Data Recipient Report for the

Janssen Clinical Trial Data Set

CNTO312C0168T41

"A Study of the Safety and Effectiveness of Infliximab (Remicade) in Patients With Rheumatoid Arthritis"

Product Name	Remicade
Active Substance	Infliximab
Dataset Type	Other
Study Code	CNTO312C0168T41
NCT Number	NCT00036387
Reporting Effort	Final
Version	2.0
Date	July 30, 2026

Contents

1 Introduction

The purpose of this project was to perform anonymization of the Janssen CNTO312C0168T41 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 consistent with the Sharing Anonymized and Functionally Effective (SAFE) Data Standard framework [1, 3]. 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 SAFE Data Standard provides a structured framework for safely sharing clinical trial data by balancing patient privacy with data usability. It uses a risk-based anonymization approach that measures the likelihood of re-identification as a function of both the data itself and the context in which it is shared, applying statistical techniques such as generalization, suppression, and date shifting to reduce this risk below acceptable thresholds.

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 CNTO312C0168T41, 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 CNTO312C0168T41 was received in a custom raw data format.

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 [2].

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:

- CNTO312C0168T41 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
Birth Date	Suppressed
Age	Generalized to 5-year intervals
Race	Suppressed

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 5-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 < 5	0
$5 \leq$ Age < 10	5
$10 \leq$ Age < 15	10
$15 \leq$ Age < 20	15
...	...
$70 \leq$ Age < 75	70
$75 \leq$ Age < 80	75
$80 \leq$ Age < 85	80
$85 \leq$ Age < 90	85

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 [4]. 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, 21 Sep 2001).

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 "CNTO312C0168T41 Transformation Summary.csv" for a catalog of all transformations applied to the dataset.

3 Conclusions

The re-identification risk of the Janssen CNTO312C0168T41 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] Stephen Bamford, Sarah Lyons, Luk Arbuckle, and Pierre Chetelat. Sharing Anonymized and Functionally Effective (SAFE) data standard for safely sharing rich clinical trial data. *Applied Clinical Trials*, 31(7/8), 2022.

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

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

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

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

[6] 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

SAFE Sharing Anonymized and Functionally Effective

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 [5, 6]:

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 CNTO312C0168T41

Dataset	Number of Rows
ADMIN	9418
AE	5509
ANY_NONE	666
COMTRSUM	929
CONPROC	3124
CRITERIA	21
DEMO	1084
DEMOG	1084
DRGACSCR	9302
DRUGACCT	9342
HAQ	5343
INFCTSUM	1074
IVRS	1084
IVRSRAND	1088
JNTSC_LL	130312
JNTSC_LR	130298
JNTSC_UL	186140
JNTSC_UR	186160
LAB	310514
MANIFST	1084
MEDHX	1084
MEDS	17847
MEDSUMM	6326
PAGE_SEQ	4629
PK_ATSA	15605
QOL	4177
RAEVAL	9317

Dataset	Number of Rows
RAINFO	1084
RANDCODE	2
TB_INFO	1084
TERMDOSE	186
TERMSTDY	154
UNBLIND	1
VISITS	14376
VITALS	1083

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