A De-identification Strategy Used for Sharing One Data Provider's Oncology Trials Data through the *Project Data Sphere*® Repository

Prepared by:

Bradley Malin, Ph.D.
2525 West End Avenue, Suite 1030
Nashville, TN 37203
Email: brad.malin@gmail.com
Phone: +1 615 775 3553

June_ 2013

Table of Contents

THIS DOCUMENT PROVIDES ONE EXPERT'S OVERVIEW OF HOW DE-IDENTIFICATION IS DEFINED IN DIRECTIVES AND REGULATIONS, THE PRINCIPLES BY WHICH THE EXPERT ADDRESSED IDENTIFIABILITY, AND AN EXPLANATION OF HOW THE EXPERT ACCOMPLISHED DE-IDENTIFICATION FOR SPECIFIC DATASETS FOR SUBMISSION BY ONE DATA PROVIDER TO THE *PROJECT DATA SPHERE*® REPOSITORY. THE DOCUMENT IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY, IS NOT OFFERED AS LEGAL ADVICE AND SHOULD NOT BE USED AS A SUBSTITUTE FOR SEEKING LEGAL OR OTHER EXPERT ADVICE. *PROJECT DATA SPHERE, LLC* MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED, WITH RESPECT TO THE USEFULNESS OR SUFFICIENCY OF THE INFORMATION FOR PURPOSES OF DE-IDENTIFYING ANY CLINICAL TRIAL DATA AND IS NOT RESPONSIBLE FOR ANY ACTION OR FAILURE TO ACT IN RELIANCE UPON INFORMATION IN THIS DOCUMENT.

Executive Summary

Project Data Sphere, LLC, an independent initiative of the CEO Roundtable on Cancer's Life Sciences Consortium, sponsors a shared platform by which clinical trial data will be available to researchers for further study, with the goal of accelerating innovation to improve cancer research.

It is anticipated that the *Project Data Sphere*® repository will receive data from private pharmaceutical companies and other entities, such as cooperative groups and academic medical centers. The trial data will be accessible to registered users of the *Project Data Sphere* repository around the world. To protect the privacy of the corresponding research subjects, *Project Data Sphere*, LLC requires data providers to share data in a de-identified manner that satisfies applicable legal requirements.

In preparation for uploading its datasets to the *Project Data Sphere* repository, one data provider engaged an expert in de-identification. Through this engagement, the expert developed a de-identification approach for, and assessed it with, a specific dataset that the data provider proposed to upload to the *Project Data Sphere* repository. The goal was to create a data privacy protection approach for this specific dataset that met or exceeded the de-identification requirements of various countries and regions, such as the EU Data Protection Directive, local European member state directives, and the Privacy Rule of the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Since certain directives, such as the EU Data Protection Directive, do not provide specific technical guidelines by which de-identification should be assessed, the expert focused primarily on the HIPAA Privacy Rule.

With respect to the HIPAA de-identification model, the expert concluded that certain aspects of traditional oncology trial data research cannot be achieved if the requirements of a *Safe Harbor* strategy (e.g., date-stamped data points) are satisfied. As such, the expert investigated the extent to which data associated with the data provider's datasets satisfied the *Expert Determination* de-identification strategy (also commonly referred to as the "Statistical Standard"). In doing so, the expert determined that a sufficient level of de-identification could be achieved for the dataset by generalizing certain aspects of trial participants' demographics (e.g., place of residence and age at time of an adverse event, such as death). The expert's analysis showed that the calculated risk associated with re-identification of the records in the data provider's dataset is less than the risk permitted under HIPAA Safe Harbor, while still preserving data utility for research purposes. The expert further assessed the extent to which the de-identification approach used for the subject datasets protect the identities of research subjects in other countries and regions of the world with different distributions of population demographics.

1. Introduction

The *Project Data Sphere*® repository is a universal platform to responsibly share oncology clinical trial datasets to accelerate cancer research. It is designed to network stakeholders in the cancer community-researchers, industry, academia, providers, and other organizations in a collaborative effort to transform “big data” into solutions for cancer patients.

This initiative will feature data collected from various environments across the globe. Because all data provided to the *Project Data Sphere* repository will not be subject to the same set of privacy laws and regulations, the provider of the dataset to which this document relates directed the expert to seek to meet or exceed the requirements of laws and regulations of the most highly regulated jurisdictions. In particular, the expert focused on how the subject dataset could be disseminated in a manner that meets or exceeds the de-identification requirements of the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). This regulation was used as a guideline because many regulatory mechanisms, such as the European Union (EU) Data Protection Directive, provide limited specific technical guidelines by which identifiability should be assessed and de-identification should be performed.

