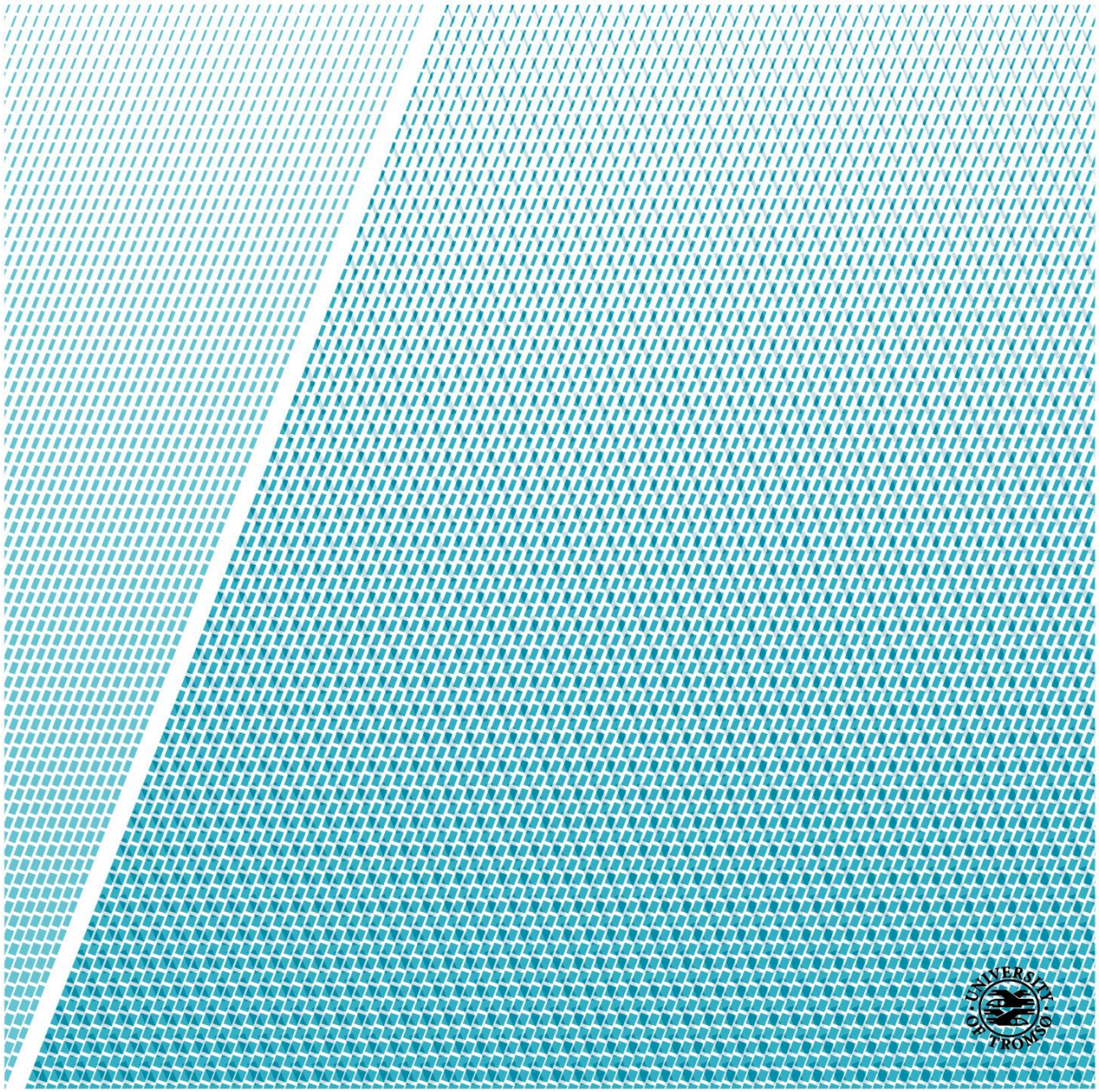
UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Department of Pharmacy

How will anonymization of simulated clinical data affect the data utility of pharmacoepidemiological studies?

Chi Kei Cheang

Master thesis in Pharmacy



Acknowledgement

This master thesis was carried out at the Department of Pharmacy research group of Clinical Pharmacy and Pharmacoepidemiology (IPSUM), University of Tromsø The Arctic University of Norway (UiT)

I would like to express my sincere gratitude to my supervisor Kristian Svendsen for all the technical help for coding and with writing this thesis. This thesis would not have been possible without your help and guidance. I deeply appreciate you for always being available in times when I needed help.

Thanks to my girlfriend, Lise. Thanks for loving and supporting me during this thesis.

I would also like to thank to all my friends and a big thanks goes to Moshen Askar and Aime Manzi. This master period would not have been the same without you.

Finally, I would like to thank my family for their continuous support and motivation during my studies.

May 2019

Chi Kei Cheang

Table of contents

Abbreviation and definition:

Anonymization: “The process of rendering data into a form which does not identify individuals and where identification is not likely to take place” (1)

Clinical data: “Clinical data can be referred as clinical reports and individual personal information collectively”(2)

Clinical report: “A complete document of clinical overviews, clinical summaries and clinical study reports together with the following appendices to the clinical study report” (2)

Clinical Study Report (CSR): “A detailed document about the methods and results of a clinical trial. It is a scientific document addressing safety and efficacy and its content is similar to an academic paper” (3)

Confidence Interval: Cl, “A confidence interval is a range of values calculated by statistical methods which includes the desired true parameter. The confidence level of 95% is usually selected. This means that the confidence interval covers the true value in 95 of 100 studies performed” (4)

Direct identifier: “Direct identifiers are elements that permit direct recognition or communication with the corresponding individuals, e.g. personal names, email addresses, telephone numbers, and national insurance numbers” (2, 5)

European Medicine Agency (EMA): “An European agency that is responsible for scientific evaluation, supervision and safety monitoring of medicines in the EU” (6)

Individual Patient Data (IPD): “The individual data separately recorded for each participant in a clinical study” (2)

Quasi identifier: “Quasi identifiers are variables representing an individual’s background information that can indirectly identify individuals, example: sex, age, race, ethnicity, height and weight.” (2, 5)

Summary

Background: The pressure to share more data and being more transparency of clinical study reports has grown and becomes an important topic in recent years. Before clinical data and clinical results can be shared they must undergo anonymization. How anonymization of clinical data affects the utility is poorly-studied, especially in pharmacoepidemiology.

Objective: The aim of the study is to describe and evaluate how anonymization of simulated clinical data will affect the data utility of pharmacoepidemiological analyses of these data.

Method: We have simulated five clinical datasets with different characteristics, associations, types of outcome and study populations. Suppression, generalization, randomization and k-anonymity were used as our anonymization approaches. These methods will be evaluated by the change in the data and statistical results before and after anonymization.

Result: K-anonymity and suppression were the methods that affected the simulated clinical data the most, while generalization and randomization affected the data least. With k-anonymity and suppression there is a risk to overestimating the clinical results due to the elimination of unique records. On the other hand, generalization and randomization preserved the most data utility but they were less effective in anonymizing the data.

Conclusion: Our study revealed that different anonymization approaches can affect the clinical results differently. The more we anonymize a record or attribute, the less utility is provided. It is therefore important to construct a balance of data utility and effectiveness of anonymization before the clinical data are published. More investigations about how anonymization of clinical data affects data utility are needed in order to maximize the benefit of using anonymized clinical data to improve public health.

1. Introduction

The pressure to share more data and being more transparency of clinical study reports has grown and becomes an important topic for pharmaceutical industry, academic research and public health (7). Greater transparency, especially sharing of clinical study has been therefore taken into account due to the EU health regulation system development and the implementation of new policies. According to European Medicine Agency (EMA)'s annual report of 2017, the number of requests for access to documents has increased significantly in recent years from 416 requests in 2014 to 865 requests in 2017, and the number of requests of pages released following access to document also has increased dramatically from 167 309 pages in 2014 to 487 092 pages in 2017 (8). Additionally, the usage of clinical data on EMA's website has also increased, and it is more than 4 times more usage of clinical data in 2017 (with 126 300 views and downloads) compares to 2016 (with 28 079 views and downloads) (8).

Sharing clinical data is thereby an important element for pharmaceutical industry, academic research and public health. Shared clinical data can be reused to perform other purposes such as meta-analyses, individual patient data meta-analyses, academic researches, pharmacoepidemiological researches, systematic reviews and reanalysis to enhance public health care (9, 10). Furthermore, sharing clinical data can benefit transparency, reliability of data extraction and reuse for new purposes in order to save time and money (11). While, having more data transparency can provide a better understanding in clinical data that can enhance innovation and scientific inquiry related to new drugs, developing a more robust regulatory system and allowing other medicine developers to learn from past successes and failures which can benefit the public health (12-14). The problem with having more transparency and sharing clinical data is clinical study reports (CSR), which contain individual patient data (IPD) that need to be protected before they publish (1).

According to policy 0070 which was established in 2014, all studies from the pharmaceutical industry in Europa including Norwegian companies due to European Economic Area (EEA) Agreement, has to be publicly available and open for everyone right after the publications of clinical reports are submitted to EMA (1). Consequently, the clinical data that are generated during a clinical study will be shared and is open for everyone to reuse for other purposes and analyses. Clinical study reports that are generated during a clinical study will be published, while individual patient data will be removed. Since we know that clinical reports and clinical

study reports contain personal information which are shared can be at risk to be re-identified by a third party, the policy 0070 has particularly emphasized that the protection of personal data and commercially confidential information are important (1). Additionally, protecting personal data is needed and fundamental because it is enshrined in general data protection regulation (GDPR) (15).

The situation today is the policy 0070 is consisted of two phases (2). The first phase is about publishing of clinical reports which has started since 1st Jan 2015. This means all clinical study reports, clinical overviews and clinical summaries will be published on EMA's website and are available to access by anyone. Anyone who creates an account on the website will be able to access all the clinical study reports and information. While the second phase is pertained to publishing of individual patient data which will be implemented in a unknown later stage (2). This means everyone in the future can access the clinical reports and IPD as long as one creates an account on the website. According to this action, the probability for attempting a re-identification will be high since everyone can access these data.

Therefore, the policy aims to ensure that the data is adequately protected and minimizes the potential for unlawful retroactive patient identification that can be conducted by a third party (1). Besides, it is also important to emphasize the objectives of this policy. The policy also aims to "benefit public health, promote better informed use of medicines, develop new knowledge in the interest of public health, secondary data analysis e.g. serious side effect and ensuring the future investment in pharmaceutical research and development"(1).

1.1. Disclosure

Another thing that must be taken into consideration is how a third party discloses a dataset or data. A disclosure from a third party or an adversary will occur normally in three different forms: Identity disclosure, membership disclosure and attribute disclosure (16, 17).

Identity disclosure is a type of disclosures that occurs when an attacker can connect a participant's record in a published dataset. For example, if an attacker knows an individual's name, zip code and gender in a data (table 1). There is a high probability that the individual can be singled out of the dataset. Two tables with dataset are presented below to describe some examples for identity- and attribute disclosure. Table 1 contains a dataset with four patients with their name, zip code, sex, age and their disease condition. Table 2 is an anonymized/redacted version of table 1.

Suppose an attacker knows Bob's zip code and his disease condition from table 1. It is likely that the attacker can re-identify Bob in table 2 which is a partially anonymized/redacted version of table 1, even if the name and zip code is redacted. "This is because **Bob** is the only male in the table who lives in zip code 124xx and has diabetes" (16).

Membership disclosure can occur when an attacker is able to determine an individual's record is whether or not contained in a published dataset (16, 17). Let us assume a dataset that contains information on only breast-cancer patients. An attacker can by finding out that a patient's record is contained in the dataset deduce the fact that the patient has breast-cancer. This can represent a threat to patient's privacy (16).

Attribute or sensitive information disclosure is a threat that occurs when an individual's attribute can be linked with their sensitive information (16, 17). This type of disclosure can mostly occur when an individual is already known. For example, assume an attacker knows Tine's age and zip code but not her disease condition from table 1. The attacker can infer or deduce the rest of information and conclude that Tine must have diabetes in table 2 (18).

Table 1: An example of a dataset of four patients

Name	**Zip code**	**Gender**	**Age**	**Disease**
Bob green	12455	M	56	Diabetes
Mark maxi	12655	M	34	Flu
Tine brown	12344	F	35	Diabetes
Maria blue	12755	F	61	Asthma

Table 2: A partially anonymized/redacted version of table 1

Name	Zip code	Gender	Age	Disease
-	124xx	M	>40	Diabetes
-	126xx	M	<40	Flu
-	123xx	F	<40	Diabetes
-	127xx	F	>40	Asthma

1.2. Anonymization

De-identification of the clinical data is therefore necessary, and different approaches are used to prevent re-identification, securing the personal privacy and reducing the risk of disclosure by a third party or an adversary in order to avoid breaching privacy laws. This de-identification process called anonymization and was used and recommended by the policy 0070 (1).

Anonymization can be simply defined as "a process of masking and de-identifying some personal sensitive data or attributes" (5, 9). Moreover, this process must be processed in an optimal and convenient way that none of these data can single out an individual or link with another identified dataset. *"So more precisely, these data shall not belikely reasonably to be used by a third party or the controller once the anonymization is applied to the data*" (5).

An identifiable data such as personal data or individual's attribute is categorized into two groups, directly or indirectly -identifiers (5, 9). A direct identifier is defined as patient's name, patient number, patient's health record, telephone number, etc. such information can ease identifying of an individual (9, 18, 19). While, an indirect identifier also termed "quasi-identifier" can be zip code, age, race, sex, background information, date of birth, clinic visit, ethnicity, etc. (9, 16, 17, 19, 20). Such information can be linked with other information, and results in a high possibility of identifying an individual (1, 2, 5, 9). Disease code, disease, diagnose, test result etc. are sensitive attributes or information that are important for an analysis and shall whether be removed or redacted (19).

The benefit of clinical data is to give an opportunity to other researchers to reuse these data for other purposes, to test out new analyses and mining new approaches that can provide a new perspective to the original data. It is thereby important to anonymize the personal information to secure the usage of these data without risk of personal information breach.

Therefore, All the personal information in a CSR must be anonymized adequately before it is published (to reduce the risk of re-identification), and that means all the direct identifier in a CSR must be completely deleted before publishing (9, 17). Therefore, the focus will be to anonymize the quasi-identifiers. The problem is that some quasi-identifiers cannot be simply removed or redacted, because these can be critical attribute(s) for analysis (19).

A well-known re-identification experiment was published by A. Narayanan and V. Shmatikov (21). They performed research on a Netflix prize dataset where customers rated on a scale 1-5 on over 18 000 movies by 500 000 subscribers of Netflix (21). More than 100 million ratings

were collected, and publicly released after being redacted. The objective in this study was to investigate the risk of re-identification for this publicly released dataset. The ratings were redacted and anonymized according to internal privacy policy at Netflix, where all customer's personal information was removed except ratings and dates. The researchers revealed that up to 99% of records could be uniquely re-identified in the dataset by using 8 movie ratings and dates that had a 14-day error. And they could re-identify 68% of records by using 2 ratings and dates with a 3-day error (5, 21). The researcher summarized that there is always a chance that some individuals can be re-identified from an anonymized dataset, and any re-identification may cause a potentially harm for study participants because it can disclose any personal information or disease record to the public.

