Anonymization Report for the Janssen Clinical Trial Data Set TOPMATMIG303

Product Name	Topamax
Active Substance	Topiramate
Study Code	TOPMATMIG303
Version	Data Recipient Report
Date	November 30, 2020

Evaluation of Re-Identification Risks

Janssen contracted Privacy Analytics Inc. to perform the assessment of the risk of re-identification and the anonymization for this trial, as well as author this Anonymization Report. The purpose of this project is to evaluate the risk of re-identification found in the data provided by Janssen for the study TOPMATMIG303.

The evaluation of re-identification risks documented in this report is subject to the Limitations and Qualifications set forth below. The risk of re-identification was determined to be below the threshold requested by the client, as well as the threshold based on the Privacy Analytics Invasion of Privacy Assessment, which is in turn based on existing precedents for the release of health information for secondary purposes.

Implementation of the Methodology:

1. The evaluation of re-identification risks was conducted by qualified professionals with appropriate knowledge of, and experience with, generally accepted statistical and scientific principles and methods for rendering information not individually identifiable;
2. To the best of their knowledge, the professionals performing the evaluation have applied generally accepted statistical and scientific principles and methods for rendering information not individually identifiable in performing this evaluation of re-identification risks; and
3. The professionals performing the evaluation have documented the methods and results of the analysis that justify the evaluation of re-identification risks described in this report.

Limitations and Qualifications:

The statement set out above is subject to the following limitations:

a) The evaluation of the risk of re-identification is based on the information provided to Privacy Analytics Inc., by Janssen, and the result of this evaluation is contingent upon the assumption that such information is complete and accurate.

b) The evaluation of the risk of re-identification is based on a series of assumptions documented in this report that Janssen has confirmed to be reasonable given its business.

c) The evaluation that the risk of re-identification is below the described threshold is based on the professional judgment of probabilities relevant to the circumstances and assumptions documented in this report. Under no circumstances should the evaluation of re-identification risk be understood as, or represented as, a guarantee, warranty, or representation that it is impossible for one or more data subjects in the database to be re-identified. Given that the risk of re-identification is defined relative to the described threshold, a residual possibility remains that one or more data subjects in the project files could be re-identified.

d) It is recommended that Janssen re-assesses the assumptions and parameters underlying the evaluation of the risk of re-identification if material changes to these assumptions or parameters warrant it. The assumptions and parameters include, but are not limited to, the context (security, privacy, and contractual controls), variable definitions, and scope of the project files.

e) This evaluation of the risk of re-identification is subject to all the terms and limitations set forth in the agreement under which this reports was produced, including all attachments that are incorporated therein by reference.

©2020 Privacy Analytics Incorporated. All Rights Reserved. **Notice:** With the exception of Janssen information contained herein, this document is protected by copyright by Privacy Analytics Incorporated.

Contents

Contents **4**

1 Executive Summary **5**

2 Introduction **7**
2.1 Clinical Trial Overview . . . 7
2.2 Data Recipients . . . 7
2.3 Coverage . . . 7
2.4 Definitions . . . 8

3 Anonymization Process **9**
3.1 Use of Eclipse . . . 9
3.2 Output Format of De-identified Datasets . . . 9

4 Risk Analysis Methods **11**
4.1 Plausible Attack Models . . . 11
4.2 Thresholds . . . 11
4.2.1 Risk Threshold . . . 11
4.2.2 Uniqueness . . . 12
4.3 Risk Measurement Algorithms . . . 12
4.4 Other Considerations . . . 12
4.4.1 Prevalent Population . . . 12
4.4.2 Adversary Power . . . 13
4.4.3 Deceased Patients . . . 14

5 Direct and Quasi-Identifiers **15**

6 Transformations **29**
6.1 Direct Identifiers . . . 29
6.2 Quasi-identifiers . . . 29

7 Risk Measurement Results **34**

8 Conclusions **35**

References **36**

A Definitions **38**
A.1 Acronyms . . . 38
A.2 Identifiers . . . 38
A.3 Glossary . . . 40

1 Executive Summary

The purpose of this project was to assess the re-identification risk of the TOPMATMIG303 clinical trial data set.

Identifying fields have been categorized as direct or quasi-identifiers. These represent data items that can be used independently (direct identifiers) or in combination with one another (quasi-identifiers) to identify an individual.

A probability of re-identification threshold of 0.09 was chosen at the request of the client. The re-identification risk of the Janssen project files, before and after de-identification and the risk threshold was:

	Re-identification Risk	**Percent Uniques**
Original Trial	0.3	1.57%
Threshold	0.09	1%
De-identified Trial	0.0898	0%

The data masking and de-identification strategy required the following modifications to the original project files:

Identifier	**Transformation**
Subject IDs	Masked
Site IDs	Suppressed
Patient dates	PHUSE shifted
Free Text	Suppressed
Age	Generalized to 12-year intervals
Date of Birth	Suppressed
Height	Suppressed
Concomitant medication dates	Suppressed

Additionally, the data value columns from supplementary tables and other parameter-value tables containing identifying information were suppressed.

This report details the re-identification risks identified, the considerations taken in choosing the risk threshold, and the de-identification and masking steps that were applied. The methodology used for risk measurement and anonymization satisfies contemporary criteria for anonymization methodologies, is consistent with available

guidance from regulators, and has been publicly documented and peer-reviewed. It is also consistent with other standards and guidelines for clinical trial data anonymization.

2 Introduction

The purpose of this project was to perform anonymization of the Janssen TOPMATMIG303 clinical trial data set.

The anonymization of this data set is part of an initiative by Janssen to share more clinical trial data with the research community for secondary purposes. 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 research participants are used to maximum effect in the creation of knowledge. The data release is subject to certain criteria being met, including a requirement to effectively anonymize the data.