The data provider's data was longitudinal in nature, and the relative time of events was critical to the analysis and reuse of such information. As such, it was vital to retain information about oncology trial participants that pertained to time periods at the level of relative weeks. As described in greater depth below, the time periods in question are smaller than that which is permitted by the Safe Harbor de-identification standard of the HIPAA Privacy Rule (details provided below). Therefore, information needed to be reported in a manner that satisfied the Expert Determination approach to de-identification according to the Privacy Rule.

The expert recognized, however, that HIPAA is a regulation in the US and that other countries have different population distributions. Thus, the expert assessed how the de-identification approach created for this data provider based on a US standard related to the populations of other countries associated with the clinical trial data.

The following sections of this document review relevant components of both US and international data protection directives that were considered by the expert and how the protections instituted for the subject dataset related to the requirements of de-identification.

2. The Expert's Summary of Privacy Rules and Regulations

2.1. Privacy Regulation in General

While individuals are not always endowed with the ability to exert control over whether or not their data is collected, the use and sharing of personal data may be regulated. Even if it not regulated by formal policy, it is often desirable to disseminate personal data in a manner that upholds their expectations of privacy.

In the United States, there is no centralized legal statute for data protection. Rather, a mixture of laws are tailored to oversee the handling and disclosure of specific types of personal data, such as the Graham-Leach-Bliley Act for financial data, the Federal Educational Rights and Privacy Act for student data, and the Health Insurance Portability and Accountability Act (HIPAA) for health data. By contrast, in other regions, such as the EU, the Data Protection Directive 95/46/EC provides a foundational set of guidelines by which all person-specific data is collected, used, and shared.

Regardless of the locale, many data protection regulations permit the sharing of de-identified data. For instance, the Data Protection Directive, which strictly prohibits secondary uses of person-specific data without individual consent, provides an exception to the ruling in Recital 26, which states that the:

> principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable.

However, what does it mean for data to be "identifiable"? How do we know when it is no longer identifiable? The Data Protection Directive, and similar directives around the world, provide limited guidelines regarding how data should be protected.

Yet, these directives and regulations often point to the de-identification standard of the Privacy Rule in HIPAA as a potential guideline.

2.2. HIPAA and De-identification

To provide appropriate context, portions of this section of the document were derived from a guidance document recently published by the Office for Civil Rights (OCR) at the US Department of Health and Human Services [OCR 2012].

Under the Health Insurance Portability and Accountability Act of 1996 (HIPAA), the Privacy Rule protects all "individually identifiable health information" held or transmitted by a covered entity or its business associate, in any form or media, whether electronic, paper, or oral. The Privacy Rule calls this information "protected health information" (PHI). By definition, "individually identifiable health information" is information, including demographic data that relates to:

- the individual's past, present or future physical or mental health or condition,
- the provision of health care to the individual, or
- the past, present, or future payment for the provision of health care to the individual,
- and that identifies the individual or for which there is a reasonable basis to believe it can be used to identify the individual.

PHI (Protected Health Information) includes many common identifiers (e.g., personal forename and surname, residential address, and Social Security Number or similar national identifier), but

there are also potential “quasi-identifiers” (e.g., birth date and specific geocode of residence) which may permit a recipient of the data to determine the identity of the corresponding subject.

When health information does not identify an individual, and there is no reasonable basis to believe that it can be used to identify an individual, it is said to be “de-identified” and is not protected by the Privacy Rule.

More specifically, 45 C.F.R., §164.514(a) of the Privacy Rule provides the standard for de-identification of individually identifiable health information:

> **§ 164.514 Other requirements relating to uses and disclosures of protected health information.**
> (a) *Standard: de-identification of protected health information.* Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.

Section 164.514(b) of the Privacy Rule contains the implementation specifications that a covered entity, or affiliated business associate, must follow to meet the de-identification standard. In particular, the Privacy Rule outlines two routes by which health data can be designated as de-identified. These alternatives are summarized in Figure 1 (adapted from [OCR 2012]).

