A MULTI-LAYER MODEL FOR PRIVACY PRESERVING POLICY MAKING FOR DISCLOSURE OF PUBLIC HEALTH DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
MIDDLE EAST TECHNICAL UNIVERSITY

MEHRDAD ALIZADEH MIZANI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

SEPTEMBER 2013

Approval of the thesis:

A MULTI-LAYER MODEL FOR PRIVACY PRESERVING POLICY MAKING FOR DISCLOSURE OF PUBLIC HEALTH DATA

submitted by **MEHRDAD ALIZADEH MIZANI** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Medical Informatics Department, Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Director, Graduate School of **Informatics Institute** ________________

Assist. Prof. Dr. Yeşim Aydın Son
Head of Department, **Medical Informatics** ________________

Prof. Dr. Nazife Baykal
Supervisor, **Informatics Institute, METU** ________________

Examining Committee Members:

Prof. Dr. Ayşen Dener Akkaya
Department of Statistics, METU ________________

Prof. Dr. Nazife Baykal
Informatics Institute, METU ________________

Prof. Dr. Banu Çakır
Department of Public Health, Hacettepe Unirverity ________________

Assist. Prof. Dr. Erhan Eren
Department of Information Systems, METU ________________

Assist. Prof. Dr. Yeşim Aydın Son
Department of Medical Informatics, METU ________________

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: MEHRDAD ALIZADEH MIZANI

Signature :

ABSTRACT

A MULTI-LAYER MODEL FOR PRIVACY PRESERVING POLICY MAKING
FOR DISCLOSURE OF PUBLIC HEALTH DATA

Mizani, Mehrdad Alizadeh
Ph.D., Department of Medical Informatics
Supervisor : Prof. Dr. Nazife Baykal

September 2013, 160 pages

Health organizations in Turkey collect ever-increasing amount of individual data are valuable source of information for public health research. However, due to privacy risks, they publish data in aggregated rather than individual forms. The lack of standardized policies regarding secondary uses of health data leads to ineffectiveness of available technical methods. As a result, access to and utilization of person-specific datasets by public health researchers become extremely cumbersome. The bias introduced by privacy protection methods also makes data inefficient for epidemiological and public health contexts. We developed a three layer model for evidence-based policy making for secondary uses of health data. The first layer covers the evaluation of anonymized datasets based on clustering analysis independent of the underlying algorithm. The second layer provides the researcher with Representability Vector (RV), which consists of information about factors affecting the interpretation of research results. RV is also a method to gather researcher requirements and context-oriented evidence. The third layer, provides a generic framework for policy making with pseudo-contents covering the issues of anonymization and RV. This framework provides a dynamic approach for disclosure of and reporting bias to the researcher while emphasizing the policy issues along with context, evidence, and regulations.

Keywords: Public health, Epidemiology, Privacy, Policy making, k-anonymity

ÖZ

HALK SAĞLIĞI VERİSİ AÇIKLANMASINDA MAHREMİYETİ KORUYANPOLİÇE OLUŞTURMA İÇİN ÇOK KATMANLI MODEL

Mizani, Mehrdad Alizadeh
Doktora, Sağlık Bilişimi Bölümü
Tez Yöneticisi : Prof. Dr. Nazife Baykal

Eylül 2013 , 160 sayfa

Türkiyedeki sağlık kurumları giderek artan oranda halk sağlığı araştırmalarında değerli bir kaynak olan bireysel verileri toplamaktadır. Ancak, mahremiyetle ilgili riskler sebebiyle, bu veriler bireysel biçim yerine toplu şekilde yayımlanmaktadır. Sağlık verilerinin ikincil kullanımlarına dair standardize edilmiş poliçelerin eksikliği mevcut teknik yöntemleri verimsiz hale getirmektedir. Sonuç olarak, bireye özel verilerin halk sağlığı araştırmacıları tarafından erişimi ve kullanımı oldukça sıkıntılı hale gelmektedir. Mahremiyeti koruyan yöntemlerin ortaya koyduğu yanlılık verilerin epidemiyoloji ve halk sağlığı bağlamlarında da verimsiz hale gelmesine sebep olmaktadır. Sağlık verilerinin ikincil kullanımına yönelik kanıta dayalı poliçe oluşturulması için üç katmanlı bir model geliştirdik. İlk katman temel algoritmadan bağımsız şekilde anonimleştirilmiş veri analizinin değerlendirilmesini kapsamaktadır. İkinci katman araştırmacıya araştırma sonuçlarının yorumlanmasını etkileyen faktörler hakkında bilgi içeren Temsil Edebilme Vektörü (RV) sunmaktadır. RV aynı zamanda araştırmacının gereksinimlerini ve bağlam temelli kanıtları elde etme yöntemidir. Üçüncü katman anonimleştirme ve RV konularını ele alan psödo içerikle poliçe oluşturmak için genel bir yapı sunmaktadır. Bu yapı ikincil sağlık verisinin açıklanmasına ve araştırmacıya sunum yanlılığına dinamik bir yaklaşım sunarken bağlam, kanıt ve düzenlemelerle ilgili poliçeleri de vurgulamaktadır.

Anahtar Kelimeler: Halk sağlığı, Epidemiyoloji, Mahremiyet, Poliçe tasarımı, *k*-anonymity

To my father, sister, and the memory of my mother.

ACKNOWLEDGMENTS

I would like to thank my supervisor Professor Nazife Baykal for her constant support, guidance and friendship. It was a great honor to work with her for the last 10 years. Our cooperation influenced my academic and world view highly. Members of my thesis committee influenced and supported this work and their contribution were most valuable for me. Prof. Banu Çakır supplied a lot of important material on public health and epidemiology. Her ideas and support helped me to build the frame of this work from an epidemiological point of view. Other members of my thesis committee Professor Ayşen Dener Akkaya, Dr. Erhan Eren, and and Assist. Prof. Yeşim Aydın Son always gave valuable feedback for the progress of this work. The first four years of my PhD studies were also supported by TUBITAK-BIDEB PhD scholarship for foreign students(2215).

My family provided invaluable support for this work and always made me feel loved and cared. My father, Majid Alizadeh Mizani, constantly reminded that he stands by me even from thousand of kilometers away. My sister, Mahasti Alizadeh Mizani, was always there to give me hope, strength, and scientific ideas.

I am also thankful for all the love and support by Arda Arıkan, Derya Şahhüseyinoğlu, Deniz Evren, Oya Çinar, Banu Diken, Asuman Korkusuz, Zinnur Vapur, Sezgi Saraç, Çiğdem Toskay, Pınar Enginsu, Neval Binler Arıkan, Sibel Gülnar, Hasan Özkan, Elif Yılal, Hale karabekir, and Çiler Buket Tosun. And, finally, very special thanks to my cats Monsieur Paprika, Viva, and Shushu who comforted me during my stressful days, played around in my joyful times, and taught me the value of silence and staying in the moment.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation for this study

Public health informatics is an emerging field and is considered as a sub-category of health informatics and health information management [1]. In Turkey, most of the studies related to information management in healthcare are related to clinical informatics, which are mostly focused on primary and clinical aspects of healthcare, with less emphasis on public health issues. Primary uses of health data refer to their usage in clinical context in order to provide direct care to a patient, or health-related services to a healthy individual. Data gathered and used for such primary purposes are called "primary" or "clinical" data. The accumulation of primary data is a valuable source for research outside clinical settings. Alternatively, secondary uses refer to the application of health data for non-clinical purposes with an aim to improve the health of population or to enhance healthcare system. Examples of secondary uses include research in epidemiology, public health, healthcare policy making, quality management, disease surveillance, and public health ethics. While the focus of secondary uses is not on clinical services, their results ultimately affect the clinical protocols and the policies they are practiced under.

In recent years, the number of organizations that gather public health data has increased considerably. Individual data, or person-specific data, gathered by healthcare providers, public health agencies, registries, and census bureaus constitute the basis for public health research. However, using and sharing person-specific data beyond the boundaries of healthcare organization pose threat to the privacy of individuals. Health-related datasets include confidential characteristics that, without required precautions, may lead to privacy leaks and undesirable consequences, including discrimination in employment, insurance, and government services [2]. Privacy concerns have led many health organizations to publish data in aggregated forms or to maintain restrictive access to their datasets in their own premises. Although aggregated data provide overall information useful for ecological studies, they diminish data integrity and affect the generalizability of the results to the whole population.

Available solutions to protect the privacy of person-specific data are mostly theoretical, technical, and context sensitive. Public health organizations are usually reluctant to utilize those techniques largely because of unknown risks, changing requirements and contexts, and complex nature of privacy protection. In fact, privacy protection of individual data requires more than implementing a certain algorithm. Due to the changing nature of data usage and privacy risks, there is a need for robust policies, dynamic procedures, collaborative inter-organizational effort, and multi-disciplinary expert intervention. Additionally, the requirements of researchers, and available evidence, are rarely included in popular privacy protection methods. This leads to uncertainty about the distortion imposed on epidemiological variables which makes the objective measurement of bias extremely difficult, if not impossible.

In the light of what is stated above, the motivation for this study was based on the following inter-related issues:

- A generic method to analyze the results of existing technical methods either in restrictive or public access to secondary health data.
- Incorporating the requirements of researchers, context of data usage, and available evidence as determinants of privacy protection methods
- Reporting the dynamics of information loss in order to facilitate the objective assessment of the bias
- A generic framework for policy making for evidence-based privacy protection of secondary health data, while reporting information loss.

1.2 Inter-related concepts of this study: Aspects of privacy protection in disclosure of secondary health data

This section reviews some of the aspects of privacy protection in disclosure of health data for secondary uses, which constitute the basics of our proposed model.

1.2.1 Access to secondary health data and associated privacy issues

The focus of this study is on secondary uses of health data, which will also be referred to as "population research" and "public health" throughout this manuscript according to the context. In traditional healthcare model, the primary source of health data was patient records maintained in hospitals or at the office of primacy care physician. While, in the past, health data were limited to disease and treatment, they are now more comprehensive to include health and lifestyle issues. They also include supporting information affecting and determining healthcare procedures, such as employment and insurance data. Many of these data elements are gathered and used primarily for clinical purposes. Such uses of health data are called "primary uses", or "disease-based research". Secondary uses of health data, at the other hand, refer to utilization of data in activities that are not directly provide health service to an individual. While the immediate results of the utilization of secondary health data do not affect an individual, they ultimately affect the procedures, treatment protocols, policies, and regulations related to healthcare. Public health uses of health data do not necessarily require a researcher to know the patient or have personal interaction with them. As a result, having access to data of a sample population, without any personally identifiable elements, is adequate to conduct public health research, as long as the data include variables required for that particular research.

The eminent problem in secondary uses of person-specific data is privacy leaks. Privacy can be controlled and protected in clinical and controlled settings with effective privacy policies. However, disclosing health data to third parties for public health research gets data out of the controlled boundaries of the original healthcare organization. Third parties are not under the effect of internal privacy policies. At the other hand, with the growth of publicly accessible datasets, it is easier to link disclosed healthcare data with other sources and infer more details about the health condition of an individual. All these factors lead healthcare providers, and individuals alike, to be unwilling to disclose their data for research purposes.

Access to secondary health data is maintained in different ways depending on the sensitivity of data. This study focuses on two main approaches for data disclosure which are (1) publicly accessible, either in aggregated or person-specific forms, and (2) research-oriented access in which dataset is prepared according to researcher requirements. Original datasets gathered by healthcare providers, and stored by public health organizations, contain attributes with different privacy sensitivity. Examples of type and nature of these attributes are directly identifiable, personal characteristics, and health-related data. Table 1.1 shows an example of an identifiable dataset. In this table, name and social security number (ssn) are directly identifiable attributes, which do not need to be known by researcher.

Table1.1: Identifiable person-specific dataset

name	ssn	age	zip	test result
John	4537645738	23	5400	87
Jack	3746837483	25	5600	83
Mary	4434556666	27	5700	79
Jane	4374688810	22	5400	89
Cindy	2008947880	32	5200	84
Joseph	9890093773	34	5400	82
Sam	5610298389	34	5800	82
Jack	5563777123	39	5300	91

The interest of public health research is to find meaningful associations between a health condition and other variables. In Table 1.1, the 'test result' is the health condition while age and zip code are other variables. Assume that the aim of using this hypothetical dataset is to find any meaningful association between the health condition, reflected by test result, and at least one of the age or zip code attributes. Unlike many public datasets, health-related datasets are considered sensitive in terms of privacy. In the context of healthcare, it is assumed that personal characteristics, such as age and zip code, are known by an outsider who has access to publicly accessible datasets. However, attributes reflecting the health condition gathered by healthcare organization, public health agency, or the epidemiologists are not known publicly. The aim of privacy protection in person-specific data disclosure is to hide the identification of the individual whose health condition is contained in a disclosed dataset. The risk to privacy arises when the relationship between the health condition and the identity of the person is revealed. It is important to maintain such privacy protections, regardless of sensitivity and stigma attached to a health condition [3].

In order to protect the privacy of individuals, data holders use different approaches.

The first seemingly viable approach is to remove directly identifiable attributes from dataset. Table 1.2 shows a de-identified version of Table 1.1 . The second approach is to aggregate health-related attribute and disclose summary statistics for groups of people who share the same characteristics. This method is shown in Table 1.3. Another category for comparison of privacy protection approaches is whether the organization discloses data to the public or maintains access in a restrictive and controlled environment.

Table1.2: De-identified person-specific dataset based on Table 1.1

age	zip	test result
23	5400	87
25	5600	83
27	5700	79
22	5400	89
32	5200	84
34	5400	82
34	5800	82
39	5300	91

The starting point of privacy protection is to handle identifiable attributes. Examples of directly identifiable attributes are name, last name, address, telephone number, email, and social security number. While removing these attributes seems to be adequate in concealing identity, it is proved to be ineffective in privacy protection. It is shown that 87 % of the population of the United States have nearly unique characteristics based on 5-digit zip code, gender and birthdate which can be used to link the dataset with other publicly accessible data in order to uniquely re-identify an individual [4].

Table1.3: Aggregated dataset based on Table 1.1

age	zip	freq	avg(value1)
21-30	54*-57*	3	84.5
31- 40	52*-58*	4	84.75

Aggregating the dataset counters the privacy risks by presenting the statistical summaries for group of people or, in extreme cases, for the whole data set. The aim of population research, however, is to find the risk factors of a large group of people using association, observational, or comparison studies conducted on multiple type data from multiple sources [5]. Regardless of size and source of the sample data, the, person-specific data are essential for association studies. Aggregated data may be used for ecological studies, yet, such uses may rise to ecological fallacy which is a wrong assumption that a group characteristic in aggregated data is the same as the characteristics of individuals [6, p. 186].

These issues are the first to motivate this study with the realization that person-specific data are inaccessible outside the controlled clinical settings leading to unavailability of datasets for public health researchers. Where secondary health data are available, they are highly distorted which leads to biased results that are non-generalizable to the whole population.

1.2.2 k-anonymity: a de-facto standard for privacy protection of person-specific data

In this study we examined k-anonymity, proposed by Latanya Sweeney [4, 7], as the core method for protecting the privacy of person-specific datasets. This method ensures that all records with same characteristics, that may be used to infer about the identity of an individual, appear at least k times in the disclosed data. In other words, the chance of successful re-identification of an individual is at most $1/k$. While this method does not completely eliminate the risk of re-identification, it adds uncertainty as to the exact identification of the record owner. Table 1.4 shows a dataset with directly identifiable attributes removed at the left side. The first row is a record belonging to a person living in zip code 34232 with birthdate of 1994. Assume the case that this zip code is assigned to a small area where mainly elderly people live. In such a case, it is quite easy to find a younger person born in 1994 using publicly accessible datasets. These characteristics, or attributes, that can be combined to link a record with externally available data are called quasi-identifiers [7]. The table shown at the right side is a 2-anonymous version of the same dataset. It can be seen that the values of quasi-identifiers are altered in this table and their different combinations count at least 2 times. This alteration is referred to as "generalization", where the values are represented with less detail until the complete concealing of the value is reached, which is called suppression[8].

Table1.4: A de-identified dataset and its k-anonymized version

Birthdate	Zip	Condition
1994	34232	Diabetes
1994	34543	Hypertension
1971	36230	Diabetes
1970	36623	Osteoporosis
1979	36541	Diabetes
1999	37900	Asthma
1992	37831	Diabetes
1984	32221	Hypertension
1982	32301	Heart disease

Birthdate	Zip	Condition
1990	34***	Diabetes
1990	34***	Hypertension
1970	36***	Diabetes
1970	36***	Osteoporosis
1970	36***	Diabetes
1990	37***	Asthma
1990	37***	Diabetes
1980	32***	Hypertension
1980	32***	Heart disease

There is a large number of proposed and applied enhancements for k-anonymity in the literature, some of which are theoretical and highly context sensitive. Each of these methods are suitable for a certain setting, nature of data, context of use, sensitivity of data, priority of the data holder, and policies. While these methods are effective in theory, many of them become ineffective or complex to use in actual implementation due to the complexity and high-dimensionality of datasets. They also fail to address some implementation issues mandated by regulations and policies. As a result, it is practically impossible to adapt the basic k-anonymity, or any of its enhancements, without prior information about the actual needs of public health organizations. At the other hand, the changing regulations and policies require a flexible approach in choosing and implementation of any of these methods. While k-anonymity is chosen as the core method in our study, we will not propose a new method based on it. We will present a method to analyze k-anonymous datasets without being limited by the technical details of the algorithms based on k-anonymity.

1.2.3 Privacy protection, information loss, evidence, and researcher needs

When applying k-anonymity on a dataset, the records whose quasi-identifiers count less than k are either removed or generalized to a higher level. Removing the records shrinks the sample size and generalizing the quasi-identifiers decreases the amount of information content. Both of these approaches lead to bias and inaccuracies in extending, or the generalizability of, the study results to the wider population. Generalization[1] of the values of quasi-identifiers distorts the variables with which the health condition might have an association. One challenge of k-anonymization is to determine the balance between anonymity and accuracy[7]. Several metrics are used to measure the information loss due to anonymization. The original metric of k-anonymity is called "precision metric" which is used to choose the least distorted anonymized dataset in the set of all anonymized datasets. These different datasets are the result of applying generalization in different levels. More details on information content metrics will be discussed in Chapter 4.

