An Evaluation of Re-identification Risks for the Janssen Clinical Trial Data Set 54767414MMY3004: Data Recipient Version

August 20, 2019

Evaluation of Re-Identification Risks

Janssen contracted Privacy Analytics Inc. to perform the anonymization and assessment of the risk of re-identification for this trial, as well as author the Risk Determination Report. The purpose of this project is to evaluate the risk of re-identification found in the data provided by Janssen for the study 54767414MMY3004.

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.

Contents

1 Executive Summary

The purpose of this project was to assess the re-identification risk of the 54767414MMY3004 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	0%
Threshold	0.09	1%
De-identified Trial	0.0841	0%

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

Identifier	Transformation
Unique subject IDs	Masked
Subject IDs, site IDs	Suppressed
Free Text	Suppressed
Patient dates	PhUSE shifted
Age	Generalized to 10-year intervals
Date of Birth	Suppressed
Country	Suppressed
Childbearing potential	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 54767414MMY3004 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.

This report describes the re-identification risk determination that was performed on the Janssen 54767414MMY3004 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, "An open label, phase III study comparing daratumumab, bortezomib and dexamethasone (DVd) vs bortezomib and dexamethasone (Vd) in subjects with relapsed or refractory multiple myeloma " (`clinicaltrials.gov` identifier NCT02136134).

2.2 Data Recipients

Clinical trial data are being made available to *bona fide* researchers with legitimate and valid research proposals. Researchers can use Clinical Study Data Request (CSDR) 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.

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	$Pr\left(re{-}id, acquaintance\right)$	$=$	$Pr\left(re{-}id \mid acquaintance\right) \times 0.0045$
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	$Pr\left(re{-}id \mid acquaintance\right) \times 0.0045$	$\leq$	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).

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.

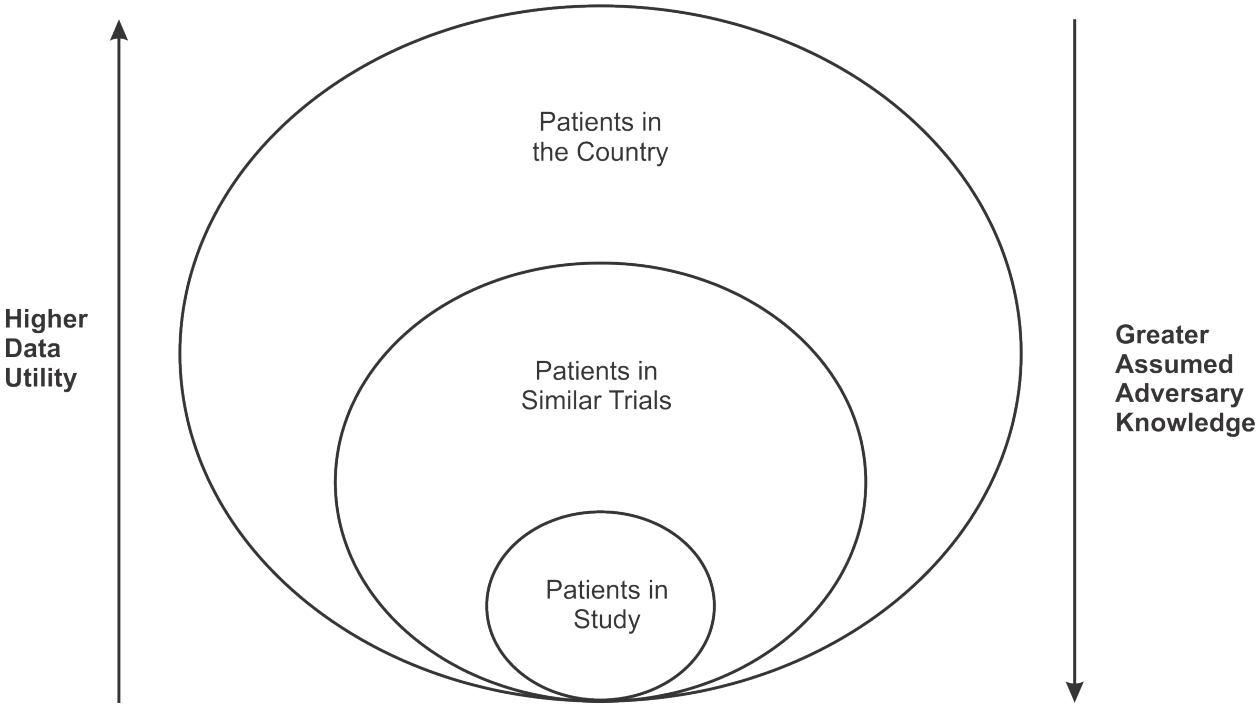

Figure 1: Different approaches to determining the prevalent population.