Anonymization can be conducted by different methods. However, randomization, suppression and generalization are the most commonly used methods (22). A record or an attribute can be redacted (removed) if suppression is applied or it can be perturbed by noises if randomization is used, or the record can be aggregated if generalization is used (table 3). Different methods can also be used together to achieve a better strategy. For example, generalization combined with suppression. A dataset's variables will first be aggregated into a group or generalized into a more general one and then unique records that stand alone will be removed.

Free-text, participant's narrative and free-text variables in a clinical study report will also be anonymized in order to minimize the risk of personal data breach.

To achieve the minimum risk of re-identification as much as possible, different anonymization-technique or a combination of different techniques must be used on clinical data. Especially, before the clinical data are published and can be accessed by anyone. Therefore, clinical study reports will normally be anonymized with suppression (redaction), generalization, randomization or/and a combination of these methods due to this concern (22).

A published clinical study report must ensure that it is impossible to single out an individual, has low possibility to link records relating to an individual and has low possibility to deduce an individual. Otherwise an analysis of re-identification risk must be performed to examine the dataset is sufficient anonymized or not (5).

Table 3: An overview of different anonymization techniques

Technique/method	Variable	Result
Redaction/suppression	Age	The record(s) will be removed from the dataset or replaced as missing value
Aggregation/generalization	Weight	The records will be combined to a more general group (e.g. 150kg, 176kg ,165kg →150-180kg)
Noise adding /randomization	Salary	Random noises will be added to an attribute (e.g. 1500kr will be adjusted +/- 500kr to 2000kr or 1000kr)

1.3. Suppression

Suppression is the easiest and most common anonymization method to protect personal information. With suppression information is completely removed from a dataset (23). All the direct identifiers and indirect identifiers are removed by using this method (17). An example is presented in table 4 where postcode and age are removed as a result of suppression, these variables are replaced by missing value or denoted (-) (23). There are two types of suppression that can be used, the first is vertical suppression such as cell suppression where an attribute or a variable is suppressed, and the second is horizontal suppression where a participant/patient is totally removed from the dataset.

The advantage of this technique is to reduce the probability to single out a unique individual or a unique record in a dataset, and to mask important information (5, 24). The problem with suppression is that some interesting findings in an analysis may be completely removed or masked.

When some indirect identifiers are removed, the study population is also reduced which leads to the power of the study may also be reduced. Furthermore, it can induce bias into the dataset if only the weak or strong associated participants are removed.

Table 4: An example of how suppression works on variables; postcode and age

Variable	Value	Suppression
Postcode	9018	-
Age	59	-

1.4. Randomization

Randomization is another of the methods to anonymize data (5). Two techniques are commonly used in randomization: noise adding and permutation. Noise adding is a simple technique where a value of an attribute is modified or adjusted. It leads to a reduction in re-identification's accuracy and the link between data and an individual. A good example (table 5) is weight and age, for the anonymized dataset an individual's weight may be adjusted +/- 3 kilos and the age may be adjusted +/-2, whilst the original dataset is measured the true kilos and age. So, the overall distribution is retained but less accurate.

Another randomization technique is permutation where the values of an attribute are changed or relocated with other values in a table. By re-locating the record, the logical relationship or statistical correlation of two or more attributes is destroyed. Consequently, "the range and the distribution of values is remained the same but the correlations between values and individuals are not*"* (5).

Table 5: An example of how noise addition work on variables (age and weight)

Variable	**Value**	**Randomization**
Age	8,55,35,67	10,53,33,69 (+/-2)
Weight	58,60,78,90	61,57,81,87 (+/-3kg)

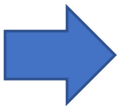
Randomization is a good anonymization method that can mask a record from the original one if an attacker doesn't know the pre-randomized distribution. When noise is added into the dataset, it can form an uncertainty for an attacker.

But the disadvantage is when an attacker knows or finds out the distribution or the permutation, it can result a potential risk of re-identification of any participant in the dataset.

1.5. Generalization

Generalization is an anonymization method that allows some quasi-identifiers to be transformed into a more generalized one. Age can be transformed into age group, date of birth can be transformed into a range of dates (months or years) and five character zip code can be transformed into three or less character zip code, a city can be transformed into a country or continent (9). Figure 1 is an example of generalization with age, zip code and disease condition. A data's precision is reduced by using generalization, and it results a lower risk of re-identification (9, 24).

Zip code	**Age**	**Disease**
12455	56	Hypertension
12655	34	COPD
12344	35	Heart failure
12755	61	Asthma



Zip code	**Age**	**Disease**
12301-12500	>40	CVD
12501-12800	<40	Lung disease
12301-12500	<40	CVD
12501-12800	>40	Lung disease

Figure 1: An example of generalized data of 4 patients. Left: original dataset, Right: generalized dataset. Chronic obstructive pulmonary disease (COPD), Cardio vascular disease (CVD)

A generalization hierarchy is normally used to describe what level a quasi-identifier can be generalized into. It is important to have a good balance between privacy protection and data utility. Therefore, a generalization hierarchy can often show us how far one shall anonymize an attribute. Figure 2 shows an example how age is performed in a generalization hierarchy.

The advantage of generalization is to mask an attribute or a record without removing it. It will be less possible for an individual to be singled out or an individual's record linked with other datasets, since generalization will aggerate the record into groups or intervals. But the problem with generalization is that these aggregated records can still be inferred and caused membership disclosure. How seriously it can harm an individual or a group of individuals will depend on what this attribute is. Another problem with generalization is that some attributes cannot be generalized e.g. sex and categorical outcome, since some analyses will be meaningless if these attributes are generalized.

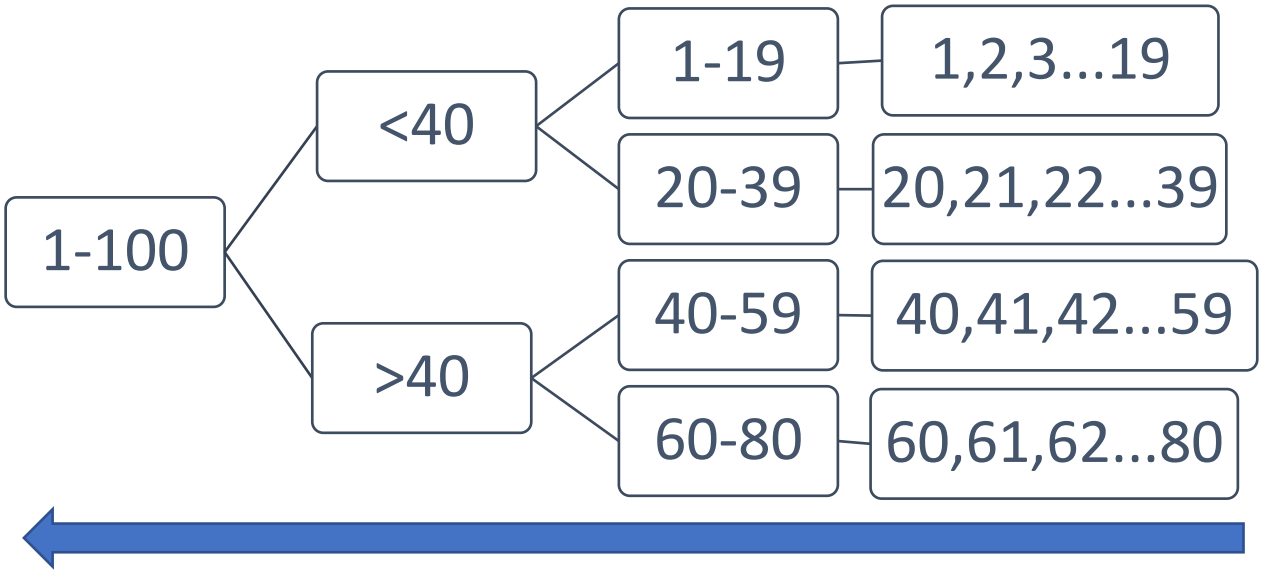


Figure 2 A generalization hierarchy of age, 1-80 was divided into 4 groups and generalized into interval (1-19) (20-39) (40-59) (60-80). (1-19) (20-39) are generalized into (<40) and (40-59) (60-80) are generalized into (>40). At last, (<40) (>40) are generalized into (1-100)

1.6. K-anonymity

A popular privacy protecting strategy is k-anonymity. K-anonymity is a more advanced combination method that transforms an original dataset into a complicated one and creates therefore difficulty for an attacker to disclose the dataset (20). An example (figure 3) where k-value is 4 (k=4) and the identifiable variables are zip code and age. The k-anonymized dataset will have at least 4 records for each value combination of zip code and age (20). This means any of the observations that has less than 4 observations in a row that doesn't contain the same attributes in the dataset would be suppressed (25). "A k-anonymized data set has the property that each record is similar to at least another k-1 other records on the potentially identifying variables" (20). This means "each equivalence class contains at least k records*"* (18).

It is common to anonymize a data with a suitable k-value, too small value of k (ea. like k=2) can lower the weight of any individual in a cluster of attributes (5). It also makes the cluster too significant and leads to a higher success rate for any inference attacks. Vice versa, the higher value of k, the stronger is the privacy secured. An attribute/a record that is anonymized under k-anonymity, will therefore have a maximum probability to be re-identified as 1/k (20, 26).

The advantage with k-anonymity is that no observations will be singled out, since the attribute(s)/record(s) will be anonymized with a k-value between 2 or more. Furthermore, a k-anonymized dataset provides difficulty for an attacker to link the dataset with other publicly available datasets.

The problem with k-anonymity is that a lot of records can be suppressed. This means the records that don't achieve the certain k-value (like the example we have in figure 3) are suppressed. The more records that do not achieve the k-value, the more records will be suppressed. This can cause a bias to the result, loss of interesting information and finding and reducing the study power as well.

Patient Nr.	Zip code	Gender	Age	Disease
1	4827	M	25	Diabetes type 2
2	9010	M	62	Flu
3	4820	M	35	Diabetes type 2
4	9015	F	55	Asthma
5	5007	F	85	Hypertension
6	9011	F	52	Diabetes type 2
7	4825	M	35	Flu
8	4821	F	28	Asthma
9	5003	M	75	Diabetes type 2
10	5034	M	80	Hypertension
11	9012	F	68	Hypertension
12	5058	F	77	Asthma
13	5015	M	15	Flu

Patient Nr.	Zip code	Gender	Age	Disease
1	482*	M	20-40	Diabetes type 2
3	482*	M	20-40	Diabetes type 2
7	482*	M	20-40	Flu
8	482*	F	20-40	Asthma
2	901*	M	41-70	Flu
4	901*	F	41-70	Asthma
6	901*	F	41-70	Diabetes type 2
11	901*	F	41-70	Hypertension
5	50**	F	≥71	Hypertension
9	50**	M	≥71	Diabetes type 2
10	50**	M	≥71	Hypertension
12	50**	F	≥71	Asthma
~~**13**~~	~~5015~~	~~**M**~~	~~15~~	~~Flu~~

Figure 3 An example of k-anonymity. Left: Original dataset, Right: Anonymized dataset where k=4. Gender and disease are not k-anonymized since these are important information

1.7. The situation today

Different pharmaceutical company is using different anonymization method to protect clinical data, and the variation of the methods and the techniques can lead to different types of impact (27). This can cause loss of usefulness clinical information, and some interesting findings might be missed e.g. serious side effects of a drug.

According to The Organization for Economic Co-operation and Development (OECD), loss of data can be considered as loss of data utility since it can cause an adverse effect that affects data's analytical completeness and analytical validity (28). Therefore, policy 0070 has highlighted the balance for the protection personal data whilst retaining scientific value of the data is important, especially when these data are considered to be reused for other purposes or further analyses (1).

The advantages of anonymization are to preserve strong privacy and prevent for any possible attack and re-identification that can cause a personal data breach. But the disadvantages are unsatisfactory for hypothesis generation, loss of data utility and sometimes it can be difficult to interpret some clinical results (16).

Several studies have measured and shown how anonymization can lead to information loss or loss of data utility. A study (9) where they have examined the probability of re-identification by using some simple anonymization methods has found that if the data is greatly anonymized, it results in a lot of information loss. However, if a dataset is not adequately anonymized, an attacker or a third party can easily disclose the dataset. Therefore, one shall always have a good balance between information loss and risk for privacy breach. Another study (20) where they have examined how much information loss k-anonymity can cause has found out that an over-anonymized dataset which is produced by k-anonymity, can result in high information loss and making the data less useful for subsequent analyses. Several studies (17, 19, 27, 29, 30) have used different anonymization method(s) or/and new anonymization algorithm(s) to examine data utility and information loss which anonymization has caused in a dataset. A study (19) that has proposed new anonymization algorithm has found out their anonymization algorithm is better to preserve data utility and provide less information loss than the existing method. Another study (17) has suggested that the existing anonymization algorithms are not optimal to use for biomedical data or other data with respect to information loss and data utility. They have therefore recommended a new anonymization algorithm to solve this problem. While several studies (27, 29, 30) have proposed that new anonymization

algorithms or new approaches to achieve higher data utility and less information loss. It is important to emphasize that none of these studies has examined how anonymization affects clinical data or clinical study report.

So far, there is only one article that pointed out the data utility can be affected if clinical study reports are anonymized as the current policy 0070 (31). The article describes that anonymized clinical study reports are difficult to use and often require expertise to read, understand and process since these documents are often complex and lengthy. The published data are often useless to be reused for other purposes like check for publication bias, check for reporting bias, systematic reviews or metanalysis, novel analysis like re-analysis with a different method or objective to the original analysis or repeat the original analysis. The authors emphasize using the anonymized clinical study report is often a time-consuming process and can often require the support from a statistician. They also emphasize that a lot of researchers need to request the unpublished data from pharmaceutical trials via data sharing platforms to access to individual patient data (IPD). By using individual patient data allows researchers to perform more complex research questions.

The anonymized clinical study report often has limited the endpoint in the published data. Research like investigation of a drug's effectiveness is often hard to perform. In such cases, using unpublished data (IPD) can be favorable.

This article has mostly focused on anonymization that is conducted in free-text, participant's narrative and free-text variables, but they also have emphasized that anonymized clinical study report can be difficult to use for subgroup analysis like treatment response across different subgroup (e.g. different age-group or different type of patients).

We can't find any of studies that have given more details for how anonymization affected the clinical data, except the last study which has given us a few details. But how the potential clinical data has been affected and what consequences it can lead to, for example for quantitative results is still a question. Therefore, in this thesis we will investigate how potential clinical data are affected by different simple anonymization approaches.

2. Aim

The main aim of the thesis is to describe how anonymization of simulated clinical data will affect the data utility of pharmacoepidemiological studies by using various methods of anonymization.

More specifically the study will:

- Find out how anonymization affects different scenarios e.g. different types of outcome? Different study populations? Different frequencies of outcome? Different strength of associations between treatment and outcomes?
- Evaluate the utility of the datasets after anonymization
- Evaluate the effectiveness of the anonymization

3. Method

3.1. Simulation study

Simulation study is new empirical experiment-study. Simulation study is generating a dataset or a study from existing study or patient data that is not real data. This method is used in different areas. For example, simulation study can be used to test bias and confounding that can result in a real study. More examples about other way to use simulation is listed in the table below (table 6).