The extend of information loss depends on the preferred algorithm and its parameters. The generally practiced approach is to anonymize a dataset with parameters determined internally by the organization that discloses the dataset. Example of such parameters are minimum k, levels of generalization, and the number of attributes to be included. The chosen values for these parameters,however, might not be of the best interest of the researchers in terms of their usability for finding a meaningful and correct association between the health condition and other characteristics. Anonymizing datasets for general users, without knowing the requirements of researchers, is what is widely practiced by data holders.

The problem with these approaches are high bias in disclosed data. Each characteristic labeled as quasi-identifier affects the outcome of k-anonymity, however, the researcher might not need many of these characteristics. The exact nature of required characteristics depends on the association between the health condition and other determinants, also known as dependent variables. Opting all characteristics in anonymized dataset regardless of the actual required determinant for the study leads to complexity and bias affecting the whole dataset. Another problem is that data holders usually calculate metrics for information loss only for internal analysis. Where information content is considered in choosing the final anonymization parameters, the resulted information loss is not reported to the third party. With unawareness about the details, including the numerical value and dynamics, of information loss it is practically impossible to ascertain the objective amount of bias. This, in turn, introduces uncertainty as to the accuracy and generalizability of the results to the population.

[1] Generalization is a term used both in epidemiology and k-anonymity domains. Refer to glossary at the end of the manuscript for more details.

1.2.4 Policy and legislative issues

One reason for reluctance to disclose person-specific data, or preferring to do so in statistically summarized forms and high levels of distortion, is the lack of uniform organizational policies and detailed national guidelines for secondary uses of health data. The reason for this is that the effectiveness of the existing technical methods and the resulted information loss depend on the nature of data. Constant changes in size and characteristics of datasets lead to changes in the nature of data in terms of their sensitivity, homogeneity, and temporality. While a technical method might exhibit desirable results for a dataset, it would probably become ineffective over time when dataset grows in size and its characteristics change. Having an effective and dynamic policy helps the organization to adopt to ever-changing and temporal nature of data with systematic decision making about the best technical method and details of its implementation.

Even the internal policies of an organization is affected by national legislations. While the end user of the disclosed data is not affected by internal policies, the organization needs to ensure that its internal policy is in line with related legislations. One limiting fact is that public health informatics is a new field in Turkey and the legislations regarding patient rights, privacy, and ethics do not cover the issues of secondary uses of health data in detail. This adds to the uncertainty about the effectiveness of internal policies. As new legislations emerge, the internal policies also become obsolete. As a result, the organizations need to have a dynamic approach for policy design and maintenance, in order to ensure conformity to legislations and to be able to disclose data with high information content. With an apparent lack of detailed legislations regarding the secondary uses of health data and with static, and mostly restrictive, organizational policies, public health organizations in Turkey limit their data disclosures to aggregated forms or to person-specific forms in controlled environments. All this leads to the unavailability of person-specific data to epidemiologists and public health researchers.

1.3 Review of problems and introduction to our proposed multi-layer model, *PARV*

As mentioned in sections 1.2.1 to 1.2.4, there are several problems that hinder the availability of secondary health data. The following list shows these problems:

- Abundance of technical privacy protection methods which are highly context sensitive. This leads to ineffectiveness of a certain method over time or due to dynamic nature of datasets.
- High information loss in disclosed data and unavailability of details about bias. This leads to uncertainty about the generalizability of the results to the whole population.
- Disclosing datasets for pubic use without considering the actual requirements of researchers and context based evidence. This leads to high information loss and diminished usability of dataset for population research.

- Lack of dynamic policies and comprehensive regulations for secondary uses of health data, especially is research oriented approach.

In order to address these problems in the context of public health informatics, we proposed a multi-layer model encompassing a framework for policy making. Our model consists of three inter-related Layers. Layer 1 of the model addresses the issue of a generic method to analyze the anonymity and to handle the records that pose risk to the privacy. Layer 2 of the model focuses on a unified method to report the characteristics of information loss to the researchers to assist them in incorporating the objective measurement of bias in their models. It also provides a means to incorporate context-based evidence in the analysis and anonymity processes. Layer 3 of the model, which is built upon layers 1 and 2, presents a generic framework for policy making in evidence-based and context-oriented disclosure of secondary health data, while protecting the privacy and reporting the information loss. We named our model *PARV* using the abbreviation of each layer. The whole model is a *Policy* making framework (layer 3) for analyzing the *Anonymization* process (layer 1) and producing *Representability Vector* as a tool to report bias and collect the context-based evidence (layer 2).

1.3.1 Primary aim of *PARV* model: context-oriented and evidence-based approach

The basis of *PARV* model is to consider the requirements of researcher, in the context of epidemiology and public health, in the anonymizing and information loss measurement processes. The widespread practiced approach by almost all health organizations is to disclose data in aggregated or person-specific forms in controlled environments, without knowing the specific requirements of the end user. Incorporating the requirements based on context of research makes it possible to choose the necessary characteristics of the interest to the researcher and to eliminate inessential characteristics. This reduces the dimensionality of data, distortions, and information loss. While the main focus is on research-oriented approach, our multi-layer model can also be applied for datasets disclosed to public, without considering the context and researcher requirements.

1.3.2 *PARV* model, layer 1: A generic method for analyzing the anonymization results

The first part of *PARV* model focuses on analyzing the results of k-anonymity based algorithms using clustering techniques. In this approach, the records that do not conform to k-anonymity are viewed as outlier points in dataset. Clustering techniques of data mining are effective tools for spotting patterns in data that appear less often than accepted margin, hence making them outliers. Clustering techniques can be applied on k-anonymized dataset without knowing the details of the actual algorithm used for anonymization. This facilitates a generic analysis method of the anonymized dataset regardless of the underlying technical algorithms. It aims to address the issues arising from the dynamic nature of data affecting the technical algorithms over time. It also provides a means for comparing different techniques using a uniform method.

The results of the clustering based analysis of anonymized datasets will be used as an input for the second layer of *PARV* model. The details of the first layer of our model will be discussed in Chapter 3.

1.3.3 *PARV* model, layer 2: Representability Vector (*RV*) for measuring the bias and collecting context-oriented evidence

The second layer of *PARV* model aims at reporting the details of information loss to the researcher with or without initial consideration of their requirements. This facilitates the objective measurement of bias and ascertaining the generalizability of the results to the whole population. This layer also have a direct connection with the first layer of the model where the anonymization might repeat according to the information loss and the revised requirements of the researcher. While there are some numerical metrics to measure information loss, these metrics are used internally by data holders to balance information loss and privacy protection. However, these metrics are not disclosed to the researcher. Additionally, each algorithm is equipped with a different metric. We proposed a method to calculate information loss based on cluster validity indices, regardless of underlying algorithm, and to report the details of the bias caused by anonymization to the researcher. We called this method "Representability Vector" (*RV*) as it shows how much the disclosed dataset resembles the original dataset. *RV* contains the numerical values for information loss and other characteristics of anonymization and information content. Furthermore, *RV* can also be used to describe the context of research and the requirements of the epidemiologists. In this case, *RV* can be used to capture the evidence-based requirements of researcher and to incorporate them in *PARV* model.

1.3.4 *PARV* model, layer 3: A framework for privacy preserving policy making

While anonymization and *RV* are technical concepts, their actual implementation and effectiveness are highly dependent on the context they are used in. The context is determined by the regulations and the priorities of the project or the organization. Both regulations and priorities might change over time as new requirements emerge and previous unknown risks and ethical concerns are identified. As a result, the technical measures, or their implementation, naturally become obsolete in the face of rapidly changing requirements and information needs. In order to address this issue, we proposed a conceptual framework for evidence-based and context sensitive policy making encompassing the layers 1 and 2 of *PARV* model. This framework emphasizes the expert intervention and importance of incorporating context and evidence in the process. It is conceptual in terms that it provides pseudo-contents, presenting high level and abstract points of policy making. Figure 1.1 depicts the relationship between the three layers of *PARV* model.

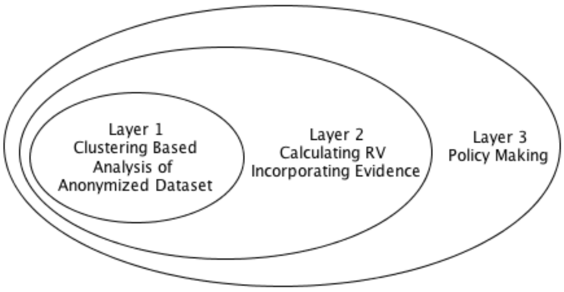

Figure 1.1: Three levels of *PARV* model

1.4 Comparision between *PARV* model and existing methods

In traditional anonymization methods the aim is to anonymize a locally held dataset and disclose it to unknown data users. This model is depicted in figure 1.2. The anonymization is performed according the the parameters determined local registry. These choices are made based on ad-hod data disclosure characteristics, priorities, or by local policies. In almost all cases the choices of parameters and methods are made without considering the evidence, context, or the requirements of the researcher.

The information content measurement is calculated as a means to optimize anonymization rather than reporting the bias to the researcher. Thus, the only information disclosed to the researcher is anonymized dataset with no information about the quantity or dynamics of bias and information loss. In this model, the choice of the anonymization method or the parameters might change over time as the characteristics of data, regulations, or the demands of the end user change. As a result, it is not uncommon to change the methods and information content measurements accordingly. This would require a change in local policies and related procedures affecting the implementation of those methods. Example of this model is original k-anonymity where the parameters are minimum k, choices of quasi-identifiers, and characteristics of generalization of each quasi-identifier. The precision metric is used in this method to measure the information loss of different disclosed tables and to choose the least distorted table. The requirements of the researcher is not included in the method and the only disclosed data is k-anonymized dataset.

In *PARV* model, as depicted in figure 1.3, the anonymization, information content measurement, and policy issues are separated into three inter-dependent layers. The anonymization handles the k-anonymity based methods as a black box and analyzes the results using clustering outlier analysis. The goal is to find the records in original data that are not conforming to k-anonymity and to analyze the approach used by the registry to handle them. This analysis comprises the layer 1 of *PARV* model. Clustering is used as a generic analysis method in this layer, hence, even if the underlying anonymity method or parameters change, the analysis procedure and the methods for interpretation the results remain the same. Additionally, this layer takes into account the requirements of the researcher and the evidence to guide the initial anonymization of data.

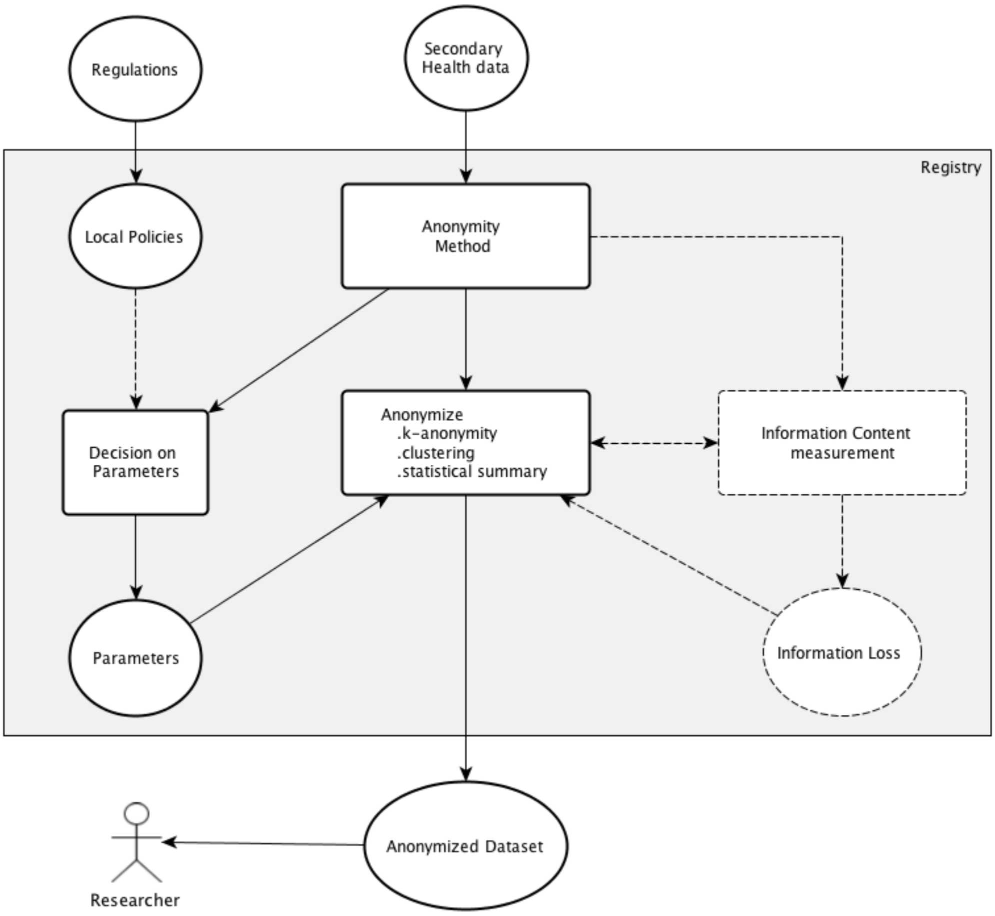

Figure 1.2: Traditional model for privacy preserving disclosure of health data

The results of the analysis performed in layer 1 undergo another cycle of analysis in layer 2 of $PARV$ model. In this layer, the cluster validity indices are used to calculate the actual information loss of the anonymized data. This information loss, along with other evidence-based and research-oriented characteristics of anonymization, are reported to the researcher in the form of Representability Vector (RV). RV shows the resemblance of the anonymized dataset to the original dataset, hence, indicates how far the disclosed dataset 'represents' the sample data. It can be used by researcher to ascertain the generalizability of their research to the population and the quantity and characteristics of bias. At the other hand, it can be used to revise the anonymization cycle and re-anonymize data to render it suitable to the research context. The choice of the parameters and methods of layers 1 and 2 of $PARV$ model are context-oriented and are affected by regulations and organizational priorities. In order to being able to dynamically manage the anonymization and bias reporting procedures in accordance with dynamic nature of data and regulations, the third layer of $PARV$ model presents a framework for policy making encompassing layers 1 and 2. It is a generic policy making roadmap including pseudo-policies adaptable to unique requirements of registries the context of secondary data usage.

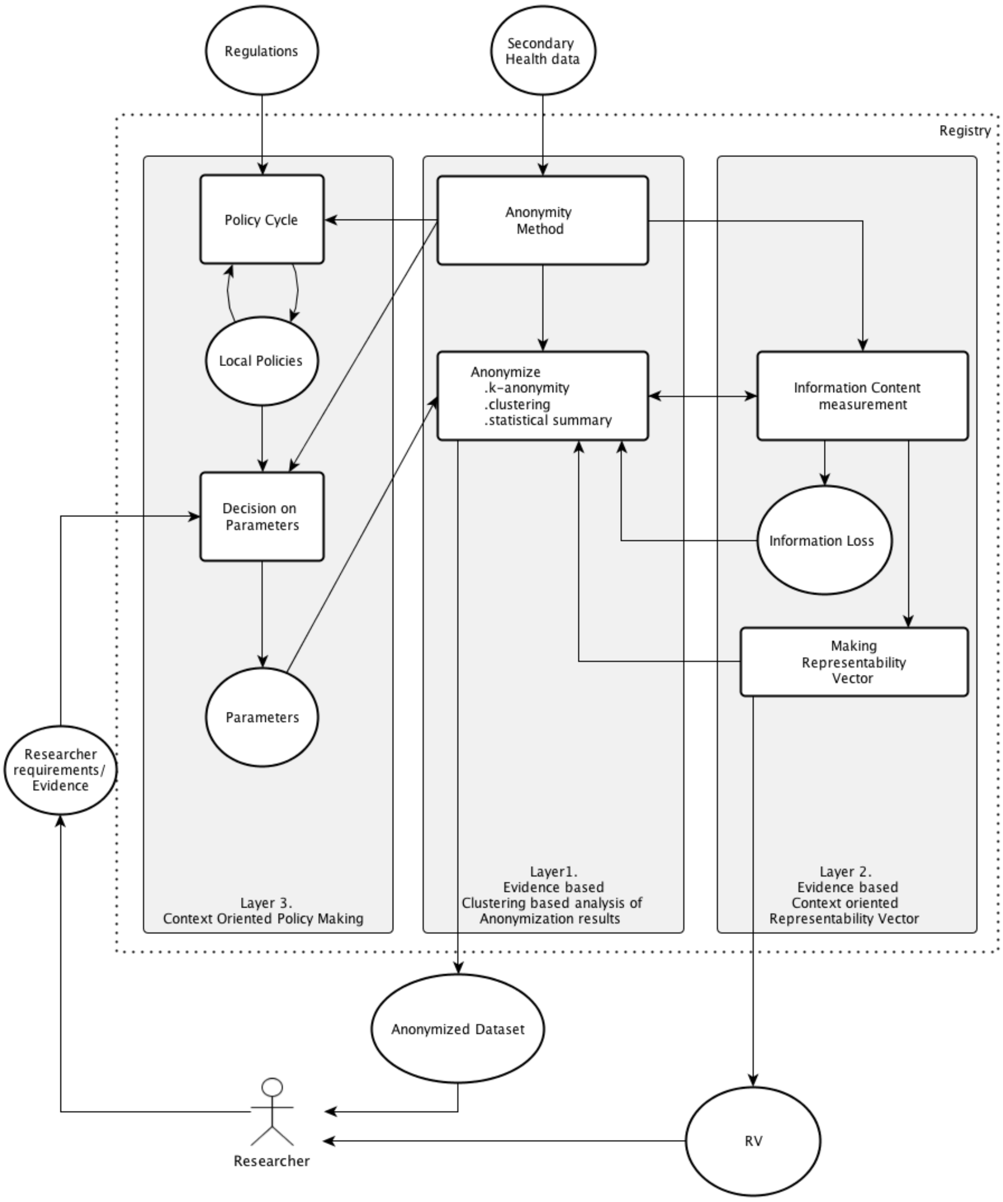

Figure 1.3: Multi-level *PARV* model for evidence-based policy making for privacy preserving disclosure of health data and *RV*

CHAPTER 2

BACKGROUND

2.1 Privacy issues of disclosing health data for secondary uses

The principal aim of our study is to highlight the importance of secondary health data for improving public health and encouraging their effective usage through policy based privacy protection. This section provides review of the issues of secondary uses of health data and associated privacy concerns pertaining to our study. All reviews and interpretations are centered around public health informatics domain which our study is highly related to. Hersh identifies four levels of medical informatics namely bioinformatics, medical imaging, clinical informatics and public health informatics [9]. Clinical aspects of informatics are related to primary uses of health data which refer to collection and uses of data to provide direct care to an individual patient [10]. A key characteristic of primary uses of health data, either in healthcare or informatics domains, is that it is patient-centered and disease based. Such approaches are focused on individuals or small groups, centered on treating diseases, have local and small samples, and are usually short term [5]. The goal of the disease based research is "to ameliorate or cure a particular disease, regardless of its membership in a particular community" [11] (as cited in [5]). Secondary uses of health data, at the other hand, refer to situations where data is used for purposes other than those intended at the time of data collection[10, 5]. Examples of secondary uses are epidemiology, public health, healthcare policy making, health service research [12], disease and public health surveillance [13, 14], public health ethics [15], and community medicine. The aim of secondary uses of health data is to improve the health of population and promote innovations that affect the health of a community [11] (as cited in [5]). Public health, for example, is concerned with health promotion, disease and disability prevention, population surveillance, multi-dimensional determinants of health, and developing multi-factor effective interventions with both biological and social aspects [16]. Although secondary health data are used for purposes not intended at the time of collection, their uses depends on and affect the clinical practice. Diagnosis, selection of appropriate treatment, and prognosis are population based [6], hence are related to secondary data. Public health researchers use the data of individuals as health determinant to spot problems and to propose solutions to resolve them [5]. The advantages of using secondary are that it's already collected reducing the cost, its representativeness, larger sample size, and reduces bias such as recall and non-response [17].

