Toolkit for Assessing and Mitigating Risk of Re-identification when Sharing Data Derived from Health Records

Gregory Simon, MD, MPH,[1] Susan M Shortreed, PhD,[1] R Yates Coley, PhD,[1] Estibaliz M Iturralde, PhD,[2] Richard Platt, MD,[3] Sengwee Toh, ScD,[3] Brian Ahmedani, MSW, PhD[4]

1. Kaiser Permanente Washington Health Research Institute, Seattle, WA

2. Division of Research, Kaiser Permanente Northern California, Oakland, CA

3. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School

4. Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, MI

Version 1.0

May 26, 2020

This report was developed for the "Effects of Medical Products on Suicidal Ideation and Behavior – Real World Evidence" research project, funded by FDA under contract number HHSF223201810201C.

Toolkit for Assessing and Mitigating Risk of Re-identification when Sharing Data Derived from Health Records

Table of Contents

History of Modifications

Version	Date	Modification	Author
1.0	05/26/2020	Original Version	Simon, et al.

1. Executive Summary

Sharing of individual-level analytic datasets extracted from health system records can have important scientific and public health value. Sharing of data supports the rigor and reproducibility of science and may enable important secondary analyses by other investigators or research teams. Stewards of those sensitive or protected datasets, however, must carefully consider risks to individuals represented those data. Even if a sensitive or protected dataset contains no explicit identifiers (like name or health plan number) or implicit identifiers (like email address), patterns in the data could allow individuals to be identified. That re-identification of a supposedly de-identified dataset could reveal sensitive health information. This document describes the conceptual basis of re-identification risk and describes a step-by-step process for understanding, assessing, and protecting against risk. Key points include:

- Understanding sources of risk:
    - Risk of re-identification exists when data elements or variables in a protected dataset can be linked to corresponding variables or data elements in some identified or identifiable external data resource.
    - Re-identification could use public (and not sensitive) elements common to both data resources to gain access to non-public (and sensitive) information in the protected dataset.
- Assessing risk
    - Risk assessment should consider the potential for either systematic linkage to publicly available (for free or for purchase) datasets or idiosyncratic linkage to public knowledge regarding a specific individual health event.
    - Any variables common to both a protected dataset and an identified or identifiable data resource could serve as linking variables or public keys.
    - When an external data resource is available for inspection, a data steward can directly assess re-identification risk and directly test strategies to mitigate risk.
    - In most cases, the external data resource is not directly available, so a data steward must follow a series of steps to estimate and (if necessary) mitigate risk.
    - Stewards of protected data should first identify the smallest cells defined by combinations or permutations of public key or potential linking variables. The smallest number of records in any such cell is sometimes referred to as the k-anonymity threshold for that combination of key variables in that protected dataset. Risk of re-identification increases with smaller k-anonymity threshold.
    - Risk or probability of re-identification also depends on the pattern of overlap between the protected dataset and an external data resource. Risk increases with greater overlap.
    - Risk of disclosing new sensitive information is also related to the patterns of sensitive data elements within small groups or cells in the protected dataset; less variation creates more risk.
- Protecting against risk
    - Determining the appropriate risk threshold for releasing or sharing any protected dataset depends on the sensitivity of the information that might be revealed, the motivations of adversaries who might try to re-identify sensitive data, and the mechanism through which data will be released or shared.
    - Data stewards can use a variety of specific strategies to reduce risk of re-identification, including deleting or altering key variables, deleting or altering sensitive variables, and deleting or altering specific high-risk records. Selection of a method that effectively

reduces risk without compromising scientific or public health value depends on the specifics of each data sharing scenario.

- Data may be shared via a range of methods, ranging from the least controlled (a public-use dataset available to any anonymous user) to the most controlled (a controlled data enclave available only to identified and trusted users). Selection of an appropriate mechanism for data sharing depends on the sensitivity of the protected dataset, the risk of re-identification (determined by the procedures described above) and the motivations of those who might access data.

2. Understanding Sources of Risk

2.1. Background and Scope

This document intends to describe specific steps in assessing and mitigating risk of re-identifying individual records when sharing data originally derived from health system records such as electronic health records (EHRs) or insurance claims. We focus on sharing of datasets originally created for research or public health surveillance.

Any health system or other entity covered by the Health Insurance Portability and Accountability Act (HIPAA) has legal and regulatory obligations to protect the privacy of health information and to guard against re-identification[1, 2]. The basic Safe Harbor[3] standard stipulates that a dataset or other data resource can be considered de-identified if it contains none of 18 specified explicit identifiers AND if the releasing entity does not have actual knowledge that other data elements or variables would allow re-identification of individual records. A more rigorous Expert Determination[3] standard calls for the releasing entity to certify that someone with relevant statistical and scientific expertise has assessed and documented that risk of re-identifying individual records is no greater than very small. Department of Health and Human Services guidance (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html) includes useful basic examples of such an assessment. The procedures described in this document could be part of such an Expert Determination process.

In many cases, research data are extracted from health system records under a waiver of HIPAA authorization and waiver of Common Rule[4] consent. In those cases, creators of the protected dataset have specific regulatory obligations to minimize risks to individual patients or participants and to protect against disclosure of identifiable information – in addition to an ethical obligations to those who have contributed data to research.

In some cases, individual patients or participants may have provided informed consent for research use of records data, including permission for future disclosure. In those cases, creators of the protected dataset may not have specific regulatory obligations to de-identify data or assess risk of re-identification prior to disclosure. Nevertheless, all researchers operating under the Common Rule have regulatory and ethical obligations to minimize risk to participants. Those obligations would include a responsibility to assess risks to privacy or disclosure of confidential information, even when individuals have provided informed consent for future data sharing.

In some cases, research datasets may include data extracted from health records that may be linked to survey data, genomic data, or other data sources. Stewards of those research datasets must then consider how linkage to health records data could lead to disclosure of those other types of information.

This document does not explicitly consider protection of data or information that health systems might consider sensitive or proprietary[5]. Examples of such sensitive information might include rates of use of specific products or services or rates of specific adverse outcomes or events. Health systems might also

have concerns regarding identification of individual providers or healthcare facilities. Some of the methods described here to assess risk of re-identifying individuals might also apply to re-identification of providers or facilities, but this document focuses on risk to individual patients or consumers of healthcare.

Neither does this document address potential risks to groups from sharing of health data. Sharing health research data could stigmatize some disadvantaged groups (such as Native Americans or transgendered people), even if individual members of those groups cannot be identified.

We describe below a step-by-step process for:

- Understanding conditions that create re-identification risk
- Systematically assessing risk
    - Identifying external data resources that might facilitate re-identification
    - Specifying possible linking variables or keys
    - Direct testing of linkage with an external data resource
    - Assessing uniqueness or anonymity thresholds within a dataset to be shared
    - Assessing or estimating overlap with an external data resource
    - Estimating risk of re-identification for individual records
- Protecting against risk
    - Determining appropriate risk thresholds
    - Mitigating risk by altering the dataset to be shared
    - Selecting mechanisms for data sharing appropriate to potential risk

As discussed in detail below, the risk of re-identification for any individual record may have different specific interpretations in different data linkage scenarios. In addition to varying interpretation, acceptable risk thresholds will vary according to the sensitivity of the information that could be revealed, the motivation of potential adversaries, and the vulnerability of those individuals subject to re-identification.

This document aims to guide research teams with moderate technical sophistication – including the ability to examine and manipulate research datasets and to use open-source statistical packages to assess joint distributions of variables in those datasets (described below). A later section of this document includes references to additional resource, including publications, relevant software packages, and training resources. Several publications provide more technical guidance regarding estimation of re-identification risk[6-13].

We illustrate each of these steps using the example of a dataset created to develop and validate statistical models predicting suicidal behavior[14]. That dataset includes approximately 20 million records from 3 million patients, representing outpatient visits to mental health or general medical providers in seven large health systems. Each record includes patient-level indicators of demographic characteristics (age, sex, race, ethnicity), approximately 150 indicators regarding medical and mental health diagnoses and treatment prior to the visit, and indicators for suicide attempt or suicide death following the visit. When we describe assessment of anonymity thresholds within this dataset, we provide specific worked examples using the public-domain Comprehensive R Archive Network (CRAN) software package sdcMicro[15].

2.2. Terminology

We use the term “protected dataset” for a data resource that might be shared. The protected dataset will often have been created from confidential health system records and will often have been created

for research use. We use this term to distinguish an analytic dataset, usually developed for research, from the original operational health system databases (EHRs, insurance claims, laboratory results, prescription dispensings, etc.) used to create that data resource. Researchers would be unlikely to consider sharing original health system records for large numbers of patients, but might hope to share analytic datasets derived from those records.

We use the term "data steward" to describe an entity that holds and hopes to share that protected dataset. We use this term both to emphasize the data steward's ethical obligations and to recognize that a data steward may or may not have been involved in the creation of the protected dataset from the original sources.

We use the term "adversary" to describe an individual or entity who might attempt to re-identify individual records. While this term may seem to imply malicious intent, we recognize that some types of re-identification may be free of any malicious intent[16]. For example, re-identification to discover sensitive information regarding celebrities be motivated by financial gain.

We use the term "re-identification attack" to describe any process that might re-identify individuals represented in a dataset and reveal sensitive or protected information about those individuals. In general, re-identification requires specific effort, but it possible that some instances of re-identification could be inadvertent.