Janssen uses multiple platforms and mechanisms to share clinical trial data. This report documents the de-identification performed for data that is being released directly to researchers through YODA Project, a data sharing platform.

Janssen will review requests for data access and will only grant access to the data based on proposals they deem legitimate and valid from *bona fide* researchers. Janssen has indicated that they will respect the scope of the initial research for which patients have consented to have their data collected and analyzed, and only research proposals studying the same study drug or disease that was studied in the original study will be accepted. The data is made available to researchers after research proposals have been reviewed by an Independent Review Panel (IRP). The IRP consists of acknowledged experts that review the scientific merit of the data request, and is managed by the Wellcome Trust.

This report describes the re-identification risk determination that was performed on the Janssen TOPMATMIG303 trial data set, the considerations taken in deciding on an appropriate risk threshold, the re-identification risk measurements made, and the recommended de-identification and masking of the data.

2.1 Clinical Trial Overview

The data considered in this project was generated from the clinical trial, "A Double-Blind, Randomized, Placebo-Controlled, Multicenter Trial to Investigate the Efficacy and Tolerability of Topiramate in Prolonged Migraine Prevention" (`clinicaltrials.gov` identifier NCT00216619).

2.2 Data Recipients

Clinical trial data are being made available to *bona fide* researchers with legitimate and valid research proposals. Researchers can use YODA to request access to anonymized patient level data from clinical studies to conduct further research. All researchers must sign a standardized data sharing agreement before they get access to the data.

2.3 Coverage

In this project, work was performed directly on the original clinical trial data set. An analysis was performed on the data that was received. This evaluation of re-identification risk acknowledges that other information from

the trial will be shared with researchers, for example, Clinical Study Reports (CSRs) and summaries. However, it is assumed that no other identifiable patient data is being shared via this release, including identifiable patient data in clinical reports, patient narratives, and any other structured or unstructured datasets.

2.4 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 [3].

3 Anonymization Process

3.1 Use of Eclipse

The analysis described in this report was performed using Privacy Analytics Eclipse, a re-identification risk measurement software application developed by Privacy Analytics Inc.

3.2 Output Format of De-identified Datasets

Privacy Analytics performs all dataset de-identification within the SAS (Statistical Analysis System) native data file format (extension ".sas7bdat"). Datasets received in SAS version 5 (V5) 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 .sas7bdat format, inherent limitations in the .xpt format may require modifications. For consistency, Privacy Analytics employs the following reporting standard regardless of the format datasets were received in.

Based on the definition of the 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, Privacy Analytics 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 "TOPMATMIG303_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.

4 Risk Analysis Methods

The methodology used for risk measurement and anonymization satisfy contemporary criteria for anonymization methodologies [4, 5]. These criteria were derived from existing standards published by regulators, government agencies, and professional groups [8–10, 15, 18]. The general methods used by Privacy Analytics Inc. to evaluate the risk of re-identification for non-public data releases are described below with the relevant citations included where more technical detail can be found.

4.1 Plausible Attack Models

Three plausible attacks on the data were considered. The plausible attacks are consistent with the modeling of threat sources used in information security [14], and cover the universe of attacks that the data disclosure needs to protect against. In evaluating the overall re-identification risk for each of the attacks, the overall data release context of the data release is taken into account.

Attack	Equation		
T1	$Pr\left(re{-}id, attempt\right)$	$=$	$Pr\left(re{-}id \mid attempt\right) \times 0.3$
T2	$Pr\left(re{-}id, acquaintance\right)$	$=$	$Pr\left(re{-}id \mid acquaintance\right) \times 0.00042$
T3	$Pr\left(re{-}id, breach\right)$	$=$	$Pr\left(re{-}id \mid breach\right) \times 0.27$

Table 1: Summary of risk of the plausible attack models.

The remaining values in these equations need to be computed from the actual data set. These are the values for $Pr\left(re{-}id \mid attempt\right)$, $Pr\left(re{-}id \mid acquaintance\right)$, and $Pr\left(re{-}id \mid breach\right)$. There are a set of standard metrics to compute the re-identification probabilities from the actual data set [3].

4.2 Thresholds

4.2.1 Risk Threshold

For this trial, a re-identification risk threshold of 0.09 was chosen at the request of the client. Of note is that the choice of 0.09 corresponds to the risk threshold recommended by the EMA [6] for the public release of clinical data[1], in which the plausible attacks are all considered to be of probability 1. That is to say, an assessment of the data sharing context, and the potential invasion of privacy, would normally result in plausible attacks that have probability less than 1, given a data release on a secure data portal. Table 2 lists the risk levels to be maintained.

[1] As per Section 5.4, Step 2 of the EMA Policy 0070 External Guidance document, "Applicants/MAHs should identify possible adversaries and plausible attacks on the data and evaluate the impact on the risk of re-identification."

Attack	Equation		
T1	$Pr\left(re{-}id \mid attempt\right) \times 0.3$	$\leq$	0.09
T2	$Pr\left(re{-}id \mid acquaintance\right) \times 0.00042$	$\leq$	0.09
T3	$Pr\left(re{-}id \mid breach\right) \times 0.27$	$\leq$	0.09

Table 2: The risk from each plausible attack model must fall below the threshold.

4.2.2 Uniqueness

The average risk by itself will not ensure that there are no uniques in the data set. Therefore, an additional criterion is to ensure that there are no more than 1% of sample uniques that are population uniques on the basis of their demographic quasi-identifiers. The 1% cutoff for uniqueness is commonly used in the disclosure control community as a reasonable threshold (given that it is not possible for an adversary to know which sample unique is a population unique). In the case that rare demographic values or a rare incidence of death are present in the study, publicly available statistics may be used as an alternative to Eclipse's calculations for estimating uniqueness.