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 prevalent population was calculated to be 38229.

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
AE	AESPID	sponsor-defined identifier
AE	USUBJID	unique subject identifier
BE	BESPID	sponsor-defined identifier
BE	USUBJID	unique subject identifier
CE	CEGRPID	group id
CE	CESPID	sponsor-defined identifier
CE	USUBJID	unique subject identifier
CM	CMGRPID	group id
CM	CMSPID	sponsor-defined identifier
CM	USUBJID	unique subject identifier
CO	USUBJID	unique subject identifier
DD	USUBJID	unique subject identifier
DM	SITEID	study site identifier
DM	SUBJID	subject identifier for the study
DM	USUBJID	unique subject identifier
DS	DSSPID	sponsor-defined identifier
DS	USUBJID	unique subject identifier
DV	DVSEQ	sequence number
DV	DVSPID	sponsor-defined identifier
DV	USUBJID	unique subject identifier
EG	EGSPID	sponsor-defined identifier
EG	USUBJID	unique subject identifier
EX	EXGRPID	group id
EX	EXSPID	sponsor-defined identifier
EX	USUBJID	unique subject identifier
FA	FAGRPID	group id
FA	FASPID	sponsor-defined identifier
FA	USUBJID	unique subject identifier
HO	HOGRPID	group id

List of Direct Identifiers (cont.):		
Table	**Column**	**Description**
HO	HOSPID	sponsor-defined identifier
HO	USUBJID	unique subject identifier
IE	IESPID	sponsor-defined identifier
IE	USUBJID	unique subject identifier
IS	ISREFID	reference id
IS	ISSPID	sponsor-defined identifier
IS	USUBJID	unique subject identifier
LB	LBGRPID	group id
LB	LBREFID	specimen id
LB	LBSPID	sponsor-defined identifier
LB	USUBJID	unique subject identifier
MH	MHSPID	sponsor-defined identifier
MH	USUBJID	unique subject identifier
PC	PCREFID	reference id
PC	PCSPID	sponsor-defined identifier
PC	USUBJID	unique subject identifier
PF	PFGRPID	group id
PF	PFREFID	reference id
PF	PFSPID	sponsor-defined identifier
PF	USUBJID	unique subject identifier
PR	PRGRPID	group id
PR	PRSPID	sponsor-defined identifier
PR	USUBJID	unique subject identifier
QS	QSSPID	sponsor-defined identifier
QS	USUBJID	unique subject identifier
RE	RESPID	sponsor-defined identifier
RE	USUBJID	unique subject identifier
RELREC	USUBJID	unique subject identifier
RS	RSSPID	sponsor-defined identifier

List of Direct Identifiers (cont.):		
Table	**Column**	**Description**
RS	USUBJID	unique subject identifier
SC	SCSPID	sponsor-defined identifier
SC	USUBJID	unique subject identifier
SE	USUBJID	unique subject identifier
SS	SSSPID	sponsor-defined identifier
SS	USUBJID	unique subject identifier
SUPPAE	USUBJID	unique subject identifier
SUPPCE	USUBJID	unique subject identifier
SUPPCM	USUBJID	unique subject identifier
SUPPDM	USUBJID	unique subject identifier
SUPPDS	USUBJID	unique subject identifier
SUPPEG	USUBJID	unique subject identifier
SUPPEX	USUBJID	unique subject identifier
SUPPFA	USUBJID	unique subject identifier
SUPPHO	USUBJID	unique subject identifier
SUPPLB	USUBJID	unique subject identifier
SUPPPF	USUBJID	unique subject identifier
SUPPPR	USUBJID	unique subject identifier
SUPPQS	USUBJID	unique subject identifier
SUPPRS	USUBJID	unique subject identifier
SUPPTR	USUBJID	unique subject identifier
SUPPTU	USUBJID	unique subject identifier
SV	USUBJID	unique subject identifier
TR	TRGRPID	group id
TR	TRSPID	sponsor-defined identifier
TR	USUBJID	unique subject identifier
TU	TUSPID	sponsor-defined identifier
TU	USUBJID	unique subject identifier
VS	USUBJID	unique subject identifier

List of Direct Identifiers (cont.):		
Table	**Column**	**Description**
VS	VSSPID	sponsor-defined identifier
XC	USUBJID	unique subject identifier
XC	XCSPID	sponsor-defined identifier
XZ	USUBJID	unique subject identifier
XZ	XZSPID	sponsor-defined identifier
ZR	USUBJID	unique subject identifier
ZR	ZRSPID	sponsor-defined identifier

Table 3: List of direct identifiers.