We use the term "external data resource" to describe any external data to which a protected dataset could be linked. That linkage could be definitive, with 1:1 matching of each individual record. Alternatively, linkage could be probabilistic, with k:m matching where both k (a number of records in a protected dataset) and m (a number of records in an identified or identifiable data resource) are relatively small. As discussed in Section 3.1, that external data resource may be a traditional dataset including discrete data elements (e.g. a voter registration database) or less structured information that might allow linkage (e.g. knowledge that an identifiable individual has experienced a specific health event at a specific time).

We use the term "public keys" to describe specific data elements in a protected dataset that might be used (typically in combination) to facilitate linkage to an external data resource. While potential linking variables may not always be public in the common sense of that term, they are public in that they are available to a potential adversary. For example, genetic information in an online ancestry database could serve as a "pubic" key. As discussed in Section 3.2, these key data elements more often concern relatively public information (age, sex, etc.), but could include data elements or information not widely available. Genomic data could sometimes serve as a public key.

2.3. Conditions that Create Risk

Re-identification creates risk or potential for harm if an adversary can gain new access to identifiable private or sensitive information. This risk depends on characteristics of both the data resource to be shared and external data resources available for linkage. Specifically:

- The protected dataset includes sensitive or protected information not already available to potential adversaries AND
- The protected dataset includes data elements or variables that could be linked (with level of probability discussed in Section 3.6 below) to corresponding data elements in some external data resource AND

- That external data resource includes explicit identifiers (e.g. names, Social Security numbers) or implicit identifiers (e.g. IP addresses, mobile device serial numbers) not present in the protected dataset

As will be discussed in detail below, linkage between a protected dataset and an external data source depends on the relative uniqueness of the information or data elements shared by the two data sources. Consequently, risk of linkage must be considered for any pair (or combination) of data sources rather than for a protected dataset in isolation. For any protected dataset, the data elements most likely to be shared with an external data resource are relatively "public" characteristics (age, sex, race, ethnicity, area of residence) rather than relatively private or sensitive information such as specific diagnoses or treatments. In a protected dataset, it is the combination of more sensitive (but usually not "linkable") information with more public (and more likely "linkable") information that creates greatest risk. For example, linkage of data regarding age, sex, race, ethnicity, and state of residence (not sensitive) could create risk if linked to information regarding substance use disorder (sensitive).

The risk of re-identification from linking a protected dataset and an external data resource also depends on the overlap between a protected dataset and an external data resource to which it might be linked. Likelihood of re-identification is greatest when a cell or subgroup in the research dataset and the corresponding cell or subgroup in an external data resource are identical or completely overlapping. Likelihood of unambiguous re-identification is decreased by significant non-overlap or discordance in either direction (a cell or subgroup in the protected dataset is a subset of the corresponding cell or subgroup in the external data resource OR a cell or subgroup in the external data resource is a subset of the corresponding cell or subgroup in the protected dataset OR corresponding cells or subgroups in the two data sources only partially overlap).

For the remainder of this this paper we will consider linking of two data sources. The extension of these ideas to linkage across three or more datasets or data sources follows the same principles, but quantitative estimates of risk will be more complex to calculate.

Below are specific examples of potential linkage scenarios (pairs of protected datasets and external data sources) that would or would not add or create risk or potential harm.

- **Identifiable, but no new information (no added risk)**: Protected dataset includes variables computed from insurance claims and demographic characteristics of individuals (e.g. age, sex, race, ethnicity, place of residence). External data resource includes the same insurance claims and explicit identifiers (e.g. insurance plan number) covering at least some of the same population. Even if individual records could be definitively linked, this linkage would NOT create risk or harm because the protected dataset does not include any additional protected or sensitive information not included in the external data resource (i.e. already available to the potential adversary). This linkage would not allow an adversary to gain NEW access to identifiable private information.
- **Identifiable with added information (added risk):** Protected dataset includes variables computed from insurance claims, demographic characteristics of individuals (e.g. age, sex, race, ethnicity, place of residence), AND variables computed from clinical text in electronic health records. External data resource includes insurance claims and explicit identifiers (e.g. insurance plan number) covering at least some of the same population. This linkage could create risk or potential harm, because the protected dataset includes private and potentially sensitive information (from clinical text) not included in the external data resource (i.e. already available

to the potential adversary). This linkage could allow an adversary to gain NEW access to identifiable private information.

- **Added information but not identifiable (no added risk):** Protected dataset includes variables computed from insurance claims, demographic characteristics of individuals (e.g. age, sex, race, ethnicity, place of residence), AND variables computed from clinical text in electronic health records. External data resource includes insurance claims covering at least some of the same population without any explicit identifiers or potential identifiers. This linkage would NOT create risk or potential harm, because linkage would not increase the risk that individuals could be identified. Either of the two data resources might contain unique records, but additional information would be necessary to identify those unique records in either dataset. This linkage could not allow an adversary to gain new access to IDENTIFIABLE private information.

2.4. Types of Re-identification Attacks

We consider re-identification attacks in two broad categories, depending on the nature of the external data resource(s) available to a potential adversary.

We use the terms "systematic attack" or "systematic re-identification" to refer to linkage between a protected dataset and an external dataset including both linking variables or keys and implicit or explicit identifiers. Potential external data resources that could support a systematic attack are discussed in detail below, but could include databases available to the general public (e.g. voter registration data) or databases available only to specific adversaries (e.g. identifiable genomic data, identifiable credit card purchase data, identifiable vital statistics data).

We use the terms "idiosyncratic attack" or "idiosyncratic re-identification" to refer to identification of specific records in a protected dataset based on linkages to individual-level information available to a potential adversary. Potential external data resources that could support an idiosyncratic attack are discussed in detail below, but could include media accounts of a specific individual experiencing a specific health event or information regarding specific health events revealed or disclosed to acquaintances such as neighbors or co-workers.

3. Assessing Risk

3.1. Identifying Potentially Linkable External Data Sources

Systematic re-identification involves intentional linkage between a protected dataset and a specific external data resource that includes explicit or implicit identifiers. For any specific protected dataset, the list of candidate external data resources an adversary could use depends on both the existence of public keys or potential linking variables and the pattern of overlap between cells or subgroups in the protected dataset and corresponding cells or subgroups in the external data resource. Specifying those candidates typically requires expertise regarding the contents of the protected dataset and expertise regarding the range of linkable datasets available to potential adversaries. Data stewards attempting to assess re-identification risk should have that relevant expertise or seek appropriate consultation. Candidates for external data resources would typically include:

- Vital statistics data – Identified state-level records of births or deaths may be available to some potential adversaries. These records may include demographic information (age, sex, race, ethnicity) and date information (i.e. birth dates or death dates) that would permit linkage to individual records in a protected dataset. Probability of linkage would be increased if both data resources include information regarding cause of death.

- Identifiable insurance claims data – An adversary with access to identifiable insurance claims data (such as from a state all-payer claims database) could accomplish linkage to a protected dataset including data elements derived from EHR or claims sources. As discussed above, this linkage would allow an adversary to identify new private information if the protected dataset contains information not included in claims data available to an adversary (e.g. associated data from clinical text or patient reported outcomes).
- Identifiable credit card or financial data – Credit card or financial records could be linked to data derived from health records using demographic information as well as information regarding dates of service. Even if a protected dataset does not include actual dates of service, relative date information (e.g. days between two prescription dispensings) could support linkage to payment data regarding dates of payment to specific organizations or facilities[17].
- Identifiable geospatial data – While data regarding locations of work, school, or residence may be important for assessment of environmental exposures, relatively precise location data in a protected research dataset may allow linkage to other identifiable sources of geospatial data, such as mobile phone location data.
- Identifiable genomic data – Identifiable genomic information (available from criminal justice or genealogy databases) could allow linkage to a research database containing both genomic data and data derived from health records. In this scenario, genomic data would allow an adversary to identify health records data they did not already possess.

Idiosyncratic re-identification involves identification of an individual record in a protected dataset based on individual-level information available to an adversary. For any protected dataset, the candidate external data sources that could facilitate re-identification may be specified in advance or only hypothesized. Candidate data sources would typically include:

- Public accounts of notable events – News reports or other public accounts may include identifiable information regarding notable health events[18]. While public accounts would not typically include information regarding specific diagnoses or treatments, they may still include adequate detail to allow linkage to individual records in a protected dataset. For example, a news report may specify that a public figure (for whom age, sex, race, and ethnicity are known) was involved in an automobile accident on a specific date and hospitalized at a specific facility.
- Information revealed to acquaintances – Acquaintances (co-workers, neighbors, family members, etc.) may have information regarding notable health events even when such information is not available through any public account. For example, an ex-spouse may be aware of a hospitalization or emergency department visit on a specific date for a specific condition.