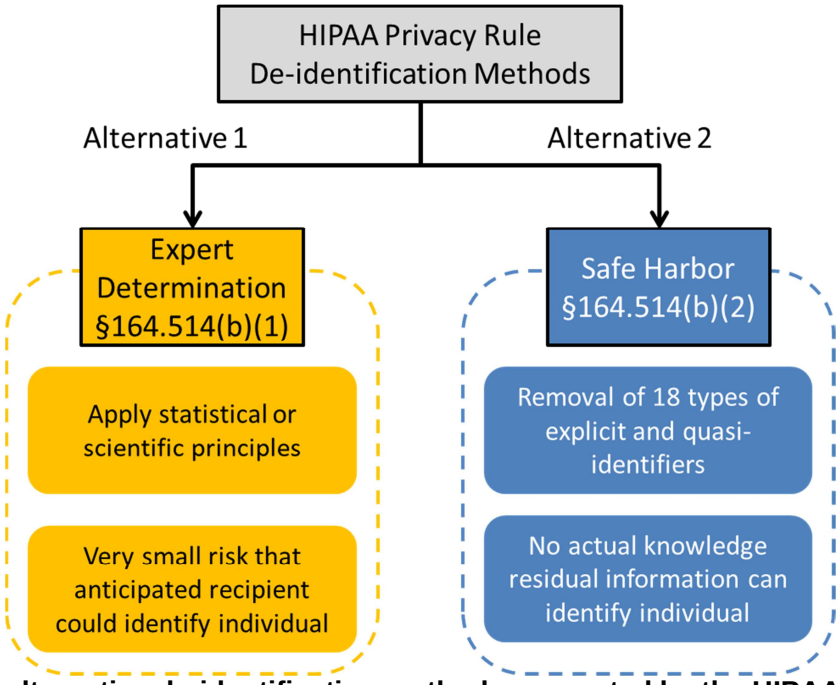

Figure 1. The alternative de-identification methods supported by the HIPAA Privacy Rule (based upon [OCR 2012]).

The first route is the “Expert Determination” method.

(b) *Implementation specifications: requirements for de-identification of protected health information.* A covered entity may determine that health information is not individually identifiable health information only if:

(1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

(i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and

(ii) Documents the methods and results of the analysis that justify such determination;

The second is the “Safe Harbor” method.

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(M) device identifiers and serial numbers
(E) Fax numbers	(N) Web Universal Resource Locators (URLs)
(F) Email addresses	(O) Internet Protocol (IP) addresses
(G) Social security numbers	(P) Biometric identifiers, including finger and voice prints
(H) Medical record numbers	
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique, identifying number, characteristic, or code
(K) Certificate / license numbers	
(L) vehicle identifiers and serial numbers, including license plate numbers	

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information

Satisfying either method will demonstrate that a covered entity, or an affiliated business associate, has met the standard in §164.514(a) above.

The HIPAA Privacy Rule goes on to provide direction with respect to “re-identification” by the covered entity in §164.514(c).

(c) Implementation specifications: re-identification. A covered entity may assign a code or other means of record identification to allow information de-identified under this section to be re-identified by the covered entity, provided

that:

(1) Derivation. The code or other means of record identification is not derived from or related to information about the individual and is not otherwise capable of being translated so as to identify the individual; and

(2) Security. The covered entity does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification.

It is important to recognize that while the re-identification provision does not permit assignment of a code or other means of record identification that is derived from identifying individual information, a covered entity may disclose such derived information if an expert determines that the data meets the de-identification requirements at §164.514(b)(1). This is particularly the case if the resulting information cannot be translated to identify the individual.

De-identified health information created following these methods is no longer protected by the Privacy Rule because it does not fall within the definition of PHI (unless the information is re-identified). This guidance provides clarification about the methodologies that can be applied to render PHI de-identified in compliance with the de-identification standard.

3. The Expert's Framework for Re-identification Risk Analysis

As will be addressed below, due to the criticality to retain certain health information attributes (such as dates) within oncology trial data sets, the Expert Determination method of the HIPAA Privacy Rule is the preferred method and thus needs to be assessed. In this regard, it is critical to clearly articulate, for specific datasets, why a risk assessment is prudent and how it should be performed.

There are an increasing number of detective-like investigations that have been published which demonstrate how health information, devoid of explicit identifiers, could be re-identified to the corresponding research subject (e.g., [Sweeney 1997; El Emam et al. 2006; El Emam & Kosseim 2009; Loukides 2010; Brown 2011; Cimino 2012; Solomon et al. 2012; Atreya et al. 2013]). However, it is important to recognize that there is a significant difference between the description of a path by which health-related information could be re-identified and the likelihood that such a path would be leveraged by an adversary in the real world. [Malin 2010] In this regard, the HIPAA Privacy Rule, is not specified in a manner that precludes the dissemination of data that could be re-identified. Rather, state that the extent to which information can be designated as de-identified must account for the context of the *anticipated recipients* who use *reasonable* means to attempt to re-identify the information. A systematic review of known attacks on de-identified health information [El Emam et al. 2011] indicates that, in practice, when data is de-identified in accordance with standards (e.g., HIPAA), it may be resilient to reasonable adversaries.