List of Quasi-Identifiers:				
Table	**Column**	**Description**	**Classification**	**Included in Risk Measurement**
AE	AEACN	action taken with study treatment	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEBDSYCD	body system or organ class code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEBODSYS	body system or organ class	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AECONTRT	concomitant or additional trtmnt given	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEDECOD	dictionary-derived term	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
AE	AEENDTC	end date/time of adverse event	Acquaintance	Yes
AE	AEENDY	study day of end of adverse event	Acquaintance	No (Correlated or repeated)
AE	AEHLGT	high level group term	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEHLGTCD	high level group term code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEHLT	high level term	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEHLTCD	high level term code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AELLT	lowest level term	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AELLTCD	lowest level term code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEOUT	outcome of adverse event	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
AE	AEPTCD	preferred term code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AEREFID	reference id	Acquaintance	No (Suppressed)
AE	AEREL	causality	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AESDISAB	persist or signif disability/incapacity	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AESDTH	results in death	Acquaintance	No (Death measured elsewhere)
AE	AESER	serious event	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AESHOSP	requires or prolongs hospitalization	Acquaintance	Yes
AE	AESLIFE	is life threatening	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AESMIE	other medically important serious event	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AESOC	primary system organ class	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
AE	AESOCCD	primary system organ class code	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
AE	AESTDTC	start date/time of adverse event	Acquaintance	Yes
AE	AESTDY	study day of start of adverse event	Acquaintance	No (Correlated or repeated)
AE	AETERM	reported term for the adverse event	Acquaintance	No (Suppressed - free text)
AE	AETOXGR	standard toxicity grade	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
BE	BESTDTC	start date/time of event	Acquaintance	No (Study date)
BE	BESTDY	study day of start of event	Acquaintance	No (Study day)
CE	CECAT	category for clinical event	Acquaintance	No (Correlated or repeated)
CE	CEDECOD	dictionary-derived term	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
CE	CEOCCUR	clinical event occurrence	Acquaintance	No (Correlated or repeated)
CE	CEOUT	outcome of event	Acquaintance	No (Death measured elsewhere)
CE	CESTDTC	start date/time of clinical event	Acquaintance	No (Correlated or repeated)
CE	CESTDY	study day of start of clinical event	Acquaintance	No (Correlated or repeated)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
CE	CETERM	reported term for the clinical event	Acquaintance	No (Suppressed - free text)
CM	CMCLAS	medication class	Acquaintance	No (Correlated or repeated)
CM	CMCLASCD	medication class code	Acquaintance	Yes
CM	CMDECOD	standardized medication name	Acquaintance	No (Correlated or repeated)
CM	CMDOSE	dose per administration	Acquaintance	No (Correlated or repeated)
CM	CMDOSFRM	dose form	Acquaintance	No (Correlated or repeated)
CM	CMDOSFRQ	dosing frequency per interval	Acquaintance	No (Correlated or repeated)
CM	CMDOSTXT	dose description	Acquaintance	No (Correlated or repeated)
CM	CMDOSU	dose units	Acquaintance	No (Correlated or repeated)
CM	CMENDTC	end date/time of medication	Acquaintance	No (Correlated or repeated)
CM	CMENDY	study day of end of medication	Acquaintance	No (Correlated or repeated)
CM	CMENTPT	end reference time point	Acquaintance	No (Study date)
CM	CMROUTE	route of administration	Acquaintance	No (Correlated or repeated)
CM	CMSTDTC	start date/time of medication	Acquaintance	Yes
CM	CMSTDY	study day of start of medication	Acquaintance	No (Correlated or repeated)
CM	CMSTTPT	start reference time point	Acquaintance	No (Study date)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
CM	CMTRT	reported name of drug, med, or therapy	Acquaintance	No (Suppressed - free text)
CO	COVAL	comment	Acquaintance	No (Suppressed - free text)
DD	DDORRES	result or finding as collected	Acquaintance	No (Suppressed)
DD	DDSTRESC	standardized result in character format	Acquaintance	No (Death measured elsewhere)
DM	AGE	age	Public	Yes
DM	BRTHDTC	date/time of birth	Acquaintance	No (Suppressed)
DM	COUNTRY	country	Public	Yes
DM	DMDTC	date/time of collection	Acquaintance	No (Study date)
DM	DMDY	study day of collection	Acquaintance	No (Study day)
DM	DTHDTC	date/time of death	Public	Yes
DM	DTHFL	subject death flag	Public	No (Death measured elsewhere)