Table 6: An overview of other aspects that simulation study can be used (32)

- *To check algebra (and code), or to provide reassurance that no large error has been made, where a new statistical method has been derived mathematically* - *To assess the relevance of large-sample theory approximations (e.g. considering the sampling distribution of an estimator) in finite samples.* - *For the evaluation of a new or existing statistical method. Often a new method is checked using simulation to ensure it works in the scenarios for which it was designed* - *For comparative evaluation of two or more statistical methods* - *For calculation of sample size or power when designing a study under certain assumptions*

In this context, we are using the simulation study in a different aspect. We use simulation to create five datasets with known distributions and associations between variables as our pre-anonymized datasets. Then the datasets will be will be anonymized by four different anonymization approaches and these distributions and associations will be compared with the pre-anonymized datasets.

3.2. Study population and design

We simulated our study dataset as a population with female and male adult participants who have chronic symptomatic pulmonary arterial hypertension. Our simulated population was inspired by a randomized controlled trial (RCT) that was published via EMA (33). Because of the published studies have limited information about patient's characteristic, comorbidity and life-style, we had to simulate additional variables not described in the study report to create our dataset.

We think hypothetically the participant in this study will be treated with two different treatments a standard treatment and a new treatment. The standard treatment is anticoagulant

therapy such as Warfarin, while the new treatment is «Riociguat». Age, hypertension, smoking status, heart failure and diabetes are also simulated at the baseline.

We decided to simulate five different cases (datasets) (table 7). In Case 1 to 4, we used a binary variable as our outcome, while in case 5 we chose to test out a continuous outcome. In Case 1, 2, 4 and 5 the study population was 10000 persons, whilst in case 3 the study population was reduced to 1000 persons. A chosen seed (set to seed= 100 in Stata) was used to provide a direct comparison across cases and to make the data reproducible. A seed is "the number with which Stata (the statistic program we used) starts its algorithm to generate the pseudo-random numbers" (34).

Table 7: An overview of case 1-5 with seed set to 100

Case	Description	Study population	Number of exposure	Number of outcome	Unadjusted Odds ratio*
1	Large effect size and frequent outcome	10000	3578	2582	6,26
2	Small effect size and frequent outcome	10000	3578	1710	1,12
3	Small sample size, small effect size and moderate frequent outcome	1000	374	143	1,67
4	Small effect size and rare outcome **	10000	3578	50	1,17
5	Moderate effect size and frequent, continuous outcome ***	10000	3578	Range: -9 – 2 Mean: -3,67 SD: 1,52	-1,55

*median odds ratio from 1000 datasets, **The outcome is a side-effect of the drug, ***The outcome is a decrease of pulmonary artery pressure (systolic blood pressure)

3.3. Simulation as our approach

All the variables that are simulated in this study are shown in table 8 and 9. The first table (table 8) shows the variables that are used in case 1-4, and their classification, value, type and how the continuous variables are distributed.

In case 5 (table 9) most of the variables remained the same, but the outcome was changed to a continuous variable. The outcome in this case is a measurement of the changing of millimeters of mercury in systolic blood pressure(mmhg) (35).

We limited the data set to a few simulated covariables and cofounders, since we wanted to focus on the anonymization part and the investigation of how anonymization affect these variables. To provide direct comparisons across cases, we used the same coding when creating identical variables across datasets except outcome.

The number of each simulated variable can be varied due to different associations were simulated. For example, hypertension in case 1 and case 2. The variable was created with same coding and same association with outcome, but the outcome was simulated differently.

Code for hypertension: gen hypertension = round((runiform()*0.7)+(0.09* exposure+0.02*outcome))

Outcome in case 1: gen outcome = round((runiform()*0.7)+(0.3* exposure))

Outcome in case 2: gen outcome = round((runiform()*0.9)+(0.025* exposure))

More detailed association and simulation for each case attached in appendix 8.1-8.5.

Table 8: An overview of variables which are used in case 1-4 with their classification, value and distribution

Variable	Classification and value	Type (Distribution)
Age	Range: 42-86 years	Continuous (Uniform)
Sex	1= male, 0= female	Essential and binomial
Weight	Range: 36-100kg	Continuous (Normal)
Treatment	1= new, 0= standard	Essential and binomial
Hypertension	1= have, 0= haven't	Binomial
Smoking status	Range: 0-9 cigars	Continuous (Normal)
Diabetes type 2	1= have, 0= haven't	Binomial
Heart failure	1= have, 0= haven't	Binomial
Outcome	1= treatment goal achieved, 0= treatment goal not achieved	Essential and binomial

Table 9: An overview of variables which are used in case 5 with their classification, value and distribution

Variable	Classification and value	Type (Distribution)
Age	Range: 42-86 years	Continuous (Uniform)
Sex	1= male, 0= female	Essential and binomial
Weight	Range: 36-100 kg	Continuous (Normal)
Treatment	1= new, 0= standard	Essential and binomial
Hypertension	1= have, 0= haven't	Binomial
Smoking status	Range: 0-8 cigars	Continuous (Normal)
Diabetes type 2	1= have, 0= haven't	Binomial
Heart failure	1= have, 0= haven't	Binomial
Outcome	Range: -9 – 2 mmhg	Essential* and continuous

* The outcome will also be suppressed in k-anonymity even it is an essential variable

3.4. Anonymization

Four anonymization methods are used in this context; suppression, randomization, generalization and k-anonymity.

3.4.1. Suppression

In this context, cell suppression is used to conduct anonymization in case 1-5 (23). Our suppression approach will be based on: eliminating the group of participants that does not have at least one observation with same attribute, which means these participants that have too many unique attributes that diverge from other participants will be suppressed. The following participant's record will be replaced with a missing value (empty cell). However, we will not suppress the essential variables since they might be useful for other purposes.

3.4.2. Generalization

For case 1-5, the dataset is generalized all the record to a more general one (table 10) (9). So, for the sensitive continuous attribute we will alter it to a range or interval. As for age, this attribute is aggregated to 5-years interval (0-45, 5-years interval to 80, then 81 and above). The same method is used for weight. All the participants are generalized into 5 kg's interval (0-45, 5kgs interval to 80, then 81 and above).

For smoking status, we categorized it into different status; non-smoker, light smoker, moderate smoker and frequent smoker. 1-3 cigars per day as light smoker, 4-6 cigars per day as moderate smoker and ≥ 7 cigars per day as frequent smoker.

For the other binomial attributes like heart failure, diabetes type 2 and hypertension, we will generalize them to a more general one which is a variable called comorbidity that stands for all the diseases. We generalized all participates who has heart failure, diabetes type 2 and hypertension into the variable (comorbidity).

In case 5, the outcome will also be generalized into different 4 categories with different intervals (group 0: 0 and above, group 1: -1 – (-3), group 2: -4 – (-6), group 3: -7 and less).

Table 10: An overview of how generalization is used as our method

Variable	Generalization
Age	0-45 years, 5-years intervals to 80, and 81 years and above
Weight	0-45kg, 5kg intervals to 80, and 81kg and above
Diabetes, heart failure, hypertension	Comorbidities
Smoking status	0: non-smoker 1-3: light smoker 4-6: moderate smoker ≥ 7: frequent smoker
Mmhg	Group 0: 0 mmhg and above (all positive value) Group 1: -1 to -3 mmhg Group 2: -4 to -6 mmhg Group 3: -7 mmhg and less

3.4.3. Randomization

For randomization, noise addition is used to reduce the precision of the sensitive records (5). We adjusted age with +/- 2 years. As for smoking status we will also do the same. We adjusted all the original records +/- 2 cigars for the smoking status and recoding all negative value to 0. In addition. We adjusted the variable "weight" with +/- 4 kg to participants. So, all the record we adjusted is no longer the true record in the dataset. The same methods are used for case 1-5. Additionally, in case 5 the outcome will also be adjusted +/- 2 mmhg.

3.4.4. K-anonymity

The general k-anonymity method will be used in our thesis, that means the records will first be generalized, then the unique records that retained will be suppressed (5). We have selected our k value to be 3 to all records. Therefore, any group of participants that didn't have at least 3 observations with same attributes in the dataset was suppressed.

First, we generalized all the records like we did on generalization (table 10). This means for case 1-5, we transformed age into age-group, weight into group of weight, smoking status into different categories and made a variable (comorbidity) that stands for all the diseases. Furthermore, in case 5 the outcome was also generalized exactly same as we did on generalization, where the outcome was generalized into 4 different categories with different intervals.

After we have generalized the variables, a new variable was created to define unique records. We used the variable to check through all records and suppressed observations that were less than 3 in a group. In order to satisfy k-anonymity as much as possible and the essential variables were also suppressed.

3.5. Statistical analysis

Data analysis was performed with the statistical software program Stata MP 15 for windows. For all cases, 1000 anonymized datasets were created from using a fixed seed from 100 to 1100. Multiple logistic regression was used to assess associations while adjusting for potential confounding factors. After multiple logistic regression, the Odds ratio (OR) was analyzed and used for further evaluations. In order to test the association between the continuous outcome and the categorical exposure while adjusting for potential, multiple linear regression was used. The regression coefficient was analyzed and used for further evaluations.

Multiple logistic regression was used to measure our outcome in case 1-4 that was adjusted for age, sex, smoking status, diabetes type 2, heart failure and hypertension, while multiple linear regression was used to measure our outcome in case 5. The median value of the result from of the 1000 analyses was calculated and used for further assessments. The unadjusted results of 1000 analyses were also calculated.

For all cases with a chosen seed (set seed=100), the adjusted pre-anonymized results were measured (case 1-4 multiple logistic regression and case 5 multiple linear regression) and used to compare with the adjusted anonymized results. This procedure was repeated for each of the different anonymization methods in order to measure the change from the correct result that was further used to evaluate the data utility. The method that we used to measure data utility was similar to one of the methods that was described in this guideline (36).

Multinomial logistic regression was used with nominal categorical dependent variables to estimate the association between outcome and exposure, and to adjust for potential confounding factors. The regression coefficient was analyzed and used for further evaluations. For generalization and k-anonymity in case 5, multinomial logistic regression was used to measure the outcome. The group that was most effective was set as the baseline group (group 3, table 10), and all other groups were compared with group 3. To be able to compare the generalized and the k-anonymized results with pre-anonymized result, we had to divide each group's result with the distance between the selected group and the baseline group. The generalized and k-anonymized result for group 0 was divided by 9 since the distance between group 0 and baseline group was 9. Group 1 was divided by 6, due to the distance between them was 6. Group 2 was divided by 3 since the distance between this group and baseline group was 3. The confidence interval for generalization and k-anonymity was not measured in case 5 since we could not define it after these results were recalculated.

We made an interval of the coverage by sorting the odds ratio (case 1-4) or the regression coefficient (case 5) of 1000 datasets ascendingly. The interval of coverage was set at 95%. The minimum value of the interval was chosen to be 2,5% which was the odds ratio or the regression coefficient of dataset 26 and the maximum value of the interval was chosen to be 97,5% which was the odds ratio or the regression coefficient of dataset 976. In addition, the median was chosen to be 50%.

This interval was made in each case in order to examine how valid each of the anonymization methods were across all cases. If a method after conducting anonymization had a result that was not included within the interval, the method had a low probability of containing the actual association.

The significance level was set at 5%.

A Directed Acyclic Graph (DAG) model (figure 4) “provided an entire graphical, and mathematical, model that can help us to minimize bias in the analysis” (37). By adjusting for confounding covariates, we can estimate the direct effect from exposure to outcome.

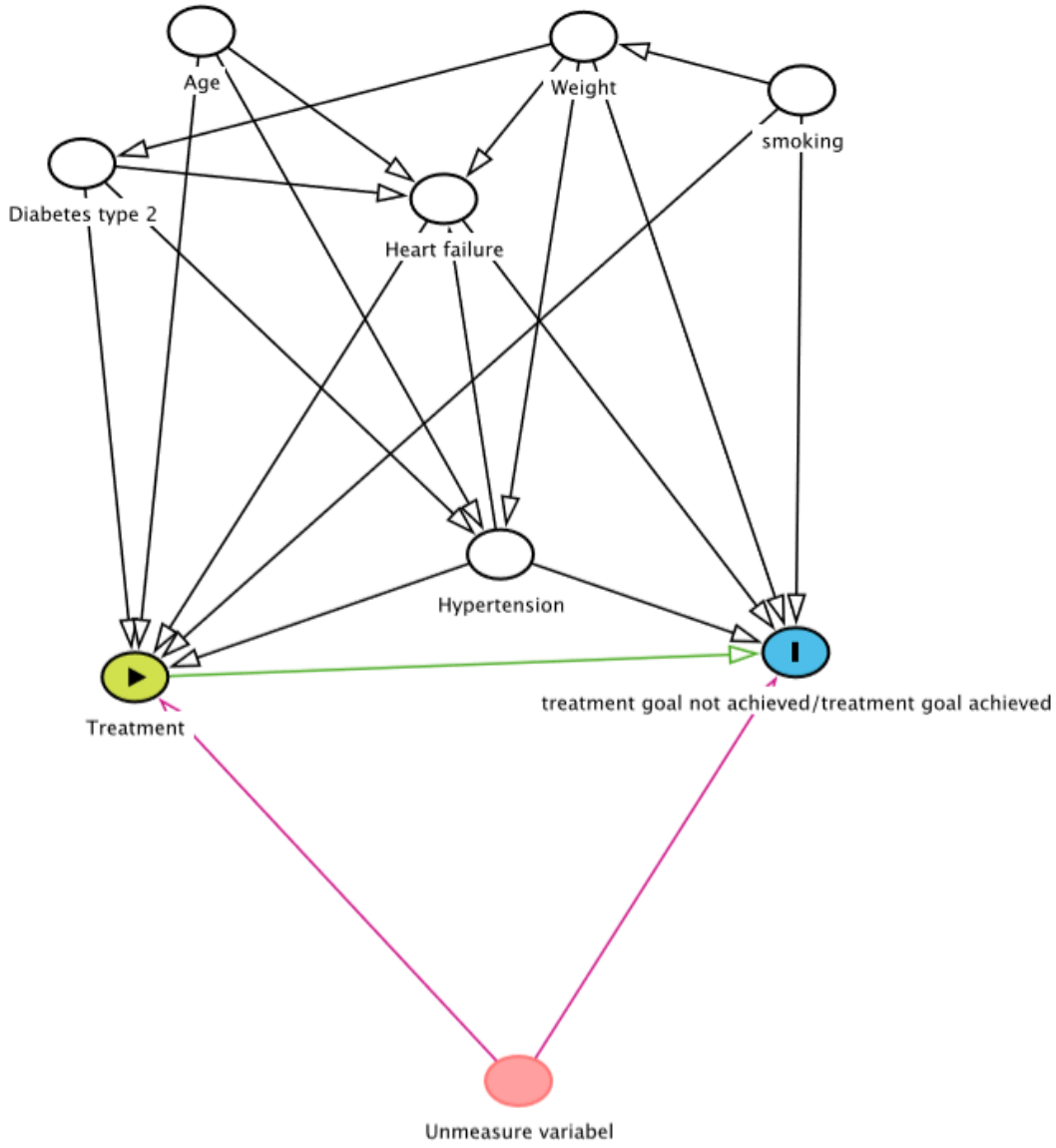


Figure 4 A DAG model for independent, covariate and dependent variable, yellow green node indicate independent variable which is exposure of interest, blue node indicate dependent variable, white nodes indicate measured variables that we need to adjust for which are confounders, and red node indicates unmeasured variable. [1]

[1] http://dagitty.net/development/dags.html?id=TLHAF2

3.6. Evaluation of the data utility after anonymization and the effectiveness of the anonymization

Finally, we used two different schemes to evaluate the performance of each method (table 11 and table 12). One for evaluating the effectiveness of the anonymization, while the other one was for evaluating the utility of data after anonymization. Different questions were used to evaluate these methods.