For purposes of the dataset being considered for deposition in the *Project Data Sphere* repository, the expert considered the broader environment in terms of how a reasonable recipient would attempt to pursue such a re-identification route. Table 1 is an adaptation of guidance in [OCR 2012] and summarizes the principles that were utilized to determine if health data is sufficiently de-identified. Table 1 is based on the model reported in [Malin et al. 2010, Malin et al. 2011], and these principles directly build on those defined by the U.S. Federal Committee on Statistical Methodology (referenced in the original publication of the Privacy Rule in 2000) [FCSM 2005]).

The expert found it helpful to separate the health information attributes, or types of data, into classes of relatively "high" and "low" risks. Although risk actually is more of a continuum, this rough partition illustrates how the context impacted the risk assessment.

Table 1. Principles used by the expert to assist in the determination of the identifiability of health information.		
Principle	**Description**	**Examples**
Replicability	Prioritize health information features into levels of risk according to the chance it will consistently occur in relation to the individual.	*Low:* Results of a patient's oral disease risk and severity
		High: Demographics of a patient (e.g. birthdate) are relatively stable
Resource Availability	Determine which external resources contain the patients' identifiers and the replicable features in the health information, as well as who is permitted access to the resource.	*Low:* The results of laboratory reports are not often disclosed with identity beyond dental environments.
		High: Patient identity and demographics are often in public resources, such as vital records -- birth, death, and marriage registries.
Distinguish	Determine the extent to which the subject's data can be distinguished in the health information.	*Low:* It has been estimated that the combination of *Year of Birth, Gender,* and *3-Digit ZIP Code* is unique for approximately 0.04% of residents in the United States [Sweeney 2007]. This means that very few residents could be identified through this combination of data alone.
		High: It has been estimated that the combination of a patient's *Date of Birth, Gender,* and *5-Digit ZIP CODE* is unique for over 50% of residents in the United States [Golle 2006, Sweeney 2002]. This means that over half of US residents could be uniquely described just with these three data elements.
Assess Risk	The greater the replicability, availability, and distinguishability of the health information, the greater the risk for identification.	*Low:* Assessment values may be very distinguishing, but they may not be independently replicable and are rarely disclosed in multiple resources to which many people have access.
		High: Demographics are highly distinguishing, highly replicable, and are available in public resources.

The expert then applied these principles with respect to the anticipated registered users of the subject dataset to determine the extent to which they could accomplish re-identification of the participants.

4. The Expert's De-identification Designation

The expert's assessment was performed in a manner that considered the context of the anticipated data recipients, who may be data providers to the *Project Data Sphere* repository, but might also be members of the general public. This evaluation determined the extent of de-identification and suppression for the subject dataset that the expert believed to be necessary to protect patient privacy.

Upon inspection of the subject dataset, the expert observed that the data consisted of over 400 distinct features. Explicit, direct identifiers (e.g. name, SSN, etc.) were removed from the dataset before any assessments of identifiability were performed. The expert then designated the remaining general demographics associated with the trials participants as belonging to the class of quasi-identifying features. These included date of birth, date of death, ethnicity, gender, and place at which the trial was conducted. The expert designated ethnicity as White, Black, Asian, or Other. All of the aforementioned attributes were designated as quasi-identifiers because they are readily available in public resources. The place of the trial was designated as a member of the quasi-identifier because it has the potential to be a proxy for general geographic area in which a trial subject resides. The quasi-identifying data fields were roughly partitioned into several types as shown in Table 2.

Table 2. Gross availability characterization of features in the sample dataset.

Type of Attribute	***Specific Attribute***	***Quasi-identifying***
Unique Identifying Numbers	Trial Participant Number	Yes
	Site of Trial	Yes
General Demographics	Date of Birth	Yes
	Date of Death	Yes
	Ethnicity	Yes
	Gender	Yes
	Place of Trial	Yes
Clinical Information	Date of Visit	Potentially
	Adverse Event	Potentially
	Treatment	No
	Diagnosis	No
	Laboratory Test Values	No
	Medications	No
	BMI	No

Unique identifying numbers, including the participant numbers for the research subjects and the site of the trial, were deemed to be potential identifiers. The participant numbers were deemed to be potential identifiers because such information may have been shared during the course of the trial and could be used to readily ascertain the identity of the corresponding subject. The site of the trial was designated as quasi-identifying because it might be used to ascertain the geographic locale in which the research subject resided.

Clinical information, such as diagnosis received or treatment provided, was not designated as quasi-identifying. This is because the expert deemed that such information was not knowledge that was readily available to potential recipients of the data.