DM	ETHNIC	ethnicity	Public	Yes
DM	INVID	investigator identifier	Acquaintance	No (Suppressed)
DM	INVNAM	investigator name	Acquaintance	No (Suppressed)
DM	RACE	race	Public	Yes
DM	RFENDTC	subject reference end date/time	Acquaintance	No (Study date)
DM	RFICDTC	date/time of informed consent	Acquaintance	No (Study date)
DM	RFPENDTC	date/time of end of participation	Acquaintance	No (Study date)
DM	RFSTDTC	subject reference start date/time	Acquaintance	No (Study date)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
DM	RFXENDTC	date/time of last study treatment	Acquaintance	No (Study date)
DM	RFXSTDTC	date/time of first study treatment	Acquaintance	No (Study date)
DM	SEX	sex	Public	Yes
DS	DSDECOD	standardized disposition term	Acquaintance	No (Suppressed)
DS	DSDTC	date/time of collection	Acquaintance	No (Study date)
DS	DSDY	study day of collection	Acquaintance	No (Study day)
DS	DSSTDTC	start date/time of disposition event	Acquaintance	No (Study date)
DS	DSSTDY	study day of start of disposition event	Acquaintance	No (Study day)
DS	DSTERM	reported term for the disposition event	Acquaintance	No (Suppressed - free text)
DV	DVSTDTC	start date/time of deviation	Acquaintance	No (Study date)
DV	DVSTDY	study day of start of deviation	Acquaintance	No (Study day)
DV	DVTERM	protocol deviation term	Acquaintance	No (Suppressed - free text)
EG	EGDTC	date/time of ecg	Acquaintance	No (Study date)
EG	EGDY	study day of ecg	Acquaintance	No (Study day)
EG	EGORRES	result or finding in original units	Acquaintance	No (Suppressed - free text)
EG	VISITDY	planned study day of visit	Acquaintance	No (Study day)
EX	EXENDTC	end date/time of treatment	Acquaintance	No (Study date)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
EX	EXENDY	study day of end of treatment	Acquaintance	No (Study day)
EX	EXSTDTC	start date/time of treatment	Acquaintance	No (Study date)
EX	EXSTDY	study day of start of treatment	Acquaintance	No (Study day)
FA	FADTC	date/time of collection	Acquaintance	No (Study date)
FA	FADY	study day of collection	Acquaintance	No (Study day)
FA	FAMETHOD	method of test or examination	Acquaintance	No (Suppressed)
FA	FAOBJ	object of the observation	Acquaintance	No (Suppressed - free text)
FA	FAORRES	result or finding in original units	Acquaintance	No (Suppressed)
FA	FASPEC	specimen type	Acquaintance	No (Suppressed)
FA	FASTRESC	character result/finding in std format	Acquaintance	No (Suppressed)
FA	VISITDY	planned study day of visit	Acquaintance	No (Study day)
HO	HOCAT	category for healthcare encounter	Acquaintance	Yes
HO	HODECOD	dictionary-derived term	Acquaintance	No (Only include AE outcome (death/hospitalization) in RM)
HO	HOENDTC	end date/time of healthcare encounter	Acquaintance	Yes
HO	HOENDY	study day of end of healthcare encounter	Acquaintance	No (Correlated or repeated)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
HO	HOSTDTC	start date/time of healthcare encounter	Acquaintance	Yes
HO	HOSTDY	study day of start of event	Acquaintance	No (Correlated or repeated)
HO	HOTERM	reported term for healthcare encounter	Acquaintance	No (Suppressed - free text)
IE	IEDTC	date/time of collection	Acquaintance	No (Study date)
IE	IEDY	study day of collection	Acquaintance	No (Study day)
IE	IEORRES	i/e criterion original result	Acquaintance	No (Suppressed)
IE	IESTRESC	i/e criterion result in std format	Acquaintance	No (Suppressed)
IE	IETEST	inclusion/exclusion criterion	Acquaintance	No (Suppressed)
IE	IETESTCD	inclusion/exclusion criterion short name	Acquaintance	No (Suppressed)
IE	VISITDY	planned study day of visit	Acquaintance	No (Study day)
IS	ISDTC	date/time of collection	Acquaintance	No (Study date)
IS	ISDY	study day of collection	Acquaintance	No (Study day)
IS	ISNAM	vendor name	Acquaintance	No (Suppressed)
IS	ISRFTDTC	date/time of reference point	Acquaintance	No (Study date)
LB	LBDTC	date/time of specimen collection	Acquaintance	No (Study date)
LB	LBDY	study day of specimen collection	Acquaintance	No (Study day)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
LB	LBENDTC	end date/time of specimen collection	Acquaintance	No (Study date)
LB	LBENDY	study day of end of specimen collection	Acquaintance	No (Study day)
LB	LBNAM	vendor name	Acquaintance	No (Suppressed)
LB	LBREASND	reason test not done	Acquaintance	No (Suppressed - free text)
LB	VISITDY	planned study day of visit	Acquaintance	No (Study day)
MH	MHBODSYS	body system or organ class	Acquaintance	No (Correlated or repeated)
MH	MHCAT	category for medical history	Acquaintance	No (Correlated or repeated)
MH	MHDECOD	dictionary-derived term	Acquaintance	Yes
MH	MHDTC	date/time of history collection	Acquaintance	No (Study date)
MH	MHENRTPT	end relative to reference time point	Acquaintance	No (Correlated or repeated)
MH	MHSCAT	subcategory for medical history	Acquaintance	No (Correlated or repeated)
MH	MHSTDTC	start date/time of medical history event	Acquaintance	Yes
MH	MHSTDY	study day of start of history event	Acquaintance	No (Correlated or repeated)