For the effectiveness of anonymization (table 11), we first evaluated which method can single out an individual's record. A new variable was made in order to identify single unique records. If a record can be single out, we rated the method with a Yes, if no records could be singled out, we rated the method with a No.

At last, we evaluated a record's linkability and possibility to be deduced by attacker. This was done by determining how many attributes of each participant we can identify. If more than three attributes of a participant except the essential attributes can be identified after anonymization, this method had a high chance to link with another dataset or be deduced by an attacker. We rated this method as high chance. If we cannot identify more than three attributes of a participant, we rated this method as low chance.

For the utility of data after anonymization (table 12). We first evaluated whether the anonymized datasets can be used to do other pharmacoepidemiological analyses. This was done by determining if each dataset can be used for analyses like sub-group analysis, group-specific analysis or disease-target analyses. If a dataset was able to be used to perform other analyses after an anonymization method was conducted, we rated the method with a Yes. If a dataset was not able to perform other analyses after an anonymization method was conducted, we rated the method with a No.

Furthermore, we identified any datasets where the result had changed the statistical conclusion. This was done by comparing the confidence interval for each anonymized result with the pre-anonymized result. If an anonymized result had changed the confidence interval from significance to nonsignificant or vice versa, we rated it as Yes, and if an anonymized result had not changed the confidence interval (from significance to nonsignificant or vice versa), we rated it as No.

As for information loss, we counted how many records or participants that was eliminated during anonymization. Only the number of participants that was used in the statistical

analyses was counted. If a method had over 20% of information loss, we rated it as Very much. If the method had less than 20% information loss, we rated it as Slightly, and if the method did not have any information loss, we rated it as None.

We also evaluated how valid the result was after anonymization. This was done by measuring the differential of odds ratio/regression coefficient between the anonymized result and the pre-anonymized result in percentage. If odds ratio/regression coefficient had changed more than 5%, we rated it as less valid. If it changed less than 5% but more than 1%, we rated it as Moderate. If the result changed less than 1%, we rated it as Very much.

Table 11 evaluation scheme for the effectiveness of the anonymization

Effectiveness of anonymization
1. Can an individual's record be singled out? Yes/No
2. Can an individual's record link with other record or deduce by an attacker? High chance /Low chance

Table 12 evaluation scheme for the utility of data after anonymization

Utility of data after anonymization
1. Is the data able to perform other analyses like subgroup analysis or other pharmacoepidemiological analyses after anonymization? Yes /No
2. Does the statistical conclusion changed after anonymization? Yes /No
3. Any loss of information after anonymization? Very much/Slightly/ None
4. How valid is the result after anonymization? Very much/Moderate/Less valid

3.7. Ethic

Simulated dataset and hypothetical participant were used in this study. Therefore, no real patient data or personal preserved data were used. No approval for the information was needed in this thesis.

4. Results

Using 1000 simulations (table 13 and table 14), we found out the odds ratio for case 1 to be 6,16 which was adjusted for sex, age, weight, hypertension heart failure, smoking status and diabetes type, and the unadjusted odds ratio was 6,26. For case 2 where the outcome was low associated, the adjusted odds ratio was 1,10 and the unadjusted odds ratio was 1,12. For case 3 where the sample size is small, the adjusted was 1,65 and the unadjusted odds ratio was 1,67. For case 4 where the outcome was an infrequent event, the adjusted odds ratio was 1,15 and the unadjusted odds ratio was 1,17. For case 5 where the outcome is a continuous variable, the adjusted regression coefficient was estimated -1,54 and the unadjusted regression coefficient was -1,55.

Table 13 result for the pre-anonymized dataset adjusted for sex, age, weight, hypertension, heart failure, smoking status and diabetes type 2

	Case 1	Case 2	Case 3	Case 4	Case 5
Median odds ratio	6,16	1,10	1,65	1,15	-1,54
Standard deviation	0,05	0,04	0,14	0,19	0,02
95% interval of coverage	5,63- 6,78	1,01- 1,20	1,25-2,16	0,76-1,64	-1,50 – (-)1,59

Table 14 unadjusted result for the pre-anonymized dataset

	Case 1	Case 2	Case 3	Case 4	Case 5
Median odds ratio	6,26	1,12	1,67	1,17	-1,55
Standard deviation	0,05	0,04	0,14	0,19	0,03
95% interval of coverage	5,70-6,90	1,03-1,22	1,29-2,17	0,79,-1,65	-1,55 – (-)1,56

4.1. Anonymization

All the anonymization methods have differently affected the result, and all the results we got from multiple logistic regression, multiple linear regression and multinomial logistic regression were adjusted for sex, age, weight, hypertension, heart failure, smoking status and diabetes type 2.

4.1.1. Case 1 - large effect size and frequent outcome

In case 1, the generalization and randomization had the smallest change (table 15). While, suppression and k-anonymity changed the result significantly. After generalization was conducted the Odds ratio was 6,19 with a change around 0,42%. For randomization the odds ratio was 6,16 with a change around -0,01%. K-anonymity and suppression affected the result most, with a change around 41,38% for k-anonymity and 336,96% for suppression. The odds ratio was 8,72 for k-anonymity and 26,93 for suppression. The confident interval of each methods had changed but did not affected the statistical conclusion.

Table 15 Pre-and anonymized result for case 1

Case 1	Odds ratio	% changed	Confidence interval
Pre-anonymized	6,16	-	5,62-6,76
Generalization	6,19	0,42	5,64-6,78
Randomization	6,16	-0,01	5,62-6,76
k-anonymity	8,71	41,38	7,85-9,66
Suppression	26,93	336,96	19,66-36,88

4.1.2. Case 2 - small effect size and frequent outcome

In this case, the association between exposure and outcome was lower. After anonymization was conducted, the result (table 16) also changed but not as much as case 1. We found out generalization and randomization had the same pattern as case 1, where these methods had a lower change on the result. The Odds ratio was around 1,09 for generalization and 1,08 for randomization. These methods changed the result with 0,23% for generalization and 0,01% for randomization. The k-anonymity and suppression didn't change the result as much as in case 1. The odds ratio was 1,10 for k-anonymity and 1,01 for suppression with a change around 1,45% for k-anonymity and -6,41% for suppression. We observed that the confident interval had changed for k-anonymity (1,01-1,20), and the statistical conclusion had changed from nonsignificant to significant.

Table 16 Pre-and anonymized result for case 2

Case 2	**Odds ratio**	**% changed**	**Confidence interval**
Pre-anonymized	1,08	-	0,99-1,18
Generalization	1,09	0,23	0,99-1,18
Randomization	1,08	0,01	0,99-1,18
k-anonymity	1,10	1,45	**1,01-1,20**
Suppression	1,01	-6,41	0,79-1,31

4.1.3. Case 3 - small sample size, small effect size and moderate frequent outcome

In case 3 the sample size was reduced to 1000. The odds ratio was 1,43 after generalization was conducted and the odds ratio was 1,41 after randomization was used (table 17). In addition, generalization and randomization had a slightly change to the result, they changed result with 1,67% for generalization and 0,01% for randomization. While, k-anonymity had a higher change to the result with 9,39% and the odds ratio was 1,54. Suppression was unable to measure due to too many unique records were suppressed. Although the essential variables were retained, the logistic regression was unable to use because too many of the adjusted variables were suppressed. K-anonymity had also changed the confident interval significantly (0,80-2,98), and the statistical conclusion had changed from significant to nonsignificant.

Table 17 Pre-and anonymized result for case 3

Case 3	**Odds ratio**	**% changed**	**Confidence interval**
Pre-anonymized	1,41	-	1,07-1,85
Generalization	1,43	1,67	1,09-1,88
Randomization	1,41	0,01	1,07-1,85
k-anonymity	1,54	9,39	**0,80-2,98**
Suppression	-	-	-

4.1.4. Case 4 - small effect size and rare event

In case 4, we had an infrequent event as our outcome. We found out that it was very hard to optimize the anonymization in this case, since the participant who had infrequent event also had a lot of unique record. Consequently, we were not able to use logistic regression after k-anonymity or suppression was conducted, because many of participant's records were suppressed. Generalization and randomization were the most optimal techniques in this case. The odds ratio was 1,29 for generalization and 1,30 for randomization (table 18). We also found out these methods had a slightly change on the result with -0,60% for generalization and 0,06% for randomization compared to the pre-anonymized dataset.

Table 18 Pre-and anonymized result for case 4

Case 4	Odds ratio	% changed	Confidence interval
Pre-anonymized	1,30	-	0,90-1,89
Generalization	1,29	-0,60	0,89-1,88
Randomization	1,30	0,06	0,90-1,89
k-anonymity	-	-	-
Suppression	-	-	-

4.1.5. Case 5 - moderate effect size and frequent, continuous outcome

We had a continuous outcome in case 5. Since the outcome can be too sensitive we anonymized them too, even it can affect the result. In this case, k-anonymity and generalization had to use multinomial logistic regression due to nominal outcome. While, multiple linear regression was used in randomization and suppression. We found out the regression coefficient was -1,55 for randomization and (-)1,61 for suppression (table 19). These methods had changed the result with -0,01% for randomization and 3,99% for suppression.

As for generalization and k-anonymity, we found out group 0 had a regression coefficient around -2,33 for generalization and -4,17 for k-anonymity. Group 1 had a regression coefficient around -0,77 for generalization and -3,56 for k-anonymity. While group 2 had a regression coefficient around -0,62 for generalization and -6,02 for k-anonymity. In this case k-anonymity had affected the result most with a change around 169,11% for group 0, 130,18% for group 1 and 288,88% for group 2. Generalization had also changed the result significantly but not as much as k-anonymity, it changed the result with 50,54% for group 0, (-)50,35% for group 1 and -60,05% for group 2. We can see that k-anonymity and generalization were not suitable to use for continuous outcome, even these methods were good to anonymize the data.

Table 19 Pre-and anonymized result for case 5

Case 5		Coefficient	% changed	Confidence interval
Pre-anonymized		-1,55	-	-1,50 - (-1,59)
Generalization				
	0	-2,33	50,54	-
	1	-0,77	-50,35	-
	2	-0,62	-60,05	-
	3	Base outcome	-	-
Randomization		-1,55	-0,01	-1,50 - (-1,59)
k-anonymity				
	0	-4,17	169,11	-
	1	-3,56	130,18	-
	2	-6,02	288,88	-
	3	Base outcome	-	-
Suppression		-1,61	3,99	-1,50 - (-)1,72

4.2. The interval of coverage

For case 1, only generalization and randomization were the methods that were included within the interval of coverage, while K-anonymity and suppression deviated from the interval (table 20).

For case 2-4, All the methods were included in the interval of coverage, except suppression in case 3 and 4 and k-anonymity in case 4.

For case 5, only randomization was included in the interval, rest of the methods were deviated from the interval. Both generalization and k-anonymity could not be defined, since different type of analysis was used.

Table 20 The results from case 1-5 and the 95% interval of coverage

	Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**
95% interval of coverage	5,63- 6,78	1,01- 1,20	1,25-2,16	0,76-1,64	-1,50 – (-)1,59
Generalization	**6,19***	**1,09***	**1,43***	**1,29***	**-**
Randomization	**6,16***	**1,08***	**1,41***	**1,30***	**-1,55***
K-anonymity	8,71	**1,10***	**1,54***	**-**	-
Suppression	26,93	**1,01***	-	-	-1,61

*Values that were included within the interval

4.3. Information loss

As for information loss, only suppression and k-anonymity had information loss across all cases (figure 5). In case 1, k-anonymity had a loss around 10,86% and 82,52%for suppression. In case 2, k-anonymity had a loss around 10,94% and 84,66% for suppression. In case 3, k-anonymity and suppression had the highest information loss among all the cases, 66,82% for k-anonymity and 98% for suppression. In case 4, k-anonymity and suppression had the least information loss among all the cases with 6,24% for k-anonymity and 74,38% for suppression. In case 5, k-anonymity had an information loss around 11,94% while suppression had an information loss around 88,81%.

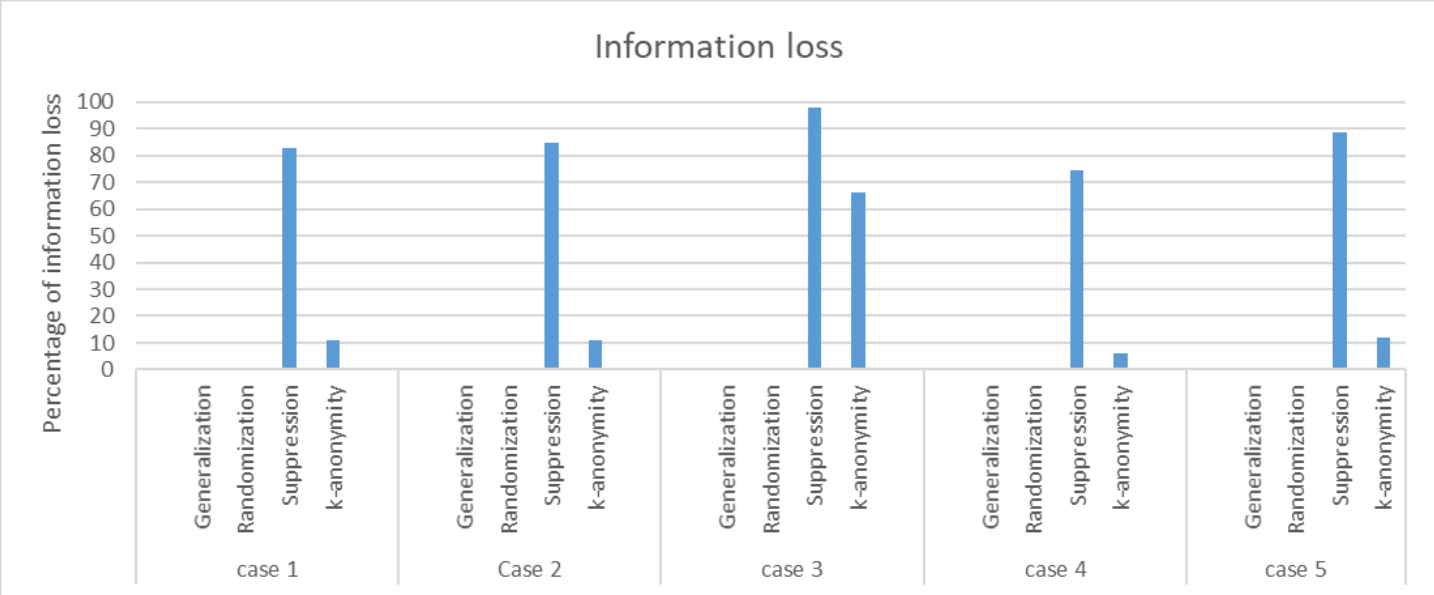


Figure 5 Information loss after anonymization

4.4. Record identified after anonymization

Not all the anonymization methods were perfect to anonymize a dataset. Randomization and generalization were the only methods that had some records that could be identified after they were used (figure 6). In case 1, 5,18% records were identified after generalization and 83,20% records were identified after randomization. In case 2, we observed 5,4% records were identified after generalization and 85,24% records were identified after randomization.

In case 3, the dataset only had 1000 records, but we were able to identify 39% records after generalization and 97,9% records after randomization. Furthermore, we were able to identify 3,64% records after generalization and 74,68% records after randomization in case 4.

In case 5, 6,11% records were identified after generalization was applied and 91,03% records were identified after randomization was applied.

