Context Assessment A questionnaire that evaluates the privacy and security controls in place for a data recipient.

Recipient Trust Context Assessment A questionnaire that evaluates the motives, capacity, and contracts in place with regard to data recipient performing a re-identification attack.

A.4 Output Format of De-identified Datasets

All dataset anonymization was performed within the SAS (Statistical Analysis System) native data file format (extension ".sas7bdat"). Datasets received in SAS version 5 (V5) or version 8 (V8) transport file format (extension ".xpt") must first be converted to .sas7bdat for processing. Following de-identification, all datasets are converted from .sas7bdat to .xpt for delivery. For datasets originally received in .xpt format, this conversion should not pose a problem. However, for datasets received in non-xpt format, inherent limitations in the .xpt format may require modifications.

Based on the definition of the format, conversion of a dataset to XPT transport file format may require modification of the following in the anonymized datasets:

1. Shortening the dataset names,
2. Shortening variable names in the datasets,
3. Shortening dataset or variable labels,
4. Splitting long character values into new variables.

B Datasets Delivered in TMC114FD2HTX3001

Table	Number of Rows
AE	2886
CM	4135
CO	1373
DA	94187
DM	866
DP	17026
DS	1816
DV	50
EC	4
EG	870
EX	9113
FA	3250
GT	28930
HO	55
IE	133
II	9698
LB	431846
MH	14999
PE	54252
PT	23730
QS	760
RELREC	6560
RF	6217
SC	225
SE	1630
SG	23
SU	11900

Table	Number of Rows
SUPPAE	3104
SUPPCM	33374
SUPPDA	94306
SUPPDM	3987
SUPPDP	262
SUPPDV	44
SUPPEG	197
SUPPEX	24958
SUPPFA	134
SUPPGT	5821
SUPPHO	55
SUPPII	9660
SUPPLB	582529
SUPPSG	19
SV	7189
TA	10
TE	6
TI	105
TS	86
TV	23
VS	27199
ZR	3625

Table 2: List tables considered and the number of rows in each.