4.3 Risk Measurement Algorithms

The value for $Pr\left(re{-}id \mid \bullet\right)$ was computed as the average re-identification risk assuming that an adversary does not know who is in the database, known as average journalist risk in the literature [3]. Journalist risk is measured with two attack directions: sample to population and population to sample. Given that it is not possible to know in advance which method of attack an adversary will use, the overall risk needs to be formulated as the maximum of the two measurements.

4.4 Other Considerations

4.4.1 Prevalent Population

There are three general options in the choice of the prevalent population for risk measurement as illustrated in Figure 1: patients in the trial, patients in similar trials, and patients in the country (or region). Assuming that the population is all patients in the country is the least conservative approach, and has been used to anonymize other public data releases such as for Project Data Sphere [13]. Assuming that the population is all patients in the study is the most conservative strategy and results in extensive transformation to the data, and in practice will have severe consequences on data utility.

The choice of prevalent population for risk measurement should consider precedents as well as the assumptions one is willing to make about adversary knowledge. Assuming that the adversary knows who participated in a specific trial is a very strong assumption and is unlikely to be true for the primary adversaries considered. An adversary knowing that a patient participated in some trial for a given indication over a particular period of time is much more plausible. These different assumptions are illustrated in Figure 1.

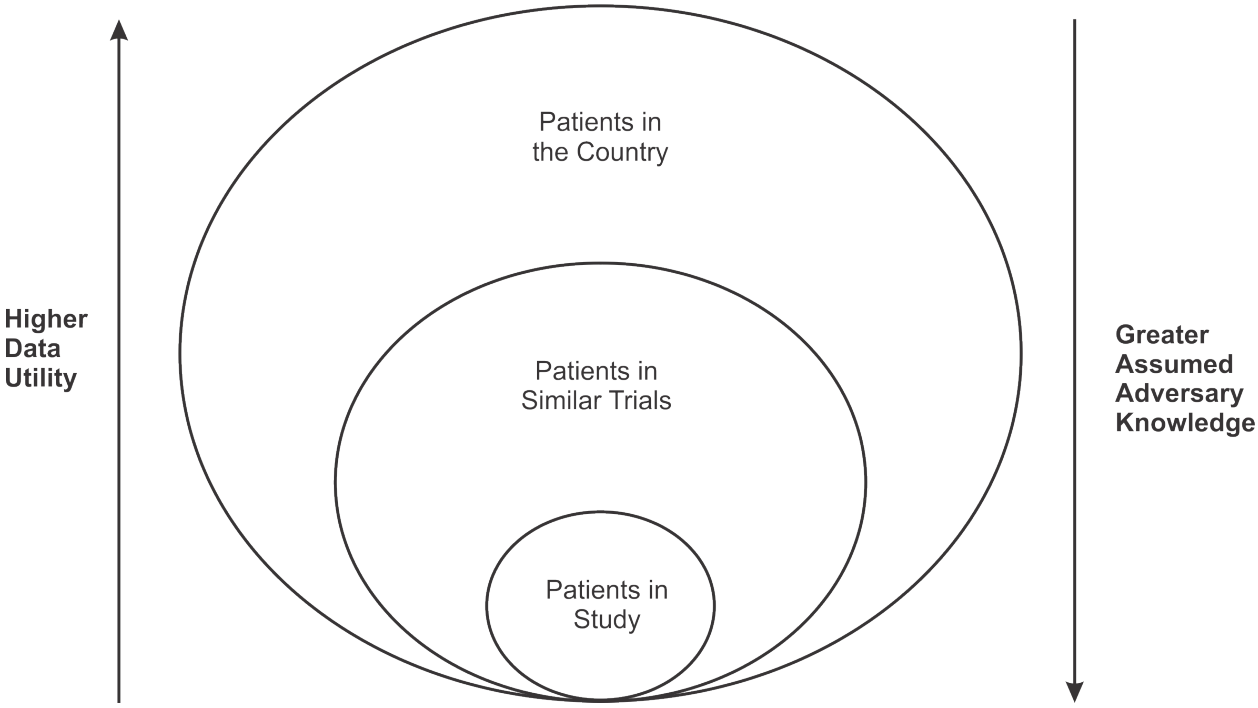

Figure 1: Different approaches to determining the prevalent population.

Therefore, in this analysis, the "patients in similar trials" assumption is used as it balances a number of factors. This is a plausible assumption about adversary knowledge consistent with the assumptions described in Section 4.1, and will ensure adequate data utility for the shared clinical report(s).

In this trial, the size of the prevalent population was calculated to be 1715.

4.4.2 Adversary Power

Plausible adversaries are modelled such that an adversary has limited knowledge of medical events, in addition to the quasi-identifiers that have minimal (or do not) change over time, and are more likely to be known (e.g., gender, date of birth, country).

The assumption of limited knowledge of the adversary is known as the power of the adversary. Previous research that considered the power of the adversary always assumed that the power is fixed for all patients [7, 12, 21–23]. The value of adversary power is denoted by p. This idea is used to model plausible adversaries and estimate the risk of re-identification.

Where medical events are described in structured datasets, they can be grouped together into events, representing related pieces of information. When a trial participant has multiple quasi-identifiers values associated with events, the concept of power of the adversary is applied to the number of events an adversary

could know about. As a result, attacks are modeled such that the adversary know about p medical events. If every medical event contains both a medical history term and a date, then it follows that the attack models an adversary knowing p medical history terms and p dates of medical events.

The value used in this risk measurement was $p = 3$. Given that known re-identification attacks on other public health data sets used only a single event in a longitudinal sequence of events [2, 20], this choice of power is more conservative than empirical attacks would suggest.

4.4.3 Deceased Patients

In some clinical trials, participants are not likely to survive for many years after the end of the study or may not survive the study. The question is whether it is necessary to protect the identities of deceased patients, at or to the same extent as live patients. Because global data from multiple jurisdictions may fall under different regulations, the lowest common denominator is the safest assumption to make (i.e., the most restrictive). In such a case it is prudent to treat all trial participants the same, whether they are deceased or not, and ensure that the risk of re-identification for all trial participants is below the computed thresholds.