There were, however, two pieces of clinical information that were deemed to be potential quasi-identifiers. These features pertained to the date of the visit and information relating to adverse events. The date of a visit is designated as one of the eighteen features in the Safe Harbor list

of the HIPAA Privacy Rule. However, in the subject dataset, the expert observed that these visits were not associated with information that was disclosed in the setting of resources that could be readily leveraged for identification purposes, such as hospital discharge databases. At the same time, while the date of a visit is not necessarily an identifier in its own right (because there is no centralized public record of when an individual participates in a trial), it could imply the date of death. The expert deemed this information to be sensitive data and therefore higher risk because death-related information can be found in numerous public resources, such as death notices and obituaries. Additionally, in the US, an individual's death date (in combination with birth date, place of birth, fore- and surname, and place of last Social Security payment) is often publicly published by the Social Security Administration to mitigate identity fraud. With respect to the subject dataset, the expert observed that the adverse event was often listed as the expiration of the patient. The expert concluded that taking this fact and combining it with the date of the visit could lead to a direct inference of the date of death.

5. The Expert's Protection Approach

The de-identification approach used by the expert for the subject dataset is akin to the generalization and suppression model adopted by rule-based de-identification policies, such as HIPAA Safe Harbor. The generalization model is outlined in Table 3.

<table>
<tr><th colspan="2">Table 3. De-identification approach for subject dataset.</th></tr>
<tr><td colspan="2">(A) Names</td></tr>
<tr><td colspan="2">(B) All geographic subdivisions were reported at a granularity no smaller than a region, whereby a region, is defined as a collection of countries in a certain geographic locale, such as North America, South America, Asia-Pac, Europe, or Africa</td></tr>
<tr><td colspan="2">(C) All ages below 85 were fixed at the point of trial registration, rounded down to the nearest integer; all ages over 85 reported as 85+, All death-related events, as well as direct indicators of such an event, were reported as relative week of the year</td></tr>
<tr><td>(D) Telephone numbers</td><td>(M) device identifiers and serial numbers</td></tr>
<tr><td>(E) Fax numbers</td><td>(N) Web Universal Resource Locators (URLs)</td></tr>
<tr><td>(F) Email addresses</td><td>(O) Internet Protocol (IP) addresses</td></tr>
<tr><td>(G) Social security numbers</td><td rowspan="2">(P) Biometric identifiers, including finger and voice prints</td></tr>
<tr><td>(H) Medical record numbers</td></tr>
<tr><td>(I) Health plan beneficiary numbers</td><td>(Q) Full-face photographs and any comparable images</td></tr>
<tr><td>(J) Account numbers</td><td rowspan="2">(R) Any other unique, identifying number, characteristic, or code</td></tr>
<tr><td>(K) Certificate / license numbers</td></tr>
<tr><td>(L) vehicle identifiers and serial numbers, including license plate numbers</td><td></td></tr>
</table>

There are several core differences between the de-identification approach used by the expert for the subject dataset and Safe Harbor.

- First, this approach is more restrictive with respect to geographic indicators and certain age ranges.

    - While Safe Harbor permits the dissemination of almost all ZIP-3 geocodes, this approach reported no such information. Rather, it only reported that the trial participant resides somewhere in the region where the trial was run.
    - Additionally, Safe Harbor permits the dissemination of one-year age groups up to 89 years old. By contrast, this approach restricted one-year age groups to an age limit of 84. The remaining ages were reported as 85+.
- Second, this approach is less restrictive with respect to certain time sensitive attributes. While Safe Harbor precludes the dissemination of dates that are more specific than one year, this approach allowed for the dissemination of the following information:
    - Death related dates were reported as the relative week of the year for the event.
    - The actual visit date to a trial facility was left as the date, provided it did not imply a birth (death) date. In the case that such an implication transpired, the date was treated as a birth (death) date.

It was also deemed that the identifying numbers mentioned in the previous section would be replaced with random values. These values would, however, be consistently substituted for the trial participants. It was deemed that the random nature of these values met the requirements of the HIPAA Privacy Rule's "re-identification" code requirement because it was not correlated with information about the participants.

6. The Expert's Re-identification Risk Analysis

6.1. Baseline Risk: Safe Harbor

To assess re-identification risk, the expert utilized public data from the U.S. Decennial Census [AFF 2012] to determine the extent to which the features {*Year of Birth-90+*, *Year of Death*, *Gender*, *Race*, 3-*digit ZIP*} (using the appropriate top-coding for individual's age 90+ and ZIP codes with small populations as described below) could have led to unique identification of clinical trials participants. The analysis was based on the attack described in [Sweeney 1997; 2002]. In particular, the analysis utilized publicly accessible demographic tables (specifically, the PCT12 tables), which report population counts by age group, gender, and race combination.[1] The expert further subdivided this information by splitting the population along the 3-digit ZCTA (related to the zip code), while grouping the ZCTAs with less than 20,000 people into a single group. For reference, the set of ZCTAs that were aggregated are shown in Table 3. For each [gender, race, age, ZCTA] population combination, the expert aggregated the counts for all age groups over 89 years old.