MH	MHTERM	reported term for the medical history	Acquaintance	No (Suppressed - free text)
PC	PCDTC	date/time of specimen collection	Acquaintance	No (Study date)
PC	PCDY	actual study day of specimen collection	Acquaintance	No (Study day)
PC	PCNAM	vendor name	Acquaintance	No (Suppressed)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
PC	PCRFTDTC	date/time of reference point	Acquaintance	No (Study date)
PF	PFDTC	date/time of specimen collection	Acquaintance	No (Study date)
PF	PFDY	study day of specimen collection	Acquaintance	No (Study day)
PF	PFNAM	vendor name	Acquaintance	No (Suppressed)
PF	VISITDY	planned study day of visit	Acquaintance	No (Study day)
PR	PRCAT	category for procedure	Acquaintance	No (Correlated or repeated)
PR	PRENDTC	end date/time of procedure	Acquaintance	No (Correlated or repeated)
PR	PRENDY	study day of end of procedure	Acquaintance	No (Correlated or repeated)
PR	PRLOC	location of procedure	Acquaintance	No (Suppressed - free text)
PR	PROCCUR	procedure occurrence	Acquaintance	No (Correlated or repeated)
PR	PRSCAT	subcategory for procedure	Acquaintance	No (Correlated or repeated)
PR	PRSTDTC	start date/time of therapeutic procedure	Acquaintance	Yes
PR	PRSTDY	study day of start of procedure	Acquaintance	No (Correlated or repeated)
PR	PRTRT	name of procedure	Acquaintance	No (Suppressed - free text)
QS	QSDTC	date/time of finding	Acquaintance	No (Study date)
QS	QSDY	study day of finding	Acquaintance	No (Study day)
QS	QSORRES	finding in original units	Acquaintance	No (Suppressed)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
QS	QSSTRESC	character result/finding in std format	Acquaintance	No (Suppressed)
QS	QSSTRESN	numeric finding in standard units	Acquaintance	No (Suppressed)
QS	VISITDY	planned study day of visit	Acquaintance	No (Study day)
RE	REDTC	date/time of measurements	Acquaintance	No (Study date)
RE	REDY	study day of measurements	Acquaintance	No (Study day)
RE	VISITDY	planned study day of visit	Acquaintance	No (Study day)
RS	RSDTC	date/time of response assessment	Acquaintance	No (Study date)
RS	RSDY	study day of response assessment	Acquaintance	No (Study day)
SC	SCORRES	result or finding in original units	Acquaintance	No (Correlated or repeated)
SC	SCSTRESC	character result/finding in std format	Acquaintance	Yes
SE	SEENDTC	end date/time of element	Acquaintance	No (Study date)
SE	SEENDY	study day of end of element	Acquaintance	No (Study day)
SE	SESTDTC	start date/time of element	Acquaintance	No (Study date)
SE	SESTDY	study day of start of element	Acquaintance	No (Study day)
SS	SSDTC	date/time of test	Acquaintance	No (Study date)
SS	SSDY	study day of test	Acquaintance	No (Study day)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
SUPPAE	QVAL	data value	Acquaintance	No (Suppressed)
SUPPCE	QVAL	data value	Acquaintance	No (Study date)
SUPPCM	QVAL	data value	Acquaintance	No (Suppressed)
SUPPDM	QVAL	data value	Acquaintance	No (Suppressed)
SUPPDS	QVAL	data value	Acquaintance	No (Suppressed)
SUPPEX	QVAL	data value	Acquaintance	No (Suppressed)
SUPPFA	QVAL	data value	Acquaintance	No (Suppressed)
SUPPHO	QVAL	data value	Acquaintance	No (Suppressed)
SUPPPR	QVAL	data value	Acquaintance	No (Suppressed)
SUPPRS	QVAL	data value	Acquaintance	No (Suppressed)
SUPPTR	QVAL	data value	Acquaintance	No (Suppressed)
SUPPTU	QVAL	data value	Acquaintance	No (Suppressed)
SV	SVENDTC	end date/time of visit	Acquaintance	No (Study date)
SV	SVENDY	study day of end of visit	Acquaintance	No (Study day)
SV	SVSTDTC	start date/time of visit	Acquaintance	No (Study date)
SV	SVSTDY	study day of start of visit	Acquaintance	No (Study day)
SV	VISITDY	planned study day of visit	Acquaintance	No (Study day)
TR	TRDTC	date/time of tumor measurement	Acquaintance	No (Study date)
TR	TRDY	study day of tumor measurement	Acquaintance	No (Study day)
TR	TRORRES	result or finding in original units	Acquaintance	No (Suppressed)
TR	TRSTRESC	character result/finding in std format	Acquaintance	No (Suppressed)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
TR	TRSTRESN	numeric result/finding in standard units	Acquaintance	No (Suppressed)
TR	VISITDY	planned study day of visit	Acquaintance	No (Study day)
TU	TUDTC	date/time of tumor identification	Acquaintance	No (Study date)
TU	TUDY	study day of tumor identification	Acquaintance	No (Study day)
VS	VISITDY	planned study day of visit	Acquaintance	No (Study day)
VS	VSDTC	date/time of measurements	Acquaintance	No (Study date)
VS	VSDY	study day of vital signs	Acquaintance	No (Study day)
VS	VSORRES	result or finding in original units	Acquaintance	No (Suppressed)
VS	VSRFTDTC	date/time of reference time point	Acquaintance	No (Study date)
VS	VSSTRESC	character result/finding in std format	Acquaintance	No (Suppressed)
VS	VSSTRESN	numeric result/finding in standard units	Acquaintance	No (Suppressed)
XC	VISITDY	planned study day of visit	Acquaintance	No (Study day)
XC	XCDTC	date/time of specimen collection	Acquaintance	No (Study date)
XC	XCDY	actual study day of specimen collection	Acquaintance	No (Study day)