In the US, information on the deceased is still considered personal health information for 50 years after death under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[2], and in general clinical sites will be covered by HIPAA [11]. In the EU, The Data Protection Directive 95/46/EC applies to "natural persons", and data on the deceased is not about natural persons any more. However, the Article 29 Working Party has noted that data on the deceased may still be personal information because it may indicate familial diseases relevant to living children or siblings, or may have other country-specific restrictions on it [1].

On this basis, the risk of re-identification has been evaluated and mitigated on all trial participants, regardless of whether or not an individual is still alive at the time of anonymization.

[2] 45 C.F.R. 160.103 General Provisions: Definitions - Protected Health Information.

5 Direct and Quasi-Identifiers

This section explains how pieces of information were determined to be direct and quasi-identifiers, and what identifiers are found in the clinical reports. Direct Identifiers (DIs) are variables in the clinical reports that would uniquely refer to a given trial participant (for example, a subject ID). Quasi-Identifiers (QIs) are variables or pieces of information that when used in conjunction with other QIs can be used to identify trial participants. Quasi-identifiers are further identified as being public and/or acquaintance quasi-identifiers. Public quasi-identifiers are those which could appear in public registries. For example, in the United States the voter registry lists contains age, ZIP code, and gender. An adversary could attempt to link information in the data set to a public registry in an attempt to re-identify someone in the data set. Acquaintance quasi-identifiers are those that only an acquaintance of the target would know; for example, event dates and diagnoses are considered to be acquaintance quasi-identifiers. Any quasi-identifier that is considered a public quasi-identifier, is also included in acquaintance risk measurements. A more thorough description of direct and quasi-identifiers is provided in Section A.

For the purpose of measuring the risk of re-identification, identifiers to be considered must be knowable, replicable, and distinguishable. For example, a serious adverse event is considered knowable, while a non-serious adverse event is not considered knowable. The identifier must be distinguishable, meaning that an adversary could use the identifier to distinguish among individuals in the clinical reports. For example, a diagnosis that is required as part of the inclusion criteria (such as a diagnosis of breast cancer in a study about breast cancer) would not fit the criterion of distinguishability. Finally, the identifier must be replicable, or sufficiently stable over time. For example, a trial participant's date of birth is replicable, but lab values are typically not sufficiently replicable to be considered identifiers.

If a piece of information does not meet the three criteria of being replicable, distinguishable, and knowable, it is not considered to be an identifier. If a piece of information is replicable, distinguishable, and knowable, then those that are uniquely identifying (for example, a name or subject ID) are considered direct identifiers. If the information is not uniquely identifying, and is analytically useful (for example, an age) the information is considered a quasi-identifier.

Lists of the direct and quasi-identifiers considered are provided in Table 3 and Table 4.

List of Direct Identifiers: Table	Column	Description
ADVERSE_EVENTS	AE_BRM_NUMBER	
ADVERSE_EVENTS	CRFID	
ADVERSE_EVENTS_FU	CRFID	crfid
CONCOMITANT_MED	CRFID	
CONCOMITANT_MED_FU	CRFID	
DEMOGRAPHIC_ANALYSIS_DATA	CRFID	crfid
DISPOSITION_DOUBLEBLIND	AE_BRM_NUMBER	
DISPOSITION_DOUBLEBLIND	CRFID	
DISPOSITION_OPENLABEL	AE_BRM_NUMBER	
DISPOSITION_OPENLABEL	CRFID	
EXPOSURE_DOUBLEBLINDPHASE	CRFID	crfid
EXPOSURE_OPENLABELPHASE	CRFID	crfid
MEDICALHISTORY	CRFID	crfid
PROPHYLACTICS_DURING_FU	CRFID	crfid

Table 3: List of direct identifiers.