Table 4. 3-digit ZCTA with populations expected to be less than 20,000 people in size.							
036	059	063	102	203	556	692	790
821	823	830	831	878	879	884	890
893							

[1] It has been noted that there exists a certain amount of error in public versions of the U.S. Census data (particularly the public use microdata samples, or PUMS files) that may influence analyses with age 65 and older. [Alexander 2010] Similar bias may exist in the tables used in this evaluation. However, the expert noted that such inaccuracy is at the cell level, while the analysis conducted for this assessment is about aggregated results. Thus, it was not expected that such errors would influence the general findings of this investigation.

After constructing the dataset, the expert derived an estimate number of uniques that were reported in the population. The estimation process was based on the strategy proposed in [Golle 2006]. The statistics published by the U.S. Census are in aggregated form. For instance, they report how many males with age 50-54 reside in a certain 3-digit ZCTA and are Caucasian. For reference, the expert called this number *n.* To continue the analysis, the expert disaggregated the five year bin into one year bins and calculated the expected number of bins with one individual. This disaggregation is a traditional balls-to-bins problem. In a generalized form, to calculate the number of bins with exactly *i* people, the problem reduces to a binomial function of the form:

$$f_i(n) = \binom{n}{i} b^{1-n}(b-1)^{n-i}$$

For this analysis, *i* is equal to one, reducing the computation to:

$$f_i(n) = n\left(\frac{b-1}{b}\right)^{n-1}$$

The value for *b* varies for different age groups in the Census data. For instance, an age group of 0-4 has 5 bins, while an age group of 60-61 has 2 bins.

After computing this value for each bin, the results were summed and normalized by the total U.S. population in the Census tables. Based on this analysis, the expert found that 0.48% of the U.S. population was estimated to be unique.

6.2. Re-identification Risk Assessment for the Protection Approach

The expert continued the analysis by assessing the re-identification risk for the de-identification protection approach using the attributes {*Year of Birth-85+*, *Week of Death*, *Gender*, *Race**, *Country of Residence*}. In this case, all races not designated as White, Black, or Asian were aggregated into a group called Other. The Safe Harbor risk served as a threshold for acceptable re-identification risk. While the de-identification approach for the subject dataset did not permit the year of birth to be readily distinguished, a 365 day window for a birthday could be inferred from a combination of the date of registration and age at the time of the event.

Further analysis used the same estimation method described above, except Year of Birth was top-coded with 85+ and all ZCTAs were aggregated into a single location. Based on this analysis, the expert found that 0.000001% of the U.S. population was estimated to be unique. This finding implies that the risk for the de-identification approach used for the subject dataset was significantly smaller than the Safe Harbor de-identification model.

6.3. Extrapolating the Risk Assessment Beyond the US

The population distribution used in the aforementioned analysis is not necessarily representative of other countries. As such, it is fair to question if such a de-identification model is appropriate for data collected from facilities outside of the U.S. To generalize the assessment, the expert performed a sensitivity analysis on identifiability with respect to the size of the population and the distribution of people to the quasi-identifying values.