However, barriers of usage of secondary data renders disclosed datasets less useful in terms of aforementioned advantages. Along with technical barriers, some of the societal and organizational barriers are healthcare systems, privacy, informed consent, and trust between organization [18]. Methods used to protect privacy usually affect the accessibility of data, reduce sample size, and increase bias. This concern is due to the sensitivity of health-related characteristics that might lead to discrimination or stigmatization against the individuals whose data are disclosed [2, 19, 20].

The main challenge using of secondary health data is balancing the privacy protection and data accessibility that would benefit the public [7, 21, 22, 23, 24, 25, 26]. Privacy is defined as "The right and desire of a person to control the disclosure of personal health information" [27]. Another definition is "protection from being brought to the attention of other" [28] (as cited in [29]).This hinders the actualization of public health objectives with regular and ongoing access to person-specific data [19]. These datasets however provides a thorough picture of health and socioeconomic status of a person which is valuable for finding unknown health determinants or to find association between known determinants. Sweeney [7] mentions that even the minimal privacy protection applied on a dataset reduces it's usefulness and quality for research. Hence, health organization who are aware of this tension refuse to disclose data or distort them beyond usefulness. HIPAA, for example, allows the use of personal health information for public health purposes, and yet many data providers are reluctant to disclose data due to concerns over inappropriate usage of data [30].

Almost all proposed solutions for this dilemma affects the secondary data and its usage. Following list shows the factors determining the value of secondary data [17]. These factors are affected by solutions for tension between privacy and utility in one way or another.

- Completeness of the registration of the individuals
- The accuracy and degree of completeness of the registered data
- The size of the data source, registration period
- Data accessibility, availability, and cost
- Data format
- Possibilities of linkage with other data sources

The solutions for the tension between privacy and utility fall into different categories including purely technical, procedural, and regulatory. One solution is conditioning the secondary uses of health data to obtaining the informed consent of the patients. This is a procedural approach affected by regulations rather than a technical one. This approach is more suitable for situation where the future uses of data are known and the number of patients are not high. In reality, however, contacting all individuals for informed consent is impossible [18] which leads to non-response or participation bias [31, 32]. Individuals who support consent based uses of their health data are also concerned about losing control over their data [33]. This might increase with patients becoming more aware of sensitivity of data and the anxiety over unauthorized access by third parties such as insurance or pharmaceutical companies [34].There is also uncertainties about the situations where consent is implied and when it can be waived [35]. Another practiced solution is to restrict the access to secondary data to the premises of the registry.One example is the policy of Turkish Statistical Association where access to certain non health-related categories of person-specific data is only allowed for a limited time in their own premises and the usage and results must go through privacy checking [36]. This approach provides controlled use of data for public health officials within the agency where monitoring the access to data minimizes the risk to privacy [37]. However, access to data by individual researchers or private organizations might not be maintained at the same level as those of public health officials. Additionally, the restrictive use of data within an organization limits the researcher to the data analysis tools provided by the organization. This will hinder the researches performed using complex tools and data mining softwares that might not be present in the data providing organization.

Another popular solution is to aggregate the data for whole dataset or for group of people. Example of this measure is the publicly accessible census data of Turkish Statistical Association. This method distorts data considerably, especially when it is applied on the whole dataset. Researchers based on aggregated data are not generalizable to the whole population and can not be used for cohort or case-control studies. This approach leads to ecological bias, also called aggregation bias [38], which is a wrong conclusion that an association observed in a grouped data holds true for individuals [39]. There are situations where ecological bias is negligible, however, identifying those situations are impossible without access to person-specific data [38]. In spite of ecological bias, aggregated data are valuable sources for epidemiological studies as they provide overall information for researches aimed at finding etiological relationships [6]. Therefore, they play a role in defining public health problems and hypothesizing the possible causes [40].

Our study is centered around, but not limited to, another family of solutions for tension between privacy and utility which is anonymization of person-specific data. The following sections provide technical, information content, researcher requirements, and policy aspects of anonymization of person-specific data.

2.2 Technical aspects of anonymization in secondary uses of helath data

Health data contains several pieces of information along with health-related information that are exclusively gathered in healthcare organizations. When data is used, or disclosed for secondary uses, any linkage between person identity and the health-related information must be concealed. This is done in order to prevent the re-identification and subsequent harm based on what the intruder learns about an individual from an unwanted re-identification [41]. The first step to achieve this goal is de-identification. Health Insurance Portability and Accountability Act, HIPAA, defines de-identified data as a dataset that does not identify a patient whether by itself or it can not be used to identify an individual [42]. Therefore, de-identification refers to the removing the explicit identifiers from a dataset [43]. Based on HIPAA [42], Identifiable attributes are either determined by experts or are choses based on safe harbor method. Based on safe harbor method some of these identifiers are name, geographical information, dates, telephone and fax number, email, social security number, account number, medical record numbers, and IP.

While removing explicit identifiers seems a straightforward and sufficient way to protect privacy, it is fall short to fully prevent re-identification. Latanya Sweeney showed that about 87% of the population of the United States show unique characteristics based some indirect identifiers such as 5-digit zip code, birthdate, and gender [4]. Anonymization techniques to counter privacy attacks can be categorized into five groups which are data suppression, data generalization, data swapping, micro-aggregation, and macro-aggregation [44]. Data suppression completely removes the value of the attribute. Data generalization groups data in equivalent classes by replacing their values by a more general hierarchical value. Data swapping replaces the values of attributes with the values of other attributes and vice versa. Micro-aggregation clusters the attribute values around a representative, such as the average value. Macro-aggregation refers to statistical summary datasets which are not person-specific. There are a huge number of algorithms and approaches proposed each with abilities to address only a part of the privacy attacks. The assumption is that the attacker has access to and background knowledge of some publicly accessible data about individuals. The four major types of attack models are linking a record to spot the identity (Record linkage), linking the records to spot the sensitive attribute (Attribute linkage), linking tables to determine a person's data is present therein (Table linkage), and a variation in posterior and prior knowledge about data (Probabilistic attacks) [45].

Our study is a generic approach to analyze the results of these methods regardless of the details of the algorithms used or the attack mode they address. The central approach used for analysis which clarifies the concepts we use in our model is best explainable with the concepts of generalization method, and specifically k-anonymity. The reasons for choosing k-anonymity is that its simple, intuitive, easy to understand and validate [46]. All above-mentioned methods are field structured methods. Other methods, such as Scrub system for de-identifying the textual data [7, 47], are beyond the scope of our study.

After de-identification, the remaining attributes are either exclusively health-related, or non-identifiable attributes. Health-related attributes are called confidential or sensitive attributes as their relation to the identity must be concealedThe non-identifier attributes may be known by outsiders and be included in publicly accessible datasets. Example of such attributes are age, gender, ethnicity, job, and education level. Assuming that the all externally accessible datasets are known, the combination of non-identifiable attributes that can be used to infer about the identity of an individual are called quasi-identifiers and the attributes with identical values for all quasi-identifiers are called equivalent classes [7]. Three basic models for protecting privacy are null-map, wrong-map, and k-map [7]. Assume that a de-identified dataset is disclosed. Also assume that an attacker tries to re-identify the record owners by linking available datasets in order to map de-identified record to an identifiable record. Null-map model refers to situations where at least on of the records in re-identification process map to a non-existent person in the population. Wrong-map model maps the record to a real person however the value of the attributes belong to someone else in the population. In k-map model, the re-identification process maps to at least k records of real or unreal persons in the population. k-anonymity protection model incorporated the k-map model. The aim of k-anonymity model is to guarantee that each equivalent class has at least k members.k-anonymity is actualized by applying generalization and suppression [7, 8]. Generalization replaces the values of the quasi-identifiers with less specific values. The values of each quasi-identifier follows along a path of generalizable values beginning from the most informative to least informative in terms of granularity and specificity. Suppression, which can be considered as the highest level of generalization, completely conceals the value of the quasi-identifier. Figure 2.1 depicts a the generalization of birthdate and zip code up to the suppression level.

$$\overbrace{1994}^{birthdate_0} \to \underbrace{\overbrace{1990}^{birthdate_1} \to \overbrace{1900}^{birthdate_2}}_{generalized} \to \underbrace{****}_{suppressed}$$

$$\overbrace{34232}^{zip_0} \to \underbrace{\overbrace{3423*}^{zip_1} \to \overbrace{343**}^{zip_2} \to \overbrace{34***}^{zip_3} \to \overbrace{3****}^{zip_4}}_{generalized} \to \underbrace{*****}_{suppressed}$$

Figure 2.1: Generalization of the birthdate and zip code including suppression

Generalizing leads to the values of combinations of quasi-identifiers to count more than those of the original table. This increases the number of occurrences of unique values of combinations of quasi-identifiers. When the minimum number of value of combinations is k the table is called as a k-anonymous table. While applying k-anonymity does not fully eradicate the risk of re-identification, it reduces the possibility of successful unique re-identification. It in fact increases the uncertainty, or entropy, regarding the identity of an individual. Higher k and generalization levels increase this entropy considerably.

Table2.1: Generalization of birthdate and zip code at levels 1 and 3 respectively

no.	**birthdate**	**gender**	**zip code**	**condition**
1	1990	Male	34***	Diabetes
2	1990	Male	34***	Hypertension
3	1970	Female	36***	Diabetes
4	1970	Female	36***	Osteoporosis
5	1970	Female	36***	Diabetes
6	1990	Female	37***	Asthma
7	1990	Female	37***	Diabetes
8	1980	Male	32***	Hypertension
9	1980	Male	32***	Heart disease

Table 2.1 shows a 2-anonymous table with birthdate, zip code and gender as quasi-identifiers. In order to realize 2-anonymity, the values of birthdate are generalized one level and the values of zip code are generalized three levels. This generalization is applied at the same level per quasi-identifier per record. As it is seen, the combinations of the values of quasi-identifiers count at least 2. Table 2.2 shows the same dataset conforming to 2-anonymity with higher levels of generalization for different records. These two tables are examples of different levels of generalization.

Generalization and suppression might be applied to in different levels [7].

- Domain generalization: where the values of a single quasi-identifier is generalized for whole attribute. For example the values for all records of birthdate attribute in table 2.2 have been generalized one level along the Domain Generalization Hierarchy. This approach is also called as attribute level generalization.

- Value generalization: where the values of a single quasi-identifier might be generalized to different levels for different records. For example, the zip code attribute is table 2.2 has been generalized to the third level in record 3 and to the fourth level in record 6. This approach is also called as cell level generalization.

- Attribute level suppression: where the values of a certain quasi-identifier attribute is suppressed for all records. Since suppression conceals the whole data, this approach is equivalent to completely removing an attribute.

- Record level suppression: where a record is completely removed from dataset.

- Cell level suppression: where the value of a quasi-identifier is suppressed for certain records. For example the values of gender attribute in table 2.2 has been suppressed for records 1, 2, 6, and 7.

Table2.2: Generalization of birthdate and zip code at different levels per record

no.	**birthdate**	**gender**	**zip code**	**condition**
1	1990	****	3****	Diabetes
2	1990	****	3****	Hypertension
3	1970	Female	36***	Diabetes
4	1970	Female	36***	Osteoporosis
5	1970	Female	36***	Diabetes
6	1990	****	3****	Asthma
7	1990	****	3****	Diabetes
8	1980	Male	32***	Hypertension
9	1980	Male	32***	Heart disease

While high levels of generalization might be applied to protect privacy, it diminishes the integrity of data. It reduces the amount of useful detail which in turn affects the models used by researcher. It can considerably affect the generalizability of results to the whole population in epidemiological studies. A highly generalized person-specific dataset might become useless for case-control or cohort studies due to high levels of distortion. As a result, there is a trade-off between privacy protection and accuracy of data.

MinGen [7, 48], is the original model of k-anonymity with an information-theoretic point of view. Actual data is multi-dimensional containing more than one quasi-identifiers. With each quasi-identifier having a different type, they have different generalization hierarchies and chosen generalization and suppression levels. These levels including manipulation at record, attribute, or cell levels. All these combinations produce different datasets all conforming to k-anonymity. Many of these tables, however, distort data unnecessarily. The aim of MinGen algorithm is to find the least distorted dataset conforming to k-anonymity. MinGen is the basic algorithm I use throughout this study to analyze other related models. MinGen uses the concept of k-minimal distortion. k-minimal requires that a dataset satisfies k-anonymity and its associated precision metric or weighted precision metric be the minimal among all k-anonymous datasets. A k-minimal distorted dataset is also satisfies k-minimal generalization where there is no unnecessary generalizations applied.

2.3 Basis for the first level of *PARV* model: k-anonymity extensions and clustering based approaches

There are many algorithms proposed that are based in k-anonymity or aim at addressing its shortcomings in terms of privacy protection, computational complexity, or complications in actual usage. This section presents the most popular models and methods based in k-anonymity and related studies to my proposed solutions.

Datafly algorithm adds the concepts of anonymity level and recipient profile to k-anonymity[7, 49, 50]. Anonymity level indicates the sensitivity of the quasi-identifier whose higher values indicate the demand for more generalization and higher k. Even with high anonymity levels, there could be records that are not conform to k-anonymity. These outlier records are deleted in Datafly algorithm. Recipient profile indicates the sensitivity of an attribute whose higher values indicate concerns over the linking to other datasets. Attributes with higher sensitivities will be chosen for generalization over the other attributes. This is used to release different datasets in accordance to the profile of the last user. If the last user is a trusted physician, then the sensitivity would be lower and the recipient profile will be closer to 0. Therefore, Datafly is suitable for conditions where the policy calls for considering the nature of the use based on the profile of the recipient. None of these parameters are reported to the researcher.

μ-Argus is another algorithm, similar to Datafly. Unlike Datafly which removes small equivalent classes, μ-Argus uses suppression at cell level or generalization at attribute level [7, 50]. Datasets intended for public access are go through tighter rules of anonymization comparing to those disclosed to researchers [51]. Another feature of μ-Argus is its emphasis on expert review which puts the last decision about generalization and suppression on human decision. This method is suitable for situations where expert review is more important that technical efficiency and for contexts where the preserving the sample size is vital.

p-sensitive k-anonymity is an enhancement proposed for k-anonymity with an aim to address the inherent shortcoming of k-anonymity called attribute disclosure. It refers to situations where the intruder learns something new about an individual without necessarily infer about their identity [41] (as cited in [52]). p-sensitive k-anonymity addresses the issue of attribute disclosure based on sensitive attribute. It requires that the number of distinct values of sensitive attributes be at least p per equivalent class. It is suitable for situations where the size of the dataset and k are high and sensitive attributes are fairly heterogeneous in each equivalent classes.

ℓ-diversity also addresses the issue of attribute disclosure by guaranteeing that each equivalent class contains at least ℓ well-represented values for corresponding sensitive attributes [53]. It is suitable for situations where sensitive attributes are not diverse enough and the registry does not have full knowledge about external datasets. t-closeness is another algorithm that covers the issues that are not fully addressed by ℓ-diversity. It requires that the distance between the distribution of sensitive attribute per equivalent class and the distribution of sensitive attributes of the whole dataset be less than a predefined threshold t [54]. One example of situations that t-closeness can be used is when removing outliers are permissible and actually smoothes the distribution of sensitive attributes[54].

The choice and behavior of each of these algorithms, chosen from many algorithms and applications [55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72], depends heavily on the nature of data, the context they are used, degree of human intervention, external data sources, sensitivity of data determined by experts, and policies. For example, the policy guides the decision based on the tradeoff between each of k-anonymity, ℓ-diversity, or p-sensitive k-anonymity algorithms that do not protect privacy completely and t-closeness which protects privacy completely but considerably distorts the relation between sensitive attributes and other attributes [73]. All these show that the context of data disclosure is a direct determinant in effective privacy protection. Our study addresses this issue by analyzing the results of these and similar algorithms independent of the actual implementation of each of them. We used clustering analysis to analyze the outliers and cluster validity indices to calculate a generic information content measurement. This approach is general in terms that it also extendible to methods not based on k-anonymity, such as aggregation and clustering based methods.

k-anonymity can be viewed as clustering problem [74]. It can be said that the aim of any anonymization technique is to handle outliers. While traditional pattern recognition approaches views the extreme patterns as noise [75], they might also be viewed as unknown information about new categories, topics, patterns, deviant behavior, or dangers [75, 76].