Example (Suicide Risk Prediction Dataset):

The protected dataset includes outpatient visits from seven health systems between 2009 and 2015, including both visits to specialty mental health providers and visits to general medical providers when a mental health diagnosis was recorded. Each health system serves members in one or two specific states. Any individual could contribute multiple visits. Data elements regarding each individual include demographic characteristics (age in 5 categories, sex, race in six groups, Hispanic ethnicity). Data elements regarding each visit include yes/no indicators regarding receipt of specific mental health diagnoses or treatments during discrete time periods (3 months, 1 year, 5 years) prior to the visit, yes/no indicators for specific general medical diagnoses during discrete time periods prior to visit, and separate

yes/no indicators for the occurrence of a suicide death or a likely suicide attempt (diagnosis of self-inflicted injury or poisoning) in the 90 days following the visit. Date of each visit is indicated by calendar year only, with all other dates expressed in relative terms (e.g. visit in year 2012 with at least one psychiatric hospitalization between 1 and 5 years prior). Given this data structure, candidate external data resources that might facilitate re-identification include:

- *State vital statistics data – Identifiable vital statistics data (including cause of death) could be available to some adversaries. Given the rarity of suicide death, identified vital statistics data could allow identification of individual records in the protected dataset, allowing an adversary to gain access to other sensitive information regarding mental health diagnoses and treatments received prior to death.*
- *Public accounts of self-harm events OR information revealed to acquaintances – An adversary with knowledge of a specific hospitalization or emergency department visit for self-harm or suicide attempt could use demographic or other information in the protected dataset to identify that specific event. The adversary could then use that linkage to learn other sensitive information about mental health diagnoses or treatments.*

3.2. Defining Potential Public Keys in a Protected Dataset

As described above, public keys are specific variables or data elements in a protected dataset that would allow linkage to some identified or identifiable external data resource, possibly facilitating re-identification of individual records in the protected dataset. Potential public keys should be considered separately for each external data resource. The term "public" implies that the information is available to a potential adversary, even if it is not available to the general public. Any variable or data element common to both data sources could serve as a public key, but the most concerning are data elements that can be combined to create unique or relatively rare combinations in each of the two data sources. When identifying potential public keys for systematic attacks, detailed information may be available regarding specific information included in specific external data resources. When identifying potential public keys for idiosyncratic attacks, data stewards must instead anticipate the types of information that might be available to potential adversaries. In some cases, a data element in a research data set may correspond exactly to a data element in an external data resource (e.g. date of birth in an electronic health record corresponding to date of birth in vital statistics data). In other cases, correspondence may only be approximate (e.g. race or ethnicity in an electronic health record may or may not be identical to race or ethnicity in vital statistics data).

When a protected dataset contains multiple records per person, then a data steward may need to consider key or linking variables defined by multiple records. For example, a protected dataset including hospital discharge diagnoses would allow a potential adversary to compute days between hospital discharges for any individual. A data steward would therefore need to consider days between discharge dates as a potential linking or public key variable.

Example (Suicide Risk Prediction Dataset):

Identification of potential public keys in this protected dataset are considered separately for each of the external data resources listed above here:

- *Linkage to state vital statistics data – Data elements found in both the protected dataset and state vital statistics data would include: age, sex, race, ethnicity, and state of residence, and year of death by suicide.*

- *Linkage to public accounts of self-harm events or to information revealed to acquaintances – Data elements in the protected dataset that might be available to a potential adversary would include: age, sex, race, ethnicity, and year of a known self-harm event or suicide attempt.*

3.3. Direct Testing of Record Linkage with an External Data Resource

In some cases of potential systematic attack, an identified or identifiable external data resource that might be available to a potential adversary is also available to the data steward. For example, a data steward may have access to state vital statistics data, drivers' license data, or voter registration data. When all potential external data resource are available to the data steward, the steps described below to estimate re-identification risk are not necessary. Instead, a data steward can link the two data sources using overlapping key variables (described above) and assess the uniqueness of linkage for each individual record.

Probability of re-identification can be directly determined for every record in the protected dataset. Any record in the protected dataset that is uniquely linked to a single record in the identified or identifiable external data resource (i.e. 1:1 match) would have a probability of re-identification equal to 1 (certainty). Any record in the protected dataset that is uniquely matched to a group of m records in the external data resource would have a probability of re-identification equal to 1/m. Any record in a group of k records in the protected dataset that are all linked to a single record in the external data resource would have a probability of re-identification equal to 1/k. Any record in a group of k records in the protected dataset that are all linked to a group of n records in the external data resource would have a probability of re-identification equal to 1/(k*n).

After ascertaining probability of re-identification for each record, a data steward can then consider appropriate thresholds for re-identification risk and undertake any necessary steps to reduce risk – as described below.

If specific external data resources are both available and of great concern, then data stewards should attempt this direct linkage. In most cases, however, all potential external data sources that could allow a re-identification attack are not available to a data steward; thus, this direct linkage is not possible. In addition, this method of direct linkage to determine risk is not possible for idiosyncratic attacks where the external data source can only be hypothesized. Consequently, data stewards could instead estimate risk in relation to potential external data resources that are not directly accessible for linkage or inspection. Specific steps in that estimation process are described below.

3.4. Testing Uniqueness or Anonymity Thresholds in a Protected Dataset

In protected datasets including small numbers of individuals, simple inspection of frequencies for potential public key variables may sometimes identify unique or nearly unique records (e.g. only two or three records with a specific value for a potential public key). As discussed in Section 3.5, however, unique or nearly unique records in small datasets do not often create significant risk of re-identification.

In large protected datasets, simple human inspection of frequencies for individual key variables will not often identify unique or nearly unique records due to the large number of potential combinations. For example, only five key variables with four categories each would create 1024 categories to be inspected. Instead it is necessary to systematically examine how possible combinations of those key variables identify unique or nearly unique records. The smallest cell or group identified by all possible combinations of public key variables is often referred to as a k-anonymity threshold[6, 19]. Smallest cells are often defined by the least common values of public key variables. For example, a large dataset derived from health records could include only a small number of females aged 18 to 29 who identify as

Native American and Hispanic and were hospitalized in a specific state in 2015. If some combination of key variables identifies any unique record, then those key variables in that dataset would be said to yield a k-anonymity of 1 (i.e. violates a k-anonymity threshold of 2). If all possible combinations of key variables do not identify any group or cell containing fewer than n, then those key variables in that dataset would be said to yield a k-anonymity n.

As discussed above, the specification of public keys or potential linking variables for any dataset depends on the specific external data resources that might be used in a re-identification attack. Consequently, k-anonymity threshold is not a fixed characteristic of a protected dataset. Instead, it is a property of any specified set of key or linking variables in that dataset. In general, a larger number of key variables (i.e. larger number of data elements shared between a protected dataset and an external data resource) will yield a lower k-anonymity threshold.

Various methods are available for assessing k-anonymity thresholds for specified key variables in any dataset. Simple n-way cross-classification procedures in standard statistical packages can be used to identify the smallest groups or cells created by n key variable, but this technique becomes cumbersome as the number of key variables and number of potential values for those variables grows large. Specialized software allows more convenient determination of k-anonymity thresholds as well as convenient testing of "what-if" scenarios regarding specification of potential key variables[8, 15]. Those specialized tools can rapidly scan all possible combinations of a specified list of key variables and accurately identify relatively small cells or groups. The public-domain CRAN package sdcMicro[15] is one such specialized tool.

This process of identifying small cells or subgroups also helps to identify the specific values of linking or key variables that define those small cells. That knowledge is necessary when estimating the likely overlap between any small cell in the protected dataset and the corresponding cell in an external data resource (discussed in Section 3.5 below).

Example (Suicide Risk Prediction Dataset):

Determination of k-anonymity thresholds for this protected dataset should separately consider each of the external data resources listed above:

- *Linkage to state vital statistics data – Using the sdcMicro statistical package, testing for anonymity threshold using age, sex, race, Hispanic ethnicity, and state of residence, and year of suicide death by suicide as potential key variables yields the following result:*

```
## The input dataset consists of 2960786 rows and 86 variables.
##   --> Categorical key variables: site, age_group, sex, Race, hisp,
visit_year, Death90
##
## Number of observations violating
##   - 2-anonymity: 583 (0.020%)
##   - 3-anonymity: 1201 (0.041%)
##   - 5-anonymity: 2576 (0.087%)
```

  Indicating that 583 records occur in cells smaller than 2 (i.e. cell size of 1 or uniquely identified by that combination of key variables in this dataset). 1201 records occur in cells smaller than 3 (i.e. cell size of 1 or 2 records).

- *Linkage to public accounts of self-harm events or to information revealed to acquaintances – Using the sdcMicro statistical package, testing for anonymity threshold using age, sex, race, Hispanic ethnicity, and state of residence, and year of suicide attempt diagnosis as potential key variables yields the following result:*

```
## The input dataset consists of 2960786 rows and 87 variables.
##   --> Categorical key variables: site, age_group, sex, Race, hisp, visit_year, Event90
##
## Number of observations violating
##   - 2-anonymity: 810 (0.027%)
##   - 3-anonymity: 1692 (0.057%)
##   - 5-anonymity: 3513 (0.119%)
```