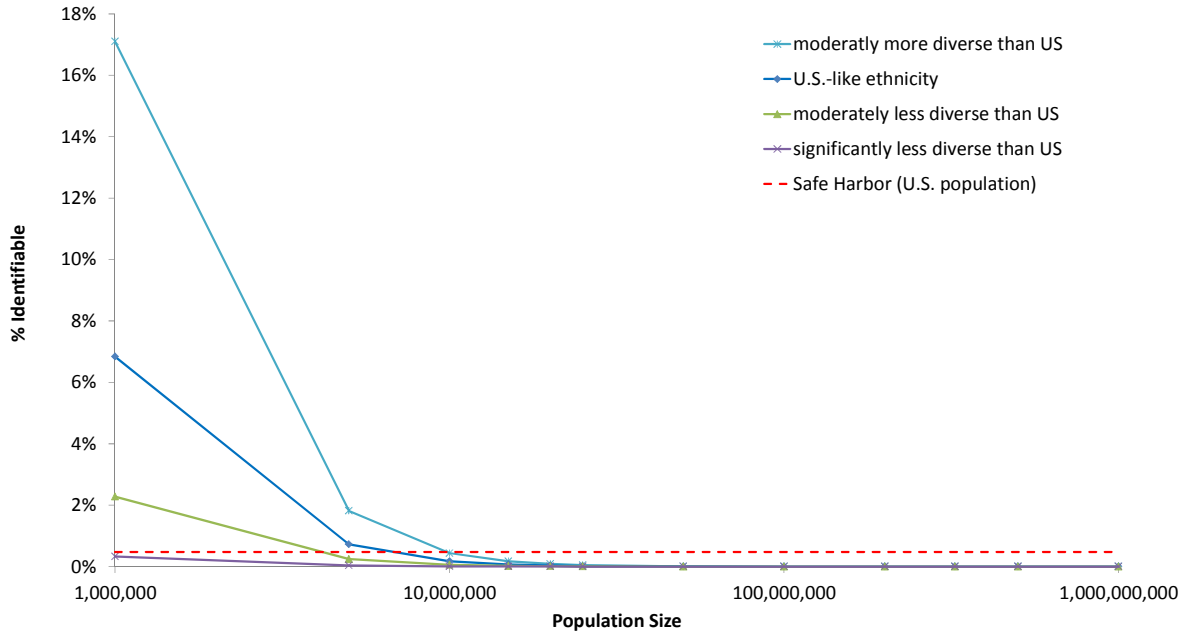

Figure 2. Re-identification rates for various countries and distribution of ethnicities.

The expert considered a situation in which the population density of the United States was similar to that of another country and varied the population size from 25 million to 1 billion people. The analysis accounted for the balance of ethnicity in other countries by considering several alternative models. Specifically, the analysis considered three situations: ethnic diversity that is, i) moderately more diverse than the U.S., ii) moderately less diverse than the U.S., and iii) significantly less diverse than the U.S. To simulate the first scenario, allocation of the White and Black populations was according to a 50-50 split. To simulate the second scenario, the analysis compressed the races into two classes, White and Other, and allocated the population into an 80-20 split. To simulate the third scenario the analysis further biased the allocation into a 95-5 split.

The results of the foregoing analysis for all four cases of identifiability assessment are depicted in Figure 2. Here, it can be seen that in populations with demographic distributions similar to the US, the expert found a significant risk of identification for populations smaller than 10 million people. At 1 million, the expert found that approximately 7% of the population was expected to be unique based on their quasi-identifier. At 5 million, this percentage dropped to around 0.72%, which was slightly larger than the Safe Harbor rate of 0.42% mentioned earlier.

As expected, the results indicated that when the ethnic diversity decreases, the rate of identifiability dropped as well. When a population is moderately less diverse than the US, the rate of identifiability for a population of size 5 million dropped to 0.03% (below Safe Harbor). For a population of size 1 million, the rate of identifiability dropped, but only to 2%, which is too risky. When the distribution becomes almost homogenous though, the risk rate of identifiability for a population of size one million dropped to 0.32%, which is acceptable.

To assess how these risks related to countries in the subject dataset, the expert considered the home countries of subjects. Based on the representation of primary ethnicity (i.e., coverage of

the population), the expert grouped populations into one of the four cases (i.e., moderately more diverse than US, similar to US, moderately less diverse than US, and significantly less diverse than US). The expert then estimated the range of identifiability for each of these countries given the applicable category of primary ethnicity in which it fell. The results are shown in Table 5 where it can be seen that the identifiability of every record contributed was considered to be below the Safe Harbor threshold.

Table 5. Re-identification risk ranges for subjects in the sample studies.

Country Represented in TROPIC Study	***Population (Millions)***	***Primary Ethnicity***	***Anticipated Identifiability Range***
Denmark	5	87%	0.03-0.24%
Finland	5	91%	0.03-0.24%
Singapore	5	95%	0.03-0.24%
Slovakia	5.5	80%	0.03-0.24%
Sweden	9.5	80%	0.008-0.05%
Czech	10	94%	0.008-0.05%
Hungary	10	93%	0.008-0.05%
Belgium	11	80%	0.008-0.05%
Netherlands	16	79%	0.003-0.02%
Chile	17	72%	0.02-0.06%
Taiwan	23	96%	0.001-0.01%
Canada	31	80%	0.001-0.007%
Argentina	41	86%	0.0001-0.007%
Spain	47	90%	0.0001-0.0009%
S. Korea	49	99%	0.0001-0.0009%
South Africa	50	80%	0.0001-0.0009%
Italy	60	96%	0.0001-0.0009%
UK	62	92%	0.0001-0.0009%
France	65	92%	0.0001-0.0009%
Turkey	75	85%	0.00002-0.00018%
Germany	82	92%	0.00002-0.00018%
Mexico	113	85%	0.000009-0.00006%
Russia	143	81%	0.000003-0.00001%
Brazil	193	48%	0.000002-0.000005%
US	315	73%	0.000001%
Europe (all countries)	711	86%	<0.00000001%
India	1220	97%	<0.00000001%

7. Conclusions

This document provided motivation and definitions for the de-identification of data derived from oncology clinical trials. In doing so, it reviewed how de-identification is described in certain regulations, with a specific focus on the technical specifications indicated in the HIPAA Privacy Rule. To illustrate how these specifications relate to specific trials data in the subject dataset, it walked through an expert's design of a de-identification approach for a specific dataset, as well as the expert's assessment of residual re-identification risk for data collected on residents in the United States and other countries around the world for which the de-identification approach was applied.