Some of the studies directly apply clustering to achieve privacy protection in general , and k-anonymity in particular. For example *r-Gather and r-Cellular* methods, cluster quasi-identifiers and present each cluster with a representative centrum, cluster size, and radius information [77]. It is shown to have a higher performance comparing to k-anonymity. It also allows the deletion of extreme outlier values. It presents the distortion either as the maximum cluster radius or as the radius of each cluster. Comparing to k-anonymity, the centrum is more representative of the members of the quasi-identifiers. Another significance of this method is that the sensitive attribute is not aggregated, making it suitable for non-ecological studies. Contextually, this method is suitable for situations where generalization and suppression is too complex to apply, extreme outlier values are not important to the researcher, and sample size manipulation is tolerated.

Another example of application of clustering to achieve k-anonymity is *k-member clustering problem* [74] which uses the distance function of clustering to calculate the information quality. This is somehow similar to our approach of hyperplane distance measurement that includes the semantics of distance in the generalization process. The difference is that our approach is heavily based on expert review such that the semantic distance between categorial attributed along the generalization tree are based on evidence and researcher requirements. Our approach also encourages the usage of any kind of distance measurement as long as the same measurement is used for comparison of different algorithms.

One advantage of using clustering viewpoint is that many epidemiological and surveillance studies use data that are naturally clustered, such that spatial data. One example is blurring the spatial location of patients in each cluster in spatial surveillance and its impact on disease detection [78]. In the same study, k-anonymity is guaranteed while skewing the location of a patient based on the density of the underlying population in that cluster. Even in less dense areas, where a higher skew is enforced, the algorithm shown to have minor effect on the performance of outbreak detection. Another example of the intersection of clustering and k-anonymization is the method proposed by Benjmain Fung, et al. which is a generalization framework to guarantee that k-anonymized dataset preserves clustering structure and remains useful enough for clustering analysis [79, 80].

Like algorithms based on k-anonymity, clustering based algorithms for k-anonymization focus on performance, in terms of applicability and privacy protection, and information content or data quality. Each of these clustering based k-anonymization algorithms [74, 77, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93], are suitable for a certain context, requirement, setting, and data type.

Our study is different from these methods in terms that it does not intend to propose a new k-anonymity or clustering approach for anonymization. Secondary uses of health data are emerging in Turkey and it is not possible to foresee the exact context of such uses or the nature of related studies. As a result, our study provides a means to analyze the aforementioned family of algorithms and approaches based on a uniform method which is clustering analysis. This will provide flexibility in choosing the algorithms, or their internal parameters, while utilizing an algorithm-independent method to analyze and compare different choices. This goal is covered in the first level of *PARV* model, presented in chapter 3, where the necessary input is prepared for a generic information content measurement.

2.4 Basis for the second level of *PARV* model: Data utility, bias, and evidence from an epidemiological perspective

The previous section presented only a handful of a large number of methods available for anonymization and privacy protection. All these methods focus on two major problems of privacy protection which are characteristics of anonymization process, such as performance, and utility of data in terms of information loss. However, the majority of these studies are not take into account the unique requirements of epidemiological and public health researches while anonymizing and assessing information content. This gives rise to lower data quality and utility in epidemiological and public health contexts. Our model aims at addressing this shortcoming by highlighting the interconnectedness of research context, privacy protection, and information content.

Epidemiology, both in clinical and population based studies, deals with spotting the distribution of diseases and health problems (descriptive epidemiology), finding their causes (analytical epidemiology), and determining suitable methods for their diagnosis, treatment, and prevention (experimental epidemiology) (Translated [94]). Privacy protection affects epidemiological studies in different ways. The first effect comes from regulations that limit data usage , especially when unfamiliar types of uses emerge, such as secondary uses of health data and joining of multi-organizational data. One example is how EU directives enforce maximum level of confidentiality levels in order to cover future cross border flow of data, which adversely impacts epidemiological research [95]. One specific example is how EU data protection directive in Estonia prohibits the linkage of data of cancer registry and death certificates, leading to biased results exhibiting higher survival rates of cancer than expected [96]. Regulatory restrictions might be obvious to the researcher who accesses the anonymized dataset. Other types of data quality issues are considered as an internal feature of the algorithms, hence, are concealed from epidemiologists and public health researchers.

The second factor is how privacy protection methods might change the nature of epidemiological research. k-anonymization achieved through clustering and generalization of quasi-identifiers tends to group individual records into semi-aggregated records. Such anonymity methods or aggregation applied on quasi-identifiers or sensitive attributes affect the entire nature of epidemiological method. Considering quasi-identifiers as independent and sensitive attributes as dependent variables, table 2.3 shows the relationship between granularity of these attributes and the nature of epidemiological study based on categorization of the type of epidemiological studies [97].

Table2.3: Effects of anonymity methods on epidemiological study types

Quasi-identifiers (independent)	**Sensitive attribute (dependent)**	**Study type**
Individual records	No aggregation	Individual base
Grouped	Aggregated	Ecological
Individual records and Grouping	No aggregation	Contextual

Third effect caused by privacy protection is related to confounding and selection bias. In non-experimental epidemiology a confounding factor is an independent variable that is not the focus of the study, hence called an extraneous variable [98], which leads to confounding bias. Confounding bias refers to mixing the effects of extraneous variable and independent variables on the effects under study [99, 100], or the dependent variables. In other words, an unknown and uncontrolled confounding factor might have a correlation with the dependent variable or be the cause of the effect under study. The suppression of attributes or high levels of grouping, such as generalization and aggregation, makes it more difficult to control the possible confounding factors. Since this bias is caused by study design rather than by chance, it gives rise to selection bias [101]. Secondary data are disclosed to epidemiologists without any details about the methods, leading to uncertainty about the nature of the study. There are, to the knowledge of the author, no privacy protection methods that inform researchers about excluded attributes, levels of manipulation, or any epidemiologically important information about attribute characteristics. This leads to inability in determining and controlling the confounding factors. Also none of the methods consider the general requirements of the epidemiologists into anonymization process, a shortcoming that hinders the controlling of confounding factors. The 'requirement of researchers' here refers to parameters and context, based on evidence, that would affect the anonymization outcome.

In this sense, our study recommends an anonymization approach based on an Evidence-

Based Public Health (EBPH). EBPH is defined as "... the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of communities and populations in the domain of health protection, disease prevention, health maintenance and improvement (health promotion)" [102]. In our study, evidence does not refer to specific sources of information or "contemporaneous research findings"[102]. It instead, follows a *"practical-operational"* approach which "suggests that temporal and contextual variation heavily influence the determination of what constitutes evidence" [103]. The evidence in *"practical-operational"* approach is more about its applicability and generalizability to a context rather than its inherent quality [103]. For this reason, our study does not present the least distorting algorithm in terms of information content. It instead puts the importance on establishing the anonymization on the context based evidence.

Last but not least, in fact more importantly, the fourth effect is the loss of information content due to privacy protection. It is crucially important and the basis of all algorithms to assess the quality of their methods and resulted datasets. Information loss due to privacy protection is a potentially useful measurement that can be used to objectively ascertain the applicability of dataset to epidemiological contexts. It can also be used to compare different methods and datasets. All the anonymization algorithms mentioned in previous sections use an information content measurement to quantify the utility or quality of the resulted dataset.

Such metrics take several parameters into account including the dynamics of generalization, semantic distance between data points, clustering based distances. A classical example is Precision Metric [7], which calculates information content by the parameters of generalization. These parameters are record based generalization level, maximum height of the generalization hierarchy for all quasi-identifiers, and the number of records and quasi-identifiers. Its is to choose the minimally distorted dataset from all k-anonymized results [7, 48]. It however does not include the semantic distance between different equivalent classes. Another example in clustering based anonymization is the radius of each cluster in *r-Gather* problem [77] as an indicator of data quality. *k-member* algorithm uses the distance metric between clusters, taking into account multi-type variables, as the data quality measurement [74].

In terms of information content, almost all methods focus on reducing information loss as a determinant of the performance of the algorithm rather than its contribution to the epidemiological models. Decisions about parameters of the algorithms are also taken based on minimizing the overall information loss ignoring the context of the data usage. From an technical point of view, the primary goal of all these metrics is to find the information content difference between the original dataset and the outcome of a certain algorithm.

Our study addresses the issue of information loss from another perspective, which is a clustering analysis approach presented in the previous chapter. Instead of applying an algorithm and calculate its absolute information loss, we compare the anonymized and the original datasets regardless of underlying anonymization algorithm or its associated information loss metric. The critical difference of this approach is that our measurement method is applicable to different algorithms and anonymized datasets. Its primary goal is to assess the resemblance of the anonymized dataset to the original dataset. In our study we call this resemblance as 'Representability' which from an epidemiological point of view would contribute to the objective assessment of how accurately the results can be generalized to the population. The details of how representability can be incorporated in epidemiological models are beyond the scope of this study.

We applied cluster validity indices on the results of the first level of *PARV* model which is a novel approach. Clustering is an unsupervised process where the number of actual classes are not previously known [104]. Normally cluster validity indices are used to evaluate the validity of outcome subgroups of clustering and the selection of the patterns that fits the underlying data [105, 106]. It is performed based in internal criteria by assessing the quantities of data itself, externally by evaluating against a pre-specified structure, or relatively by comparing the results of the same algorithm on two different datasets [106]. None of the anonymization techniques used for k-anonymization has used cluster validity indices as an indicator of information content. In our study, the core contribution of the seconds part of *PARV* model representability where the quantitative information content is calculated using cluster validity index. In this approach, we consider the anonymized dataset as the result of a clustering process. The original dataset with all quasi-identifiers are at ground zero of generalization rarely conforms to k-anonymity yet it completely representative of the actual patterns of interest to an epidemiologist. Applying the original dataset as the basis of interpreting the cluster validity index, gives an idea as to how far the final anonymized clusters resemble the original equivalent classes.

The advantage of our method is that it can be used to analyze the results of anonymization regardless of the algorithms used. It is independent of the complexity and the parameters of the algorithm. It also can be used to compare the results of two different algorithms with a generic and unified method. It is also beneficial considering the multi-dimensionality of health datasets. The problem of multi-dimensionality in clustering is a known and well defined issue and many cluster validity indices already handle the mixed type data. It should be noted that our goal is not to minimize information content. Our aim is to conceptualize the application of cluster validity indices as a unified method to ascertain information content and representability.

2.5 Basis for the third level of *PARV* model: Dealing with ambiguities of policy, regulatory, and ethical issues

As mentioned in previous sections, the first layer of our model focuses on technical aspects of analyzing anonymized datasets with a generic method based on clustering. The second layer discusses the importance of the involvement of researcher in the anonymization process by reporting *RV* and providing the epidemiological based evidence to the first and seconds layers of the model. *PARV* model , therefore, highlights the human factor and expert review in the process of anonymizing and *RV* calculation. Also our model examines the evidence from a public health point of view, which emphasizes the mutually inclusiveness nature of evidence and context. The context surrounding the topics discussed in the first and second layers of *PARV* model covers, and is affected by, regulations, organizational priorities, nature of the informed consent (opt-in vs. opt-out), expert decisions, risk analysis results, and public health policies. As a result, our approach is to cover the first and second layers of *PARV* model with a third level that provides a policy based approach. It is a conceptual and high level abstraction of policy points and issues surrounding the anonymization and *RV* preparation for an evidence-based secondary data disclosure. It is a holistic approach stressing the interconnectedness of technical, procedural, regulatory, and policy aspects. It is especially important in Turkey, or similar countries, where public health data disclosure is a new topic which is under debate and prone to interpretations and considerable changes. Problem arises when technical algorithms are chosen without considering the context and future requirements, and ignoring the dynamic nature of public heath data and research. With the existence of countless technical and procedural options, such a holistic approach leads to sustainable privacy preserving disclosure of secondary health data.

Structured procedures are essential for answering many questions that cannot be addressed by technicians only [107]. Policies and regulations are also essential in balancing confidentiality measures and the ability to improve public health [108]. In order to effectively counter the privacy risks, it is important for public health professionals to incorporate updated policies and staff education along with technical data security and confidentiality methods [109]. Latanya Sweeney argues that the mere technical methods for privacy protection is not effective without policy [7]. One of the major shortcomings mentioned in the same study is the failure to construct a policy framework that does not depend on prior knowledge about data sources, users, and usages. Dennis Deapen argues that the balance of privacy and public health requirements, specifically cancer surveillance, requires consideration of policies, regulations, and technology [108]. The same article points out the actions needed for maintaing such balance. Examples of such actions are development of policies and technical recommendation based on the context of data usage, consultation from other industries, documentation of best practices, and establishing a board of experts to provide viewpoint on ethical, regulatory, public health, and government issues.

One of the best examples of studies that consider other non-technical issues is PIA (Privacy Impact Assessment) tool of B.I.R.O. (Best Information Through Regional Outcomes) project [107]. The goal of B.I.R.O. project is to make a shared evidence-based system for a common cross-border European infrastructure to share diabetes data across regional registries. The PIA part of the project aims at assessing the privacy risks of such cross-border information system. It emphasizes the considering of privacy risks in designing phase of the infrastructure in order to avoid breaching current and future legislations and concerns. It also aims at reaching consensus on themes such as legislation, information need, and architecture. In order to achieve this flexible viewpoint, PIA of B.I.R.O. project was performed in four main steps namely structured literature search, analysis of data flow scenarios, designing a questionnaire, and performing a Delphi analysis to spot the best architecture. The chosen architecture was "aggregation by group of patients". As a result, B.I.R.O. project stresses the expert review and available evidence for risk analysis. Our study is different from B.I.R.O. project in that it is focused on analyzing the results of risk mitigation methods independent of the justification for their implementation.

Another aspect of privacy protection is the existing regulations. Examples of widely used regulations are listed below:

- The EU Data Protection Directive, officially known as Directive 95/46/EC, is in effect since 1995 and covers the protection of individuals while processing and free movement of their personal data [110]. Although it emphasizes the privacy protection, it also encourages the free movement and accessibility of data. It indicated that "Member States shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data....Member States shall neither restrict nor prohibit the free flow of personal data between Member States for reasons connected with the protection afforded under paragraph 1." [110]

- Health Insurance Portability and Accountability Act (HIPAA), which is enacted in the USA and is in effect since 1996. It "...provides federal protections for personal health information ... and gives patients an array of rights with respect to that information. At the same time, the Privacy Rule is balanced so that it permits the disclosure of personal health information needed for patient care and other important purposes." [111].

- Data Protection Act 1998, which is in effect in UK and is "an act to make new provision for the regulation of the processing of information relating to individuals, including the obtaining, holding, use or disclosure of such information." [112].

These regulations are vague, and open to interpretation, when it comes to secondary uses of health data. HIPAA, for example, permits several categories for disclosure of health data. Two of the categories explicitly refer to public health purposes. The first allows the disclosure of limited dataset, which refers to highly de-identified data, for public health purposes expecting covered entities to rely on professional ethics and judgment in deciding such usages [113]. The second category allows disclosure of health data without notifying the patient according to national priorities, including research, public health activities, preventing and controlling of disease [113]. In a survey distributed to epidemiologist about the effects of HIPAA privacy rule on human subject research, the majority of the participants stated the problems such as variability in interpretation of the rule and no substantial improvement in privacy protection [23]. Ambiguities, confusions over, and contradictory guidance in Data Protection Act 1998, also hinders secondary uses of health data, which needs consensus within health service and academic communities [114], and must be addressed by different experts including study designers and ethics committee [115]. The effectiveness of policies based on regulations depends on the context also. For example, the estimation of likelihood of re-identification of data following the HIPAA privacy rule depends on the context and nature of data, such as real demographic characteristics [43].

Such ambiguities and subjectivity of interpretations of risks and regulations are apparent in Turkey. Our model encourages a dynamic and context-oriented approach for policy making in order to assist the organization to adapt to inevitable regulation updates in the future.

2.5.1 Public health data in Turkey: Privacy, Regulatory,and Ethical aspects

Public health informatics is an emerging field in Turkey. The most important source of census and public data in Turkey is Turkish Statistical Institute(TurkStat) [116]. TurkStat provides access to person-specific data on certain conditions, such as restrictive access in the premise of the TurkStat, upon signing protocols, and through tight review processes[36]. Documents of TurkStat, provided on their website ,states that micro-data has been de-identified by removing personally identifiable attributes [36] . No information is provided as to whether any modification similar to k-anonymity is applied in this level. TurkStat also publishes statistically summarized datasets to the public which are highly aggregated. Although a similar approach to k-anonymity is applied in statistical datasets, where one of the rules requires tagging the attributes with less that three unique values as confidential [117], Information about introduced information loss is not provided to data users. Other organizations in Turkey also provide data that can be used for secondary purposes. Turkish Institute of Family Medicine, for example, provides statistical summary of health-related data to the public [118].

From a regulatory point of view, issues surrounding the patient privacy in Turkey are argued as an essential part of patient rights. Although patient rights are conceptually universal, their practical implementation varies from one country to another [119]. In Turkey, patient privacy is mentioned in the statute of patient rights [120] which was finalized at 1998, and is still in effect. This regulation is used by many public health organizations. One example is the Turkish Institute of Public health whose regulation about patient rights is actually the same as the statute mentioned above [121, 122].

Statements about privacy in the Turkish statute of patient rights are generally related to clinical uses of data such as protecting confidentiality and informed consent. There is only one general statement that addresses the issue of using health data for research purposes. This general statements does not indicate any risks involved in possible re-identification using linkage and inferences. The following is the translation of this statement which is the last part of point 23 of the statute.

"Point 23: ... For research and educational purposes also, the patient identity can not be disclosed without consent" [120].

Privacy, from an ethical point of view, is mentioned in the deontology of medical practice in Turkey[119, 123]. It contains a general statement indicating that the patient identity should not be disclosed for reporting and publishing purposes. The following show the translation of the 4th statement of the deontology of medical practice in Turkey.

"Point 4: ... Patient identidy can not be disclosed in cases presented in medical meetings and publications." [123].

The only specific example that might be interpreted as secondary uses of health-related data is the requirement to remove patient's name from abortion reports. The 22nd statement states that the reason for decisions about abortion must be justified by at least two specialists. It also mandates the documenting of those justifications and decisions in form of reports. One copy of this report must be sent to the medical chamber with the name of the patient removed. It is, however, allowed to include the date and place where the abortion took place in the report. [123]

2.5.2 A descriptive survey on the secondary uses of health data in Turkey

We conducted an e-survey in order to elicit the problems surrounding the access to secondary health data, awareness, and policy issues. A comprehensive prevalence study with high generalizability was beyond the scope of our study. The questionnaire is presented in appendix A. The Checklist for Reporting Results of Internet E-surveys (CHERRIES), which was prepared for submitting a related paper to the Journal of Medical Ethics, is presented in appendix B.

