Data Recipient Report for the

Janssen Clinical Trial Data Set

212082PCR4001

"A Registry to Observe the Treatment of Prostate Cancer Under Routine Medical Care"

Product Name	ZYTIGA
Active Substance	Abiraterone acetate
Dataset Type	OMOP
Study Code	212082PCR4001
NCT Number	NCT002236637
Reporting Effort	Final
Version	1.0
Date	March 28, 2023

Contents

1 Introduction

The purpose of this project was to perform anonymization of the Janssen 212082PCR4001 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 212082PCR4001, 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 212082PCR4001 was received in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard. For more information on this standard see https://www.ohdsi.org/data-standardization

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 [redacted]

[redacted]

2.2 Supporting Documentation

The following documents were provided to assist with the analysis:

- [redacted] Transformation Summary 212082PCR4001
- Annotated CRF
- Annotated CRF Index and ref to ds_v3.xlsx

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 (person_id)	Masked
Free-text	Suppressed
Patient dates	PHUSE shifted
Country	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.

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, 14 June 2013).

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 [redacted] Transformation Summary 212082PCR4001.csv" for a catalog of all transformations applied to the data set.

3 Conclusions

The re-identification risk of the Janssen 212082PCR4001 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

CDM Common Data Model

FPE Format-Preserving Encryption

OMOP Observational Medical Outcomes Partnership

PHUSE Pharmaceutical Users Software Exchange

SAS Statistical Analysis System

SQL Structured Query Language

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.

B Output Format of De-identified Datasets

The original data were received from Janssen primarily as Statistical Analysis System (SAS) files; datasets received in SAS version 5 (V5) transport file format (extension ".xpt") must first be converted to .sas7bdat for processing. Each .sas7bat file was exported as a Structured Query Language (SQL) table to a Microsoft SQL server database using the Microsoft SQL Server Native Client driver for Open Database Connectivity (ODBC). Information was extracted from these SQL tables for the purpose of risk measurement with [redacted]. The anonymization strategy determined from the measurements was applied to the tables. The anonymized SQL tables were exported back to SAS files using the same interface. The SAS variable labels in the original SAS files were added back to the variables in the anonymized files. After performing quality control checks on the anonymized SAS files, all datasets are converted from .sas7bdat to .xpt for delivery to Janssen to be shared with the data recipient(s).

As a result of data anonymization, the attributes of an anonymized variable in an anonymized SAS file may differ from the attributes of the variable in the original file. For non-anonymized variables, the following SAS attributes may differ between the anonymized and original SAS files as a result of the SAS to SQL data conversion process.

1. Internal precision of a numeric variable: When a SAS numeric variable is exported to a SQL table, its values are rounded to d decimal digits, where $d \geq 0$ is the decimal scaling factor of the SAS display format $w.d$. As a result, the internal precision of the variable in the anonymized SAS file is also limited to d digits.

2. SAS formats: The SAS format (for example, "DATE", "DATETIME", "$" or no format) of a variable in an anonymized file may differ from the original file. These differences were all identified and reviewed during quality control checks. No additional data processing was performed on the anonymized files to reconcile variable formats with those of the original files except in rare occasions where the difference in formats between the anonymized and original file was deemed important enough to potentially impact data analysis. For example, this would be the case if a SAS "TIME" variable file was changed to "DATETIME" during the anonymization process.

3. Length of a numeric variable: The length (in bytes) of a numeric variable in the anonymized file may differ from the original file. In such cases, no additional data processing was done on the SAS anonymized file to reconcile the variable lengths in the anonymized SAS file with those in the original file.

4. Format width of a numeric variable: When the width w specified by the SAS format $w.d$ of a numeric variable is too small to properly display all values taken by the variable, an error may occur when exporting the SAS file to a SQL table. In such case, the format width of the variable was increased before performing the export to ensure that the operation is successful and as a result, the variable's width in the anonymized SAS file differs from the original file.

For datasets originally received in .xpt format, the final conversion from .sas7bdat to .xpt for delivery should not pose any additional problems. However, for datasets received in .sas7bdat format, inherent limitations in the .xpt format may require modifications.

Based on the definition of the .xpt format, conversion of a SAS dataset to V5 transport file format fails when one or more of the following requirements is not met:

1. The dataset name must not have more than 8 characters,
2. Variable names in the dataset must not have more than 8 characters,
3. The character variables in the dataset must not include values that are more than 200 characters long.

If any of these requirements are not met, [redacted] has modified de-identified datasets as follows:

1. Values exceeding 200 characters in character variables were split among new variables so that any value contained in a variable did not exceed 200 characters long,
2. Datasets with names exceeding 8 characters were renamed,
3. Variables with names exceeding 8 characters were renamed.

Characters not permitted in XPT variable or dataset identifiers were remapped or omitted, and identifiers conflicting with keywords were replaced.

If modifications were made to deliver V5 transport files, these are summarized in the spreadsheet "212082PCR4001_SAStoXPTMapping_*.xlsx" delivered with the de-identified datasets. This spreadsheet can be used to map the names of the variables and datasets in the V5 transport files to their original names in the datasets shared by Janssen. If no spreadsheet is present, no modifications were required.

The spreadsheet contains the following two worksheets:

Dataset Summary : Lists the original and new name of the datasets that were renamed for conversion to the V5 transport file format. The original dataset names are captured in column"SAS_DATASET" and the new names are captured in column"XPT_DATASET". The worksheet is empty (except for column headers) if no dataset was renamed.

Variables Summary : Lists the original and new name of variables that were renamed and the new character variables that were added to a dataset after splitting a value exceeding 200 characters. The original variable names are captured in column"SAS_VARIABLE" and the new names are captured in column"XPT_VARIABLE". The worksheet is empty (except for column header) if no variable was renamed or no new variable was added to a dataset.

The dataset and variable names used in this document are the original names found in the initial datasets received from Janssen.

C Datasets Delivered in 212082PCR4001

Table	Number of Rows
CONDITION_ OCCURRENCE	19018
DEATH	1767
DRUG_ EXPOSURE	16732
FACT_ RELATIONSHIP	27316
LOCATION	16
MEASUREMENT	62939
NOTE	1827
OBSERVATION	825750
OBSERVATION_ PERIOD	3159
PERSON	3159
PROCEDURE_ OCCURRENCE	12200
VISIT_ OCCURRENCE	25754

Table 1: List tables considered and the number of rows in each.