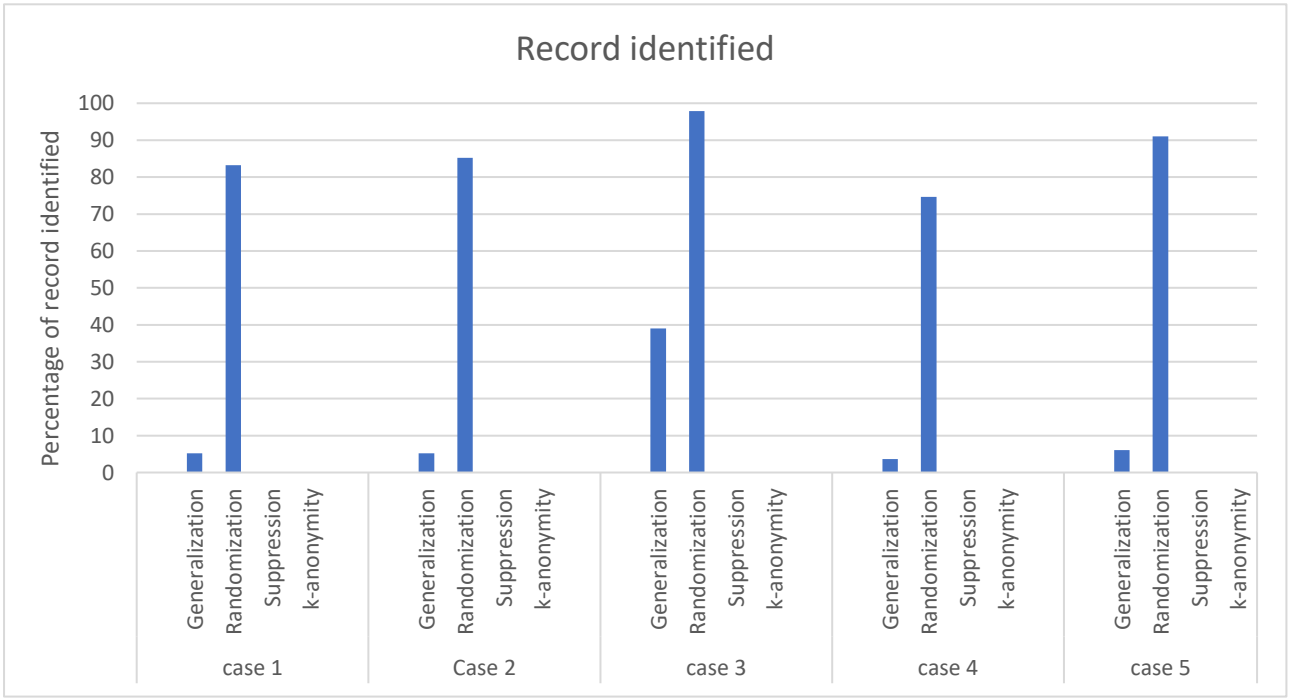


Figure 6 Percentage of record identified after anonymization

4.5. Data utility and Effectiveness of anonymization

According to the evaluation scheme of data utility, randomization was the method that preserved most utility. Generalization was the next method that preserved most utility after randomization (table 21). We also found out suppression was the worst method that can be used to a dataset since it affected a data very much. All methods behaved similarly in all cases for the effectiveness of anonymization, so therefore only one table was made to illustrate the effectiveness of each method (table 22).

K-anonymity was the most effective method to anonymize the data, since this method could not single out any records and had low chance to be deduced by an attacker or linked to other dataset (table 22). However, this method had low utility compared to generalization and randomization (table 21). Generalization, randomization and suppression were the worse methods to anonymize a dataset, since these methods could not fulfill the criteria.

First look at generalization and randomization that performed similarly (table 22). These methods had low chance to be deduced by an attacker or linked to other dataset, but they still can single out individual records. While suppression was opposite, this method could not single out individual records but had high chance to be deduced or linked to another dataset by an attacker. It is important to emphasize that records that we were able to single out in randomization were the randomized records, not original records.

Table 21 Evaluation for the utility of dataset after anonymization

	Is the data able to perform other analyses like subgroup analysis or other pharmacoepidemiological analyses after anonymization?	Does the statistical conclusion change after anonymization?	Any loss of information after anonymization?	How valid is the result after anonymization?
Case 1:				
Generalization	Yes	No	None	Very much
Randomization	Yes	No	None	Very much
K-anonymity	Yes	No	Slightly	Less valid
Suppression	Yes	No	Very much	Less valid
Case 2:				
Generalization	Yes	No	None	Very much
Randomization	Yes	No	None	Very much
K-anonymity	Yes	Yes	Slightly	Moderate
Suppression	Yes	No	Very much	Less valid
Case 3:				
Generalization	Yes	No	None	Moderate
Randomization	Yes	No	None	Very much
K-anonymity	Yes	Yes	Very much	Less valid
Suppression	Yes	Can't define	Very much	Can't define
Case 4:				
Generalization	Yes	No	None	Very much
Randomization	Yes	No	None	Very much
K-anonymity	Yes	Can't define	Slightly	Can't define
Suppression	Yes	Can't define	Very much	Can't define
Case 5:				
Generalization	Yes	Can't define	None	Less valid
Randomization	Yes	No	None	Very much
K-anonymity	Yes	Can't define	Slightly	Less valid
Suppression	Yes	No	Very much	Moderate

Table 22 Evaluation for the effectiveness of dataset after anonymization:

	Can an individual's record be singled out?	Can an individual's record link with other record or deduce by an attacker?
All cases		
Generalization	Yes	Low chance
Randomization	Yes	Low chance
K-anonymity	No	Low chance
Suppression	No	High chance

5. Discussion

Our results showed k-anonymity and suppression had the most impact on the result, while generalization and randomization had the least impact on the result. In case 1 (large effect size and frequent outcome) k-anonymity changed the odds ratio (OR) more than 41%, while suppression changed the odds ratio most with 337%. Furthermore, k-anonymity and suppression were the only methods that were outside the interval of coverage in case 1 and they had an information loss around 11% for k-anonymity and 83% for suppression. The result after k-anonymity and suppression was less valid due to the change of the OR. Generalization and randomization changed the OR least with less than +/- 1%.

In case 2 (small effect size and frequent outcome), only suppression had a significant impact on the OR with approximately -6% change. Generalization and randomization had the least impact on the OR with less than 1% change. K-anonymity had a lower impact on the OR compared to suppression with approximately 1% change, but k-anonymity changed the confidence interval from nonsignificant conclusion to significant conclusion (from 0,99-1,18 to 1,01-1,20). Besides, k-anonymity had an information loss around 11%, while suppression had an information loss around 85%. The OR after suppression was less valid due to the change of the OR, while the OR after k-anonymity was moderate valid.

In case 3 (small sample size, small effect size and moderate frequent outcome), suppression was unable to conduct due to the elimination of the unique records. K-anonymity had the most impact on the result with 9% change and a change in the confidence interval from significant conclusion to nonsignificant conclusion (from 1,07-1,85 to 0,80-2,98). While generalization and randomization had least impact on the result with almost 2% for generalization and 0,01% for randomization. K-anonymity also had an information loss around 67% and suppression had an information loss around 98%.

In case 4 (small effect size and rare outcome), k-anonymity and suppression were unable to conduct due to the elimination of unique records. while generalization and randomization had a small impact on the OR with a change less than +/- 1%. K-anonymity and suppression had the least information loss in this case compared to other cases. K-anonymity had an information loss around 6% and suppression had an information loss around 74%.

In case 5 (moderate effect size and frequent, continuous outcome), randomization had the least impact on the result with a change around -0,01%. Suppression had a lower impact on the result than generalization and k-anonymity with a change around 4%, while generalization

and k-anonymity had most impact on the result due to a different analysis method was used. For generalization, the result had a change around 50% for group 0, -50% for group 1 and -60% for group 2. For k-anonymity, the result had a change around 170% for group 0, 130% for group 1 and 289% for group 2. For the information loss, k-anonymity had a loss around 12%, while suppression had a loss around 89%. The result was less valid after k-anonymity and generalization were conducted, while after suppression was conducted the result was moderate valid.

For the effectiveness of anonymization, k-anonymity was the best method to anonymize the data, while generalization, randomization and suppression were similarly ineffective to anonymize the data.

First looking at k-anonymity and suppression. These methods had provided less data utility due to changes in statistical conclusion, overestimation of result and information loss. This concern is most likely due to the elimination of the unique records. For example, k-anonymity changed the statistical conclusion in case 2 (small effect size and frequent outcome) from nonsignificant to significant conclusion (from 0,99-1,18 to 1,01-1,20), and case 3 (small simple size, small effect size and moderate frequent outcome) from significant to nonsignificant conclusion (from 1,07-1,85 to 0,80-2,98). The k-anonymity OR changed significantly compared to the pre-anonymized OR. For example, in case 1 (large effect size and frequent outcome) we can see an overestimation of the result, since the odds ratio changed nearly 41% (from OR: 6,16 to 8,71). The overestimation could be a result of eliminating cases of non-exposure or eliminating the variables that were combined with non-exposure in order to achieve k-anonymity. The same pattern of overestimation was observed in case 3 with nearly 10% higher than the pre-anonymized OR after k-anonymity was achieved.

A study that examines a new method to preserve the utility better and less information loss in 2014 (38), has found the similar trend as our result. The study has used k-anonymity (k=100), condensation algorithm, two fixed reference points method (TERP) and improved microaggregation (as their algorithm) to anonymize the data. They have examined the change between the anonymized data and the pre-anonymized data in three statistical analyses (linear regression, logistic regression and Cox's proportional hazards model). These anonymization approaches are evaluated by measuring the change in the parameters of these statistical analyses (% change of coefficients) before and after anonymization. They have found the coefficients in linear regression model, logistic regression model and the exponential values

of the coefficients in cox's proportional hazards model have changed most after k-anonymity was used. K-anonymity had highest percentage change of coefficients compared to other anonymization approaches, which was similar to our cases where the k-anonymity also had high changes.

The same overestimation problem could be observed in suppression. This method changed the result and dataset significantly. The sample size decreased dramatically and caused an overestimation like what k-anonymity did in case 1 (large effect size and frequent outcome). Suppression changed the OR with approximately 337% (OR from 6,16 to 26,93). Although, suppression was a good method to anonymize single unique records, the method was not better than generalization and randomization to prevent linkage attack or inference attack (see table 22). According to EMA's clinical data publication report for Oct 2016-Oct 2017(39), suppression was the most used method to anonymize data. The problem with suppression is variables or/and observations that are eliminated can be essential and critical for the data. In other words, eliminating these variables or/and observations can make the data no longer to be used for other purposes or/and analyses. Therefore, other methods like randomization and generalization or k-anonymity are recommended to use to anonymize data rather than suppression (22).

We have expected k-anonymity and suppression would cause an underestimation due to many records were removed, but the result showed these methods had provided overestimation. We did not suppress all essential variables in suppression, but still the results were significantly affected. All the regression analyses require the completeness of the variables to conduct the analysis. Therefore, in suppression the multiple logistic regression analysis and multiple linear regression analysis did not include the entire study population. This is because some of participant's variables were suppressed and not included in the analyses.

A bias might be induced if only the selected records are anonymized and caused an overestimation of benefit. As for pharmacoeconomic or health technology assessment, an overestimation in a cost-effectiveness analysis can result a better incremental cost-effectiveness ratio (ICER) for new treatment than the current treatment which leads to higher chance for the new treatment to be approved (40).

A recent simulation study (41) has shown that post-anonymized data or report can lead to false conclusions or biases in analyses. This is a study where they have used simulated time to event data to examine different methods to improve the accuracy or the validity of the result

and reducing the missing time bias that anonymization has created. They have emphasized that anonymization has impacted the study results and especially for the time to event data. It is therefore important to identify the bias that anonymization can cause and try to adjust them.

On the other hand, generalization and randomization had the similar performance in preserving the utility of clinical data and the effectiveness of anonymization, both had overall higher utility than suppression and k-anonymity. We expected these methods would not affect the utility very much, but we did not expect that generalization had such a small effect on the data, with maximum around 2% changes compared to the pre-anonymized OR across all cases. This is because when a variable was generalized in the dataset, only the variable will be aggregated, but the overall distribution will be retained. Randomization was expected to have a small effect on the data, since all values and distributions were retained, only noise was added on the data to reduce the accuracy.

However, generalization was not good enough to use as a single method to anonymize data, since some records were not adequately secured and could be easily singled out an individual by an attacker. In our study about 5-7% of unique records were able to be identified after generalization was applied. Moreover, when the sample size is lower such as case 3, about 39% of unique records were able to be identified after generalization was applied. It was impossible for generalization to achieve no unique records in our study since not all the variables (essential variables) were generalized.

The same was for randomization too. This method was also not good enough to anonymize the record. Records that were transformed by randomization may still have a high risk to be re-identified by an attacker, even if an attacker does not know the pre-randomized distribution. Assume an attacker knows an individual's information like age, gender, one of the co-morbidities (hypertension, heart failure or diabetes) and outcome. The attacker can use the information to predict the rest of data or link them to another dataset to completely identify this individual. Therefore, it is not recommended to use randomization as a single method to anonymize data.

Among all cases, case 5 (moderate effect and frequent, continuous outcome) was the only scenario where the outcome was also anonymized by different methods. Generalization and k-anonymity were the methods that had the most impact on the data utility after they were used in case 5. The aggregated outcome needed a different type of regression, and therefore multinomial logistic regression was used to measure the result. After the analysis was

conducted, the results were not directly comparable with the pre-anonymized coefficient, since different type analyses were used. The k-anonymity result and generalization result needed to be re-calculated to similar coefficients that were comparable to the pre-anonymized coefficient. Due to this problem, k-anonymity and generalization seemed to be the inappropriate methods to anonymize continuous data especially continuous outcome.

For the utility part. The result showed all methods were still able to conduct pharmacoepidemiological analyses, but not all the methods were suitable for every pharmacoepidemiological analysis. If a result is aggregated, it can be hard to use on a subgroup analysis or target specific analysis like examination of the effect in specific age group like subgroup analysis for elder participant or subgroup analysis for specific disease patient (e.g. cardiovascular disease patient). The problem with aggregated result and aggregated data is more apparently in meta-analyses and systematic reviews. To perform a meta-analysis or systematic review, all the results and data from the studies that are included must be comparable, corresponding or correlated to each other (42). However, using aggregated data to conduct these analyses might require highly skilled researcher (31). Advantages of using aggregated data are less time consuming and cheap to perform a meta-analyze compares to use individual patient data (IPD), which are the unpublished data (43). Meta-analysis of IPD might be time consuming and expensive, but it allows researchers to answer more complex and detailed research questions, additionally to achieve a more valid estimation (31). The result in case 5 showed the pre-anonymized outcome provided more utility compared to aggregated outcome respectively, to conduct other analysis like meta-analysis. In addition to this concern, aggregated data and results can be difficult to perform a re-analysis of a study to verify the result or the conclusion of the study (31).

An interval of coverage can indicate which estimate contained the actual result. In our results, most of the methods were included in the interval of coverage across the cases, except suppression in case 3 and case 4, k-anonymity in case 4 and case 5, and generalization in case 5.

The interval of coverage for k-anonymization and generalization in case 5 could not be measured, since a different type of regression analysis was used to measure the outcome. But for suppression in case 3 and case 4 and k-anonymity in case 4, too many records were suppressed and therefore the interval of coverage could not be defined in these methods. Across all cases, the validity or accuracy of the anonymized result seemed to be highest or

had highest probability to be included within the interval when a dataset had moderate effect size, big sample size and frequent outcome to be anonymized without being suppressed.