2.5.2.1 Methods

We prepared the questions of the survey aiming at distributing it to epidemiologists, public health specialists, and medical informaticists. The Theme of the questions were as follows:

- Demand for person-specific secondary health data
- Difficulties in accessing to person-specific secondary health data
- General experience in accessing and using person-specific and aggregated data
- General awareness about privacy issues and protection in secondary uses of health data
- How policies could be effective in accessing person-specific data

The ethical aspects of our questionnaire was approved by the Research Center for Applied Ethics of the Middle East Technical University. The questionnaire was checked by four experts in the field of epidemiology, public health, and medical informatics. It also pre-tested by 7 experts and 6 students of the fields mentioned above. The e-survey was prepared in Google Docs. Informed consent was presented as the first paragraph of the questionnaire indicating the voluntary nature of the questionnaire, the purpose of the research, and the fact that no personal information will be gathered from participants. We sent the link to the e-survey to the mailing lists of health informaticists in Middle East Technical University, Epidemiologists in Hacettepe university, and a widely used mailing list among public health professionals. Health informaticists groups contained mainly students and faculty members of health informatics departments. This members of this group are mainly from engineering backgrounds. The link was not shared with individuals outside these mailing lists. The individuals participating in pre-testing were asked to refuse filling the survey. The questionnaire had 15 main questions with a total of 45 sub-questions. The questions were closed-ended with one optional open-ended box for further feedbacks. The e-survey was available from 20 Dec 2011 to 30 Jan 2012.

2.5.2.2 Results

We analyzed the results in a basic descriptive way, using cross-tabulation between the profession and the given answers, due to the limitations of e-surveys. One major limitation was uncertainty about the view,participation, and completion rates. The reason is that the exact number of people who clicked on the e-survey link, refused to fill it, or refused to submit the form are unknown due to unavailability of such information in Google docs at the time of conducting this e-survey. As a result, a full scale statistical analysis with high generalizability was not achievable. Therefore, the results of the e-survey are presented here as suggestive rather than conclusive assumptions. Among all participants which were 45 people, 33 were single out based on their reference to secondary usage of health data. Among these, 25 were public health specialists and 8 were health informaticists. Missing values for each question were handled by filling them with the mode of values for each profession group. The two main profession group used in the analysis were medical informaticists and public health professionals which also included epidemiologists.

54.5% of participant prefer to use aggregated datasets over person-specific data. 48% of the participants find it very difficult to obtain person-specific data while the difficulties of obtaining aggregated data are mentioned as moderate. When asked about ranking the difficulties in accessing, is a 5 scale ranking, the mostly chosen rank (more than 37%) was "major difficulty' for difficulties resulted from unknown missing values and low usability of aggregated datasets for research. In open-ended section, some difficulties were mentioned as "unknown methods for aggregation, exclusion of attributes necessary for our research, wrong classification and unknown preprocessing methods applied on data, and the lack of standardization if multi-organizational datasets."

One question asked about the most important barrier in accessing or using person-specific data. 42.2% of participants mentioned the 'fixed policies of data holders to release data in aggregated from' as the main barrier. Among 12% of participants who has chosen 'the concerns over legal aspects of privacy leaks' non were from medical informaticist profession.

Another question asked the participants to rank the problems that have affected their research. Among all the questions, only those with 'manageable', 'major', and 'extremely complicated' where analyzed. The problems, and the maximum percentage of participants that were as follows:

- Data fragmentation. 35.7% of all participants ranked this problem as 'manageable'. However, all of them are from public health domain. 50% of medical informaticists consider this problem as 'major'.
- Lack of standardization. 40% of all participants ranked this problem as 'major'. In (34.8% in public health group and 57.1% in medical informaticists group)
- 37.5% in public health group considers the missing data as a 'major' problem, while 42.9% in medical informaticist group considers it a 'manageable' problem.
- The problem in obtaining person-specific data is mentioned as a 'major' problem by 38.1% in public health domain. 57.1% in medical informaticist group mentioned it as 'manageable'.

- 52% of all participants have mentioned the 'concerns over the privacy of individuals' as a 'minor' problem affecting their research. Among those who rank this problem as 'major', 40%, and those who rank it as 'extremely complicated', 100% were from public health group.
- The majority (40%) had 'minor' problems regarding 'legal aspects of using individual data', among which 86.7% were from public health and 13.3% from medical informatics group.
- 'The lack of coordination between organizations' were ranked 37% both as 'manageable' and 'major'. However, among those who ranked it as 'manageable', 80% were from public health group.
- The majority of all participants (33.3%) ranked 'the lack of legislations as 'manageable', and 29.2% as 'major'. Among those who ranked it as 'major', 85.7% were from public health and 14.3% were from medical informatics group.

41.9% of all participants consider 'accessing to person-specific data' very important for conducting research, among which 69.2% were from public health group. For the 'seriousness of the privacy problems of person-specific data', on a scale of 1 to 5, 64.3% of participants ranked it 5 and 21.4% ranked it 4. 77.8% who believe that it is a very serious problem are from public health group. We asked the participants whether it is enough to remove directly identifiable attributes in order to protect privacy. In public health group, 56% considers it as an adequate measure comparing to 12.5% in medical informatics group.

Another question asked the participants to rank the effect of barriers in maintaining uniform policies for publishing secondary health data. The problems, and the maximum percentage of participants that were as follows:

- 37.5% of participants find the cost of designing and maintaining of policies as not a very important factor.
- 54.5% of the participants, among which 72.2% are in public health group, consider the unawareness about the importance of secondary uses of health data as a major barrier.
- 42.4% of participants, among which 78.6% are in public health group, consider the need for a centralized effort to design and implement the policy as an extremely serious barrier.
- 37.5% of participants, among which 66.7% are in public health group, consider prohibiting state regulations as the main barrier.

Additional comments in the last open-ended question contains statements about "the lack of standardization", and "problems arising from using different Turkish terms for the same concept".

The results suggest that the demand for individual data is higher among public health specialists. The main difficulty described in accessing individual data is the existence of strict organizational policies limiting data disclosure to only aggregated forms. Where individual data were available, partitioned data and lack of standards and coordination between organizations were mentioned as main problems in effective data usage. There is an apparent misconception in participants in public health group about the inadequacy of removal of directly identifiable attributes to protect privacy. While medical informaticists are generally neutral toward policy issues, public health specialists are more aware of policy needs and the effect of legal aspects. The majority of public health specialists suggest the need for a centralized effort to design policy and a central multi-disciplinary team to act as the liaison between organizations.

CHAPTER 3

PARV MODEL, LAYER 1 OF 3: A GENERIC CLUSTERING BASED METHOD FOR ANALYSIS OF *K*-ANONYMIZED DATASETS

3.1 Outline of *PARV* model, layer 1

The first level of *PARV* model is related to analysis of anonymized datasets for secondary uses. The focus is on disclosing person-specific dataset based on the requirements of researcher while protecting the privacy of individuals. Aggregated datasets and disclosing to unknown end users are considered as special cases in this study. The main method used for person-specific data disclosure in *PARV* model is based on k-anonymity [7, 4], which is a well-known standard method for privacy protection of individual data.

This chapter presents a generic method based on clustering to analyze the results of k-anonymized datasets in order to prepare the input necessary for the second layer of the model. Therefore, rather than presenting a new method or algorithm, this chapter examines the application of clustering outlier analysis techniques on k-anonymity results in order to calculate information loss. Being informed about objective measurements of information loss is necessary in determining the accuracy of epidemiological and public health studies.

In this chapter, set theory notations are used to present the basic idea of anonymization and our proposed concepts of clustering analysis of k-anonymized datasets and RV in chapter 4. Throughout this manuscript, we will formally describe basic and derived concepts related to k-anonymity. The notations and terminology will be different from the original definitions of these concepts as represented by Latanya Sweeney [7]. Where a concept is same as or similar to the basic concepts of k-anonymity, or have been represented with a different name in this manuscript, it will be mentioned.

This chapter presents the following concepts and issues:

- Formal description of basic concepts of k-anonymity. These concepts are same as those in original work of Latanya Sweeney [7] with different formal notions.
- Multi-step clustering analysis, which is our proposed clustering based analysis of the anonymized datasets regardless of the underlying k-anonymity based algorithm or its parameters.

- k-outlier handling approaches, which are using clustering based analysis in order to handle the records that are considered as outliers in terms of their conformity to k-anonymity restriction.

3.2 Advantages of clustering based analysis applied on k-anonymity

There are several algorithms proposed in the literature for anonymizing person-specific datasets. The focus of this study is on the algorithms based on k-anonymity. Each of these algorithms address a specific issue related to k-anonymity or its implementation in different contexts. The characteristics of dataset considerably affect the behavior of anonymization algorithms. For example p-sensitive k-anonymity algorithm [52] anonymizes datasets with homogeneous sensitive attributes across groups of identical quasi-identifiers. It however leads to high number of record deletion for heterogenous sensitive attributes. As a result, the choice of the algorithm is highly context sensitive. Additionally, organizational polices are influenced and guided by national regulations. In case of public health informatics, existing regulations do not cover the privacy issues in depth and are expected to change considerably in near future. Consequently, sticking to a certain algorithm might cause major changes in internal procedures with the growing nature of data, changing priorities, and dynamic regulations.

This chapter provides clustering analysis technique as a means to analyze the anonymized datasets regardless of the underlying algorithm. k-anonymized datasets can be viewed as the results of clustering [74]. Clustering refers to finding points in data that are similar to each other and are dissimilar to the points in other clusters[74, 124]. Records having the same values for quasi-identifiers can be considered as the points in the same cluster [74]. The advantage of this method is that clustering analysis techniques are suitable for large datasets. There are also clustering methods suitable for high-dimensional datasets. High dimensionality, in particular, makes k-anonymization complex and ineffective in protecting privacy [70]. As a result, clustering can be used as an analysis technique to make the results of the anonymization ready for bias measurement. This method can be used to minimize bias by re-anonymizing dataset, to compare the results of different methods or parameters, and to improve existing policies.

3.3 Definition of the basic concepts related to k-anonymity

This section presents the concept of anonymizing person-specific datasets based on k-anonymity algorithm proposed by Latanya Sweeney [7, 4]. The dataset shown in table 1.1 contains personal identifiable characteristics. Tables 1.2 and 1.3 show examples of de-identified person-specific and aggregated datasets respectively, based on table 1.1 . Suppose that we have a set of secondary health data belonging to a large group of people. A 'person' in this group refers to an individual whose health-related data is included in this set. It is supposed that each person is identified by a local, national or universal unique identifier.

Definition. Person: p
A person represents a unique individual whose data are gathered by a public health

registry. Each person belongs to the whole population whose secondary data is gathered. The set of all persons p is a subset of the set of "entities" as defined in original k-anonymity [7, p. 69]

Definition. Record: r
A record r is a single unit of data belonging to a person. It is possible to have records that do not belong to a real person. This concept is equivalent to the term "tuple" in original k-anonymity [7, p. 69].

- r_r is a record corresponding to a real person.
- r_u is a record referring to an unreal person.
- r_f is the set of all fabricated records of unreal individuals

Definition. Population: P
Population P is a set of records r such that:

- P^r is the set of all real records r_r of real persons in the population. P^r is the same as the set "population", which is a superset of "subjects" in [7, p. 63]
- P^u is the set of all unreal records r_u in the population. P^u is the same as the set ("universal" − "population") in [7, p. 63].

Definition. Sample Population: P^s
Is a subset of P and contains secondary health records r. Sample population P^s is the same as the set "subjects" in [7, p. 63].

Definition. Registries: Reg=$\{reg_1, reg_2, \ldots, reg_m\}$
A registry reg_i is an organization involving in collecting, storing, transforming, sharing, anonymization, joining, or disclosing of secondary health data. Reg is the collection of all the registries affected by same public health and data protection regulations. This term is equivalent to the concept of "data holder" used in [7].

Definition. Registry Sample Population: P^{rs}

P^{rs} is a subset of P^s and contains data about persons collected by a registry. Non-equivalent P^{rs} may include records belonging to the same person. This term is equivalent to the concept of "subjects" in [7, p. 63].

Definition. Dataset: D=$\{r_1, r_2, \ldots, r_d\}$
Let D_{reg_1} be the data collected and held internally by registry reg_1. D_{reg_1} is a set of d records such that $\forall\ i \leq d, r_i \in P^{rs}_{reg_i}$ and $\forall\ i, j \leq d\ and\ i \neq j \mid r_i \neq r_j$. In other words, each record r is unique and belongs to one and only one person p whose data are collected by registry reg_i. It also means that there is no duplication of records belonging to a person in a single dataset maintained by a registry. Two distinct dataset, however, may contain the record belonging to the same person. Dataset concept is the same as "table" in [7].

Definition. Attribute: A
Attributes are characteristics that are gathered about a sample population each with a

specific value changing from one person to another. Let A_D=$\{a_1, a_2, \ldots, a_n\}$ be the set of n attributes gathered about a sample population. For each $r_i \in D$, X_{r_i}=$\{x_{i1}, \ldots, x_{in}\}$ are the values associated with the attributes in A_D, such that $x_{ij} \in X_{r_i}$ is the value associated with attribute $a_j \in A_D$. With attributes defined, we can redefine a record as the following:
r_{ij}=$\{< a_j, \upsilon_j >| \; a_j \in A_D, \upsilon_j \; is \; the \; associated \; value \; of \; a_j \; for \; record \; r_i \; of \; D\}$. This term is the same in the [7, p. 69] with a different notation.

For example, table 3.1 shows a dataset gathered by several organization which contains records belonging to real persons r_r, unreal persons r_u due to missing and wrong records, and fabricated persons r_f added for privacy protection.

Table3.1: Example dataset containing records of real, unreal, and fabricated persons

set and no.	name	ssn	age	zip	other attributes	test result
r_{r1}	...	...	...	...	...	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r_{r100}	John	4537645738	23	34232	...	79
r_{r101}	Jack	3746837483	23	34232	...	83
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r_{u1}	Mike	998367736748	82	51409	...	–
r_{u2}	———	———————-	23	—-	...	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r_{f1}	XXXXX	000000000000	34	54093	...	34.5
r_{f2}	XXXXX	000000000000	23	34542	...	76.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table3.2: Example dataset D of a P^{rs} gathered by reg_1 with sample size of 9 containing only real records r_r

no.	name	ssn	age	zip	gender	test result
1	John	4537645738	23	34232	male	79
2	Jack	3746837483	23	34232	male	83
3	Peter	4434346666	23	34240	male	82
4	Jacob	4374688810	24	34232	male	89
5	Cindy	2008947880	25	51409	female	75
6	Joseph	9890093773	25	51409	male	81
7	Sam	5610298389	25	51409	male	83
8	Jack	5563777123	45	51409	male	87
9	Isabel	5563777123	24	51409	female	78

Table 3.2 is an example of a dataset D gathered by reg_1 and represents sample population P^{rs} with nine members. Each row in this dataset is a record r_r belonging to a real person identified uniquely by the combination of their name and social security numbers. Each column of this dataset is an attribute a_j belonging to A_D. Attribute 'no.' is used here to identify the records in the text and is not a real attribute to be disclosed.

3.3.1 Attribute types and basic functions

Let D be a dataset held by registry reg_i. A_D is the set of attributes of D and A_D=I_D∪Q_D∪S_D where I_D denotes identifiable attributes, Q_D denotes quasi-identifier attributes, S_D denotes sensitive attributes and $I_D \cap Q_D = \emptyset$, $I_D \cap S_D = \emptyset$, and $Q_D \cap S_D = \emptyset$.

Identifiable attibute:
Identifiable attribute, I_D is the set of attributes that can lead to direct or indirect re-identification of a person $p \in P$ or $p \in P^{rs}$. When an attribute is directly identifiable without any additional information it is the same as an 'explicit identifier' in [7, p. 72]. Examples of identifiable attributes in dataset shown in table 3.2 are 'name' and 'ssn'.

De-identification function:
f_d:$D \rightarrow D^0$, refers to the act of removing directly identifiable attributes such as name, last name, address, telephone, email, and social security number. D^0 is the de-identified dataset. For example applying f_d on dataset shown in table 3.2 removes the 'name' and 'ssn' attributes resulting in a dataset shown in table 3.3.

Table3.3: De-identified dataset $D^0 = f_d(D)$ where D is a dataset shown in table 3.2

no.	age	zip	gender	test result
1	23	34232	male	79
2	23	34232	male	83
3	23	34240	male	82
4	24	34232	male	89
5	25	51409	female	75
6	25	51409	male	81
7	25	51409	male	83
8	45	51409	male	87
9	24	51409	female	78

It is also possible to replace the value with a unique value that can be re-identify the person using a mechanism known only by registry. We call this method as 'legitimate re-identification'. The justification for and the use of this method is beyond the scope of *PARV* model.

Re-identification function:
f_r:$D^{k^*} \rightarrow P$, refers to uniquely re-identify a person p from a de-identified or anonymized dataset not by legitimate re-identification . f_r refers to re-identification, accidental leaks, or privacy attacks using the combination of non-identifiable attributes or linking them with externally available datasets. For example in dataset shown in table 3.3, the re-identification might be actualized based on a combination of 'age', 'zip', and 'gender' to link this dataset to external datasets to find persons with infrequent characteristics.

Quasi-identifier attibute:
Let $P^{rs}_{reg_i} \in P^s$ be the registry sample population of reg_i and D=$\{r_1, r_2, \ldots, r_d\}$ be a subset dataset of P^{rs} with A_D=$\{a_1, a_2, \ldots, a_n\}$ of n attributes. Let quasi-identifiers Q_D be a subset of attributes $Q_D \subset A_D$ and r^q denote the set of records with identical values for quasi-identifiers. The following is the formal definition of quasi-identifier.