This document does not represent specific guidance from *Project Data Sphere, LLC* for any data provider. Instead, it is meant to serve as an explanation of how de-identification of specific data was achieved for one data provider whose data has been uploaded to the *Project Data Sphere* repository. Any data provider planning to submit datasets to the *Project Data Sphere* repository should consult with an expert individual or company who is well-versed and experienced in the generation of de-identified data.

8. References

[Alexander et al. 2010] J.T. ALEXANDER, M. DAVERN, B. STEVENSON. Inaccurate age and sex data in the Census PUMS files: evidence and implications. Working Paper 15703, National Bureau of Economic Research. January 2010.

[Atreya et al. 2013] R. ATREYA, J. SMITH, A. MCCOY, B. MALIN, R. MILLER. Reducing patient re-identification risk for laboratory results within research datasets. Journal of the American Medical Informatics Association. 2013; 20: 95-101.

[AFF 2012] American Fact Finder. Available online: http://www.factfinder.census.gov

[Brown 2011] I. BROWN, L. BROWN, D. KORFF. Limits of anonymisation in NHS data systems. British Medical Journal. 2011; 342: d973.

[Cimino 2012] J.J. CIMINO. The false security of blind dates: chronomization's lack of impact on data privacy of laboratory data. Applied Clinical Informatics. 2012; 3: 392-403.

[El Emam et al. 2006] K. EL EMAM, S. JABBOURI, S. SAMS, Y. DROUET, M. POWER. Evaluating common de-identification heuristics for personal health information. Journal of Medical Internet Research. 2006; 8(4): e28.

[El Emam & Kosseim 2009] K. EL EMAM, P. KOSSEIM. Privacy interests in prescription data, part 2: patient privacy. IEEE Security and Privacy Magazine. 2009; 7(1): 75-78.

[El Emam et al. 2011] K. EL EMAM, E. JONKER, L. ARBUCKLE, B. MALIN. A systematic review of re-identification attacks on health data. PLoS One. 2011; 6: e28071.

[FCSM 2005] SUBCOMMITTEE ON DISCLOSURE LIMITATION METHODOLOGY, FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. Report on statistical disclosure limitation methodology. Statistical Policy Working Paper 22, Office of Management and

Budget. May 1994. Revised by the Confidentiality and Data Access Committee. 2005. Available online: http://www.fcsm.gov/working-papers/wp22.html

[Golle 2006] P. GOLLE. Revisiting uniqueness of simple demographics in the US population. Proceedings of the 5th ACM Workshop on Privacy in the Electronic Society. 2006: 77-80.

[Loukides et al. 2010] G. LOUKIDES, J. DENNY, B. MALIN. The disclosure of diagnosis codes can breach research participants' privacy. Journal of the American Medical Informatics Association. 2010; 17(3): 322-327.

[Malin et al. 2010] B. MALIN, D. KARP, R. SCHEUERMANN. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. Journal of Investigative Medicine. 2010; 58(1): 11-18.

[Malin et al. 2011] B. MALIN, G. LOUKIDES, K. BENITEZ, E. CLAYTON. Identifiability in biobanks: models, measures, and mitigation strategies. Human Genetics. 2011; 130(3): 383-392.

[OCR 2012] OFFICE FOR CIVIL RIGHTS, US DEPT. OF HEALTH AND HUMAN SERVICES. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. November 26, 2012. Available online: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

[Sweeney 1997] L. SWEENEY. Weaving technology and policy together to maintain confidentiality. Journal of Law, Medicine, and Ethics. 1997; 25(2-3): 98-110.

[Sweeney 2002] L. SWEENEY. K-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems. 2002; 10(5): 557-570.

[Sweeney 2007] L. SWEENEY. Testimony before the National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information. August 23, 2007.

[Solomon et al. 2012] A. SOLOMON, R. HILL, E. JANSSEN, S. SANDERS, J. HEIMAN. Uniqueness and how it impacts privacy in health-related social science datasets. Proceedings of the 2nd ACM International Health Informatics Symposium. 2012: 523-532.