In general, most of the utility was preserved when we did not eliminate or suppress any records. This might be an important thing to consider when one is considered to use k-anonymity or suppression to anonymize personal information or important patient record. A good example was case 4 where the outcome was an infrequent event, it was impossible to conduct any pharmacoepidemiological analysis when most of the participants were suppressed. Furthermore, the result could not be measured after suppression or k-anonymity were used.

On the other hand, for the effectiveness of anonymization, k-anonymity was the most optimal method to anonymize the data compared to generalization, randomization and suppression. Suppression could not avoid high chance of inference attack or link attack since only the unique records are suppressed. Assume an attacker has all information to a participant in our study, after suppression is applied the attacker still have a 50% chance to identify the participant, since two of the participants can have similar information. This is considered to have a high probability to successfully re-identify a participant, but if the group of two participants that have similar information also be suppressed, the dataset may retain less than 5-10% study population to use for other purposes or analyses that can consider as low data utility. Furthermore, due to the decrease in study population the result might also be less reliable and valid.

Our methodology was based on simple simulation and coding that provided an insight for how anonymization of simulated clinical data affects the analysis result and a better understanding of anonymization in pharmacoepidemiology. To provide direct comparisons across cases, we used the same seed when creating all datasets and same coding when creating identical variables across datasets. For the anonymization part, some anonymization methods had limitations to anonymize the data and not all anonymization methods were suitable for every type of variables. For example, randomization was more suitable to use on numeric and continuous variable than categorized variable (44). While, generalization, suppression and k-anonymity were suitable for most type of variables.

Despite, there is always a tradeoff between utility of data and effectiveness of anonymization that one needs to consider, no matter how the clinical data will be anonymized. More anonymization will preserve less data utility and vice versa. The biggest question due to this

concern is what shall pharmaceutical companies do? According to the current situation, different pharmaceutical companies have different policies to anonymize the data, and the data transparency is therefore very vary (45). A better and global standard for how to anonymize clinical data is needed in the future in order to achieve better data utility and more transparency.

An important thing that should be into consideration is open data access. The current problem is anyone can access the CSR that are published by EMA. The more people get access, the higher is the number to attempt a re-identification. In other words, no matter the available data is more or less anonymized, the probability for a third party to perform a re-identification is high. Besides, if any pharmaceutical company breaks the general data protection regulation (GDPR), they can be fined up to 20 million euros due to the penalty of personal data breach (46). A stricter anonymization as the current situation is therefore used to ensure no personal data breach, which has also provided low data utility. To achieve better data utility due to this problem, a better data access security system must be implemented to regulate the individual who accesses these data. This might lead to reduced strictness of anonymization and facilitate a better balance between data access and data utility.

Another important thing to be taken into consideration is publishing of individual patient data (IPD) in the future due to phase 2 of policy 0070. IPD might be a better resource to use for studying other purposes compares to clinical study report (CSR), since they can provide more useful and reliable information. These data might benefit in many aspects. For example, pharmacoepidemiological studies like meta-analysis and systematic review since IPD might be easier to compare or combine data from different studies. Besides, IPD can provide a better understanding and interpretation of a study's result and conclusion. On the other hand, individual patient data are more sensitive than CSR and contain patient's information and important commercially confidential information that can be abused by a third party. Due to this problem, pharmaceutical companies might create an anonymization procedure as strict as possible in order to protect the personal data, which also can provide low data utility.

According to the current policy (1), this process is under construction and various aspects need to be reviewed and clarified. The policy has stated that their first target is "to undertake public consultation with all concerned stakeholders on the various aspects in relation to IPD to provide a clarification"(1). What might also need to be clarified is how to share these data. A new guideline and policy for how to anonymize these data and who can access these data is needed with respect to the patient privacy and commercial confidence. Besides, A new

guideline for how to use this type of data is also needed in order to provide a better understanding of the disease(s) or/and the treatment(s).

In addition to data sharing of clinical reports and individual patient data, transparency in clinical report and clinical data might also be an important object to taken into account in order to benefit public health and pharmaceutical industry. We think a new policy and standard that pertain to transparency is also needed in future in order to maximize the utility of using clinical data. A better transparency might benefit pharmaceutical industry by making the regulatory process clearer and predictable (14). Furthermore, this article (14) has emphasized that "transparency might also benefit the public health by allowing medicine developers to learn from past successes and failures or enable the wider scientific community to make use of detailed clinical data to develop new knowledge. It might also allow other researchers to verify original analysis and conclusions, to conduct further analyses, and to examine the positions of the regulator and challenge them where appropriate".

5.1. The strength and limitation of the study

The strength of this study was we have created a simulation study which allowed us to know the correct result of our analyses and how these gold-standard results change when different anonymization techniques are used. The flexibility of simulation study allowed us to examine different type of results, adding or changing any variable and its distribution.

Another strength we had in this study was that by using simulated data, we had no ethical challenges in this study. That is because simulated clinical data was used, not real clinical data. Otherwise it can be time consuming to get approve for the usage of clinical data and solving other ethical problems. Additionally, it was hard to find a good dataset to perform our study.

We have stored the code that we used for the creating the simulated datasets and for the anonymization methods that can be re-used or tested for other analyses. This is an advantage that allowing other researchers to re-use them for other purposes. Furthermore, our study is reliable and transparent since each case can be reproduced.

As far as we know, there are almost no studies that have examined similar or same objective as ours. Thereby, our study is the one of the first studies that investigated how anonymization of simulated clinical data affects the analysis and the data utility. Our study can benefit other future investigations that examine the data utility after anonymization.

There are a few limitations in our study. First at all, only simple anonymization methods were used, and no advanced methods like l-diversity and t-closeness were used in this study. The reason for why only the simple methods were used is because we want to imitate how clinical study report is normally anonymized. Clinical study reports are normally anonymized with redaction (suppression), generalization, randomization or/and a combination of these methods.

Secondly, we did not use a real clinical dataset. A real clinical dataset will normally have more variables and probably more or less participant than our datasets. In addition, a real clinical dataset has often more essential variables that may have vary correlations to each other. Therefore, anonymizing these variables might lead to loss of data utility and loss of validity.

Thirdly, we were not able to access all the information from the clinical study report. The current clinical study reports have limited useful information to be used for our study (33). Therefore, hypothetical and fixed variables and values are created to perform our study which can reduce the reliability of the study.

It is important to emphasize that our evaluations of effectiveness and utility were based on our subjective judgement. This can vary from person to person. Since there are no standard guidelines to evaluate the effectiveness of anonymization and the data utility, we had to use subjective judgement to evaluate the cases. The reliability of the results is therefore reduced. More studies in this concern with different evaluations is needed to increase the reliability of our results.

There are a few things we could do differently in this study, for example testing out different k-value to see which k preserves most utility. However, we do not assume a higher k will give us better utility, and neither could we see a lower k will be more effective to anonymize the records. Therefore, A k= 3 is often recommended in studies, but in practice k=5 is often used instead (47). However, A k value between 2 and 15 is needed to ensure the data is secured (47). A k=5 should be tested to see how much it differ from a k=2. What we could also do were instead of using the standard k-anonymity where the record is first generalized and then suppressed, we could first add noises to the records then generalize them. This approach may have high probability to preserve more utility due to no records will be eliminated. Since the records are randomized and generalized the risk of disclosure may also be minimized.

Furthermore, we could generalize bigger the intervals or the groups, so more record might be included in the interval or group. An advantage with a bigger generalized group is the risk of re-identification is reduced since the variable will be less specific (9). But at the same time, the utility will also be reduced since the aggregated group or interval is bigger and can be more difficult to conduct other pharmacoepidemiological analyses like group specific analysis or target specific analysis.

Besides, we could have some other professionals to test our dataset for linkage attack or inference attack to increase the validity or/and reliability for our evaluations. We could give 10-participant's information to 10 random persons and let them try to identify those participants in the anonymized dataset. Thereafter, we measure and evaluate how many participants they could identify to see how effective our anonymization methods were.

The same thing could be done to examine the re-usability. We could gather 10 random persons and let them check the dataset are still usable for other pharmacoepidemiological analyses after different anonymization methods are used. We give them the anonymized datasets and ask them to perform some specific analyses, then we measure and evaluate how many of the analyses they were able to conduct and what kind of analyses they were capable to perform.

Imputation of random value or pre-defined value such as mean or median to replace the record we suppressed could be used in this study. Imputing random or pre-defined value may retain more utility of the attribute than just suppressing them and leaves the cell empty. The most important thing to consider is not imputing all the suppressed record perfectly. Since a perfect imputation will just result high chance to recover the pre-suppressed value(s)/record(s). A study that examines different strategies to anonymize taxonomic data (48) has found that multiple imputation can preserve a high level of data utility and minimize the level of disclosure risk. This study has used different anonymization methods like data shuffling, microaggregation and multiple imputation to investigate which of them gives most utility.

We could also make a re-identification scale or perform a re-identification analysis to evaluate the level of re-identification risk for each variable after anonymization. A study has performed an evaluation for the risk of re-identification of patients from hospital prescription records (49). The objective of this study was to evaluate the ability to re-identify patients from prescription records. They have measured and quantified each anonymized variable with a

level of re-identification risk. However, the study has only shown the evaluation of re-identification risk for generalized or/and suppressed variable.

Other utility-based measurements could be used in our simulation study. There are a lot of approaches for different type of variables that can be used to measure the utility after anonymization (36). For example, we could measure the mean or median of each anonymized numeric variable and calculate how much they derivate from the pre-anonymized variables. Different simulation approaches could also be used in the study like creating more variables or covariates, to have bigger sample size, stronger effect size or more frequent outcome.

5.2. Future investigation

I would suggest for the future investigation to examine the data utility of using advanced anonymization methods like t-closeness and l-diversity or advanced anonymization algorithms combining more than one method. It would be interesting to see how much utility the advanced methods can preserve and are they still usable for pharmacoepidemiological analyses after conducting compare to our finding. Another thing that might be interesting to investigate is the utility of real clinical data after anonymization. It can be interesting to investigate the differences between real clinical data and our simulated data.

What might also be interesting to investigate is the data utility of time to event data after different anonymization methods were conducted. One can create an immediate outcome like immediately treated and a final outcome like death. Then anonymizing these data and measuring the utility after anonymization and the effectiveness of anonymization.

Overall, more investigations about how anonymized clinical data (CSR and IPD) affects data utility are needed to maximize the benefit of data sharing and data utility while minimizing the risk of identification (50).

6. Conclusion

When we simulated datasets using different anonymization methods they affected the analysis results differently. K-anonymity and suppression were the methods that affected the analysis results most, while randomization and generalization were the methods that affected the result least. There is always a tradeoff between data utility and effectiveness of anonymization. Better anonymization will preserve less data utility and reverse.

For the data utility after anonymization, randomization and generalization were the methods that preserved most utility. While for the effectiveness of anonymization, k-anonymization was the most effective method to anonymize data.

Therefore, it is important to construct a good balance before the clinical data are published. More investigations about how anonymized clinical data (CSR and IPD) affects data utility are needed for conducting other pharmacoepidemiological analyses, and to maximize the benefit of using anonymized clinical data to improve public health.

7. References

1. European Medicines Agency. Policy 0070; European Medicines Agency policy on publication of clinical data for medicinal products for human use. Amsterdam: European Medicines Agency; 2014.
2. European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical date a for medicinal products for human. London: European Medicines Agency; 2017.
3. European Medicines Agency. Clinical data available [Internet] Amsterdam: European Medicines Agency; 2016 [cited 2019 Apr 30]. Available from: https://clinicaldata.ema.europa.eu/web/cdp/background#1.
4. du Prel J, Hommel G, Rohrig B, Blettner M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Deutsches Arzteblatt international. 2009;106(19):335-9.
5. Article 29 Data Protection Working Party. Opinion 05/2014 on anonmymisation techniques. Brussels: Directorate-General for Justice and Consumers; 2014.
6. European Medicines Agency. Who we are [Internet] Amsterdam: European Medicines Agency; 2017 [cited 2019 Jan 12]. Available from: https://www.ema.europa.eu/en/about-us/who-we-are.
7. Manamley N, Mallett S, Sydes M, Hollis S, Scrimgeour A, Burger H, et al. Data sharing and the evolving role of statisticians. BMC Medical Research Methodology. 2016;16 (Suppl 1):75.
8. European Medicines Agency. Annual Report 2017 [Internet] London: European Medicines Agency; 2018 [updated 2018 May 02; cited 2019 Apr 18]. Available from: https://www.ema.europa.eu/en/documents/annual-report/2017-annual-report-european-medicines-agency_en.pdf.
9. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. British Medical Journal. 2015;350:h1139.
10. Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk Washington (DC): The National Academies Press; 2015.
11. Shokraneh F, Adams CE, Clarke M, Amato L, Bastian H, Beller E, et al. Why Cochrane should prioritise sharing data. British Medical Journal. 2018;362:k3229.
12. Grigg SE, O'Sullivan JW, Goldacre B, Heneghan C. Transparency of the UK medicines regulator: auditing freedom of information requests and reasons for refusal. British Medical Journal Evidence-Based Medicine 2019;24:20-5.
13. US Food and Drug Administration. FDA Commissioner Scott Gottlieb, M.D., on new steps FDA is taking to enhance transparency of clinical trial information to support innovation and scientific inquiry related to new drugs [Internet]: U.S. Food and Drug Administration; 2018 [updated 2018 Jan 16; cited 2019 Apr 26]. Available from: https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm592566.htm.
14. Papathanasiou P, Brassart L, Blake P, Hart A, Whitbread L, Pembrey R, et al. Transparency in drug regulation: public assessment reports in Europe and Australia. Drug Discovery Today. 2016;21(11):1806-13.
15. EUGDPR.org. The EU General Data Protection Regulation [Internet]: EUGDPR.org; 2018 [updated 2018 May 24; cited 2019 Apr 20]. Available from: https://eugdpr.org/.
16. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of Biomedical Informatics. 2014;50:4-16.
17. Kohlmayer F, Prasser F, Kuhn KA. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. Journal of Biomedical Informatics. 2015;58:37-48.