List of Quasi-Identifiers (cont.): Table	Column	Description	Classification	Included in Risk Measurement
XC	XCORRES	result or finding in original units	Acquaintance	No (Suppressed)
XC	XCSTRESC	character result/finding in std format	Acquaintance	No (Suppressed)
XZ	XZDTC	date/time of specimen collection	Acquaintance	No (Study date)
XZ	XZDY	study day of specimen collection	Acquaintance	No (Study day)
XZ	XZSTRESN	numeric result/finding in standard units	Acquaintance	No (Suppressed)
ZR	VISITDY	planned study day of visit	Acquaintance	No (Study day)
ZR	ZRDTC	date/time of randomization	Acquaintance	No (Study date)
ZR	ZRDY	study day of randomization	Acquaintance	No (Study day)
ZR	ZRNAM	vendor name	Acquaintance	No (Suppressed)
ZR	ZRORRES	result or finding in original units	Acquaintance	No (Suppressed)
ZR	ZRSTRESC	character result/finding in std format	Acquaintance	No (Suppressed)
ZR	ZRSTRESN	numeric result/finding in standard units	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 meausrement. 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 encyption. This type of encryption creates an encrypted value that has the same length as the original ID.

Direct Identifier Transformations:		
Table	**Column**	**Transformation Applied**
AE	AESPID	Suppressed
AE	USUBJID	Masked
BE	BESPID	Suppressed
BE	USUBJID	Masked
CE	CEGRPID	Suppressed
CE	CESPID	Suppressed
CE	USUBJID	Masked
CM	CMGRPID	Suppressed
CM	CMSPID	Suppressed
CM	USUBJID	Masked
CO	USUBJID	Masked
DD	USUBJID	Masked
DM	SITEID	Suppressed
DM	SUBJID	Suppressed
DM	USUBJID	Masked
DS	DSSPID	Suppressed
DS	USUBJID	Masked
DV	DVSEQ	Suppressed
DV	DVSPID	Suppressed
DV	USUBJID	Masked