List of Quasi-Identifiers: Table	Column	Description	Classification	Included in Risk Measurement
ADVERSE_EVENTS	AE_ACTION_TAKEN		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS	AE_COTHERAPY_STARTED		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
ADVERSE_EVENTS	AE_END_DATE		Acquaintance	Yes
ADVERSE_EVENTS	AE_MEDDRA_CODE		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS	AE_OUTCOME		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS	AE_SERIOUS		Acquaintance	Yes
ADVERSE_EVENTS	AE_SEVERITY		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS	AE_START_DATE		Acquaintance	Yes
ADVERSE_EVENTS	AE_VERBATIM		Acquaintance	No (Suppressed - free text)
ADVERSE_EVENTS	MEDDRA_SOC_TEXT		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS	MEDDRA_FIELD1_PTCODE		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
ADVERSE_EVENTS	MEDDRA_FIELD1_SOCCODE		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS	MEDDRA_LOWERLEVELTERM		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS	MEDDRA_PREFERREDTERM		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	DRUG	drug	Acquaintance	No (Suppressed)
ADVERSE_EVENTS_FU	FIELD1_PTCODE	ae meddra-pt code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	FIELD1_SOCCODE	ae meddra-soc code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	FIELD2_LLTTEXT	ae meddra-llt text	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
ADVERSE_EVENTS_FU	FIELD2_PTTEXT	ae meddra-pt text	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	FIELD2_SOCTEXT	ae meddra-soc text	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	GENERIC	generic	Acquaintance	No (Suppressed)
ADVERSE_EVENTS_FU	AE_ACT	action	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	AE_COTH	ct start	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	AE_END_D	ae end date	Acquaintance	No (Suppressed)
ADVERSE_EVENTS_FU	AE_MED_NAME	medication name	Acquaintance	No (Suppressed)
ADVERSE_EVENTS_FU	AE_MED_NAME_WHONO	who med number	Acquaintance	No (Suppressed)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
ADVERSE_EVENTS_FU	AE_MEDDRA	ae meddra-llt code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	AE_OUT	outcome	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	AE_SEV	severity	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
ADVERSE_EVENTS_FU	AE_START_D	ae from date	Acquaintance	No (Suppressed)
ADVERSE_EVENTS_FU	AE_V	ae verbatim	Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED	ATC_TXT0		Acquaintance	Yes
CONCOMITANT_MED	DRUG		Acquaintance	No (Suppressed)
CONCOMITANT_MED	FIELD2_PTTEXT		Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED	GENERIC		Acquaintance	No (Suppressed)
CONCOMITANT_MED	CT_END_D		Acquaintance	No (Correlated or repeated)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
CONCOMITANT_MED	CT_INDICATION		Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED	CT_ROUTE		Acquaintance	No (Correlated or repeated)
CONCOMITANT_MED	CT_ROUTE_SP		Acquaintance	No (Correlated or repeated)
CONCOMITANT_MED	CT_START_D		Acquaintance	Yes
CONCOMITANT_MED	CT_V		Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED	CT_VALUE		Acquaintance	No (Correlated or repeated)
CONCOMITANT_MED	CTROUTE		Acquaintance	No (Correlated or repeated)
CONCOMITANT_MED	INDIC_AE		Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED_FU	ATC_TXT0		Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED_FU	FIELD2_PTTEXT		Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED_FU	GENERIC		Acquaintance	No (Suppressed)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
CONCOMITANT_MED_FU	CM_END_D		Acquaintance	No (Suppressed)
CONCOMITANT_MED_FU	CM_START_D		Acquaintance	No (Suppressed)
CONCOMITANT_MED_FU	CT_END_D		Acquaintance	No (Suppressed)
CONCOMITANT_MED_FU	CT_INDICATION		Acquaintance	No (Suppressed - free text)
CONCOMITANT_MED_FU	CT_START_D		Acquaintance	No (Suppressed)
CONCOMITANT_MED_FU	CT_V		Acquaintance	No (Suppressed)
CONCOMITANT_MED_FU	CT_VALUE		Acquaintance	No (Suppressed)
CONCOMITANT_MED_FU	CT_WHONO		Acquaintance	No (Suppressed)
CONCOMITANT_MED_FU	PMP_LASTVISIT_D		Acquaintance	No (Suppressed)
DEMOGRAPHIC_ANALYSIS_DATA	AGE		Public	Yes
DEMOGRAPHIC_ANALYSIS_DATA	BIRTH_D	date of birth	Public	Yes
DEMOGRAPHIC_ANALYSIS_DATA	ENDDATE_DOUBLEBLIND_TREATMENT		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	ENDDATE_OPENLABEL_TREATMENT		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	ENDDATE_PROSPECTIVE_BASELINE		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	FIRSTVISITDATE		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	GENDER		Public	Yes

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
DEMOGRAPHIC_ANALYSIS_DATA	HEIGHT_BASELINE		Acquaintance	Yes
DEMOGRAPHIC_ANALYSIS_DATA	LASTVISITDATE		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	RACE		Public	Yes
DEMOGRAPHIC_ANALYSIS_DATA	RACE_SPECIFICATION		Public	No (Suppressed - free text)
DEMOGRAPHIC_ANALYSIS_DATA	STARTDATE_DOUBLEBLIND_TREATMENT		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	STARTDATE_OPENLABEL_TREATMENT		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	STARTDATE_PROSPECTIVE_BASELINE		Acquaintance	No (Study date)
DEMOGRAPHIC_ANALYSIS_DATA	WEIGHT_BASELINE		Acquaintance	Yes
DISPOSITION_DOUBLEBLIND	AE_COTHERAPY_STARTED		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_DOUBLEBLIND	AE_END_DATE		Acquaintance	No (Correlated or repeated)
DISPOSITION_DOUBLEBLIND	AE_MEDDRA_CODE		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
DISPOSITION_DOUBLEBLIND	AE_OUTCOME		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_DOUBLEBLIND	AE_SERIOUS		Acquaintance	No (Correlated or repeated)
DISPOSITION_DOUBLEBLIND	AE_SEVERITY		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_DOUBLEBLIND	AE_START_DATE		Acquaintance	No (Correlated or repeated)
DISPOSITION_DOUBLEBLIND	AE_VERBATIM		Acquaintance	No (Suppressed - free text)
DISPOSITION_DOUBLEBLIND	MEDDRA_SOC_CODE		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_DOUBLEBLIND	MEDDRA_SOC_TEXT		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
DISPOSITION_DOUBLEBLIND	MEDDRA_LOWERLEVELTERM		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_DOUBLEBLIND	MEDDRA_PREFERREDTERM		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_DOUBLEBLIND	REASON11_SP	specification why subject withdrew consent	Acquaintance	No (Suppressed)
DISPOSITION_DOUBLEBLIND	REASON13_SP	specification of other reason	Acquaintance	No (Suppressed - free text)
DISPOSITION_DOUBLEBLIND	REASON9	pregnancy	Acquaintance	No (Correlated or repeated)
DISPOSITION_FOLLOWUP	DATE_LAST_FU_VISIT		Acquaintance	No (Suppressed)
DISPOSITION_FOLLOWUP	FUVIS1	date 1st fu visit	Acquaintance	No (Suppressed)
DISPOSITION_FOLLOWUP	FUVIS2	date 2nd fu visit	Acquaintance	No (Suppressed)
DISPOSITION_FOLLOWUP	MAIN_REASON3_SP	specification other reason	Acquaintance	No (Suppressed - free text)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
DISPOSITION_OPENLABEL	AE_COTHERAPY_STARTED		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_OPENLABEL	AE_END_DATE		Acquaintance	Yes
DISPOSITION_OPENLABEL	AE_MEDDRA_CODE		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_OPENLABEL	AE_OUTCOME		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_OPENLABEL	AE_SERIOUS		Acquaintance	Yes
DISPOSITION_OPENLABEL	AE_SEVERITY		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_OPENLABEL	AE_START_DATE		Acquaintance	Yes
DISPOSITION_OPENLABEL	AE_VERBATIM		Acquaintance	No (Suppressed - free text)
DISPOSITION_OPENLABEL	MEDDRA_SOC_CODE		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
DISPOSITION_OPENLABEL	MEDDRA_SOC_TEXT		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_OPENLABEL	MEDDRA_LOWERLEVELTERM		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_OPENLABEL	MEDDRA_PREFERREDTERM		Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
DISPOSITION_OPENLABEL	REASON11_SP	specification why subject withdrew consent	Acquaintance	No (Suppressed - free text)
DISPOSITION_OPENLABEL	REASON13_SP	specification of other reason	Acquaintance	No (Suppressed - free text)
DISPOSITION_OPENLABEL	REASON9	pregnancy	Acquaintance	Yes
EXPOSURE_DOUBLEBLINDPHASE	DOUBLEBLINDMED_ENDDATE		Acquaintance	No (Study date)
EXPOSURE_DOUBLEBLINDPHASE	DOUBLEBLINDMED_STARTDATE		Acquaintance	No (Study date)
EXPOSURE_OPENLABELPHASE	OPENLABELMED_ENDDATE		Acquaintance	No (Study date)
EXPOSURE_OPENLABELPHASE	OPENLABELMED_STARTDATE		Acquaintance	No (Study date)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
MEDICALHISTORY	MH_CUR		Acquaintance	No (Correlated or repeated)
MEDICALHISTORY	MH_CUR_SP	active	Acquaintance	No (Suppressed - free text)
MEDICALHISTORY	MH_HIS		Acquaintance	No (Correlated or repeated)
MEDICALHISTORY	MH_HIS_SP	history	Acquaintance	No (Suppressed - free text)
MEDICALHISTORY	MH_NONE		Acquaintance	No (Correlated or repeated)
MEDICALHISTORY	MH_SYSTEM	system	Acquaintance	Yes
PROPHYLACTICS_DURING_FU	CM_END_D		Acquaintance	No (Suppressed)
PROPHYLACTICS_DURING_FU	CM_START_D		Acquaintance	No (Suppressed)
PROPHYLACTICS_DURING_FU	ENDPERIOD		Acquaintance	No (Suppressed)
PROPHYLACTICS_DURING_FU	GENERIC	drug: generic name	Acquaintance	No (Suppressed)
PROPHYLACTICS_DURING_FU	STARTPERIOD		Acquaintance	No (Suppressed)