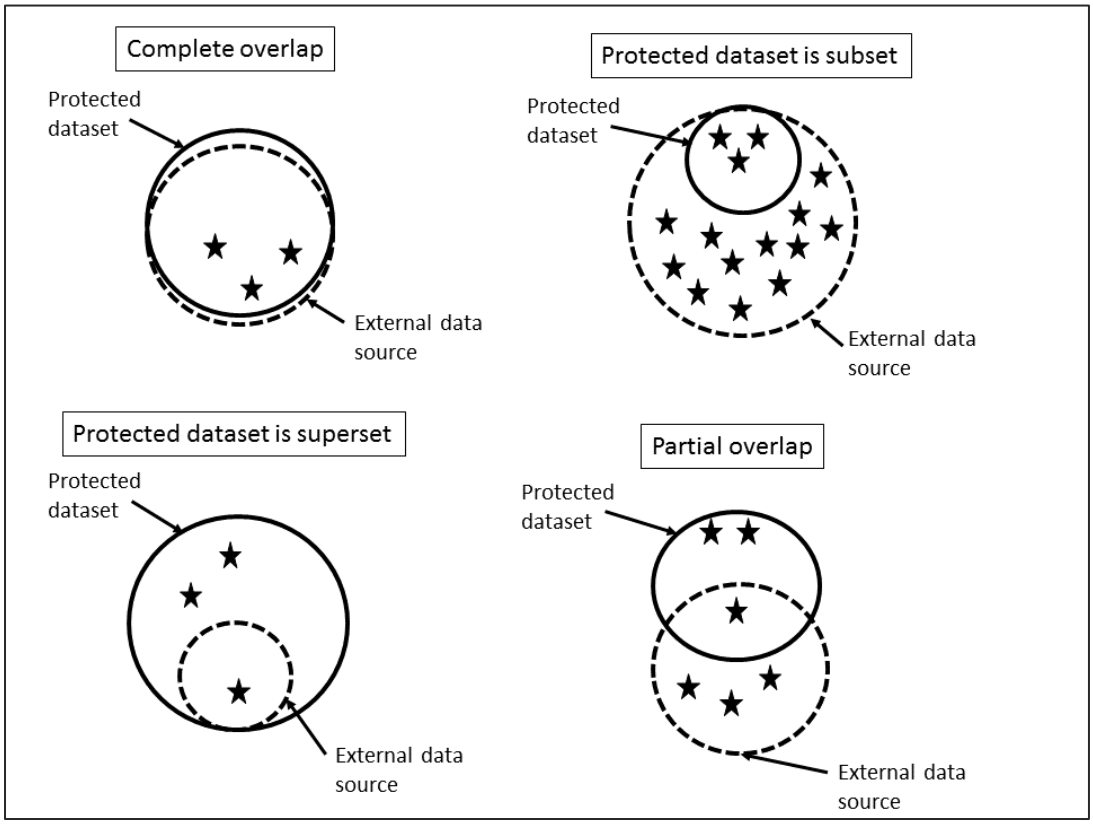
  Indicating that 810 records occur in cells smaller than 2 (i.e. are uniquely identified by this combination of key variables) and 1692 records occur in groups or cells smaller than 3 (i.e. cells including only 1 or 2 records).

3.5. Estimating Overlap with External Data Resources

As discussed above, the estimated risk of re-identification depends both on the relative uniqueness of records in the protected dataset (i.e. k-anonymity threshold) and the pattern of overlap between cells or subgroups in the protected dataset and corresponding cells or subgroups in an external data resource to which it might be linked. Greater congruence or overlap between corresponding cells or subgroups in the two data sources increases risk of re-identification. Estimating overlap is clearest regarding systematic re-identification attacks where the source of the external data resource can be clearly identified.

The figure below represents various patterns of overlap, specifically considering a group or cell in the protected dataset including only three records.

Figure 1. Overlap between protected dataset and external data resource



In the simplest scenario (upper left portion of figure), a cell or subgroup of interest in the protected dataset and the corresponding cell or subgroup in an external data resource are completely or almost completely overlapping. For example, a protected dataset could be drawn from a de-identified statewide all-payer insurance claims database and an external dataset might include all traffic accidents in that same state. In this case, any record appearing in the protected dataset would be expected to appear in the external data resource and vice versa. Given any specified set of key variables, the two data sources would have equal k-anonymity thresholds (i.e. k = m). Each of the small groups or cells defined by combinations of key variables would be expected to include exactly the same individuals in each database.

In another scenario (upper right portion of figure), a cell or subgroup of interest in the protected dataset is expected to be a subset of the corresponding cell or subgroup in the external data resource. In this case, any record in any group or cell within the protected dataset would be expected to appear in the corresponding group or cell of the external data resource, but only a proportion of records in any group or cell the external data resource would be expected to appear in the corresponding group or cell of the protected dataset. For any cell or group of records defined by a combination of linking or key variables (i.e. group *i*), we can define p_i as the proportion of records in group *i* in the external data resource that also appear in group *i* the protected dataset. In general, we assume that all p_i for different groups or cells of interest are approximately equal, but we acknowledge both the potential for biased estimation (i.e. the distributions of key variables that define cells or groups of records differ between the protected dataset and the external data resource) and random variation (i.e. given the small number of records in any cell or group of records, individual values of p_i may vary substantially across cells). As discussed

below, the potential for both biased estimation and random error argues for conservatism in using estimates of p_i to estimate re-identification risk.

In another scenario (lower left portion of figure), a cell or subgroup of interest in the protected dataset is a superset of the corresponding cell or subgroup in the external data resource. In this case, any record in any group or cell of the external data resource would be expected to appear in the protected dataset, but only a proportion of records in any group or cell of the protected dataset would be expected to appear in the external data resource. For any cell or group of records defined by a combination of linking or key variables (group *i*), we can define q_i as the proportion of that group of records in the protected dataset that also appear in the external data resource. In general, we assume that all q_i are approximately equal across subgroups, but we again must acknowledge the potential for both biased estimation and random variation.

In the most complex scenario case (lower right portion of figure), a cell or subgroup of interest in the protected dataset and the corresponding cell or subgroup in population from which the external data resource are partially overlapping. For any group of records or cell (group *i)*, we can define the corresponding quantities p_i (the proportion of group *i* records in the external data resource that appear in group *i* in the protected dataset) and q_i (the proportion of group *i* records in the protected dataset that appear in group *i* the external data resource).

When the external data resource can be directly examined (discussed above), it is not necessary to estimate p_i and q_i. If a data steward can attempt linkage directly, that process will accurately and directly determine p_i and q_i for all cells or groups in the protected dataset.

Estimation of p_i and q_i typically begins with generally knowledge regarding the sources of the protected dataset and the external data resources. For example, if health plan X serves approximately one-third of the population of a specific state, then we would initially estimate that a person appearing in an identified state motor vehicle accident database would have a one-third chance of also appearing in a de-identified health insurance claims database from that health plan (i.e. p_i=1/3). But we may have also determined that small cells or groups in the health claims database are defined by membership in less common racial or ethnic groups. If health plan X serves a disproportionately higher proportion of people in those racial and ethnic groups, then we would estimate p_i to be greater than 1/3 for cells defined by those racial and ethnic groups and lower for others.

These concepts still apply, in a modified way, to estimating overlap regarding idiosyncratic re-identification. When we consider information that might be idiosyncratically available to a potential adversary, we must consider how the sources of that information might affect overlap between a specific cell or subgroup in the protected dataset and identifying information available to the potential adversary. In some cases of idiosyncratic attacks, information available to an adversary may allow re-identification to focus on very specific subgroups within a protected dataset. This is best illustrated by concrete examples. Consider the scenario of a public account regarding a prominent person being hospitalized following a motor vehicle accident. An adversary might attempt to use public information regarding that person (age, sex, race, ethnicity, date of hospitalization, broad category of diagnosis) to link that public account to a research database derived from electronic health records. That research database is derived from records of all enrolled members of health system X. System X serves approximately one fourth of residents of state Y where that hospitalization occurred and serves a similar proportion of residents of two neighboring states, with all three states having approximately equal numbers of residents. Depending on the specific information available, this case could map to any of the four scenarios described above.

- Complete overlap – This would occur if a potential adversary knew that the hospitalization occurred within health system X AND if the research data allowed the adversary to limit linkage to records from state Y. Consequently, the potential adversary would know that a person with those characteristics in state Y has 100% probability of appearing in the research dataset (p_i=1) *and* that a person with those characteristics in records from health system X would have received care in state Y (q_i=1).
- Protected dataset is subset – This would occur if a potential adversary did not know if the hospitalization occurred within health system X or some other health system, but the protected dataset did allow a potential adversary to identify records from state Y. Consequently, the potential adversary would know that a person with those characteristics in state Y has 1/4 probability of appearing in the research dataset because health system X covers 25% of people residing or receiving care in state Y (p_i=1/4) and that a person with those characteristics in records from health system X would have received care in state Y (q_i=1).
- Protected dataset is superset – This would occur if a potential adversary knew that the hospitalization occurred within health system X, but the protected dataset did not allow a potential adversary to specifically identify records from state Y. Consequently, the potential adversary would know that a person with those characteristics in state Y has a 100% probability of appearing in the research dataset (p_i=1) and that a person with those characteristics in records from health system X would have 1/3 probability of receiving care in state Y (q_i=1/3).
- Partial overlap – In this case, the potential adversary does not know if this hospitalization occurred within X healthcare system, and the protected dataset does not allow the potential adversary to specifically identify hospitalizations occurring in state Y. Consequently, the potential adversary would know that a person with those characteristics in state Y has a 1/4 probability of appearing in the research dataset (p_i=1/4) and that a person with those characteristics in records from health system X would have 1/3 probability of receiving care in state Y (q_i=1/3).