$\exists P^* \subset P^{rs}, |P^*| \geq 1,\ r^q \in D \mid (f_r \circ f_d)(r^q){=}P^*$. In other words, the combination of the values of quasi-identifiers may lead to re-identification of a single person or map to a subgroup of sample population. For example in dataset shown in table 3.3, 'zip', 'age', and 'gender' attributes are quasi-identifiers. In this table, records in rows 1, 2, and 3 and in rows 6 and 7 share the same value for quasi-identifiers. In this example $P_1^*{=}\{r_1, r_2, r_3\}$ and $(f_r \circ f_d)(r^q)$ refers to three distinct records in rows 1, 2, and 3. For row 5, re-identification function applied on de-identified dataset would map to a single person in sample population P^{rs}.

3.3.2 Furthur concepts and re-definitions of k-anonymity in the context of *PARV* model

In this section, a different approach to define k-anonymity is presented. The goal of these concepts are to prepare a base to analyze k-anonymized datasets regardless of actual anonymity algorithm used.

Sensitive attibute:
Sensitive attributes $S^D \subset A_D$ where $A_D{=}\{a_1, a_2, \ldots, a_n\}$ is the set of all attributes of dataset D representing health-related data, which can be considered as dependent or independent variable in research. The following list describes the properties of sensitive attributes.

- All members of S^D are health-related attributes

- The relation between S^D and I^D is unknown to researcher and public. In other words S^D is exclusively collected by health organization or registry and its relation to the identity of persons are not to be known by third parties.

- The aim of k-anonymity based algorithms in general, and *PARV* model in particular, is to prevent inferences about any relation between a person p and his/her sensitive attribute in anonymized dataset.

- The aim of using an anonymized dataset for population research is to find a causal relationship between an element of S^D as 'effect' and an element of Q^D or another element S^D as 'cause'.

Example: The attribute 'test result' in dataset shown in table 3.2 is a sensitive attribute. The relation between the value of 'test result' in row 1 and 'John' is only known by the healthcare organization gathering this dataset or by registry reg_i that gathered it as a sample.

Definition. Sub-record
Given a record $r_i \in D$ with $X_{ri}{=}\{x_{i1}, x_{i2}, \ldots, x_{in}\}$ of associated values for n attributes of r_i, a sub-record φ_i of r_i is a set of attribute values which are subset of X_{r_i}. Sub-record $\varphi_i \subseteq r_i$ is a subset of r_i with m attributes such that $m \leq n$ and $X_{\varphi_i}{=}\{x_{ip}, \ldots, x_{iq}\}$ of associated values for m attributes such that $q{-}p{=}m$ and $X_{\varphi_i} \subseteq X_{r_i}$.

For example, assume that D^0 is de-identified dataset with 'age', 'zip', and 'gender' as quasi-identifiers and 'test result' as sensitive attribute. Let $r_i = < 23, 34232, male, 79 >$ be the associated values of attributes. $\varphi_i =< 23, male, 79 >$ is a sub-record based on r_i with associated values for <'age', 'gender', 'test result'>.

Definition. Quasi Sub-record
Let φ_i be a sub-record of $r_i \in D$. A quasi sub-record, denoted by φ_i^Q, is a sub-record with quasi-identifiers as the only attribute types. In other words, $\varphi_i^Q \subseteq r_i$ and $X_{\varphi_i^Q} = \{x_{ij} \mid x_{ij}\ is\ associated\ value\ of q \in Q_D\ and\ 1 \leq j \leq |Q_D|\}$.
For example assume that D^0 is de-identified dataset with 'age', 'zip', and 'gender' as quasi-identifiers and 'test result' as sensitive attribute. Let $r_i = < 23, 34232, male, 79 >$ be the associated of attributes. $\varphi_i =< 23, 34232, male >$ is a quasi sub-record based on r_i with associated values for <'age', 'zip', 'gender'>. Another example is $\varphi_i =< 34232, male >$ with associated values for <'zip', 'gender'>.

3.3.3 Functions used in levels 1 and 2 of *PARV* model

In addition to the functions of basic k-anonymity presented so far, *PARV* model uses some other functions as shown in the following list:

- Data gathering function, f_g:$P \rightarrow P^s$, refers to the the act of gathering secondary health data from a population P and creating a sample population P^s. f_g can be a direct gathering function collecting non-clinical data, such as public health data and census data, or already collected clinical data. Is it similar to the 'collection' function in [7].

- Local sampling function, f_l:$P \rightarrow P^{rs}$, refers to the act of gathering secondary health data by a local registry reg. For example dataset shown in table 3.2 is the result of applying f_l on population dataset, shown in table 3.1, by registry reg_1.

- Clustering based analysis function f_c^i:$D^0 \rightarrow D^*$, refers to the act of analyzing de-identified dataset $f_d(D)$=D^0 using clustering approach. This functions is the basis levels 1 and 2 of *PARV* model.

- k-anonymization function f_k^m:$D^* \rightarrow D^{k^0}$, refers to the act of anonymization of D^* using k-anonymization algorithm as the basic technique. It can be applied in different stages to achieve other k-anonymized datasets with higher k or different clustering results. D^{k^n} denotes a dataset derived from applying f_k^n on D^* for n times. D^{k^1} denotes the first dataset, derived from f_k^1, which conforms to k-anonymity in the chain of anonymization.

- Multistep cluster analysis functions f_{msc}
In the context of *PARV* model, secondary health datasets in all forms, including original, de-identified, or anonymized, are considered as clustering results. While k-anonymization can be actualized by applying f_k in one step, it is also possible to revise and re-apply f_k to change the resulted dataset D_* to meet the policy constraints or context-oriented evidence. The following notation show the chain of multi-step clustering analysis.

$$\overbrace{D \xrightarrow{f_d} D^0 \xrightarrow{f_c^1} D^1 \ldots D^{i-1} \xrightarrow{f_c^*} D^* \ldots \xrightarrow{f_k^1} D^{k^1} \ldots \xrightarrow{f_k^*} D^{k^*} \xrightarrow{f_s} D_s^{k^*} \xrightarrow{f_h}}^{f_{msc}} RV$$

The following list describes the notations in this chain:

- D is the identifiable dataset.
- f_d is the de-identification function.
- D^0 is the de-identified dataset.
- f_c^i is the clustering based analysis functions applied for the i^{th} iteration.
- D^i is the dataset resulted from f_c^i.
- f_k^j is the k-anonymization functions applied for the j^{th} iteration.
- D^{k^j} is the dataset resulted from f_k^j.
- Sensitive attribute clustering analysis function, f_s:$D^{k^*} \rightarrow D_s^{k^*}$, refers to the act of analyzing the sub-clusters of sensitive attribute in each of the clusters of D^{k^*}.
- RV is the Representability Vector[1]. .
- Representability function f_h:$D^{k^*} \rightarrow RV$, refers to the act of deriving RV from the last dataset resulted from multistep clustering analysis function.
- f_{msc} is a composite function of de-identification, clustering analysis, k-anonymization, and RV calculator functions collectively denoting the multi-step clustering analysis of *PARV* model

- The equivalent class derivation function $f_e : Q_{D^c} \rightarrow \Phi^{D^c}$ derives equivalent classes, which will be defined in section 3.3.4, from Q_{D^c}. The equivalent class derivation function $f_e : Q_{D^c} \rightarrow \Phi^{D^c}$ is a total, non-injective, and surjective function.

3.3.4 Equivalent class and k-anonymity property defined in the context of *PARV* model

This section provides the concept of equivalent class which is used by many studies based on k-anonymity. The definitions and notations however are based on the multi-step clustering analysis function f_{msc}.

Definition. Equivalent class
Let D^c=$\{r_1, r_2, \ldots, r_m\}$ be a dataset with A_D=$\{a_1, a_1, \ldots, a_n\}$ attributes, $Q_D \subset A_D$, and Q_D=$\{q_i, \ldots, q_j\}$ such that $j-i < n$. D^c is the dataset derived from multi-step clustering function f_c with at least one key, identifiable attribute or sensitive attribute. Let X_{r_l}=$\{x_{l_i}, \ldots, x_{l_j}\}$ be the associated values of Q_D for quasi sub-record $\varphi_l^Q \subseteq r_l$. Equivalent class Φ is the set of sub-records Φ=$\{\varphi_c^Q : \forall\alpha, \beta \leq c, \alpha \not= \beta, X_{\varphi_\alpha}=X_{\varphi_\beta}\}$. Dataset shown in table 3.4 shows the dataset in table 3.3 with each equivalent class colored differently. In this example quasi sub-records are records with 'age', 'zip', and 'gender' attributes and the set of equivalent classes are as follows:

Φ^{D^0}=$\{\{r_1, r_2, r_3\}, \{r_4\}, \{r_5\}, \{r_6, r_7\}, \{r_8\}, \{r_9\}\}$.

[1] Refer to chapter 4 for the second level of *PARV* model

With Φ^D=$\{\phi_1, \phi_2, \dots, \phi_\mu\}$ be the set of all equivalent classes over dataset D and $f_e : Q_D \to \Phi^D$ be the equivalent class derivation function. The followings are the properties of f_e and Φ^D:

- f_e is a total function: $\forall q_i \in Q_D, \exists \phi_j \in \Phi^D \mathit{such\ that}\ f_e(q_i)=\phi_j$
- Φ^D is a finite set: $|\Phi^D| < \mathbb{N}$.
- Φ^D is the set of all equivalent classes over D.
- Each equivalent class is a finite set: $\forall \phi_i \in \Phi^D,\ |\phi_i| < \mathbb{N}$.
- All the members of an equivalent class are identical:
  ϕ_l=$\{\varphi_1, \dots, \varphi_s\}$, $\forall\ i \neq j\ \mathit{and}\ i < s\ \mathit{and}\ j \leq s,\ \varphi_i$=$\varphi_j$.

Table3.4: $D^0 = f_d(D)$ shown with equivalent classes in different colors

no.	age	zip	gender	test result
1	23	34232	male	79
2	23	34232	male	83
3	23	34240	male	82
4	24	34232	male	89
5	25	51409	female	75
6	25	51409	male	81
7	25	51409	male	83
8	45	51409	male	87
9	24	51409	female	78

Basic k-anonymity
Given D^c and de-identified dataset D^0=$D^c - I_{D^c}$, let Φ^{D^0} be the set of all equivalent classes over D^0. In other words Φ^{D^0}=$\{\phi_1, \phi_2, \dots, \phi_\mu\}$ such that $|\phi_1| + |\phi_2| + \cdots + |\phi_\mu|$=$|\Phi^{D^0}|$. D^0 is k-anonymized if and only if $\forall \phi_i \in \Phi^{D^0}$, $|\phi_i| \geq k$.

Comforming and non-conforming eqiovalent classes
Let Φ^{D^α} be the set of equivalent classes over $D^\alpha \subseteq D^1$. Sub-sets of D^α with cardinalities less than k are called non-conforming equivalent classes denoted by Δ_Φ where: Δ_Φ=$\{\phi_i :\ \phi_i \in \Phi^{D^\alpha} \bigwedge |\phi_i| < k\}$. I also use δ_0=$|\Delta_0|$ to denote the number of non-conforming equivalent classes. Subsets of Φ^{D^α} with cardinalities equal or greater than k are called conforming equivalent classes denoted by Γ_Φ=$\{\phi_j :\ \phi_j \in \Phi^{D^\alpha} \wedge |\phi_i| \geq k\}$. I also use γ_Φ=$|\Gamma_\Phi|$ to denote the number of conforming equivalent classes.

For example, assume that k is 2 for k-anonymity. In dataset shown in table 3.4 the set of non-conforming classes is $\Delta_{\Phi D^0}$=$\{\{r_4\}, \{r_5\}, \{r_8\}, \{r_9\}\}$ and the set of conforming classes is $\Gamma_{\Phi D^0}$=$\{\{r_1, r_2, r_3\}, \{r_6, r_7\}\}$. Table 3.5 shows a 2-anonymous dataset based on dataset shown in table 3.3. Here, $\Gamma_{\Phi D^{K*}} = \{\{r_1, r_2, r_3, r_4\}, \{r_5, r_9\}, \{r_6, r_7, r_8\}\}$ is the set of all equivalent classes, which at the same time are conforming equivalent classes. Since there is no non-conforming class, this table is called a 2-anonymous table.

k^--anonymity property
The aim of k-anonymity is to reduce the possibility of successful re-identification to

$1/k$. Let D^k=$f_k(D^0)$ be a k-anonymized dataset. Having known all available related datasets, re-identification of D^k ideally maps to at least k records. In other words, $|(f_r \circ f_k)(D^0)| \leq k$ or $f_r(D^k)$=$\{p_i \in P, i \leq k\}$. k^--anonymity property states that if at least one record is removed from conforming equivalent class $\phi_j \in \Phi^{D^\alpha}$ of D^k, possibility of successful re-identification of the resulting equivalent class remains $1/k$.

Table3.5: $D^{K^*} = f_k^*(D)$ as the k-anonymized dataset with $k = 2$, based on D^0 shown in table 3.3

no.	age	zip	gender	test result
1	20-30	342**	male	79
2	20-30	342**	male	83
3	20-30	342**	male	82
4	20-30	342**	male	89
5	20-30	51409	female	75
6	20-50	51409	male	81
7	20-50	51409	male	83
8	20-50	51409	male	87
9	20-30	51409	female	78

3.4 Multi-step cluster analysis as the basis of the first layer of *PARV* model

This section describes the concept of applying clustering analysis on k-anonymous datasets. While this method can be used to anonymize datasets, the goal of applying clustering is fundamentally different. Instead of presenting a new anonymization technique, the aim of *PARV* model is to map the clustering analysis on already k-anonymized datasets. The underlying algorithm can be any of the methods based on original k-anonymity or clustering based approaches. All these underlying methods can be seen as the results of clustering as they treat equivalent classes as the unit of anonymization and counting k. These conforming equivalent classes are similar to clusters with the values of quasi-identifiers as centroids representing independent variables and the values of sensitive attributes as the dependent variable of at least k points. In other words D^k is the set of all $< x_i, y_i >$ data points where x_i represents conforming equivalent classes, each with at least k identical elements, and y_i represents sensitive attribute per record. Each conforming equivalent class is clustered around x_i with at least k members with codomain y_i.

Multi-step clustering analysis refers to applying clustering algorithms applied on original dataset along with all intermediate datasets up until D^{k*} and comparing the original clustering with the last stage clustering. Multi-step clustering analysis consists of two main steps as shown and compared to traditional anonymity methods in figure 3.1.

1. The first step is to map the clustering on original dataset to find the records that break the k-anonymity constraint. In this context, such records are considered

as outliers. As a result, the first step is to apply clustering as an outlier handling method. We call these outlier records 'k-outliers'.

2. The second step is to classify the methods used by underlying anonymity algorithm to handle k-outliers and to analyze how it would affect the information loss metics and resulted RV as the seconds step of $PARV$ model.

Given a multi-step clustering function $f_{msc}:\ D^c \rightarrow \Xi^0 \rightarrow \Xi^1 \rightarrow \cdots \rightarrow \Xi^k \rightarrow \ldots \Xi^{k^*}$, Ξ^k is the set of clusters each conforming to k-anonymity or k^--anonymity while Ξ^{k-1} includes at least one cluster which does not conform to k-anonymity.

Equivalent cluster

Let Ξ^{k^*}=$\{\xi_1^{k^*},\xi_2^{k^*},\ldots,\xi_n^{k^*}\}$ be a k-anonymized dataset. Each $\xi^* \in \Xi^*$ is a set of at least k members denoted by ξ^*=$\{\varphi_i^* : \varphi_i^* \subset \phi_j \mid \phi_j \in \Phi^{D^0}$ *and* $\forall \alpha, \beta \le i,\ \alpha \neq \beta,\ X_{\varphi_\alpha} = X_{\varphi_\alpha}\}$. We call ξ_i^* an equivalent cluster. When ξ_i^* conforms to k-anonymity or k^--property it will be called k-anonymous or k^--anonymous cluster respectively.

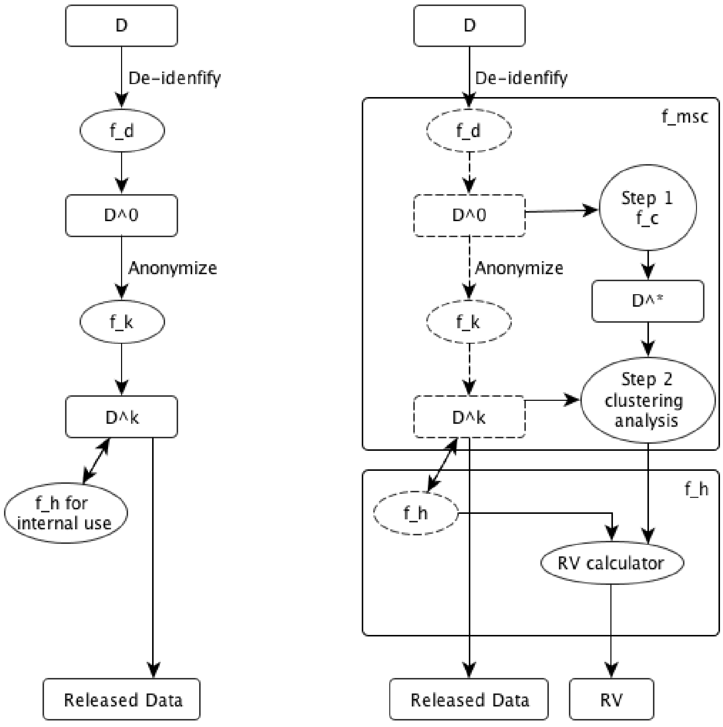

Figure 3.1: Two main steps of clustering based analysis f_{msc} applied on the results of traditional anonymity methods

3.5 First step cluster analysis of *PARV* model

First step clustering analysis aims at spotting the patterns of data that have guided the anonymization. This analysis is necessary in order to construct a basis for comparison between the original dataset and the anonymized dataset for calculating RV in the second layer of *PARV* model. The original dataset can be the same as the sample dataset. In this case the the values of non-identifiable attributes are the same as the values in the records of sample data. However in some cases, the registry might decide to change the values in order to actualize the anonymization. This decision could have internal justification based on policies or external reasons based on researcher requirements. For example, assume that the original dataset has the base values for zip code representing an area. Based on researcher requirements, it is possible to begin at a higher level, such as district or city code, if geographic data is not a health determinant in that particular study. This initial decision reduces the complexity of anonymization and affects the measurement of information content .