Table 4: List of quasi-identifiers, the classification as public or acquaintance quasi-identifiers, and their handling for risk measurement.

6 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 risk measurement. In all cases, modifications to these fields are applied to all other linked fields.

6.1 Direct Identifiers

The patient direct identifier transformations are described in Table 5. 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.

Direct Identifier Transformations: **Table**	**Column**	**Transformation Applied**
ADVERSE_EVENTS	AE_BRM_NUMBER	Suppressed
ADVERSE_EVENTS	CRFID	Masked
ADVERSE_EVENTS_FU	CRFID	Masked
CONCOMITANT_MED	CRFID	Masked
CONCOMITANT_MED_FU	CRFID	Masked
DEMOGRAPHIC_ANALYSIS_DATA	CRFID	Masked
DISPOSITION_DOUBLEBLIND	AE_BRM_NUMBER	Suppressed
DISPOSITION_DOUBLEBLIND	CRFID	Masked
DISPOSITION_OPENLABEL	AE_BRM_NUMBER	Suppressed
DISPOSITION_OPENLABEL	CRFID	Masked
EXPOSURE_DOUBLEBLINDPHASE	CRFID	Masked
EXPOSURE_OPENLABELPHASE	CRFID	Masked
MEDICALHISTORY	CRFID	Masked
PROPHYLACTICS_DURING_FU	CRFID	Masked

Table 5: Explanation of transformations applied to direct identifiers.

6.2 Quasi-identifiers

The following data transformations have been applied to quasi-identifiers in this dataset:

Generalization Reduce the precision of a field. For this specific project, the age of subjects was generalized to 12-year intervals using Eclipse. Table 6 summarizes the mapping of age to generalized age by Eclipse for

age greater or equal to 0 years.

Age (Years)	Generalized Age Value (Years)
$0 \leq$ Age < 12	0
$12 \leq$ Age < 24	12
$24 \leq$ Age < 36	24
$36 \leq$ Age < 48	36
$48 \leq$ Age < 60	48
$60 \leq$ Age < 72	60
$72 \leq$ Age < 84	72

Table 6: 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 [16]. 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 November 2003).

As a limitation, when PHUSE date shifting a datetime variable, Eclipse replaces the time part with an appropriate number of zeros.

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

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

Table 7 describes the transformation that was carried out on each quasi-identifier. Identifiers that were not transformed are not listed in the Table.