Example (Suicide Risk Prediction Dataset):

Assessing pattern of overlap for this protected dataset should separately consider each of the external data resources listed above:

- *Linkage to state vital statistics data – Each of the health systems contributing to the protected dataset serves fewer than 12% of state residents in each system's state of operations. If a potential adversary had access to identified or identifiable state vital statistics data, then any cell or subgroup in the protected dataset would be a subset of the corresponding cell or subgroup in the external data resource, and we would estimate p_i to be 0.12. Alternatively, if the research dataset did not include information regarding state of residence, then any cell or subgroup in the research dataset would have a partial overlap relationship to the corresponding cell or subgroup in the state vital statistics database (only some records from the dataset could be represented in any state database and only some records from any state database could be represented in the research dataset).*
- *Linkage to public accounts of self-harm events or to information revealed to acquaintances – If an adversary were aware of an identified individual treated for self-harm or suicide attempt in a state served by one of the health systems contributing to the protected dataset, the corresponding cell or subgroup in the protected dataset would be a subset of that external data resource. We would again estimate p_i to be 0.12. If the adversary were aware that an identified*

individual was treated by one of the health systems contributing to the protected dataset, then the protected dataset would completely overlap with the external data resource.

3.6. Estimating Re-identification probability

As discussed above, it may sometimes be possible to directly examine uniqueness or probability of linkage between a protected dataset and an identified or potentially identifiable data resource. In most cases, however, that external data resource is not available for direct interrogation. Consequently, uniqueness or probability of linkage (and re-identification) must be estimated using the known group sizes or k-anonymity threshold(s) of the protected dataset and the estimated overlaps (proportions p_i and q_i) with the identified or identifiable data resource with which the protected dataset could be linked.

In the scenario of complete overlap, the probability of re-identifying any record at any level of k-anonymity is inversely proportional to that k-anonymity threshold. Records with k-anonymity of 1 (i.e. only one record with this specific combination of values for public key or potential linking variables) can be re-identified with complete certainty. Records with k-anonymity greater than one (i.e. a group or cell of k records have this specific combination of values for public key or potential linking variables) can be re-identified with probability of 1/k.

In other overlap scenarios, probability of re-identification depends on both k-anonymity thresholds in the protected dataset and pattern overlap with an external data resource. We use those overlap proportions to estimate the size of the corresponding group or cell in the external data resource and the proportion of that corresponding group who would also be contained in the protected dataset. When a cell or subgroup if interest in the protected dataset is a subset of the corresponding cell or subgroup in the external data resource, we estimate probability of re-identification as p_i/k). When a cell or subgroup of interest in the protected dataset is a superset of the corresponding cell or subgroup in the external data resource, we estimate probability of re-identification as q_i/k. When cells or subgroups in the two data resources are partially overlapping, we estimate probability of re-identification as (p_i*q_i)/k.

The above calculations regarding the subset and partial overlap scenarios illustrate an important principle regarding re-identification risk in small research datasets. In a sensitive or protected dataset including a small number of people, a larger proportion of records will fall into unique or relatively small cells (i.e. more small values of k). But any cell or subgroup of interest in that dataset will usually also represent a small proportion of the corresponding cell or subgroup in some identified or identifiable external data resource (i.e. small value of p_i). Those two factors (uniqueness within the protected dataset and overlap with corresponding group in an external resource) typically have competing or counter-balancing effects on probability of re-identification. The balance of those two effects depends on the specific characteristics of each protected dataset, so we cannot assume that smaller datasets always carry higher or lower levels of re-identification risk. We should recognize, however, that the protective effect of a small overlap proportion (i.e. small value of p_i) can be rapidly and completely cancelled by a relatively innocuous disclosure. For example, a study of people with hypertension might be expected to include fewer than 1 in 1000 people with hypertension in a particular state (i.e. estimated p_i=.001 with respect to some statewide information or database). Even for people in unique cells (k=1), estimated probability of re-identification is only 1 in 1000. If, however, an individual reveals that they participated in that specific study, p_i is now equal to 1 for this individual (i.e. they are known to be represented in both the research dataset and the identifiable data resource), dramatically increasing risk of re-identification.

Example (Suicide Risk Prediction Dataset):

Estimating probability of re-identification for this protected dataset should separately consider each of the external data resources listed above:

- *Linkage to state vital statistics data – As discussed above, the protected dataset would be considered a subset of the external data resource with p_i generally estimated to be 0.12 (i.e. health system cares for approximately 12% of state residents). Consequently, records in unique cells would have estimated probability of re-identification equal to 0.12 and records in groups or cells of size 2 would have estimated probability of re-identification equal to 0.12/2 = 0.06.*
- *Linkage to public accounts of self-harm events or to information revealed to acquaintances – If an adversary were aware of an emergency department visit or hospitalization for self-harm for a resident of the state, then records in unique cells would have an estimated probability of re-identification of 0.12. If an adversary were aware of an emergency department visit or hospitalization for self-harm within one of the participating health systems, then records in unique cells (k-anonymity=1) would have an estimated probability of re-identification of 1.0.*

3.7. Practical Interpretation of Re-identification Probability

The calculations above regarding probability of re-identification are symmetric regarding direction of overlap (i.e. *p* and *q* have equal effects on probability). But different overlap patterns have different practical implications or interpretations. This is illustrated by translating the calculations above into plain-language descriptions of re-identification risk for any record in the protected dataset falling within a group with k-anonymity threshold equal to n.

For a systematic re-identification attack, those probabilities or certainties of re-identification can be expressed as follows. Below we assume that there are n people in the external data resource that have the same values on the key variable as an individual in the protected data set who is unique in the data set in those variable values.

- Complete overlap: This person in the protected dataset could be linked to n identified or identifiable records available to a potential adversary.
- Protected dataset is subset: This person in the protected dataset could be linked to an estimated n/p_i identified or identifiable records available to a potential adversary.
- Protected dataset is superset: This person in the protected dataset has q_i probability of appearing in the external data resource and, if they do appear, could be linked to an estimated n identified or identifiable records available to a potential adversary.
- Partial overlap: This person in the protected dataset has q_i probability of appearing in the external data resource and, if they do appear, could be linked an estimated $n/(q_i*p_i)$ identified or identifiable records available to a potential adversary.

These plain-language descriptions illustrate the distinction between the two sources of uncertainty in re-identification: the likelihood that an adversary attempting linkage would access records for an individual AND the number of names (or other identifiers) that an adversary could link to that individual's sensitive information.

For an idiosyncratic re-identification attack, public keys or linking variables typically include notable health events (e.g. hospitalizations) that may be identifiable to acquaintances or via public accounts. Consequently, probabilities or certainties of re-identification can be expressed in relation to those notable events, as follows:

- Complete overlap: This person in the protected dataset could be linked to n events that might be known to a potential adversary.
- Protected dataset is subset: This person in the protected dataset could be linked to n/p_i events that might be known to a potential adversary.
- Protected dataset is superset: This person in the protected dataset has q_i probability of having experienced an event known to a potential adversary. If so, this person could be linked to n such events.
- Partial overlap: This person in the protected dataset has q_i probability of having experienced an event known to a potential adversary. If so, this person could be linked to n/(q_i*p_i) such events.

Those plain-language descriptions illustrate the distinction between two contributors to probability or certainty of re-identification: the probability that a person represented in a protected dataset will appear in an external data resource and the number of identified or identifiable records to which that person could be linked. Even if these two parameters contribute equally to an overall probability or certainty of re-identification, different stakeholders may attach different values or weights to those two contributors to risk. As discussed in Section 3.8 below, further assessment of heterogeneity within high-risk groups is only relevant to the latter issue – the number of records in a protected dataset in any group or cell defined by a specific combination of key variables.

Example (Suicide Risk Prediction Dataset):

Using specific information regarding anonymity thresholds and patterns of overlap, we can create plain-language descriptions of re-identification risk for each of the two re-identification scenarios relevant to this protected dataset:

- *Linkage to state vital statistics data – An individual in the protected dataset dying by suicide appearing in a group or cell of size 3 could be linked to approximately 3/0.12= 24 identified or identifiable records in state vital statistics data.*
- *Linkage to public accounts of self-harm events or to information revealed to acquaintances – (If not known to be treated by participating health system) An individual in the protected dataset dying by suicide appearing in a group or cell of size 3 could be linked to approximately 3/0.12 = 24 events that might be known to a potential adversary. (If known to be treated by participating health system, an individual in the protected dataset dying by suicide appearing in a group or cell of size 3 could be linked to 3 events that might be known to a potential adversary).*

3.8. Assessing Heterogeneity within High-risk Groups

As discussed above, re-identification creates harm when an adversary gains new access to sensitive information regarding identifiable individuals. As illustrated by the diagrams and calculations above, risk for any sensitive protected dataset is related to k-anonymity or the size of the smallest cells or groups defined by public key or linking variables. But we must also consider the possibility that groups or cells within the protected dataset are homogeneous with respect to sensitive data elements or variables. For example, the calculations above might determine that identified persons in a commercial credit database could be linked to n records in a protected dataset that includes data regarding an especially sensitive or stigmatized diagnosis. We could then say that probability of linkage is equal to 1/n. But we would interpret that risk quite differently if either all or most of the records in that specific group within the protected dataset had indicators of that sensitive condition. Evaluating this additional determinant of risk requires:

- Identifying specific cells or groups in the protected dataset with low k-anonymity thresholds

- Specifying variables (other than public key or linking variables) in the protected dataset that might be especially sensitive
- Examining the distributions of those sensitive variables within the specific groups with low k-anonymity thresholds in the protected dataset

The term I-diversity[20] is sometimes used to refer to heterogeneity of sensitive data elements in a small group of records. Interpretation of I-diversity, however, depends on the nature of a specific sensitive data element. For a dichotomous data element (e.g. presence or absence of a specific diagnosis), I-diversity of 2 is both adequate and the maximum possible. For multi-categorical or data elements, consideration of individual values may be necessary to determine a level of I-diversity necessary to prevent disclosure of sensitive or stigmatized information.

