Data Recipient Report for the

Janssen Clinical Trial Data Set

212082jpn202

"A Phase 2 Study to Evaluate Safety and Efficacy of Abiraterone in Participants With Prostate Cancer Who Have Received Docetaxel"

Product Name	Zytiga
Active Substance	Abiraterone Acetate
Dataset Type	SDTM
Study Code	212082jpn202
NCT Number	NCT01795703
Reporting Effort	Final
Version	2.0
Date	September 17, 2025

Contents

Contents **2**

1 Introduction **3**
1.1 Data Set Model . . . 3
1.2 Definitions . . . 3

2 Anonymization Process **4**
2.1 Use of Software . . . 4
2.2 Supporting Documentation . . . 4
2.3 Output Format of Anonymized Datasets . . . 4
2.4 Transformations . . . 4
2.5 Implemented Transformation Types . . . 5

3 Conclusions **7**

References **8**

A Definitions **9**
A.1 Acronyms . . . 9
A.2 Identifiers . . . 9
A.3 Glossary . . . 9

B Datasets Delivered in 212082jpn202 **10**

1 Introduction

The purpose of this project was to perform anonymization of the Janssen 212082jpn202 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 212082jpn202, 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 212082jpn202 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:

- 212082jpn202 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	PHUSE shifted; Generalized to 10-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 year of birth of subjects was first PHUSE date shifted then generalized to 10-year intervals. Table 1 summarizes the mapping of year of birth to generalized year of birth for values greater or equal to 1900.

Year of Birth	Generalized Year of Birth (BRTHDTC)
$1900 \leq$ YOB < 1910	1900
$1910 \leq$ YOB < 1920	1910
$1920 \leq$ YOB < 1930	1920
$1930 \leq$ YOB < 1940	1930
$1940 \leq$ YOB < 1950	1940
$1950 \leq$ YOB < 1960	1950
$1960 \leq$ YOB < 1970	1960

Table 1: Year of birth 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, 15 Jun 2012).

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 "212082jpn202 Transformation Summary.csv" for a catalog of all transformations applied to the dataset.

3 Conclusions

The re-identification risk of the Janssen 212082jpn202 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 212082jpn202

Dataset	Number of Rows
AE	443
BE	61
CD	122
CM	2179
CO	328
DATADEF	23
DC	1088
DM	64
DS	128
DV	31
EG	13881
EX	1906
IE	21
LB	30572
MH	313
PC	1336
QS	9547
SG	0
STUDYDEF	1
TR	1585
VALDEF	210
VARDEF	489
VS	4920
XC	206
XE	456
XO	270
XP	35

Dataset	Number of Rows
XZ	211

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