The Role of the Statistician in Data Anonymisation

Parveen Kumar, Kalpesh Prajapati, Jeremy Wheeler

GCE Solutions Ltd.

GCE SOLUTIONS

Derive Value from Excellence...

INTRODUCTION

Data transparency through clinical data sharing has the potential to strengthen the integrity of clinical trial systems as well as benefiting academic research and medical practise. This topic is gathering awareness amongst pharmaceutical companies, but it raises concerns around the potential to leak the personal health information of patients participating in clinical trials. The European Medicines Agency (EMA) has published Phase 1 of their Policy 0070 around requirements for anonymization of clinical documents for all studies submitted to EMA for market authorization, signaling a commitment towards data sharing.The industry is looking for efficient ways to anonymise clinical data to reduce the risk of re-identification and still maintain data utility.

This poster will discuss the role of the Statistician in anonymising clinical data, the calculation of risk of re-identification and the quantitative assessment of data utility after anonymisation.

Need for Anonymisation

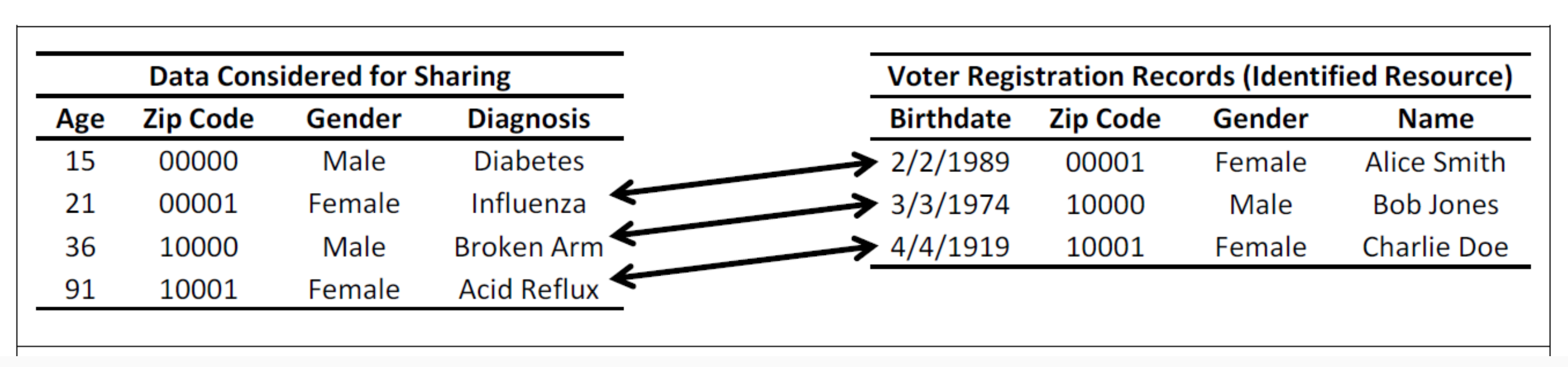
Anonymised/De-identified/Redacted data:

Data in a form that does not identify individuals and where identification through its combination with other data is not likely to take place

Direct Identifiers in a dataset include Subject ID, Investigator Name, etc, and could lead to identification of subjects. This violates Data Protection Acts worldwide to protect personal information of individuals.

Quasi-identifiers are dataset variables that by themselves do not identify a specific individual but can be aggregated and “linked” with other information to identify subjects.

For example:



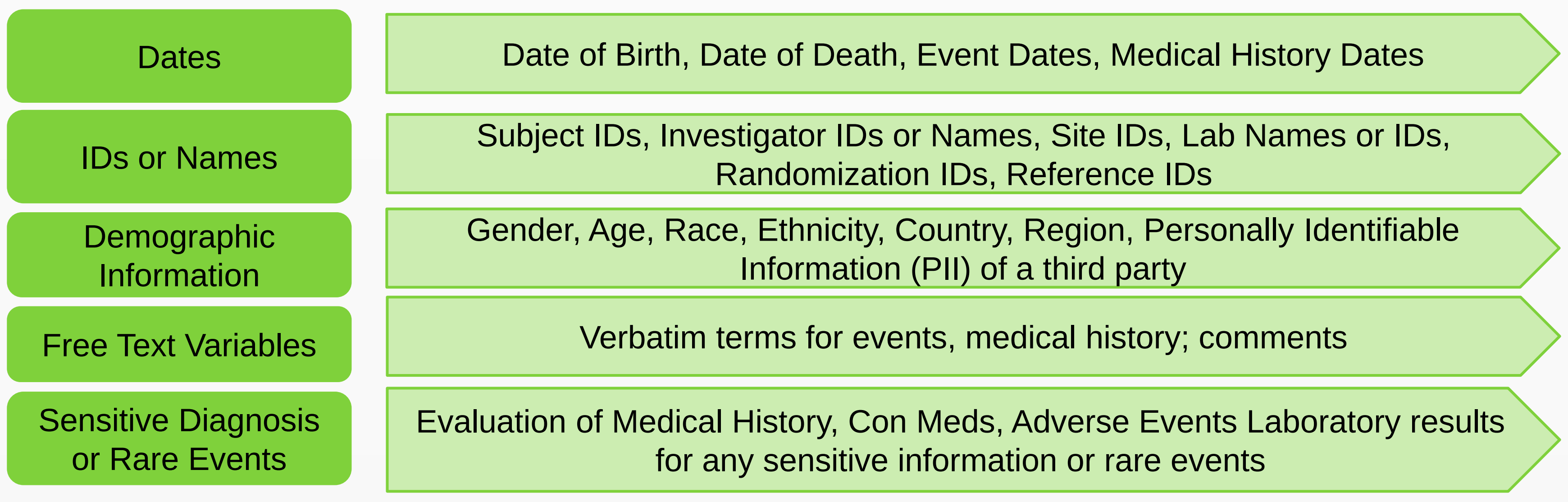
Data Considered for Sharing

Age	Zip Code	Gender	Diagnosis
15	00000	Male	Diabetes
21	00001	Female	Influenza
36	10000	Male	Broken Arm
91	10001	Female	Acid Reflux

Voter Registration Records (Identified Resource)

Birthdate	Zip Code	Gender	Name
2/2/1989	00001	Female	Alice Smith
3/3/1974	10000	Male	Bob Jones
4/4/1919	10001	Female	Charlie Doe