Quasi-Identifier Transformations:		
Table	**Column**	**Transformation Applied**
ADVERSE_EVENTS	AE_END_DATE	PHUSE date shifting
ADVERSE_EVENTS	AE_START_DATE	PHUSE date shifting
ADVERSE_EVENTS	AE_VERBATIM	Global suppression
ADVERSE_EVENTS_FU	DRUG	Global suppression
ADVERSE_EVENTS_FU	GENERIC	Global suppression
ADVERSE_EVENTS_FU	AE_END_D	Global suppression
ADVERSE_EVENTS_FU	AE_MED_NAME	Global suppression

Quasi-Identifier Transformations (cont.): Table	Column	Transformation Applied
ADVERSE_EVENTS_FU	AE_MED_NAME_WHONO	Global suppression
ADVERSE_EVENTS_FU	AE_START_D	Global suppression
ADVERSE_EVENTS_FU	AE_V	Global suppression
CONCOMITANT_MED	DRUG	Global suppression
CONCOMITANT_MED	FIELD2_PTTEXT	Global suppression
CONCOMITANT_MED	GENERIC	Global suppression
CONCOMITANT_MED	CT_END_D	Global suppression
CONCOMITANT_MED	CT_INDICATION	Global suppression
CONCOMITANT_MED	CT_START_D	Global suppression
CONCOMITANT_MED	CT_V	Global suppression
CONCOMITANT_MED	INDIC_AE	Global suppression
CONCOMITANT_MED_FU	ATC_TXT0	Global suppression
CONCOMITANT_MED_FU	FIELD2_PTTEXT	Global suppression
CONCOMITANT_MED_FU	GENERIC	Global suppression
CONCOMITANT_MED_FU	CM_END_D	Global suppression
CONCOMITANT_MED_FU	CM_START_D	Global suppression
CONCOMITANT_MED_FU	CT_END_D	Global suppression
CONCOMITANT_MED_FU	CT_INDICATION	Global suppression
CONCOMITANT_MED_FU	CT_START_D	Global suppression
CONCOMITANT_MED_FU	CT_V	Global suppression
CONCOMITANT_MED_FU	CT_VALUE	Global suppression
CONCOMITANT_MED_FU	CT_WHONO	Global suppression
CONCOMITANT_MED_FU	PMP_LASTVISIT_D	Global suppression
DEMOGRAPHIC_ANALYSIS_DATA	AGE	Generalization of age to 12-year intervals
DEMOGRAPHIC_ANALYSIS_DATA	BIRTH_D	Global suppression
DEMOGRAPHIC_ANALYSIS_DATA	ENDDATE_DOUBLEBLIND_TREATMENT	PHUSE date shifting
DEMOGRAPHIC_ANALYSIS_DATA	ENDDATE_OPENLABEL_TREATMENT	PHUSE date shifting

Quasi-Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
DEMOGRAPHIC_ANALYSIS_DATA	ENDDATE_PROSPECTIVE_BASELINE	PHUSE date shifting
DEMOGRAPHIC_ANALYSIS_DATA	FIRSTVISITDATE	PHUSE date shifting
DEMOGRAPHIC_ANALYSIS_DATA	HEIGHT_BASELINE	Global suppression
DEMOGRAPHIC_ANALYSIS_DATA	LASTVISITDATE	PHUSE date shifting
DEMOGRAPHIC_ANALYSIS_DATA	RACE_SPECIFICATION	Global suppression
DEMOGRAPHIC_ANALYSIS_DATA	STARTDATE_DOUBLEBLIND_TREATMENT	PHUSE date shifting
DEMOGRAPHIC_ANALYSIS_DATA	STARTDATE_OPENLABEL_TREATMENT	PHUSE date shifting
DEMOGRAPHIC_ANALYSIS_DATA	STARTDATE_PROSPECTIVE_BASELINE	PHUSE date shifting
DISPOSITION_DOUBLEBLIND	AE_END_DATE	PHUSE date shifting
DISPOSITION_DOUBLEBLIND	AE_START_DATE	PHUSE date shifting
DISPOSITION_DOUBLEBLIND	AE_VERBATIM	Global suppression
DISPOSITION_DOUBLEBLIND	REASON11_SP	Global suppression
DISPOSITION_DOUBLEBLIND	REASON13_SP	Global suppression
DISPOSITION_FOLLOWUP	DATE_LAST_FU_VISIT	Global suppression
DISPOSITION_FOLLOWUP	FUVIS1	Global suppression
DISPOSITION_FOLLOWUP	FUVIS2	Global suppression
DISPOSITION_FOLLOWUP	MAIN_REASON3_SP	Global suppression
DISPOSITION_OPENLABEL	AE_END_DATE	PHUSE date shifting
DISPOSITION_OPENLABEL	AE_START_DATE	PHUSE date shifting
DISPOSITION_OPENLABEL	AE_VERBATIM	Global suppression
DISPOSITION_OPENLABEL	REASON11_SP	Global suppression
DISPOSITION_OPENLABEL	REASON13_SP	Global suppression
EXPOSURE_DOUBLEBLINDPHASE	DOUBLEBLINDMED_ENDDATE	PHUSE date shifting
EXPOSURE_DOUBLEBLINDPHASE	DOUBLEBLINDMED_STARTDATE	PHUSE date shifting
EXPOSURE_OPENLABELPHASE	OPENLABELMED_ENDDATE	PHUSE date shifting
EXPOSURE_OPENLABELPHASE	OPENLABELMED_STARTDATE	PHUSE date shifting
MEDICALHISTORY	MH_CUR_SP	Global suppression
MEDICALHISTORY	MH_HIS_SP	Global suppression

Quasi-Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
PROPHYLACTICS_DURING_FU	CM_END_D	Global suppression
PROPHYLACTICS_DURING_FU	CM_START_D	Global suppression
PROPHYLACTICS_DURING_FU	ENDPERIOD	Global suppression
PROPHYLACTICS_DURING_FU	GENERIC	Global suppression
PROPHYLACTICS_DURING_FU	STARTPERIOD	Global suppression

Table 7: Explanation of transformations applied to quasi-identifiers.

7 Risk Measurement Results

Re-Identification risk analysis was conducted using the methods described in Section 4.

Once all direct identifiers are masked, they no longer contribute to re-identification risk.

The residual risk from quasi-identifiers is described in Table 8. Observe that after anonymization, the re-identification risk falls below the threshold for both the re-identification risk and uniqueness.

	Re-identification Risk	Percent Uniques
Original Trial	0.3	1.57%
Acceptable Threshold	0.09	1%
De-identified Trial	0.0898	0%

Table 8: Summary of risk measurement results before and after anonymization.