For dichotomous variables, concern about I-diversity will often be asymmetric. Risk or potential harm is greater when all members of a group or cell share some sensitive or stigmatized characteristic than when all members do not have that sensitive or stigmatized characteristic.

The qualitative process of identifying specific sensitive information should help data stewards to determine appropriate quantitative thresholds for re-identification risk (discussed in Section 4.2 below).

Because these procedures identify specific variables and specific records that carry greater risk of disclosing sensitive information, this procedure may also guide data stewards toward specific risk mitigation strategies (also discussed in Section 4.2).

Example (Suicide Risk Prediction Dataset):

This dataset contains a large number of variables indicating presence/absence of potentially sensitive aspects of mental health history (e.g. history of suicide attempt, history of psychiatric hospitalization, diagnosis of mood or substance use disorder). Consequently, we are concerned that a relatively small cell or group (i.e. 3 to 5 records) would be homogeneous with respect to presence of one of these sensitive events. If, for example all 5 members of a group had a history of psychiatric hospitalization, then linkage of an identified record in an external data source to that group of five would reveal with 100% certainty that the identified individual had a history of psychiatric hospitalization. Probability that all members of a small cell will share some sensitive characteristic is related to the overall prevalence of that characteristic in the sample. Consequently, we examine the diversity of small cells (sized 3, 4, or 5) with respect to diagnosis of depressive disorder (overall prevalence in our protected dataset >50%) and diagnosis of alcohol use disorder (overall prevalence in our protected dataset <10%). Regarding potential linkage to state mortality data, we consider cells defined by the key variables sex, age group, race, ethnicity, state of residence, and year of suicide death. Among cells sized 3, 4, or 5 approximately 18% are homogeneous with respect to presence of a depressive disorder diagnosis but none are homogenous with respect to presence of an alcohol use disorder diagnosis. Consequently, an adversary who linked identifiable records to cells sized 3, 4, or 5 would gain certain knowledge about depression diagnosis in a significant number of records but would not gain certain knowledge about presence of an alcohol use disorder diagnosis. In evaluating the risk to people represented in the research dataset, we should consider that an adversary could only accomplish this linkage with prior knowledge of death by suicide. Because depression diagnosis would often be presumed in people dying by suicide, gaining that knowledge would likely pose little additional risk.

4. Protecting Against Risk

4.1. Determining Appropriate Risk Thresholds

The procedures described above can yield a quantitative estimate of re-identification risk, a plain-language description of that quantitative estimate, and a further assessment of risk regarding specific groups of records and specific sensitive data elements. None of these procedures, however, determine a specific acceptable threshold of risk.

Determination of an acceptable threshold should consider the perspectives of various stakeholders, including patients or consumers affected by the health condition(s) of interest, health systems contributing potentially sensitive data, and researchers or others who might make use of shared data. Stakeholders can be informed by the quantitative estimates of risk described above but should also consider the sensitivity of the information that might be revealed and the vulnerability of those at risk of inadvertent disclosure. Stakeholders can also help to balance risks of re-identification against the potential scientific or public health benefits of data sharing. People affected by the health condition of interest can contribute valuable perspective on both risks and potential benefits.

As discussed in Section 3.7 above, two types of uncertainty can contribute to probability of definite re-identification. From the perspective of a person represented in a protected dataset including sensitive information, these two can be expressed as:

- The probability my records would be included in data available to an adversary
- The number of names my records could be linked to

While those two sources of uncertainty contribute equally to quantitative estimates of re-identification probability, stakeholders may have different qualitative evaluations of them. Consequently, engaging stakeholders in discussions regarding appropriate risk thresholds may need to distinguish those sources of uncertainty.

Determination of an appropriate risk threshold should also consider the motivation of potential adversaries who might attempt re-identification. Motivation may be greater when research data might include sensitive information regarding public figures. An adversary's motivation may also be greater when the protected dataset concerns especially controversial treatments or policies.

In general, determination of appropriate risk thresholds will consider maximum risk to any person represented in the protected dataset (i.e. risk to those represented in the smallest cells or groups defined by potential key variables). In some cases, average risk across the entire protected dataset may be considered. When risk of re-identification is confined to a very small number of records, then efforts to mitigate risk (discussed in detail below) may focus on reducing risk for those specific records rather than modifications to the entire protected dataset.

Previous discussions of re-identification risk have proposed thresholds of 0.2 (1 in 5) for general health information and 0.05 (1 in 20) for especially sensitive or stigmatized information when sharing public-use datasets[8, 11, 21, 22]. As discussed in Section 4.3 below, higher thresholds (even up to 1.0 probability of linkage) may be appropriate for more tightly controlled data sharing mechanisms.

Example (Suicide Risk Prediction Dataset):

If an adversary were able to identify individual records in the protected dataset, they could gain access to information regarding a range of mental health diagnoses and treatments (e.g. history of psychiatric hospitalization, use of specific psychiatric medications). But if re-identification required that a potential adversary already had access to identifiable information regarding suicide attempt and/or suicide death,

which is both strongly associated with and typically more sensitive than, having a mental health diagnosis or receiving mental health treatment, then selection of an appropriate threshold for re-identification risk or certainty should consider the sensitivity of the NEW information that an adversary would gain, not information already available.

4.2. Risk Mitigation Strategies

If the methods described above do not identify any records in a protected dataset with a probability of re-identification exceeding a reasonable threshold (i.e. reasonable to affected stakeholders given the information that might be revealed and the motivations of potential adversaries), then no additional efforts to reduce risk are necessary.

In some cases, the methods described above will identify some number of records with a probability of identification exceeding a reasonable threshold. In those cases, data stewards can consider a range of options for reducing or mitigating risk. Selection of specific options will depend on the risks identified as well as the sources of scientific or public health value in the data to be shared. Preferred strategies are those that adequately mitigate risk with minimal loss of scientific or public health value.

Traditional strategies for mitigating risk involve limited redactions or alternations of the research dataset to address specific risks. Potential strategies of this type are listed below, in general order of smaller to greater potential loss of scientific or public health value, allowing that loss of value must be considered on a case-by-case basis. These strategies are not mutually exclusive, and a combination of strategies may sometimes be optimal.

- Coarsening potential public key or linking variables – Fine-grained specification of potential public key variables (e.g. representing age in narrow bands) may lead to unique or low k-anonymity thresholds in some protected datasets. Coarsening the representation of those key variables (e.g. using broader age bands, combining rare racial or ethnic groups) may significantly reduce re-identification risk without sacrificing significant scientific or public health value.
- Perturbing potential public key or linking variables – A data steward could perturb values of key or linking variables in specific small groups or cells. For example, a data steward could perturb the age value for a single record uniquely identified by a specific age band and a specific less common racial or ethnic group. While this strategy may allow data stewards to retain more detailed representation of key variables in other records, it does intentionally introduce error into one or more records.
- Deleting potential public key or linking variables – This strategy may be preferable if a potential public key variable contributes significantly to re-identification risk for large numbers of records AND does not contribute significant scientific or public health value. This might occur if a potential public key significantly influences knowledge of overlap. For example, deleting state of residence may transform a complete overlap scenario (higher risk) to a "protected dataset is superset" scenario (typically much lower risk).
- Deleting individual records – When significant risk of re-identification is confined to one or a small number of records, deleting those records prior to sharing may be an appropriate method for reducing risk without sacrificing public health value. This may be a desirable strategy when specific characteristics (e.g. delivery of quadruplets) create high-risk outliers in a protected dataset with a generally low level of risk.
- Perturbing sensitive data elements – If the distributions of sensitive data elements within specific cells (described above) creates significant risk within a small number of records, then

altering distribution of those specific data elements within those specific records may be appropriate[22]. Since these sensitive data elements are often included due to their scientific importance, careful thought must be given to the impact of perturbing or altering those data elements on scientific or public health value.

- Deleting sensitive data elements – When a significant proportion of records have an unacceptable risk of re-identification and other options (described above) are not adequate, then deletion of specific sensitive data elements (e.g. diagnosis of or treatment for substance use disorder) could be deleted prior to data sharing. This strategy could be reasonable if deletion of those sensitive data elements did not compromise scientific or public health value.

Alternative privacy-preserving strategies involve systematic alterations of the entire dataset. Potential strategies of this type include:

- Summarizing individual covariates using propensity scores or risk scores – When potential users of a sensitive research dataset can clearly identify the covariates or confounders relevant to a specific research question, then it may be possible to share only propensity scores or risk scores without any meaningful loss of scientific or public health value[23, 24]. In this scenario, the scientific question must be clearly articulated before data sharing so that appropriate propensity scores or summary measures can be calculated.
- Homomorphic encryption – Development of prediction models or other applications of machine learning typically require access to full detail regarding individual covariates or predictors. In those cases, homomorphic encryption may allow analysts to train machine learning models using encrypted data without loss of prediction accuracy when models are applied to original data[25]. Few data stewards and users of shared data, however, will have the technical capacity to implement these tools.