Identifiers in Clinical Trials Data

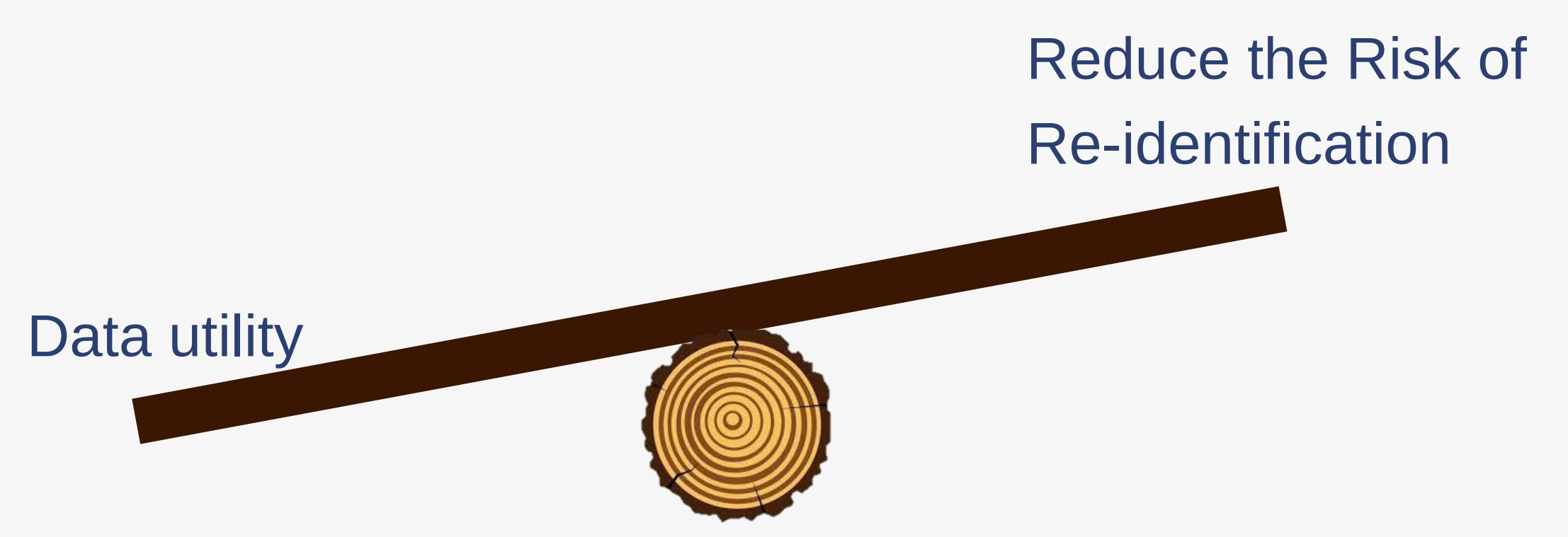


Dates	Date of Birth, Date of Death, Event Dates, Medical History Dates
IDs or Names	Subject IDs, Investigator IDs or Names, Site IDs, Lab Names or IDs, Randomization IDs, Reference IDs
Demographic Information	Gender, Age, Race, Ethnicity, Country, Region, Personally Identifiable Information (PII) of a third party
Free Text Variables	Verbatim terms for events, medical history; comments
Sensitive Diagnosis or Rare Events	Evaluation of Medical History, Con Meds, Adverse Events Laboratory results for any sensitive information or rare events

Statistician’s role?

- Understand the anonymisation methods employed
- Able to assess the probability of re-identification
- Able to assess data utility across alternative anonymisation approaches
- Able to assess the optimum balance between risk and data utility

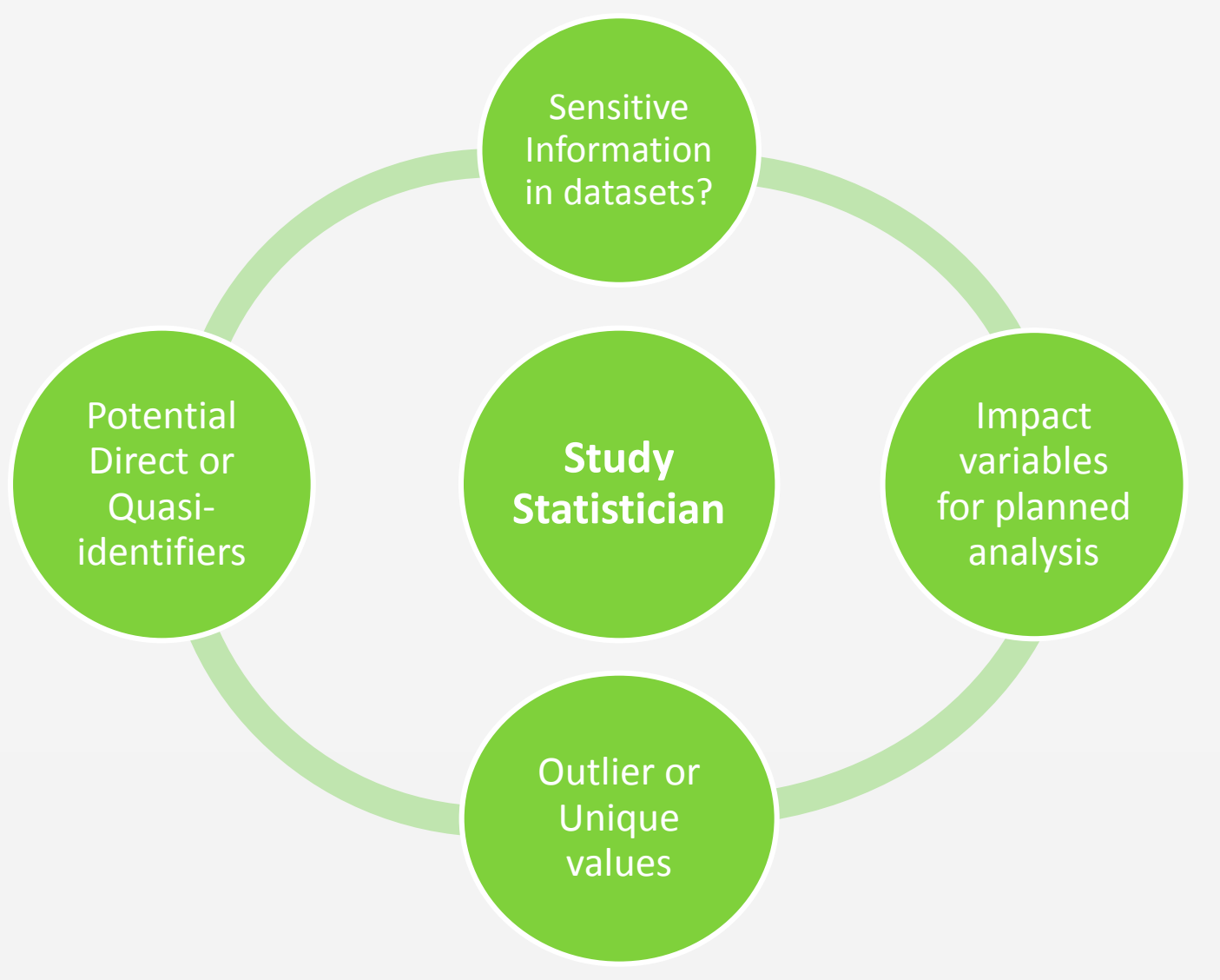
Anonymisation vs Utility



The objective of data sharing is to facilitate meaningful re-analysis, cross industry analysis, evaluation of trends and developments in therapies