First step clustering analysis, denoted by function $f_c^* : \Phi^{D^0} \rightarrow \Xi^D$, divides D^1 into λ intermediate clusters. The set of clusters is denoted as $\Xi^0 {=} \{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$. The following are the properties of f_c^1:

Property 1: Given $\Phi^{D^0} {=} \{\Phi_1, \Phi_2, \ldots, \Phi_\mu\}$ as the set of equivalent classes over D^1, $\lambda \leq \mu$ and $\forall i \leq \mu$, $\Phi_i \in \xi_p^0$, $\Phi_i \in \xi_q^0 \mid \xi_p^0 {=} \xi_q^0$. In other words, an equivalent class maps to one and only one cluster.

Property 2: $\forall r, s \in \Phi_i \; and \; r \cap s {=} \emptyset$, if $(r \in \xi_p^0 \wedge s \in \xi_q^0) \Rightarrow \xi_p^0 {=} \xi_q^0$. In other words, two distinct records in an equivalent class are not divided between separate clusters.

The first step clustering analysis is performed in order to spot records and equivalent clusters that do not conform to k-anonymity which are called k-outliers in *PARV* model. f_c^* can be a composite function applied several times. The reason for this is that the starting point to analyze the clustering might not be the original data as mentioned earlier. A registry might manipulate, generalize, or change the number of records and consider the results of such manipulations as the original dataset. The choice of the original dataset is context-oriented and might change based on strategic decisions, evidence, or researcher requirements. As a result, f_c^* shows different additional properties based on the strategies, parameters,and the context under which it is implemented. There are four possible modes for the implementation of f_c^* as shown in the following list:

1. **Bijective mode**. Where each and all equivalent classes are mapped to a separate cluster with same elements.
2. **Surjective mode**. Where an initial grouping of equivalent classes are performed.
3. **Partial mode**. Where some of the equivalent classes are not considered to be disclosed.
4. **Injective mode**. Where there are clusters in co-domain whose elements are not included in any of the equivalent classes of the domain.

f_c^* in bijective mode

f_c^* is called to be in bijective mode if $f_c^* : \Phi^{D^0} \rightarrow \Xi^0$ is an injection and surjection. In other words, given $\Phi^{D^0} {=} \{\phi_1, \phi_2, \ldots, \phi_\mu\}$ and $\Xi^0 {=} \{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$, the following properties hold true:

- μ=λ.
- $\forall i < \mu : \xi^0_i$=$\phi_i$.
- $\forall i \neq j$ and $i, j \leq \mu$ we have $f^*_c(\phi_i)$$\neq$$f^*_c(\phi_j)$.

Figure 3.2 depicts f^*_c in bijective mode. This mode is the basic mode for first step cluster analysis. It is basically mapping each of the equivalent classes onto an identical cluster. This mode is suitable for strategies where the goal is to disclose the whole sample with least amount of alteration imposed on original distribution. Is is also suitable when k-anonymity is naturally satisfied due large sample size and homogeneity of dataset.

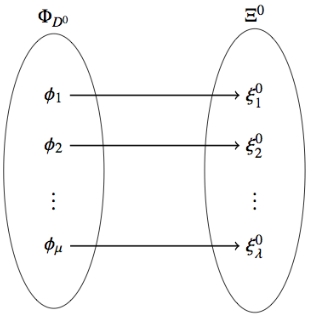

Figure 3.2: First step cluster analysis function in bijective mode

f^*_c in surjective mode

f^*_c is called to be in surjective mode when at least two equivalent classes in $domain(f^*_c)$ map to the same cluster in $range(f^*_c)$. In other words, given Φ^{D^0}=$\{\phi_1, \phi_2, \ldots, \phi_\mu\}$ and Ξ^0=$\{\xi^0_1, \xi^0_2, \ldots, \xi^0_\lambda\}$, the following properties hold true:

- $\lambda < \mu$.
- $\exists\, \xi^0_i \in \Xi^0$ where ξ^0_i=$\{\phi_p, \ldots, \phi_q\}$ and $q - p \geq 2$.
- $\exists\, \xi^0_i \in \Xi^0 \mid \xi^0_i = f^*_c(\phi_\alpha) = f^*_c(\phi_\beta)$ where $\phi_p \not\models \phi_q$.

Figure 3.3 depicts f^*_c in surjective mode. This mode is suitable for strategies where the goal is to disclose maximum number of records with a tradeoff of losing information content of unified clusters.

f_c^* in partial mode

f_c^* is called to be in partial mode when at least one of the equivalent classes in $domain(f_c^*)$ do not map to any clusters in $range(f_c^*)$. Given Φ^{D^0}=$\{\phi_1, \phi_2, \ldots, \phi_\mu\}$ and Ξ^0=$\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$, the following properties hold true:

- $\exists \phi_i \in \Phi^{D^0} \mid f_c^*(\phi_i) \notin \Xi^0$.
- $\exists\ \xi_i^0 \in \Xi^0 \mid \xi_i^0 = f_c^*(\phi_p) = f_c^*(\phi_q)$ where $\phi_p \not= \phi_q$

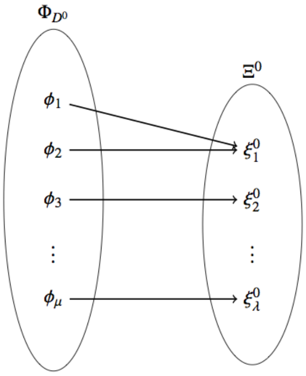

Figure 3.3: First step cluster analysis function in surjective mode

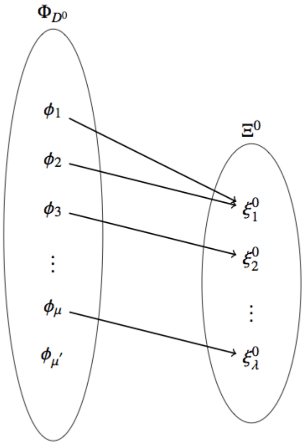

Figure 3.4: First step cluster analysis function in partial mode

Figure 3.4 depicts f_c^* in partial mode. This mode is suitable for strategies where the goal is to conceal some of the clusters in original data due to sensitivity or extreme low conformity to k-anonymity with a tradeoff of reducing the sample size.

f_c^* in non-surjective (injective or non-injective) mode

f_c^* is called to be in non-surjective mode when at least one of the equivalent clusters in $range(f_c^*)$ do not have a pre-image in $domain(f_c^*)$. In other words, given Φ^{D^0}=$\{\phi_1, \phi_2, \ldots, \phi_\mu\}$ and Ξ^0=$\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$, the following property hold true:

- $\mu < \lambda$
- $\forall \phi_i \in \Phi^{D^0}, \exists \xi_j^0 \in \Xi^0 \mid \xi_i^0 \neq f(\phi_i)$
- f_c^* can be injective or non-injective.

Figure 3.5 depicts f_c^* in non-surjective mode. This mode is suitable for strategies where the goal is to disclose a minimum number of records with the tradeoff of introducing records belonging to non-real persons or records belonging to persons not included in original sample.

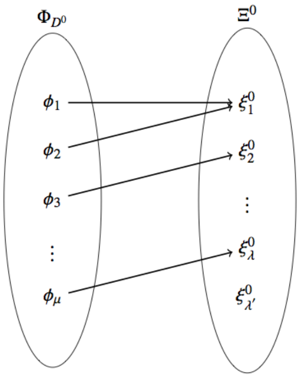

Figure 3.5: First step cluster analysis function in non-surjective mode

3.5.1 *k*-outlier as the outcome of the first step cluster analysis

The results of the first step clustering analysis results depend on the strategies surrounding the disclosure of dataset. These strategic points define the modes in which first step clustering analysis is applied, which in turn, define and affect the results. The most crucial decision is minimum k. The last procedures of fist step clustering analysis is to identity the clusters that break the constraints of the policy regarding privacy. These are called non-conforming clusters where the constraints in *PARV* model is k-anonymity.

Let Ξ^0=$\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$ be the first step clustering analysis result. The following is the definition of conforming and non-conforming clusters:

- A cluster ξ_i^0 is called a conforming cluster if all its subsets are k-anonymous equivalent classes. In other words, ξ_a^0=$\{\phi_j^0 \mid \phi_j^0 \in \Phi^{D^0}$ *and* $|\phi_j^0| \geq k\}$. The set of all conforming equivalent clusters is denoted by $\Gamma_{\xi}=\{\xi_j^0 : \xi_j \in \Xi^0 \ and \ |\xi_j^0| \geq k\}$. The number of conforming clusters is denoted by $\gamma_{\xi}=|\Gamma_{\xi}|$. I also use the term k-anonymous clusters to refer to a conforming cluster. It should be noted that a conforming cluster also have the k^{-}-anonymity property.

- A cluster ξ_i^0 is called a non-conforming cluster if at least one of its subsets do not satisfy k-anonymity property. In other words $\xi_a^0=\{\phi_j^0 \mid \phi_j^0 \in \Phi^{d^0} \ and \ \exists \phi_j^0 \in \Phi^{D^0} \ such \ that \ |\phi_j^0| < k\}$. The set of non-conforming equivalent clusters is denoted by $\Delta_{\xi}=\{\xi_j^0 : \xi_j \in \Xi^0 \ and \ |\xi_j^0| < k\}$.The number of non-conforming clusters is denoted by $\delta_{\xi}=|\Delta_{\xi}|$.

k-outlier. Privacy wise, non-conforming clusters apparently pose threat to the privacy of individuals. From a clustering point of view, these clusters exhibit a different pattern comparing to conforming clusters. One of the applications of clustering is to spot outliers in data. Outliers are data points that exhibit different pattern from the distribution of the majority of data points. Although records contained in a non-conforming cluster are not extreme data points or outliers in terms of data distribution, their collection in a non-conforming cluster characterizes them as outlier patterns. Traditional pattern recognition techniques consider outlier as noise and aim at removing them from dataset. Nevertheless, outliers can be significant in terms of revealing new patterns or threats to an individual [75]. From a clustering point of view, all k-anonymity methods aim at handling non-conforming clusters which will be called k-outliers in *PARV* model. The approach used in *PARV* model is not to detect extreme clusters and remove them. The aim is to use clustering techniques to analyze the outlier handling techniques of underlying anonymity algorithms as a generic method for preparing inputs for cluster validity indices in the second level of *PARV* model.

3.6 Second step of the multi-step cluster analysis: k-outlier handling

The second phase of the multi-step clustering analysis function is to analyze k-outlier handling approach applied by underlying algorithm. This step provides inputs for cluster validity indices used in the second level of *PARV* model. All the methods and algorithms based on k-anonymity have different approaches to handle k-outliers. This section provided the most general and top-level approaches based on hard clustering. In hard clustering, a datapoint belongs to one and only one cluster while in soft clustering a data point belongs to more than one cluster with different probabilities [125].

Given $\Xi^0=f_c^*(\Phi^{D^0})$ as the outcome of first step clustering analysis, the second step is to apply the function $f_k^* : \Xi^0 \rightarrow \Xi^*$. It is the function that maps k-outlier clusters to conforming k-anonymous or k^{-}-anonymous clusters. The following is the property of the second step analysis function:

- The final results of analyzing the outcome of first step clustering is k-anonymous. In other words, given $\Xi^0=\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$ and $\Xi^*=\{\xi_1^*, \xi_2^*, \ldots, \xi_{\lambda'}^*\}$:
  - $\exists \ \xi_i^0 \in \Xi^0$ such that $|\xi_i^0| < k$ and
  - $\nexists \ \xi_j^* \in \Xi^*$ such that $|\xi_j^*| < k$.

This section describes approaches for handling k-outlier clusters enforced by underlying algorithms in order to render the dataset k-anonymous. In reality these approaches might be used together.

3.6.1 Deletion approach

In deletion approach, k-outlier clusters are removed from the dataset. This approach is suitable for datasets with low number of small sized k-outlier clusters where their inter-cluster distances from other conforming clusters are high. For homogenous datasets, this approach causes large number of record removals. The advantage of this approach is its simplicity and minimum amount of alteration of the conforming clusters. The disadvantage of this method is reduction in sample size. Another disadvantage is the exclusion of extreme values that can be of interest to the researcher. Figure 3.6 depicts the delete approach for k-outlier handling.

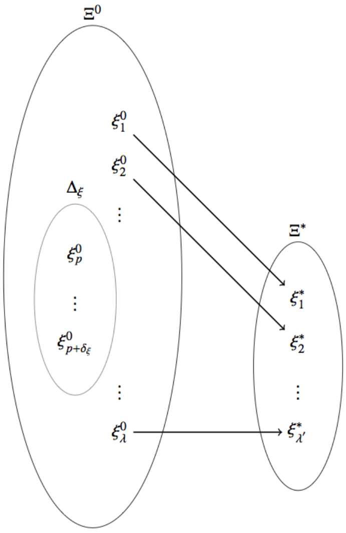

Figure 3.6: Deletion approach $f_k^* : \Xi'^0 \rightarrow \Xi^*$ where Ξ'^0=$(\Xi^0 - \Delta_\xi)$

Given Ξ^0=$\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$ and Δ_ξ=$\{\xi_p^0, \ldots, \xi_{p+\delta_\xi}^0\}$ as the set of k-outliers where $\Delta_\xi \subset \Xi^0$ and δ_ξ=$|\Delta_\xi|$, deletion approach function is defined as $f_k^* : \Xi'^0 \rightarrow \Xi^*$ where Ξ'^0=$(\Xi^0 - \Delta_\xi)$.In an exclusive mode analysis, each $\xi_i \in \Delta_\xi$ is a singleton cluster. f_ξ^* with Ξ'^0 as its domain is a bijective function. Let D denote the original dataset gathered from sample population P^{rs} and let $|D|$ denote the sample size. Assume that $|D| \geq S_{min}$, where S_{min} denotes the minimum sample size required. The following shows the sample size and the condition for applying deletion approach:

- $|D|$=$|\xi_1^0| + |\xi_2^0| + \cdots + |\xi_\lambda^0|$.
- $|\Xi^0|$=$|D| - (|\xi_p^0| + \cdots + |\xi_{p+\delta_\xi}^0|)$.
- Deletion approach is applicable if $|\Xi^0| \geq S_{min}$.

Table 3.6 shows the results of applying deletion approach on table 3.4 where k is 2. As a result of applying deletion approach, the records at rows 3, 4, 5, 8, and 9 are deleted as they are 2-outliers. This example shows how deletion approach can cause major distortion and information loss in datasets with heterogeneous values of quasi-identifiers and large number of k-outliers.

Table3.6: Deletion approach applied on table 3.4 with $k = 2$

no.	age	zip	gender	test result
1	23	34232	male	79
2	23	34232	male	83
6	25	51409	male	81
7	25	51409	male	83

3.6.2 Generalization approach

Generalization approach is the basis of original k-anonymity algorithm [7, 8, 48]. Generalization refers to the mapping of the values of quasi-identifiers to less specific and informative forms. The following list shows the properties of generalization according to [48, 8] with different notations:

- Each quasi-identifier attribute, denoted here by a, in the original dataset is in ground domain (a^0 to denote ground level)with values that are specific at their highest levels.
- An attribute in a^i can be mapped to a more general and less specific value in domain a^{i+1}. This mapping is a partial order and each value in a^i has only one associated value in a^{i+1}. The mapping between two domains are shown as $a^i \overset{a}{\rightarrow}^{i+1}$.
- Generalization is a function $a^i \xrightarrow{f_{i+1}^G} a^{i+1}$ which maps the whole values in a^i to generalized values in a^{i+1}.
- This mapping is called Domain Generalization Hierarchy with a maximal level a^{max} which is a singleton.
- Generalization hierarchy applied on values in each domain instead of the whole domain is called Value Generalization Hierarchy
- There is an additional level on top of the maximal level a^{max} where the values are completely concealed or become semantically non-informative. This action is called suppression and its level is denoted as a^s.

Given the domain generalization hierarchy $a^0 \xrightarrow{f_1^G} a^1 \xrightarrow{f_2^G} a^2 \xrightarrow{f_3^G} \dots \xrightarrow{f_{max}^G} a^{max} \xrightarrow{f^s} a^s$ where a_i^0 is a quasi-identifier attributes of D^0, f^G is the generalization function and f^s is the suppression function. The associated values of a^0 are in ground domain and have the maximum information content. Associated values of a^1=$f_1^G(a^0)$ represent less detailed values comparing to a^0. a^s=$f^s(a^{max})$ has no information content about quasi-identifier and is the same as completely obscuring the values of quasi-identifiers. At some point along the the generalization hierarchy, more than one equivalent classes at level i may become identical at level $i + 1$. In other words given $F_G=\{f_1^G, f_2^G, \ldots, f^s\}$ and $a^i=f_i^G(a^{i-1})$, $\exists \phi_\alpha, \phi_\beta \in a^i, \phi_\alpha \neq \phi\beta \mid f_{i+1}^G(\phi_\alpha) \in a^{i+1}, f_{i+1}^G(\phi_\beta) \in a^{i+1}$, and $f_{i+1}^G(\phi_\alpha)=f_{i+1}^G(\phi_\beta)$. This is similar to agglomerative hierarchical clustering where a set of n equivalent classes are clustered in a set containing less than n elements, the hierarchy is non-overlapping, and is pair-group where at each iteration at least two distinct clusters are grouped in the same cluster [126].

The following list shows an example of domain generalization hierarchy of age, and gender attributes and value generalization hierarchy for zip attribute as shown in table 3.4.

- $\{1, 2, \ldots, 100\} \xrightarrow{f_{age1}^G} \{\text{1-10}, \ldots, \text{91-100}\} \xrightarrow{f_{age2}^G} \{\text{1-25}, \text{26-50}, \text{51-75}, \text{76-100}\} \xrightarrow{f_{age*}^s} \{*\}$

- $\{male, female\} \xrightarrow{f_{gender1}^G} \{unknown\} \xrightarrow{f_{gender*}^s} \{*\}$

- $\text{23232} \xrightarrow{f_{zip1}^G} \text{23230} \xrightarrow{f_{zip2}^G} \text{23200} \xrightarrow{f_{zip3}^G} \text{23000} \xrightarrow{f_{zip4}^G} \text{20000} \xrightarrow{f_{zip*}^s} *$

The details of generalization, and its effect on the values of the attribute, are based on human intervention and organizational policies. For example, the age attribute can be generalized with 5 years interval or 10 years interval as shown in the previous example. The maximal level also depends on the context. For example, the maximal level for gender attribute in previous example in 'unknown' which in terms of information content is equivalent to '*'. Ground level also depend on organizational policies. For example, a registry might set the ground level for zip codes beginning with 232 as 2320.