Direct Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
EG	EGSPID	Suppressed
EG	USUBJID	Masked
EX	EXGRPID	Suppressed
EX	EXSPID	Suppressed
EX	USUBJID	Masked
FA	FAGRPID	Suppressed
FA	FASPID	Suppressed
FA	USUBJID	Masked
HO	HOGRPID	Suppressed
HO	HOSPID	Suppressed
HO	USUBJID	Masked
IE	IESPID	Suppressed
IE	USUBJID	Masked
IS	ISREFID	Suppressed
IS	ISSPID	Suppressed
IS	USUBJID	Masked
LB	LBGRPID	Suppressed
LB	LBREFID	Suppressed
LB	LBSPID	Suppressed
LB	USUBJID	Masked
MH	MHSPID	Suppressed
MH	USUBJID	Masked
PC	PCREFID	Suppressed
PC	PCSPID	Suppressed
PC	USUBJID	Masked
PF	PFGRPID	Suppressed
PF	PFREFID	Suppressed
PF	PFSPID	Suppressed
PF	USUBJID	Masked

Direct Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
PR	PRGRPID	Suppressed
PR	PRSPID	Suppressed
PR	USUBJID	Masked
QS	QSSPID	Suppressed
QS	USUBJID	Masked
RE	RESPID	Suppressed
RE	USUBJID	Masked
RELREC	USUBJID	Masked
RS	RSSPID	Suppressed
RS	USUBJID	Masked
SC	SCSPID	Suppressed
SC	USUBJID	Masked
SE	USUBJID	Masked
SS	SSSPID	Suppressed
SS	USUBJID	Masked
SUPPAE	USUBJID	Masked
SUPPCE	USUBJID	Masked
SUPPCM	USUBJID	Masked
SUPPDM	USUBJID	Masked
SUPPDS	USUBJID	Masked
SUPPEG	USUBJID	Masked
SUPPEX	USUBJID	Masked
SUPPFA	USUBJID	Masked
SUPPHO	USUBJID	Masked
SUPPLB	USUBJID	Masked
SUPPPF	USUBJID	Masked
SUPPPR	USUBJID	Masked
SUPPQS	USUBJID	Masked
SUPPRS	USUBJID	Masked

Direct Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
SUPPTR	USUBJID	Masked
SUPPTU	USUBJID	Masked
SV	USUBJID	Masked
TR	TRGRPID	Suppressed
TR	TRSPID	Suppressed
TR	USUBJID	Masked
TU	TUSPID	Suppressed
TU	USUBJID	Masked
VS	USUBJID	Masked
VS	VSSPID	Suppressed
XC	USUBJID	Masked
XC	XCSPID	Suppressed
XZ	USUBJID	Masked
XZ	XZSPID	Suppressed
ZR	USUBJID	Masked
ZR	ZRSPID	Suppressed

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 10-year intervals using Eclipse. Table 6 summarizes the mapping of age to generalized age by Eclipse for age greater or equal to 18 years.

PhUSE date shifting Offset a date value according to the scheme defined in the 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, 04 September 2014).

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

Age (Years)	Generalized Age Value (Years)
$18 \leq$ Age < 28	18
$28 \leq$ Age < 38	28
$38 \leq$ Age < 48	38
$48 \leq$ Age < 58	48
$58 \leq$ Age < 68	58
$68 \leq$ Age < 78	68
$78 \leq$ Age < 88	78
$88 \leq$ Age ≤ 98	88

Table 6: Age generalization

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**
AE	AEENDTC	PhUSE date shifting
AE	AEREFID	Global suppression
AE	AESTDTC	PhUSE date shifting
AE	AETERM	Global suppression
BE	BESTDTC	PhUSE date shifting
CE	CESTDTC	PhUSE date shifting
CE	CETERM	Global suppression
CM	CMDECOD	Global suppression
CM	CMENDTC	PhUSE date shifting
CM	CMENTPT	PhUSE date shifting
CM	CMSTDTC	PhUSE date shifting
CM	CMSTTPT	PhUSE date shifting
CM	CMTRT	Global suppression
CO	COVAL	Global suppression