18. Li N, Li T, Venkatasubramanian S, editors. t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. 2007 IEEE 23rd International Conference on Data Engineering; 2007 15-20 April 2007 Istanbul, Turkey: IEEE.
19. Lee H, Kim S, Kim JW, Chung YD. Utility-preserving anonymization for health data publishing. BMC Medical Informatics and Decision Making. 2017;17(1):104.
20. El Emam K, Dankar FK. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association. 2008;15(5):627-37.
21. Narayanan A, Shmatikov V, editors. Robust de-anonymization of large sparse datasets. 2008 IEEE Symposium on Security and Privacy (sp 2008); 2008 18-22 May 2008; Oakland, CA, USA: IEEE
22. Dias M. Guidance on the anonymisation of clinical reports for the purpose of publication. Guidance on the anonymisation of clinical reports for the purpose of publication in accordance with policy 0070; 6 July 2015; London. London, United Kingdom: European Medicines Agency; 2015. p. 12.
23. Ohno-Machado L, Vinterbo SA, Dreiseitl S. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance. Proceedings AMIA Symposium. 2001:503-7.
24. El Emam K, Moher E. Privacy and Anonymity Challenges When Collecting Data for Public Health Purposes. The Journal of Law, Medicine & Ethics. 2013;41(1_suppl):37-41.
25. Anspal S, Kaska M, Seppo I. Using k-anonymization for registry data: pitfalls and alternatives. Acta Comment Univ Ta. 2017;21(1):65-78.
26. Kim S, Chung YD. An anonymization protocol for continuous and dynamic privacy-preserving data collection. Future Generation Computer Systems. 2019;93:1065-73.
27. Kohlmayer F, Prasser F, Eckert C, Kuhn KA. A flexible approach to distributed data anonymization. Journal of Biomedical Informatic. 2014;50:62-76.
28. Organisation for Economic Co-operation and Development. DATA UTILITY [Internet]: OECD; 2005 [updated 2005 November 9; cited 2018 Oct 12]. Definition of data utility]. Available from: https://stats.oecd.org/glossary/detail.asp?ID=6905.
29. Goldberger J, Tassa T, editors. Efficient anonymizations with enhanced utility. 2009 IEEE International Conference on Data Mining Workshops; 2009 6-6 Dec. 2009; Miami, FL, USA: IEEE.
30. Loukides G, Gkoulalas-Divanis A. Utility-preserving transaction data anonymization with low information loss. Expert Systems with Applications. 2012;39(10):9764-77.
31. Ferran J-M, Nevitt S. EMA Policy 0070: Data Utility in Anonymised Clinical Study Reports (CSRs). Data Utility in Anonymised Clinical Study Reports (CSRs); London: European Medicines Agency; 2017. p. 23.
32. Morris T, R White I, Crowther M. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2017;38(11):2074-102.
33. Bayer Health Care. Clinical Study Report No. A62510, BAY 63-2521 [Internet] United Kingdom: European Medicines Agency; 2012 [cited 2018 Nov 05].
34. UCLA: Statistical Consulting Group. How can I draw a random sample of my data? | Stata FAQ [Internet]: UCLA: Statistical Consulting Group; [cited 2019 Apr 29]. Available from: https://stats.idre.ucla.edu/stata/faq/how-can-i-draw-a-random-sample-of-my-data/.
35. Institute for Quality and Efficiency in Health Care. InformedHealth.org [Internet] Germany: Institute for Quality and Efficiency in Health Care; 2006 [updated Sep 25, 2016; cited 2019 Apr 05]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279251/.
36. Benschop T, Machingauta C, Welch M. Statistical Disclosure Control: A Practice Guide [Internet]: The World Bank; 2018 [cited 2019 Apr 20]. Available from: https://sdcpractice.readthedocs.io/en/latest/#.
37. Textor J, Hardt J, Knüppel S. DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology. 2011;22(5):745.
38. Gal TS, Tucker TC, Gangopadhyay A, Chen Z. A data recipient centered de-identification method to retain statistical attributes. Journal of Biomedical Informatics. 2014;50: 32-45.
39. European Medicines Agency. Clinical data publication (Policy 0070) report Oct 2016-Oct 2017 [Internet] London: European Medicines Agency; 2018 [updated 2018 Jul 16; cited 2019 Feb 12].

Available from: https://www.ema.europa.eu/en/documents/report/clinical-data-publication-policy-0070-report-oct-2016-oct-2017_en.pdf.
40. Muennig P, Bounthavong M. The Average and Incremental Cost-Effectiveness Ratio. Cost-Effectiveness Analysis in Health : A Practical Approach. 3rd ed. New York: John Wiley & Sons, Incorporated; 2016. p. 9-10.
41. Caetano S-J, Dawe D, Ellis P, Earle CC, Pond GR. Methods to improve the estimation of time-to-event outcomes when data is de-identified. Statistics in Medicine. 2019;38(4):625-35.
42. The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [Internet]: Cochrane Handbook for Systematic Reviews of Interventions; 2011 [updated 2011 Mar cited 2019 Mar 18]. Available from: http://handbook-5-1.cochrane.org/.
43. Tudur Smith C, Marcucci M, Nolan SJ, Iorio A, Sudell M, Riley R, et al. Individual participant data meta-analyses compared with meta-analyses based on aggregate data. Cochrane Database of Systematic Reviews. 2016(9).
44. Templ M, Meindl B, Kowarik A, Chen S. Introduction to Statistical Disclosure Control (SDC) [Internet]2018 [cited 2019 Apr 21]:[1-31 pp.]. Available from: https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf.
45. Goldacre B, Lane S, Mahtani KR, Heneghan C, Onakpoya I, Bushfield I, et al. Pharmaceutical companies' policies on access to trial data, results, and methods: audit study. British Medical Journal. 2017;358:j3334.
46. Article 29 Data Protection Working Party. Guidelines on the application and setting of administrative fines for the purposes of the Regulation 2016/679. Belgium: Directorate-General for Justice and Consumers; 2017.
47. El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E. A globally optimal k-anonymity method for the de-identification of health data. Journal of the American Medical Informatics Association. 2009;16(5):670-82.
48. Domingo-Ferrer J, Muralidhar K, Rufian-Torrell G, editors. Anonymization Methods for Taxonomic Microdata. Privacy in Statistical Databases; 2012; Berlin, Heidelberg: Springer Berlin Heidelberg.
49. El Emam K, Dankar FK, Vaillancourt R, Roffey T, Lysyk M. Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records. The Canadian Journal of Hospital Pharmacy. 2009;62(4):307-19.
50. Lo B. Sharing Clinical Trial Data Maximizing Benefits, Minimizing Risk. The Journal of the American Medical Association. 2015;313(8):793–4.

8. Appendices

8.1. Syntax for case 1 - large effect size and frequent outcome

```
// Case 1
//Generating 10000 observations and implanting seed
clear
set seed 100
set obs 10000

//Generating pasientid to each patients
gen patientid = _n

//Generating the variable age
generate age = round(45+(rnormal(20,6)))

//Generating the variable sex with a affection of age
generate sex = round(runiform() + (0.01*age)-0.55)

//Generating the variable weight
gen weight= round(55+(rnormal(12,8)))

//Generating an exposure with an association with sex
gen exposure = round((runiform()*0.7)+(0.09* sex))
label var exposure "Treatment"
label define exposure_label 0 "Standard" 1 "New treatment"
label value exposure exposure_label

//Generating the outcome with an association exposure
gen outcome = round((runiform()*0.7)+(0.3* exposure))

//Generating the confounder hypertension with an association with exposure and outcome
gen hypertension = round((runiform()*0.7)+(0.09* exposure+0.02*outcome))
label var hypertension "Hypertension"
label define hypertension_label 0 "have not" 1 "have"
label value hypertension hypertension_label

//Generating smoking status
gen smoking = round(1+(rnormal(0,3)*0.6)+(0.09* exposure+0.05*outcome))
recode smoking min/0=0

//Generating diabetes patient
gen diabetes2= round(runiform() + (0.01*age+0.55*hypertension)-0,1)
recode diabetes2 min/1=0 // recode all 1=0 "doesn't have"
recode diabetes2 2=1 //recode all 2 =1 "have"
label var diabetes2 "Diabetes type 2"
label define diabetes2_label 0 "have not" 1 "have"
label value diabetes2 diabetes2_label

//Generating heart failure patient
gen hf= round(runiform() + (0.005*age+0.01*hypertension+0.001*diabetes2)-0.725)
label var hf "heart failure"
label define hf_label 0 "have not" 1 "have"
label value hf hf_label

//Multiple logistic regression for case 1
logistic outcome exposure sex age weight hypertension smoking hf diabetes2
```

8.2. Syntax for case 2 - small effect size and frequent outcome

```
//Case 2
//Generating 10000 observations and implanting seed
clear
set seed 100
set obs 10000

//Generating pasientid to each patients
gen patientid = _n

//Generating the variable age
generate age = round(45+(rnormal(20,6)))

//Generating the variable sex with an association with age
generate sex = round(runiform() + (0.01*age)-0.55)

//Generating the variable weight
gen weight= round(55+(rnormal(12,8)))

//Generating an exposure with an association with sex
gen exposure = round((runiform()*0.7)+(0.09* sex))
label var exposure "Treatment"
label define exposure_label 0 "Standard" 1 "New treatment"
label value exposure exposure_label

//Generating the outcome with an association with exposure
gen outcome = round((runiform()*0.9)+(0.025* exposure))

//Generating the confounder hypertension with an association with exposure and outcome
gen hypertension = round((runiform()*0.7)+(0.09* exposure+0.02*outcome))
label var hypertension "Hypertension"
label define hypertension_label 0 "have not" 1 "have"
label value hypertension hypertension_label

//Generating smoking status
gen smoking = round(1+(rnormal(0,3)*0.6)+(0.09* exposure+0.05*outcome))
recode smoking min/0=0

//Generating diabetes patient
gen diabetes2= round(runiform() + (0.01*age+0.55*hypertension)-0,1)
recode diabetes2 min/1=0 // recode all 1=0 "doesn't have"
recode diabetes2 2=1 //recode all 2 =1 "have"
label var diabetes2 "Diabetes type 2"
label define diabetes2_label 0 "have not" 1 "have"
label value diabetes2 diabetes2_label

//Generating heart failure patient
gen hf= round(runiform() + (0.005*age+0.01*hypertension+0.001*diabetes2)-0.725)
label var hf "heart failure"
label define hf_label 0 "have not" 1 "have"
label value hf hf_label

//Multiple logistic regression for case 2
logistic outcome exposure sex age weight hypertension smoking hf diabetes2
```

8.3. Syntax for case 3 - small sample size, small effect size and moderate frequent outcome

```
//Case 3
//Generating 10000 observations and implanting seed
clear
set seed 100
set obs 1000

//Generating pasientid to each patients
gen patientid = _n

//Generating the variable age
generate age = round(45+(rnormal(20,6)))

//Generating the variable sex with an association with age
generate sex = round(runiform() + (0.01*age)-0.55)

//Generating the variable weight
gen weight= round(55+(rnormal(12,8)))

//Generating an exposure with an association with sex
gen exposure = round((runiform()*0.7)+(0.09* sex))
label var exposure "Treatment"
label define exposure_label 0 "Standard" 1 "New treatment"
label value exposure exposure_label

//Generating the outcome with an association with exposure
gen outcome = round((runiform()*0.7)+(0.08* exposure))

//Generating the confounder hypertension with an association with exposure and outcome
gen hypertension = round((runiform()*0.7)+(0.09* exposure+0.02*outcome))
label var hypertension "Hypertension"
label define hypertension_label 0 "have not" 1 "have"
label value hypertension hypertension_label

//Generating smoking status
gen smoking = round(1+(rnormal(0,3)*0.6)+(0.09* exposure+0.05*outcome))
recode smoking min/0=0

//Generating diabetes patient
gen diabetes2= round(runiform() + (0.01*age+0.55*hypertension)-0,1)
recode diabetes2 min/1=0 // recode all 1=0 "doesn't have"
recode diabetes2 2=1 //recode all 2 =1 "have"
label var diabetes2 "Diabetes type 2"
label define diabetes2_label 0 "have not" 1 "have"
label value diabetes2 diabetes2_label

//Generating heart failure patient
gen hf= round(runiform() + (0.005*age+0.01*hypertension+0.001*diabetes2)-0.725)
label var hf "heart failure"
label define hf_label 0 "have not" 1 "have"
label value hf hf_label

//Multiple logistic regression for case 3
logistic outcome exposure sex age weight hypertension smoking hf diabetes2
```

8.4. Syntax for case 4 - small effect size and rare event

```
//Case 4
//Generating 10000 observations and implanting seed
clear
set seed 100
set obs 10000

//Generating pasientid to each patients
gen patientid = _n

//Generating the variable age
generate age = round(45+(rnormal(20,6)))

//Generating the variable "sex" with an association with age
generate sex = round(runiform() + (0.01*age)-0.55)

//Generating the variable weight
gen weight= round(55+(rnormal(12,8)))

//Generating an exposure with an association with sex
gen exposure = round((runiform()*0.7)+(0.09* sex))
label var exposure "Treatment"
label define exposure_label 0 "Standard" 1 "New treatment"
label value exposure exposure_label

//Generating the outcome withan association with exposure
gen outcome = round((runiform()*0.5058)+(0.0009* exposure))

//Generating the confounder hypertension with an association with exposure and outcome
gen hypertension = round((runiform()*0.7)+(0.09* exposure+0.02*outcome))
label var hypertension "Hypertension"
label define hypertension_label 0 "have not" 1 "have"
label value hypertension hypertension_label

//Generating smoking status
gen smoking = round(1+(rnormal(0,3)*0.6)+(0.09* exposure+0.05*outcome))
recode smoking min/0=0

//Generating diabetes patient
gen diabetes2= round(runiform() + (0.01*age+0.55*hypertension)-0,1)
recode diabetes2 min/1=0 // recode all 1=0 "doesn't have"
recode diabetes2 2=1 //recode all 2 =1 "have"
label var diabetes2 "Diabetes type 2"
label define diabetes2_label 0 "have not" 1 "have"
label value diabetes2 diabetes2_label

//Generating heart failure patient
gen hf= round(runiform() + (0.005*age+0.01*hypertension+0.001*diabetes2)-0.725)
label var hf "heart failure"
label define hf_label 0 "have not" 1 "have"
label value hf hf_label

//Multiple logistic regression for case 4
logistic outcome exposure sex age weight hypertension smoking hf diabetes2
```

8.5. Syntax for case 5 - moderate effect size and frequent, continuous outcome

```
//Case 5
//Generating 10000 observations and implanting seed
clear
set seed 100
set obs 10000

//Generating pasientid to each patients
gen patientid = _n

//Generating the variable age
generate age = round(45+(rnormal(20,6)))

//Generating the variable sex with an association with age
generate sex = round(runiform() + (0.01*age)-0.55)

//Generating the variable weight
gen weight= round(55+(rnormal(12,8)))

//Generating an exposure with an association with sex
gen exposure = round((runiform()*0.7)+(0.09* sex))
label var exposure "Treatment"
label define exposure_label 0 "Standard" 1 "New treatment"
label value exposure exposure_label

//Generating the outcome with an an association with exposure and age
gen outcome = round((rnormal(2,1))+(0.5* exposure)-(0.11*age))
replace outcome= round((rnormal(6.5,1))+(0.7* exposure)-(0.148*age)) if exposure ==0

//Generating the confounder hypertension with an association with exposure and outcome
gen hypertension = round((runiform()*0.7)+(0.09* exposure+0.02*outcome))
label var hypertension "Hypertension"
label define hypertension_label 0 "have not" 1 "have"
label value hypertension hypertension_label

//Generating smoking status
gen smoking = round(1+(rnormal(0,3)*0.6)+(0.09* exposure+0.05*outcome))
recode smoking min/0=0

//Generating diabetes patient
gen diabetes2= round(runiform() + (0.01*age+0.55*hypertension)-0,1)
recode diabetes2 min/1=0 // recode all 1=0 "doesn't have"
recode diabetes2 2=1 //recode all 2 =1 "have"
label var diabetes2 "Diabetes type 2"
label define diabetes2_label 0 "have not" 1 "have"
label value diabetes2 diabetes2_label

//Generating heart failure patient
gen hf= round(runiform() + (0.005*age+0.01*hypertension+0.001*diabetes2)-0.725)
label var hf "heart failure"
label define hf_label 0 "have not" 1 "have"
label value hf hf_label

//Multiple linear regression for case 5
regress outcome exposure sex age weight hypertension smoking hf diabetes2
```