8 Conclusions

The re-identification risk of the Janssen TOPMATMIG303 clinical trial database, after the anonymization as described in this report is 0.0898, which is below the data risk threshold of 0.09 imposed at the request of the client.

Note, however, that the re-identification risk above is founded on accurate estimates of the probabilities of re-identification attacks, given the level of mitigating controls and motives and capacity in the context of the data sharing environment.

References

[1] Article 29 Data Protection Working Party. Opinion 4/2007 on the concept of personal data. Technical Report WP136, June 2007.

[2] Daniel C. Barth-Jones. The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. SSRN Scholarly Paper ID 2076397, Social Science Research Network, Rochester, NY, July 2012.

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

[4] Khaled El Emam. *Risky Business: Sharing Health Data while Protecting Privacy*. Trafford, Untied States of America, 2013.

[5] Khaled El Emam. The twelve characteristics of a de-identification methodology. Technical report, Privacy Analytics Inc., 2016.

[6] European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use, November 2018.

[7] Yeye He and Jeffrey F. Naughton. Anonymization of Set-valued Data via Top-down, Local Generalization. *Proceedings of the VLDB Endowment*, 2(1):934–945, August 2009.

[8] Health System Use Technical Advisory Committee and the Data De-Identification Working Group. Best Practice Guidelines for Managing the Disclosure of De-Identified Health Information. Technical report, Canadian Institute for Health Information, 2010.

[9] Information Commissioner's Office. Anonymisation: Managing Data Protection Risk Code of Practice. Technical report, Information Commissioner's Office, 2012.

[10] Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Technical report, Washington, D.C., 2015.

[11] Jennifer Kulynych. HIPAA Compliance in Clinical Trials. *Journal of Oncology Practice*, 4(1):9–10, January 2008.

[12] Junqiang Liu and Ke Wang. Anonymizing Transaction Data by Integrating Suppression and Generalization. In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, *Advances in Knowledge Discovery and Data Mining*, volume 6118 of *Lecture Notes in Computer Science*, pages 171–180. Springer Berlin / Heidelberg, 2010.

[13] Bradley Malin. A De-identification Strategy Used for Sharing One Data Provider's Oncology Trials Data through the Project Data Sphere Repository. Technical report, Project Data Sphere, June 2013.

[14] National Institute of Standards and Technology. Guide for Conducting Risk Assessments. Special Publication SP - 800-30 Rev 1, NIST, Gaithersburg, MD, 2012. NIST Manuscript Publication Search.

[15] Office for Civil Rights. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Technical report, Department of Health and Human Services, Washington, DC, 2012.

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

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

[18] Subcommittee on Disclosure Limitation Methodology. Working paper 22: Report on statistical disclosure control. Technical report, Office of Management and Budget, 1994.

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

[20] Latanya Sweeney. Matching Known Patients to Health Records in Washington State Data. Technical report, Harvard University. Data Privacy Lab, 2013.

[21] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving Anonymization of Set-valued Data. *Proc. VLDB Endow.*, 1(1):115–125, August 2008.

[22] Yabo Xu, B. Fung, Ke Wang, A.W.C. Fu, and Jian Pei. Publishing Sensitive Transactions for Itemset Utility. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08*, pages 1109 –1114, December 2008.

[23] Yabo Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. Anonymizing Transaction Databases for Publication. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 767–775, New York, NY, USA, 2008. ACM.

A Definitions

A.1 Acronyms

CSR Clinical Study Report

DI Direct Identifier

FPE Format-Preserving Encryption

HIPAA Health Insurance Portability and Accountability Act

IRP Independent Review Panel

PHUSE Pharmaceutical Users Software Exchange

QI Quasi-Identifier

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 [17, 19], and these are illustrated in Table 9:

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.

For example, Table 9 shows the patient sex and year of birth (from which an age can be derived) as quasi-identifiers.

Other variables. These are the variables that are not really useful for determining an individual's identity. They may be clinically relevant or not. Examples of clinical variables are laboratory test results and drug dosage information. In Table 9 the lab test that was ordered and the test results are the clinical variables.

SUBJID	Sex	Year of Birth	Lab Test	Lab Result
1	Male	1959	Albumin, Serum	4.8
2	Male	1969	Creatine kinase	86
3	Female	1955	Alkaline Phosphatase	66
4	Male	1959	Bilirubin	Negative
5	Female	1942	BUN/Creatinine Ratio	17
6	Female	1975	Calcium, Serum	9.2
7	Female	1966	Free Thyroxine Index	2.7
8	Female	1987	Globulin, Total	3.5
9	Male	1959	B-type natriuretic peptide	134.1
10	Male	1967	Creatine kinase	80
11	Male	1968	Alanine aminotransferase	24
12	Female	1955	Cancer antigen 125	86
13	Male	1967	Creatine kinase	327
14	Male	1967	Creatine kinase	82
15	Female	1966	Creatinine	0.78
16	Female	1955	Triglycerides	147
17	Male	1967	Creatine kinase	73
18	Female	1956	Monocytes	12
19	Female	1956	HDL Cholesterol	68
20	Male	1978	Neutrophils	83
21	Female	1966	Prothrombin Time	16.9
22	Male	1967	Creatine kinase	68
23	Male	1971	White Blood Cell Count	13.0
24	Female	1954	Hemoglobin	14.8
25	Female	1977	Lipase, Serum	37
26	Male	1944	Cholesterol, Total	147
27	Male	1965	Hematocrit	45.3

Table 9: Example data used to illustrate the terminology used in this report.

A.3 Glossary

power of the adversary As the number of quasi-identifiers or events within a data set increases, the overall risk value also typically increases. Adversary power gives an upper bound on the number of values which could be known by an adversary since in many cases it is unlikely that an adversary would know the value associated with every quasi-identifier and event.

project files A collection of datasets (tables) associated with a structured IPD data release.