Data Recipient Report for the

Janssen Clinical Trial Data Set

212082COU-AA-302

"Abiraterone Acetate in Asymptomatic or Mildly Symptomatic Patients With Metastatic Castration-Resistant Prostate Cancer"

Product Name	ZYTIGA
Active Substance	Abiraterone Acetate
Dataset Type	SDTM
Other Study Identifier	COUGAR-302
Study Code	212082COU-AA-302
NCT Number	NCT00887198
Reporting Effort	Final
Version	3.0
Date	February 6, 2024

Contents

1 Introduction

The purpose of this project was to perform anonymization of the Janssen 212082COU-AA-302 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 212082COU-AA-302, 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 212082COU-AA-302 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:

- 212082COU-AA-302 Transformation Summary
- Annotated CRF

2.3 Output Format of Anonymized 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.

2.4 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 indirect identifiers used in the risk measurement. In all cases, modifications to these indirect identifiers are applied to all other linked fields, e.g. where country is suppressed, fields containing brand- or region-specific drug names will also be suppressed as they are linked to geography.

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
Date of Birth	Suppressed
Age	Generalized to 4-year intervals

2.5 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 4-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 < 4	0
$4 \leq$ Age < 8	4
$8 \leq$ Age < 12	8
$12 \leq$ Age < 16	12
...	...
$80 \leq$ Age < 84	80
$84 \leq$ Age < 88	84
$88 \leq$ Age < 92	88
$92 \leq$ Age < 96	92

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, 28 April 2009).

Suppression The original value is replaced with an empty cell. The following type of suppression was 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.

Please see the file "212082COU-AA-302 Transformation Summary.csv" for a catalog of all transformations applied to the dataset.

3 Conclusions

The re-identification risk of the Janssen 212082COU-AA-302 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. The indirect identifiers are attributes that, together with other attributes that can be in the dataset or external to it, enable unique identification of a data subject within a specific operational context.

Examples of indirect 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.

B Datasets Delivered in 212082COU-AA-302

Dataset	Number of Rows
ABMOD	1403
AE	18229
AMD3CNST	153
AMD4CNST	53
ANL	21119
ATRISK	7616
BNSCAN	7412
BPI	20573
CHEM	11362
CMEDS	27633
COAG	246
COMPL	824
CONT	19664
CRITERIA	95
DEATH	741
DEMO	1088
DISC	777
DOV	33795
DRGCOMP	20388
ECG	8893
ECHO	1153
ECOG	26278
FACT	8575
FU	5678
HEMA	5160
HORMTX	3606
HPI	1086

Dataset	Number of Rows
INCLEXCL	1099
LAB	882433
LREVAL	3837
MEDHX	10483
MEDPROCS	4469
MRU	4469
NLA	662
NTRG	11255
OTHTEST	36
OTX	3463
PCRS	1197
PE	21213
PEB	1216
PK	619
POPULATN	1088
PRDMOD	1272
PREMAE	13
PREMAENA	13
PREMAERP	13
RADTX	1193
RAND	1088
SAENARR	18244
SAERPT	18244
SCRBONE	1086
SCREEN	1088
SUBTHER	5630
THERHX	1167
TISSUE	1055
TLS	4427

Dataset	Number of Rows
TRG	6646
TSTN	428
TXCOMP	1053
UA_TRANS	10970
UNBLD	906
URIN	1097
VSLC	25288

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