Example (Suicide Risk Prediction Dataset):

In this protected dataset, data regarding potentially sensitive mental health diagnoses and treatments are central to scientific value (i.e. evaluating the associations between suicidal behavior and prior mental health diagnoses or treatments). Consequently, deleting or perturbing those potentially sensitive data elements would not be a viable strategy for reducing risk. For the same reason, deleting or perturbing information regarding suicidal behavior would not be a useful strategy. Age, sex, race, and ethnicity are all strongly related to risk of suicidal behavior, so deleting or altering those key or linking variables would also reduce scientific or public health value. Consequently, evaluation of risk mitigation strategies would consider deleting or altering other key variables (state and year) before deleting or altering variables of greater interest or importance. We can evaluate alternative strategies separately for two linkage scenarios:

- *Linkage to state vital statistics data – Using the sdcMicro package, we can re-evaluate k-anonymity thresholds when excluding state, year, or both state and year from the list of public keys or linking variables.*

  Excluding state from the protected dataset (i.e. eliminating it as a potential key variable) yields the following regarding k-anonymity in the protected dataset:

```
## Number of observations violating
##   - 2-anonymity: 113 (0.004%)
##   - 3-anonymity: 185 (0.006%)
```

```
##    - 5-anonymity: 275 (0.009%)
```

Excluding year from the protected dataset yields the following regarding k-anonymity in the protected dataset:

```
## Number of observations violating
##    - 2-anonymity: 92 (0.003%)
##    - 3-anonymity: 156 (0.005%)
##    - 5-anonymity: 383 (0.013%)
```

Excluding state AND year from the protected dataset yields the following regarding k-anonymity in the protected dataset:

```
## Number of observations violating
##    - 2-anonymity: 21 (0.001%)
##    - 3-anonymity: 37 (0.001%)
##    - 5-anonymity: 78 (0.003%)
```

Highest risk of re-identification lies within the 58 records that violate 2-anonymity or 3-anonymity (i.e. fall into cells sized 2 or smaller), equivalent to fewer than 0.001% of all records. Consequently, we might consider deleting or altering that small number of records. Closer inspection of those small-cell records, however reveals that they disproportionately include adolescents (aged 13-17) and people from less common racial and ethnic groups. These 58 small-cell records account for 8 suicide deaths among peopled aged 13-17 out of 16 total suicide deaths in that age group. Given public health importance of suicide in adolescents deleting or altering even those few small-cell records in this dataset might significantly reduce scientific or public health value.

- *Linkage to public accounts of self-harm events or to information revealed to acquaintances – Using the sdcMicro package, we can re-evaluate k-anonymity thresholds when excluding state, year, or both state and year from the list of public keys or linking variables.*

  Excluding state as a potential key variable yields the following regarding k-anonymity thresholds:

  ```
  ##    - 2-anonymity: 129 (0.004%)
  ##    - 3-anonymity: 241 (0.008%)
  ##    - 5-anonymity: 561 (0.019%)
  ```

  Excluding year as a potential key variable yields the following regarding k-anonymity thresholds:

  ```
  ## Number of observations violating
  ##    - 2-anonymity: 112 (0.004%)
  ##    - 3-anonymity: 256 (0.009%)
  ##    - 5-anonymity: 588 (0.020%)
  ```

  Excluding state and year as potential key variables yields the following regarding k-anonymity thresholds:

  ```
  ## Number of observations violating
  ##    - 2-anonymity: 14 (0.000%)
  ##    - 3-anonymity: 30 (0.001%)
  ```

```
##    - 5-anonymity: 58 (0.002%)
```

Highest risk of re-identification lies within the 44 records violating 2-anonymity or 3-anonymity (i.e. falling into cells sized 2 or smaller), equivalent to fewer than 0.001% of all records. Consequently, we might consider deleting or altering that small number of records. Closer inspection reveals a similar scenario to that described above, that small-cell records disproportionately include adolescents and people from less common racial and ethnic groups. Give the much higher prevalence of non-fatal suicide attempts, however, small-cell records account for only a small proportion of events. For example, these 44 small cell records account for 9 of 1275 self-harm or suicide attempt events among adolescents in the entire sample. Deleting or altering this small number of small-cell records might not significantly reduce scientific or public health value.

4.3. Mechanisms for Data Sharing

The descriptions of risk assessment and mitigation procedures above presume that data would be widely shared – and therefore available to any potential adversary. There are, however, other mechanisms that allow greater control over access, use, and re-disclosure. Use of more restrictive data-sharing mechanisms are therefore an alternative strategy for mitigating risk. More restrictive data-sharing mechanisms may be especially appropriate when risk mitigation strategies that require altering a protected dataset (described above) would significant compromise scientific or public health value. Alternative mechanisms for data sharing are described in detail elsewhere[5, 26-28] and will be briefly summarized here.

Data sharing mechanisms can be divided into data archives (which allow users direct access to data) and data enclaves (which allow users to interact indirectly with data by submitting queries or analytic programs). Data archives have been described as mechanisms for "sending data to questions", while data enclaves have been described as tools for "sending questions to data."

Data archives can allow completely unrestricted access (e.g. any anonymous user, including potential adversaries, can download a public use dataset from a completely public website) or can place restrictions on permitted users and permitted uses. Potential users could be required to identify themselves or to pass an approval or vetting process. Users might also be expected to propose specific uses of data or agree to boundaries on data use. A formal data use agreement is an instrument to define and restrict authorized users and uses. Data use agreements or other legal agreements restricting data use cannot completely prevent inappropriate use (such as attempts to re-identify individuals) or re-disclosure, but they do allow data stewards some legal recourse if inappropriate use or re-disclosure is detected.

Data enclaves allow data stewards various levels of control over access to data by potential adversaries[29-31]. A public enclave might allow any user (including potential adversaries) to submit queries while a private enclave could require potential users to identify themselves or to pass an approval or vetting process. In either a public or private enclave, queries could be limited to include specific numbers or combinations of potential public key variables, therefore reducing risk of re-identification for specific records in the enclave. If users are permitted to submit analytic programs via an enclave, then human review could be required, including review of programs prior to execution and/or review of output prior to release.

Even when deletion of sensitive records or perturbation of sensitive data elements significantly reduces scientific or public health value, a redacted or altered dataset may still allow potential users to develop and test analytic programs that could then be submitted to a data enclave. Consequently, stewards of a data enclave may make redacted or altered datasets publicly available as test beds.

Example (Suicide Risk Prediction Dataset):

Given the findings described above regarding re-identification risk and alternative strategies for mitigating risk, it would not be prudent to share data regarding suicide death through any public archive (i.e. public-use dataset not governed by a formal data use agreement). If such data were to be shared, either a formal data use agreement (specifying allowable uses and specifically prohibiting any attempts at re-identification) or a data enclave structure (technically preventing re-identification) would be necessary.

Sharing data regarding suicide attempts through a public archive could be prudent if a small number of high-risk records were removed or altered. If removal or alteration of those records was considered to significantly degrade scientific or public health value, then data could be shared via a data use agreement or a more protective technical structure.

4.4. Summary and Review

Understanding sources of risk:

- Stewards of sensitive or protected data derived from health records have both regulatory and ethical obligations to assess and address risk of re-identification prior to releasing or sharing those data.
- Risk of re-identification exists when data elements or variables in a protected dataset can be linked to corresponding variables or data elements in some identified or identifiable external data resource.
- Re-identification could use public (and not sensitive) elements common to both data resources to gain access to non-public (and sensitive) information in the protected dataset.

Assessing risk:

- Stewards of protected data should consider a range of external data sources that might allow or facilitate such a linkage, including both systematic linkage to publicly available datasets and idiosyncratic linkage to public knowledge regarding a specific individual health event.
- For any potential external data resource, a data steward should identify specific variables that could serve as public keys (i.e. variables common to both a protected dataset and an identified or identifiable external data resource).
- When an external data resource is available for inspection, a data steward can directly assess re-identification risk and directly test strategies to mitigate risk. In most cases, the external data resource is not directly available, so a data steward must follow a series of steps to estimate and (if necessary) mitigate risk.
- Stewards of protected data should first identify the smallest cells defined by combinations or permutations of key variables. The smallest number of records in any such cell is sometimes referred to as the k-anonymity threshold for that combination of key variables in that protected dataset. A small k-anonymity threshold is one key determinant of re-identification risk.
- Stewards of protected data should also examine or estimate the pattern and degree of overlap between the cells or subgroups of interest in the protected dataset and the corresponding cells or subgroups in the external data resources to which it might be linked. Overlap between corresponding cells in the two data sources is another key determinant of re-identification risk.