Challenges to anonymization include the risk of re-identification by data-mining, merging with external data sources, and collusion of internal participants

Planning Stage


Sensitive Information in datasets?
Potential Direct or Quasi-identifiers
Study Statistician
Impact variables for planned analysis
Outlier or Unique values

- The Study Statistician can identify the Direct or Quasi identifiers collected in the study
- The Study Statistician is familiar with key endpoints and associated analysis, and can gauge the potential impact on these endpoints after de-identification
- The Study Statistician may be aware of any sensitive information collected. This should be either masked or generalized (see ‘Review’ section)
- The Study Statistician is familiar with the dataset and aware of any unique or outlier values.

Execution

There are multiple methods available to de-identify direct and quasi identifiers. Examples are shown below for an age variable, and the Statistician needs to evaluate the method to maintain a balance between risk and data utility:

- Masking/suppression: remove or hide the value!
- Randomization:
  - Noise addition/perturbative masking/offsetting
  - Permutation
- Generalisation/aggregation – replace a value by a range or rounded value

Anonymised Method	Age (years)	Anonymised
Masking	36	.
Add Noise	36	36 + ε ~ N(0,s)
Random Offsetting	36	43
Generalization	36	30-40

Inputs to Anonymisation Methods

- **Dates**: Date Offset (fixed or random noise), Mask with Relative Days
- **Synthetic data generation**: Create data with similar attributes to the source data; statistical methods such as bootstrap, multiple imputation, Cholesky decomposition
- **Add Noise**: For continuous data, it is desirable to maintain the statistical properties of the original data. For example,
  - Uncorrelated noise addition (Means are preserved but neither variances nor correlations are preserved.)

    $z_j = x_j + e_j$ where random variable $e_j \sim N(0, \sigma^2_{e_j})$, such that

    Cov(e_k , e_l)=0 for all k≠l
  - Correlated noise addition (Means and correlations can be preserved by choosing appropriate α)

    As above, but $e_j \sim N(0, \alpha\ \Sigma)$, with Σ being the covariance matrix of the original data.

Review

Risk Analysis

- Risk of re-identification is the probability of an adversary being able to identify a subject in the data.
- Risk is a combination of two probabilities
  1. Probability of attack by an adversary
  2. Probability based on the uniqueness of the combination of quasi-identifiers (attribute group). A simple approach to evaluating this risk is based on the reciprocal of the size of the attribute group
- The Statistician needs to evaluate different factors, to calculate the above probabilities:
  - Extent of sharing: eg, public disclosure, regional release or controlled release to a group of researchers
  - Risk Metric: Overall or Average i.e. alternative methods to combine individual risks
  - Overall disease population in the region where the study is conducted
  - Ability to combine similar studies to calculate risk

Data Utility

Original Dataset vs Anonymized Dataset

To what extent can an analyst replicate specific study analysis on the anonymised data?

The Statistician can provide additional input via modelling and simulation:

- Comparison of alternative anonymisation strategies
- Within or across multiple studies
- Use of generalised loss measures to assess utility
  - Eg, mean squared error, or equivalent measures based on the variance or other statistics
  - Measures based on a propensity score model (logistic model to predict redacted vs original records per subject, aiming for similar propensity)

Summary

- As we move closer to Phase 2 of EMA 0070, Statisticians should be ready for greater involvement in anonymisation methods
- Be ready to evaluate strategies to achieve data privacy
- Be ready to evaluate the data utility of alternative strategies

References

External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. EMA/90915/2016. 7 December 2016

K. El Emam, Kald Abdallah. “De-identifying Clinical Trials Data”. Applied Clinical Trials, Mar 20, 2015

M-J. Woo, J.P. Reiter, A. Oganian and A.F. Karr (2009) “Global measures of data utility for microdata masked for disclosure limitation”, J. of Priv. and Conf., 1(1):111-124

Matthews GJ and Harel O. Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. Statistical Surveys. Vol 5 (2011) 1-29

Karr AF, et al. A Framework for evaluating the utility of data altered to protect confidentiality. The American Statistician (2006) 60;3