8.6. Syntax for anonymization in case 1-4

8.6.1. Generalization

```
//Generalization
//Age
 generate gen_age = age
 recode gen_age 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
 label var  gen_age "generalized age into group"
 label define gen_age_label 0 "under or equal 45" 1 "46-50" 2 "51-55" 3 "56-60" 4 "61-65" ///
 5 "66-70" 6 "71-75" 7 "76-80" 8 "81 or above"
 label value gen_age gen_age_label

//Smoking
 gen gen_smoke= smoking
 recode gen_smoke 0/0=0 1/3=1 4/6=2 7/20=3
 label var gen_smoke "generalized smoking status into group"
 label define gen_smoke_label 0 "non-smoker" 1 "low to moderate smoker" 2 "moderate to frequent smoker" 3 "frequent smoker"
 label value gen_smoke gen_smoke_label

//Weight
 gen gen_wht= weight
 recode gen_wht 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
 label var gen_wht "generalized weight into group"
 label define gen_wht_label 0 "under or equal 45kg" 1 "46kg-50kg" 2 "51kg-55kg" 3 "56kg-60kg" 4 "61kg-65kg" ///
 5 "66kg-70kg" 6 "71kg-75kg" 7 "76kg-80kg" 8 "81kg or above"
 label value gen_wht gen_wht_label

//Comorbidity
gen co_morb =hypertension
replace co_morb=1 if hypertension ==1 | diabetes2 ==1 | hf==1
replace co_morb=0 if hypertension ==0 & diabetes2 ==0 & hf==0
label var co_morb "Comorbidity"
label define co_morb 0 "have not" 1 "have"
label value co_morb co_morb_label

 //Multiple logistic regression for generalization
logistic outcome exposure sex gen_age gen_smoke gen_wht co_morb

//Checking the unique record
egen composit_y = group( gen_age gen_wht sex exposure outcome co_morb gen_smoke)
bysort composit_y : egen grp_uniquel = count( patientid)
tab grp_uniquel
```

8.6.2. Randomization

```
//Randomization

//Generating noise to the dataset with new variables
//Age
gen age_r=age
gen age_n=round(rnormal(0,0.5))
replace age_r= age_r+age_n

//Smoking
gen smoke_r= smoking
gen smoke_n=round(rnormal(0,0.5))
replace smoke_r = smoke_r+smoke_n
recode smoke_r min/0=0

//Weight
gen weight_r=weight
gen weight_n =round(rnormal(0,+1.04))
replace weight_r = weight_r + weight_n

//Multiple logistic regression for randomization with the new variables
logistic outcome exposure sex weight_r age_r smoke_r hypertension diabetes2 hf

//Checking the unique record
egen composit_u = group( age_r weight_r sex exposure outcome smoke_r hypertension diabetes2 hf)
bysort composit_u : egen grp_unique2 = count( patientid)
tab grp_unique2
```

8.6.3. K-anonymity

```
//K-anonymity
//Start with generalization
//Age
generate gen_age = age
recode gen_age 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
label var  gen_age "generalized age into group"
label define gen_age_label 0 "under or equal 45" 1 "46-50" 2 "51-55" 3 "56-60" 4 "61-65" ///
5 "66-70" 6 "71-75" 7 "76-80" 8 "81 or above"
label value gen_age gen_age_label

//Smoking
gen gen_smoke= smoking
recode gen_smoke 0/0=0 1/3=1 4/6=2 7/20=3
label var gen_smoke "generalized smoking status into group"
label define gen_smoke_label 0 "non-smoker" 1 "low to moderate smoker" ///
2 "moderate to frequent smoker" 3 "frequent smoker"
label value gen_smoke gen_smoke_label

//Weight
gen gen_wht= weight
recode gen_wht 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
label var gen_wht "generalized weight into group"
label define gen_wht_label 0 "under or equal 45kg" 1 "46kg-50kg" 2 "51kg-55kg" 3 "56kg-60kg" 4 "61kg-65kg"///
5 "66kg-70kg" 6 "71kg-75kg" 7 "76kg-80kg" 8 "81kg or above"
label value gen_wht gen_wht_label

//Comorbidity
gen co_morb =hypertension
replace co_morb=1 if hypertension ==1 | diabetes2 ==1 | hf==1
replace co_morb=0 if hypertension ==0 & diabetes2 ==0 & hf==0
label var co_morb "Comorbidity"
label define co_morb 0 "have not" 1 "have"
label value co_morb co_morb_label

//Checking unique records (our K is set to be 3, so all unique records more 3 will be suppressed)
egen composit_k = group( gen_age gen_wht  sex exposure outcome co_morb gen_smoke)
bysort composit_k : egen grp_unique = count( patientid)
tab grp_unique

//Suppress the unique records
replace patientid=. if grp_unique <=2
replace exposure=. if grp_unique <=2
replace outcome=. if grp_unique <=2
replace sex =. if grp_unique ==1
replace gen_age =. if grp_unique <=2
replace gen_smoke =. if grp_unique <=2
replace gen_wht =. if grp_unique <=2
replace co_morb =. if grp_unique <=2

//Multiple logistic regression for K-anonymity
logistic outcome exposure sex gen_age gen_smoke gen_wht co_morb
```

8.6.4. Suppression

```
//Suppression

//Making a variable that identify unique records
egen composit_suppress = group( age weight sex exposure outcome smoking hypertension diabetes2 hf)
bysort composit_suppress : egen grp_unique = count( patientid)
tab grp_unique

//Eliminating participant's record that does not have at least one observation with same attribute
replace smoking =. if grp_unique ==1
replace weight =. if grp_unique ==1
replace hypertension=. if grp_unique ==1
replace diabetes2=. if grp_unique ==1
replace hf=. if grp_unique ==1

//Multiple logistic regression for suppression
logistic outcome exposure sex age smoking weight hypertension diabetes2 hf  //

//Check the unique record
egen check_uniq = group( age weight sex exposure outcome smoking hypertension diabetes2 hf)
bysort check_uniq : egen grp_check4 = count( patientid)
tab grp_check4
```

8.7. Syntax for anonymization in case 5

8.7.1. Generalization

```
//Generalization
//Age
generate gen_age = age
recode gen_age 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
label var  gen_age "generalized age into group"
label define gen_age_label 0 "45 or below" 1 "46-50" 2 "51-55" 3 "56-60" 4 "61-65" ///
5 "66-70" 6 "71-75" 7 "76-80" 8 "81 or above"
label value gen_age gen_age_label

//Smoking
gen gen_smoke= smoking
recode gen_smoke 0/0=0 1/3=1 4/6=2 7/20=3
label var gen_smoke "generalized smoking status into group"
label define gen_smoke_label 0 "non-smoker" 1 "low to moderate smoker" 2 "moderate to frequent smoker" ///
3 "frequent smoker"
label value gen_smoke gen_smoke_label

//Weight
gen gen_wht= weight
recode gen_wht 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
label var gen_wht "generalized weight into group"
label define gen_wht_label 0 "under or equal 45kg" 1 "46kg-50kg" 2 "51kg-55kg" 3 "56kg-60kg" 4 "61kg-65kg" ///
5 "66kg-70kg" 6 "71kg-75kg" 7 "76kg-80kg" 8 "81kg or above"
label value gen_wht gen_wht_label

//Comorbidity
gen co_morb =hypertension
replace co_morb=1 if hypertension ==1 | diabetes2 ==1 | hf==1
replace co_morb=0 if hypertension ==0 & diabetes2 ==0 & hf==0
label var co_morb "Comorbidity"
label define co_morb 0 "have not" 1 "have"
label value co_morb co_morb_label

//Outcome
gen g_outcome= outcome
recode g_outcome (2=0)(1=0)(0=0) (-3/-1=1) (-6/-4=2) (-30/-7=3)
label var g_outcome "generalized outcome"
label define g_outcome 0 "0 or above" 1 "(-)1-3" 2 "(-)4-6" 3 "(-)7-below"
label value g_outcome g_outcome_label

//Multinomial logistic regression for generalizaiton (baseline group: group 3)
mlogit g_outcome exposure sex gen_age gen_smoke gen_wht co_morb,b(3)

//Checking unique record
//making a variable that identify unique records
egen composit_y = group( gen_age gen_wht  sex exposure g_outcome co_morb gen_smoke)
bysort composit_y : egen grp_uniquel = count( patientid)
tab grp_uniquel
```

8.7.2. Randomization

```
//Randomization
//generating noise to the dataset with new variables
//Age
gen age_r=age
gen age_n=round(rnormal(0,0.5))
replace age_r= age_r+age_n

//Smoking
gen smoke_r= smoking
gen smoke_n=round(rnormal(0,0.5))
replace smoke_r = smoke_r+smoke_n
recode smoke_r min/0=0

//Weight
gen weight_r=weight
gen weight_n =round(rnormal(0,+1.04))
replace weight_r = weight_r + weight_n

//Outcome
gen outcome_r=outcome
gen outcome_n=round(rnormal(0,0.5))
replace outcome_r= outcome_r - outcome_n

//Multiple linear regression for randomization
regress outcome_r exposure sex weight_r age_r smoke_r hypertension diabetes2 hf

//Checking the unique record
egen composit_u = group( age_r weight_r sex exposure outcome_r smoke_r hypertension diabetes2 hf)
bysort composit_u : egen grp_unique2 = count( patientid)
tab grp_unique2
```

8.7.3. K-anonymity

```
//K-anonymity
//Start with generalization
//Age
generate gen_age = age
recode gen_age 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
label var  gen_age "generalized age into group"
label define gen_age_label 0 "under or equal 45" 1 "46-50" 2 "51-55" 3 "56-60" 4 "61-65" ///
5 "66-70" 6 "71-75" 7 "76-80" 8 "81 or above"
label value gen_age gen_age_label

//Smoking
gen gen_smoke= smoking
recode gen_smoke 0/0=0 1/3=1 4/6=2 7/20=3
label var gen_smoke "generalized smoking status into group"
label define gen_smoke_label 0 "non-smoker" 1 "low to moderate smoker" 2 "moderate to frequent smoker" ///
3 "frequent smoker"
label value gen_smoke gen_smoke_label

//Weight
gen gen_wht= weight
recode gen_wht 0/45=0 46/50=1 51/55=2 56/60=3 61/65=4 66/70=5 71/75=6 76/80=7 81/100=8
label var gen_wht "generalized weight into group"
label define gen_wht_label 0 "under or equal 45kg" 1 "46kg-50kg" 2 "51kg-55kg" 3 "56kg-60kg" 4 "61kg-65kg" ///
5 "66kg-70kg" 6 "71kg-75kg" 7 "76kg-80kg" 8 "81kg or above"
label value gen_wht gen_wht_label

//Comorbidity
gen co_morb =hypertension
replace co_morb=1 if hypertension ==1 | diabetes2 ==1 | hf==1
replace co_morb=0 if hypertension ==0 & diabetes2 ==0 & hf==0
label var co_morb "Comorbidity"
label define co_morb 0 "have not" 1 "have"
label value co_morb co_morb_label

//Outcome
gen g_outcome= outcome
recode g_outcome (2=0)(1=0)(0=0) (-3/-1=1) (-6/-4=2) (-20/-7=3)
label var g_outcome "generalized outcome"
label define g_outcome 0 "0 or above" 1 "(-)1-3" 2 "(-)4-6" 3 "(-)7-below"
label value g_outcome g_outcome_label

//checking unique records (our K is set to be 3, so all unique records more 3 will be suppressed)
egen composit_k = group( gen_age gen_wht sex exposure g_outcome co_morb gen_smoke)
bysort composit_k : egen grp_unique = count( patientid)
tab grp_unique

//Suppress the unique records
replace patientid=. if grp_unique <=2
replace exposure=. if grp_unique <=2
replace g_outcome=. if grp_unique <=2
replace sex =. if grp_unique <=2
replace gen_age =. if grp_unique <=2
replace gen_smoke =. if grp_unique <=2
replace gen_wht =. if grp_unique <=2
replace co_morb =. if grp_unique <=2

//Multinomial logistic regression for K-anonymity (baseline group: group 3)
mlogit g_outcome exposure sex gen_age gen_smoke gen_wht co_morb, b(3)

//Checking unique record after anonymization
egen check_uni= group (gen_age gen_wht  sex exposure g_outcome co_morb gen_smoke)
bysort check_uni: egen check_unique3= count (patientid)
tab check_unique3
```

8.7.4. Suppression

```
//Suppression
//Making a variable that identify unique records
egen composit_suppress = group( age weight sex exposure outcome smoking hypertension diabetes2 hf)
bysort composit_suppress : egen grp_unique = count( patientid)
tab grp_unique

//Eliminating participant's record that does not have at least one observation with same attribute
replace age =. if grp_unique ==1
replace smoking =. if grp_unique ==1
replace weight =. if grp_unique ==1
replace hypertension=. if grp_unique ==1
replace diabetes2=. if grp_unique ==1
replace hf=. if grp_unique ==1

//Multiple linear regression for suppression
regress outcome exposure sex age smoking weight hypertension diabetes2 hf

//Check the unique record
//Making a variable that identify unique records
egen check_uniq = group( age weight sex exposure outcome smoking hypertension diabetes2 hf)
bysort check_uniq : egen grp_check4 = count( patientid)
tab grp_check4
```

8.8. Syntax for generating 1000 -datasets and -analyses adjusted for co-variates

```
//This is the code for 1000 datasets adjusted with other variables
clear //run 1000 times multiple logistic regression and put it into excel
putexcel set "testfordataset1_justified.xlsx", replace
foreach val of numlist 1/1000 { //loop for set seed 100-1100
    display "val = `val'"
    local seed = 100 + `val'
    set seed `seed'
    set obs 10000

    //Generating pasientid to each patients
    gen patientid = _n

    //Generating the variable age
    generate age = round(45+(rnormal(20,6)))

    //Generating the variable sex with a affection of age
    generate sex = round(runiform() + (0.01*age)-0.55)

    //Generating the variable weight
    gen weight= round(55+(rnormal(12,8)))

    //Generating an exposure with an association with sex
    gen exposure = round((runiform()*0.7)+(0.09* sex))
    label var exposure "Treatment"
    label define exposure_label 0 "Standard" 1 "New treatment"
    label value exposure exposure_label

    //Generating the outcome with an association exposure
    gen outcome = round((runiform()*0.7)+(0.3* exposure))

    //Generating the confounder hypertension with an association with exposure and outcome
    gen hypertension = round((runiform()*0.7)+(0.09* exposure+0.02*outcome))
    label var hypertension "Hypertension"
    label define hypertension_label 0 "have not" 1 "have"
    label value hypertension hypertension_label

    //Generating the confunder smoking
    gen smoking = round(1+(rnormal(0,3)*0.6)+(0.09* exposure+0.05*outcome))
    recode smoking min/0=0

    //Generating diabetes patient
    gen diabetes2= round(runiform() + (0.01*age+0.55*hypertension)-0,1)
    recode diabetes2 min/1=0 // recode all 1=0 "doesn't have"
    recode diabetes2 2=1 //recode all 2 =1 "have"
    label var diabetes2 "Diabetes type 2"
    label define diabetes2_label 0 "have not" 1 "have"
    label value diabetes2 diabetes2_label

    //Multiple logistic regression for the outcome and other variables
    logistic `model' s
    matrix b = e(b)
    matrix list b
        forvalues num = 1/`count' { //loop for taking all value to excel in correct row
        scalar drop _all
        display "num = `num'"
        scalar r_`num' = b[1,`num']
        scalar list _all
        local cell = char(64+`num')+string(`val')
        display "cell = `cell'"
        putexcel `cell' = r_`num'
        }
    clear
}
```