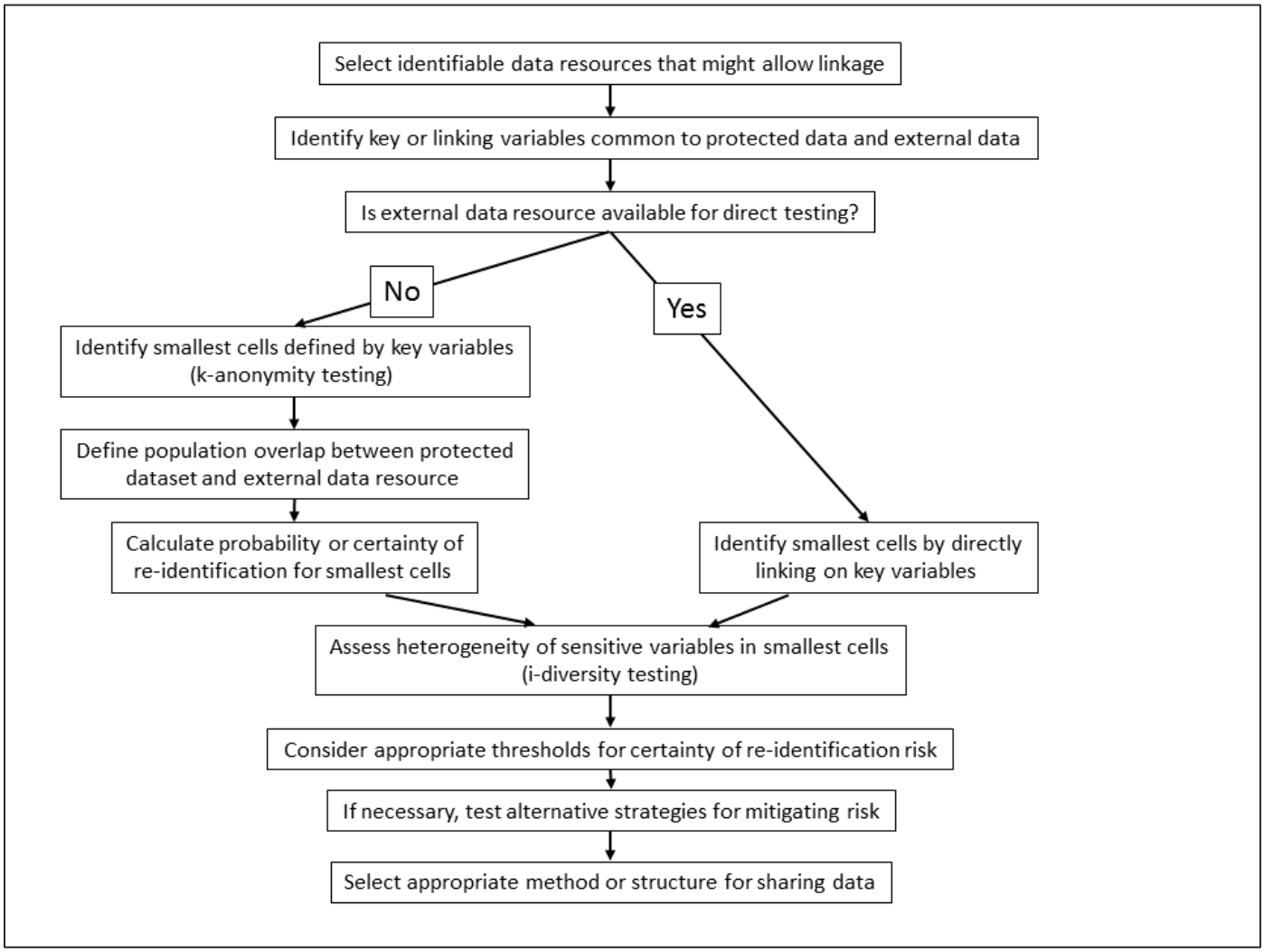
- For any record in the protected dataset, risk or certainty of re-identification can be estimated using the k-anonymity threshold for the cell in which that record resides and the estimated overlap between the protected dataset and the external data resource.
- Risk of disclosing new sensitive information is also related to the homogeneity of sensitive data elements within small groups or cells in the protected dataset.

Protecting against risk:

- Determining the appropriate risk threshold for releasing or sharing any protected dataset depends on the sensitivity of the information that might be revealed, the motivations of adversaries who might try to re-identify sensitive data, and the mechanism through which data will be released or shared.
- Data stewards can use a variety of specific strategies to reduce risk of re-identification, including deleting or altering key variables, deleting or altering sensitive variables, and deleting or altering specific high-risk records. Selection of a method that effectively reduces risk without compromising scientific or public health value depends on the specifics of each data sharing scenario.
- Data may be shared via a range of methods, ranging from the least controlled (a public-use dataset available to any anonymous user) to the most controlled (a controlled data enclave available only to identified and trusted users). Selection of an appropriate mechanism for data sharing depends on the sensitivity of the protected dataset, the risk of re-identification (determined by the procedures described above) and the motivations of those who might access data.

The figure below illustrates the steps in this process:

Figure 2. Understanding sources of risk, assessing risk, and protecting against risk



5. Additional Resources

Department of Health and Human Services guidance regarding data de-identification (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html) provides a non-technical description of re-identification risk and a description of "expert determination" methods for assessing risk.

The public-domain sdcMicro R package (https://github.com/sdcTools/sdcMicro) includes tools for a variety of specific tasks for assessing re-identification risk in a protected dataset, including: assessment of k-anonymity thresholds, counts and rates of records falling within small groups or cells, identification of key variables with greatest influence on k-anonymity thresholds, and assessment of heterogeneity of sensitive data elements within small groups or cells.

HITRUST® (https://hitrustalliance.net/) offers training and certification in data de-identification methodology.

The UK Anonymisation Network (https://ukanon.net/) has published an anonymization decision-making framework, including applications to specific examples and instructions for assessing uniqueness or risk using Microsoft Excel® spreadsheets and the SPSS® software package.

6. References

1. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 2010;17:169-77.

2.Simon GE, Shortreed SM, Coley RY, et al. Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records. EGEMS (Wash DC) 2019;7:6.

3. Guidance Regarding Methods for De-identification of Protected Health Information. Department of Health and Human Services, 2015. (Accessed May 31, 2018,

4. Federal Policy for the Protection of Human Subjects. In: Services DoHaH, ed. 45 CFR Part 462017.

5. Simon GE, Coronado G, DeBar LL, et al. Data Sharing and Embedded Research. Ann Intern Med 2017;167:668-70.

6. El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc 2008;15:627-37.

7. El Emam K, Dankar FK, Neisa A, Jonker E. Evaluating the risk of patient re-identification from adverse drug event reports. BMC Med Inform Decis Mak 2013;13:114.

8. Dankar FK, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak 2012;12:66.

9. Manrique-Vallier D, Reiter JP. Estimating Identification Disclosure Risk Using Mixed Membership Models. J Am Stat Assoc 2012;107:1385-94.

10. Chen YL, Cheng BC, Chen HL, et al. A privacy-preserved analytical method for ehealth database with minimized information loss. J Biomed Biotechnol 2012;2012:521267.

11. Lin WY, Yang DC, Wang JT. Privacy preserving data anonymization of spontaneous ADE reporting system dataset. BMC Med Inform Decis Mak 2016;16 Suppl 1:58.

12. Lee H, Kim S, Kim JW, Chung YD. Utility-preserving anonymization for health data publishing. BMC Med Inform Decis Mak 2017;17:104.

13. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-ideintifications in incomplete datasets using generative models. Nature Communications 2019;10.

14. Simon GE, Johnson E, Lawrence JM, et al. Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. Am J Psychiatry 2018:appiajp201817101167.

15. Templ M, Kowarik A, Meindl B. Statistical disclosure control for micro-data using R package sdcMicro. Journal of Statistical Software 2015;67.

16. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6:e28071.

17. de Montjoye YA, Radaelli L, Singh VK, Pentland AS. Identity and privacy. Unique in the shopping mall: on the reidentifiability of credit card metadata. Science 2015;347:536-9.

18. Matching Known Patients to Health Records in Washington State. Data Privacy Lab, 2014. (Accessed July 2, 2019, at https://dataprivacylab.org/projects/wa/1089-1.pdf.)

19. Sweeney L. k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems 2002;10:557-70.

20. Machanavajjhala A, Gehrke J, Kifer M, Venkitasubramaniam M. I-diversity: Privacy beyond k-anomymity. 22nd International Conference on Data Engineering; 2006; Atlanta, GA, USA.

21. Measuring Re-identification Risk. Data Loss Prevention API, 2018. (Accessed May 31, 2018,

22. Ursin G, Sen S, Mottu JM, Nygard M. Protecting Privacy in Large Datasets-First We Assess the Risk; Then We Fuzzy the Data. Cancer Epidemiol Biomarkers Prev 2017;26:1219-24.

23. Toh S, Reichman ME, Houstoun M, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data. Pharmacoepidemiol Drug Saf 2013.

24. Toh S, Shetterly S, Powers JD, Arterburn D. Privacy-preserving analytic methods for multisite comparative effectiveness and patient-centered outcomes research. Med Care 2014;52:664-8.

25. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform 2014;50:234-43.

26. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care 2010;48:S45-51.

27. Her QL, Malenfant JM, Malek S, et al. A Query Workflow Design to Perform Automatable Distributed Regression Analysis in Large Distributed Data Networks. eGEMs 2018;6:11.

28. Holmes JH, Brown J, Hennessy S, et al. Developing a distributed research network to conduct population-based studies and safety surveillance. AMIA Annu Symp Proc 2008:973.

29. Vogel J, Brown JS, Land T, Platt R, Klompas M. MDPHnet: secure, distributed sharing of electronic health record data for public health surveillance, evaluation, and planning. Am J Public Health 2014;104:2265-70.

30. Malenfant JM, Hochstadt J, Nolan B, et al. Cross-Network Directory Service: Infrastructure to enable collaborations across distributed research networks. Learn Health Syst 2019;3:e10187.

31. Ahlbrandt J, Brammen D, Majeed RW, et al. Balancing the need for big data and patient data privacy--an IT infrastructure for a decentralized emergency care research database. Stud Health Technol Inform 2014;205:750-4.