Quasi-Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
DD	DDORRES	Global suppression
DM	AGE	Generalization of age to 10-year intervals
DM	BRTHDTC	Global suppression
DM	COUNTRY	Global suppression
DM	DMDTC	PhUSE date shifting
DM	DTHDTC	PhUSE date shifting
DM	INVID	Global suppression
DM	INVNAM	Global suppression
DM	RFENDTC	PhUSE date shifting
DM	RFICDTC	PhUSE date shifting
DM	RFPENDTC	PhUSE date shifting
DM	RFSTDTC	PhUSE date shifting
DM	RFXENDTC	PhUSE date shifting
DM	RFXSTDTC	PhUSE date shifting
DS	DSDECOD	Global suppression
DS	DSDTC	PhUSE date shifting
DS	DSSTDTC	PhUSE date shifting
DS	DSTERM	Global suppression
DV	DVSTDTC	PhUSE date shifting
DV	DVTERM	Global suppression
EG	EGDTC	PhUSE date shifting
EG	EGORRES	Global suppression
EX	EXENDTC	PhUSE date shifting
EX	EXSTDTC	PhUSE date shifting
FA	FADTC	PhUSE date shifting
FA	FAMETHOD	Global suppression
FA	FAOBJ	Global suppression
FA	FAORRES	Global suppression
FA	FASPEC	Global suppression

Quasi-Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
FA	FASTRESC	Global suppression
HO	HOENDTC	PhUSE date shifting
HO	HOSTDTC	PhUSE date shifting
HO	HOTERM	Global suppression
IE	IEDTC	PhUSE date shifting
IE	IEORRES	Global suppression
IE	IESTRESC	Global suppression
IE	IETEST	Global suppression
IE	IETESTCD	Global suppression
IS	ISDTC	PhUSE date shifting
IS	ISNAM	Global suppression
IS	ISRFTDTC	PhUSE date shifting
LB	LBDTC	PhUSE date shifting
LB	LBENDTC	PhUSE date shifting
LB	LBNAM	Global suppression
LB	LBREASND	Global suppression
MH	MHDTC	PhUSE date shifting
MH	MHSTDTC	PhUSE date shifting
MH	MHTERM	Global suppression
PC	PCDTC	PhUSE date shifting
PC	PCNAM	Global suppression
PC	PCRFTDTC	PhUSE date shifting
PF	PFDTC	PhUSE date shifting
PF	PFNAM	Global suppression
PR	PRENDTC	PhUSE date shifting
PR	PRLOC	Global suppression
PR	PRSTDTC	PhUSE date shifting
PR	PRTRT	Global suppression
QS	QSDTC	PhUSE date shifting

Quasi-Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
QS	QSORRES	Global suppression
QS	QSSTRESC	Global suppression
QS	QSSTRESN	Global suppression
RE	REDTC	PhUSE date shifting
RS	RSDTC	PhUSE date shifting
SC	SCORRES	Global suppression
SC	SCSTRESC	Global suppression
SE	SEENDTC	PhUSE date shifting
SE	SESTDTC	PhUSE date shifting
SS	SSDTC	PhUSE date shifting
SUPPAE	QVAL	Global suppression
SUPPCE	QVAL	PhUSE date shifting
SUPPCM	QVAL	Global suppression
SUPPDM	QVAL	Global suppression
SUPPDS	QVAL	Global suppression
SUPPEX	QVAL	Global suppression
SUPPFA	QVAL	Global suppression
SUPPHO	QVAL	Global suppression
SUPPPR	QVAL	Global suppression
SUPPRS	QVAL	Global suppression
SUPPTR	QVAL	Global suppression
SUPPTU	QVAL	Global suppression
SV	SVENDTC	PhUSE date shifting
SV	SVSTDTC	PhUSE date shifting
TR	TRDTC	PhUSE date shifting
TR	TRORRES	Global suppression
TR	TRSTRESC	Global suppression
TR	TRSTRESN	Global suppression
TU	TUDTC	PhUSE date shifting

Quasi-Identifier Transformations (cont.):		
Table	**Column**	**Transformation Applied**
VS	VSDTC	PhUSE date shifting
VS	VSORRES	Global suppression
VS	VSRFTDTC	PhUSE date shifting
VS	VSSTRESC	Global suppression
VS	VSSTRESN	Global suppression
XC	XCDTC	PhUSE date shifting
XC	XCORRES	Global suppression
XC	XCSTRESC	Global suppression
XZ	XZDTC	PhUSE date shifting
XZ	XZSTRESN	Global suppression
ZR	ZRDTC	PhUSE date shifting
ZR	ZRNAM	Global suppression
ZR	ZRORRES	Global suppression
ZR	ZRSTRESC	Global suppression
ZR	ZRSTRESN	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	0%
Acceptable Threshold	0.09	1%
De-identified Trial	0.0841	0%

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

8 Conclusions

The re-identification risk of the Janssen 54767414MMY3004 clinical trial database, after the anonymization as described in this report is 0.0841, 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

CSDR Clinical Study Data Request

CSR Clinical Study Report

DI Direct Identifier

HIPAA Health Insurance Portability and Accountability Act

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.