Table3.7: Two examples of generalization appraoch applied on table 3.4 with $k = 2$

no.	age	zip	gender	test result
1	1-50	34000	unknown	79
2	1-50	34000	unknown	83
3	1-50	34000	unknown	82
4	1-50	34000	unknown	89
5	1-50	51000	unknown	75
6	1-50	51000	unknown	81
7	1-50	51000	unknown	83
8	1-50	51000	unknown	87
9	1-50	51000	unknown	78

no.	age	zip	gender	test result
1	21-30	34200	male	79
2	21-30	34200	male	83
3	21-30	34200	male	82
4	21-30	34200	male	89
5	21-30	51409	female	75
6	*	51409	male	81
7	*	51409	male	83
8	*	51409	male	87
9	21-30	51409	female	78

Table 3.7 shows two examples of generalization applied on dataset shown in table 3.4 with $k = 2$. Based on the number of quasi-identifiers and domain generalization hierarchy, several k-anonymous datasets can be derived. MinGen algorithm [7, 8] aims at finding the least distorted dataset from the set of all possible k-anonymous datasets. The methods used in MinGen method will be discussed in chapter 4.

Given Δ_{ξ}={$\xi^0_p, \ldots, \xi^0_{p+\delta\xi}$} as the set of k-outliers, several strategies can be applied in generalization approach for k-outlier handling as shown in the following list and in figure 3.7

(a) All conforming and non-conforming clusters are generalized into a single cluster. In other words $\forall \xi^0_i, \xi^0_j \in \Xi^0, f^G(\xi^0_i)$=$f^G(\xi^0_i)$.

(b) Only k-outliers are generalized until no k-outlier remains. In this strategy, the resulting clusters contain at least two generalized k-outliers, where non of the k-outliers are split between different clusters.

(c) Each k-outlier is aggregated and generalized with semantically closest conforming cluster.

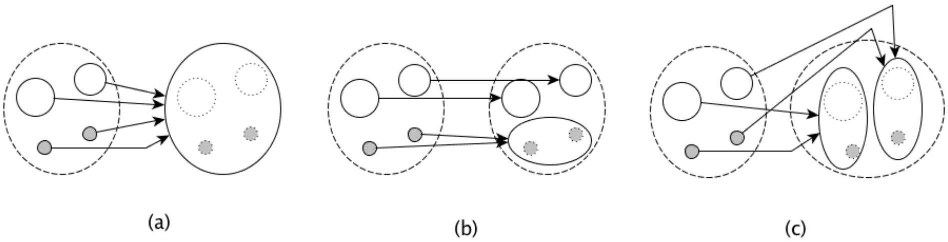

Figure 3.7: Strategies of generalization approach for k-outier handling

As a result, $f^*_k : \Xi^0 \to \Xi^*$ in generalization approach is a surjective function where $\lambda^{'} < \lambda$ as shown in figure 3.8.

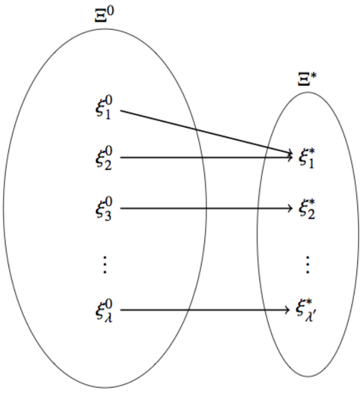

Figure 3.8: Generalization approach $f^*_k : \Xi^0 \to \Xi^*$

3.6.3 Merging approach

In merging approach, each k-outlier cluster is aggregated with the closest conforming or non-conforming cluster with the value of quasi-identifiers replaced by those of the host cluster. Figure 3.9 depicts the merging approach. Given Δ_{ξ^0}=$\{\xi^0_p, \ldots, \xi^0_{p+\delta_{\xi^0}}\}$ and $Dist(\xi^0_i, \xi^0_j)$ denoting the distance between ξ^0_i and ξ^0_j clusters, we have the following properties in merging approach:

- $\forall\xi^0_\alpha \in \Delta_{\xi^0}, \exists\xi^0_i \in \Xi^0$such that$Dist(\xi^0_\alpha, \xi^0_i)$ is minimal, which is called the closest cluster.
- Given ξ^0_α as the k-outlier cluster and ξ^0_i as the closest cluster, all the associated values $X_{\xi^0_\alpha}$ of quasi sub-records belonging to ξ^0_α are replaced by $X_{\xi^0_i}$
- If the host cluster ξ^0_i is a k-outlier and $|\xi^0_\alpha \cup \xi^0_i| < k$, the previous step is repeated for $(\xi^0_\alpha \cup \xi^0_i)$ and the next closest cluster until the resulted cluster becomes a conforming cluster.
- f^*_k in merging approach is a surjective function.
- $\lambda < \lambda^{'}$

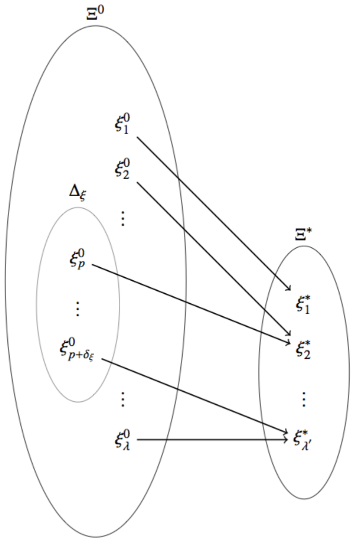

Figure 3.9: Merging approach $f^*_k : \Xi^0 \rightarrow \Xi^*$

Table 3.8 shows an example of the merging approach applied on dataset shown in table 3.4. The result of merging approach depends on the distance metric, included quasi-identifiers, and semantics of distance measurement.

3.6.4 Fabricationa approach

In fabrication approach, some records not included in the original sample are added to k-outliers. These added records may belong to real persons not included in the sampled population or unreal record. While this adds bias, it might be preferable to other approaches where the number of k-outlier clusters are few and their cardinality is very close to k. The difficulty with this approach is the value of sensitive attribute for records of unreal persons. These records can be considered as missing values and any methods for handling them can be applied to fill the sensitive attributes.

Table3.8: Merging approach applied on dataset 3.4 with $k = 2$

no.	age	zip	gender	test result
1	23	34232	male	79
2	23	34232	male	83
3	23	34232	male	82
4	23	34232	male	89
5	25	51409	female	75
6	25	51409	male	81
7	25	51409	male	83
8	25	51409	male	87
9	25	51409	female	78

Given $\Delta_{\xi^0}=\{\xi^0_p, \dots, \xi^0_{p+\delta_{\xi^0}}\}$ and $\xi^0_i \in \Delta_{\xi^0}$ and $\xi^0_i=\phi_i$, assume that $b=|\phi_i|$ and $\upsilon=\langle\phi_i\rangle$ denoting the values of sub quasi-records of ϕ_i. Also assume that $\xi^*_j=f^*_k(\xi^0_i)$ where f^*_k is the k-outlier handling function in fabrication approach. In order to solve the problem of non-conformity, $(k - b)$ records are added to ξ^*_j with υ as the value of their sub quasi-identifiers. In other words:

- $\xi^*_n=\{\overline{\phi} \mid \overline{\phi}=\phi_i + \phi^{'}_i\}$
- $\phi^{'}_i=\{\varphi^{'}_1, \dots, \varphi^{'}_{k-b}\}$
- $\forall\ \varphi^{'} \in \phi^{'}_i : \varphi^{'}_i=\langle\phi_i\rangle$
- $\forall\phi^{'}_i \in \xi^*_n \mid \phi^{'}_i \in ((P^s - P^{rs}) \cup P^u)$
- $\forall\phi^{'}_i \in \xi^*_n, S^D = \{V\}$ where V is mean, median, or mode of one of the following options
  - sensitive attribute of the receiving k-outlier cluster.
  - sensitive attribute of the whole dataset .
  - sensitive attribute of the sample population of the registry in cases where $D \subset P^{rs}$ and $D \neq P^{rs}$.
  - sensitive attribute of larger population $(P^s - P^{rs})$.
  - the value accepted as the central value of the whole population based on evidence.
  - sensitive attribute of the larger sample population.

CHAPTER 4

PARV MODEL, LAYER 2 OF 3: REPRESENTABILITY VECTOR (*RV*) AND THE ISSUES OF BIAS, RESEARCH CONTEXT, AND EVIDENCE

4.1 Outline of the seconds level of *PARV* model

This chapter discusses our proposed concept of Representability Vector (*RV*) which provides a means to report the details of information loss to researcher. The information loss is an inevitable outcome of applying anonymization on dataset which is analyzed in the first layer *PARV* model in chapter 3. *RV* is a vector containing several fields of numeric and non-numeric elements which can be used in assessing the degree of generalizability of results to the whole population and the accuracy of research. It can be used as a mechanism to provide feedback to researcher about the applicability of data to their research methods and tailoring the disclosed data according to their requirements. It can also be used to collect researcher requirements, which reflect the research context and related evidence.

As mentioned in chapter 3, anonymized datasets can be viewed as clustering results. Therefore, the first layer of *PARV* model, aims at applying clustering based k-outlier analysis on anonymized datasets regardless of underlying algorithm. This chapter focuses on applying cluster validity indices on the results of the first layer, as the second step of *PARV* model, in order to calculate the information loss. The advantage of applying cluster validity indices is that they can be used as a generic approach to calculate information loss and the resemblance of the anonymized data to the original data. As dataset is altered for anonymity, its resemblance will be deviated from original dataset, decreasing its accuracy in generalizing the results to the whole population. Regardless of the internal information loss metric used by underlying technical anonymization algorithm, cluster validity indices provide a unified method for *RV* calculation.

4.2 k-anonymization of epidemiological variables

The 'generalization' term refers to two distinct and completely different concepts in this manuscript. The first usage related the domain of k-anonymization which refers to presenting the values of quasi-identifier with less detail in order to increase the number of records with similar values for quasi-identifiers [1]. The second meaning of 'generalization' relates to the epidemiology domain. The aim of the epidemiological research is to derive conclusions about population from data derived from a sample of that population. However, it is impossible to test a new treatment or method on the whole population. The reason is that it is practically impossible, risky, and costly to reach to, and apply the new method on, everyone. Researchers instead apply their new methods on a sample population and try to make inferences about the whole population in studies such as case-control and cohort. In many of these studies the aim is to find a meaningful and statistically significant relationship between a cause and effect. In the context of epidemiology, the 'effect' usually refers to a health-related situation, or hazard, and the 'cause' is an agent, lifestyle, or socioeconomic characteristic that might have a link with the effect. The cause is called independent variable and the effect is called dependent variable. Example of an independent variable is a bacterial agent and the dependent variable is a certain disease under study. Even if a statistically meaningful causal relationship is found between dependent and independent variables, it still remains a question as to whether the researchers can infer the same relationship between a dependent variable in population and independent variable under study. The ability to extend the results of a research conducted on a sample to the whole population, or a larger group, is called 'generalizability'. It is " ...characterized by the relevance of a study's results when applied to a larger population" [127]. The accuracy of generalization would in turn affect the public health protocols or clinical procedures regarding the subject of the study. Tables 4.2 and 4.3 show datasets with suppressed quasi-identifiers as independent variables and aggregated and non-aggregated health-related attribute as dependent variable. When dependent attributes are aggregated, the study becomes an ecological study [128]. The results of these studies are suggestive rather than conclusive. In other words, ecological studies are not generalizable to the whole population due to ecological fallacy. Causal relationship can only be studied with non-aggregated values of dependent variable.

In the context of this study, the assumption is that the sensitive attribute, which is the health-related data gathered or collected by a registry, can act both as dependent and independent variables. It is also assumed that quasi-identifiers are independent variables [2]. The researcher might be interested in the causal relationship between a sensitive attribute and a quasi-identifier, or another sensitive attribute as, shown in figure 4.1.

There are three main outcomes of anonymization, based on the alteration of quasi-identifiers and sensitive attributes. In all of these cases, it is assumed that quasi-identifiers are not generalized.

- Case 1. Where sensitive attribute is not changed. In this case the data is in person-specific form and dependent variables are not aggregated. This dataset can be used for causal relationship studies such as case-control or cohort.

- Case 2. Where sensitive attribute is aggregated for some equivalent cluster. In this case the equivalent clusters with aggregated sensitive attributes can only be used for ecological studies.

[1] Refer to section 3.6.2 for details.

[2] Refer to 3.3.1 for more details about attribite types.

- Case 3. Where sensitive attributes is aggregated for all equivalent clusters. In this case dataset can only be used for ecological studies.

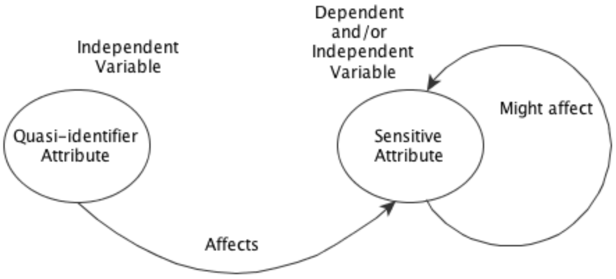
Independent Variable
Dependent and/or Independent Variable
Quasi-identifier Attribute
Sensitive Attribute
Might affect
Affects

Figure 4.1: Dependent and independent variables and attiribute types of anonymization algorithms

Apart from alteration of sensitive attributes, generalization of the values of quasi-identifiers also affect the ability to extend the results of the study to the population. For example, assume that age attribute, as a quasi-identifier, is generalized to 25 years intervals. In this example, it is not possible to accurately generalize the causal relationship between the dependent variable and the exact values of age attribute to the whole population. Tables 4.1-4.5 show different cases for quasi-identifiers and sensitive attribute alteration.

Another issue is the choice of quasi-identifiers to be included in the disclosed dataset. This problem arises when a quasi-identifier is a hidden or possible confounding factor[3]. For such quasi-identifiers, suppression or high levels of generalization hinder the determining it as a definite confounding factor.

4.3 Representability concept as a measurent for the resemblance of anomymized and sample records

Figure 4.2 shows the spectrum of manipulation applied on dependent and independent variables at attribute level in anonymized datasets. It is assumed that these datasets conform to k-anonymity. It is also assumed that no deletion is applied at record level. At extreme points of the spectrum, there are either person-specific or completely aggregated datasets with altered quasi-identifiers and sensitive attributes. Each record in completely person-specific data, represents dependent and independent variables, and epidemiological characteristics of a person without revealing their identity. As the dataset moves toward the higher manipulations of dependent and independent variables, each record becomes less informative about and less representative of the data belonging to an individual.

Table 4.1 shows a simple dataset in its original form. 'Age', 'zip code', and 'gender' attributes are quasi-identifiers and 'condition' is a numerical sensitive attribute. It is supposed that this dataset will be used in epidemiological studies. Also assume that the

[3] Refer to section 2.4 for more details about confounding factor

minimum k for k-anonymity is 2. This dataset does not satisfy 2-anonymity, however it contains original data gathered from the sample population P^s (3.3). This dataset can be used for causal relationship analysis in case-control and cohort studies where the dependent variable is 'condition' attribute and independent variable is the combination of the quasi-identifiers 'age, 'zip', and 'gender'. Since there is no aggregation applied on sensitive attribute, this dataset can be used for non-ecological studies making it generalizable to a larger population if any meaningful causal-relationship is found.

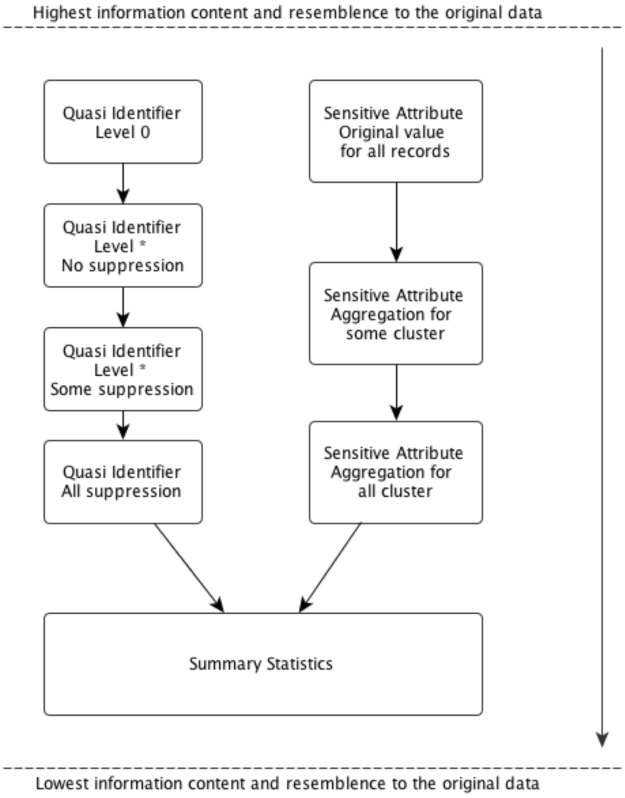
Highest information content and resemblence to the original data
Quasi Identifier Level 0
Quasi Identifier Level * No suppression
Quasi Identifier Level * Some suppression
Quasi Identifier All suppression
Sensitive Attribute Original value for all records
Sensitive Attribute Aggregation for some cluster
Sensitive Attribute Aggregation for all cluster
Summary Statistics
Lowest information content and resemblence to the original data

Figure 4.2: Spectrum of alterations applied on dependent and independent variables

Table4.1: Person-specific original dataset not conforming to 2-anonymity

age	zip	gender	condition
18	210	M	12
20	210	M	18
34	210	M	14
21	340	F	15
31	340	F	16

Another extreme, shown in Table 4.2, is when all the values of quasi-identifiers are suppressed and the sensitive attribute is aggregated for all records. This is the same as disclosing the frequency of individuals with a common aggregated value representing the central tendency of sensitive attribute. This central tendency can be the mean, mode, median, or weighted mean to count a few. This dataset can only be used for ecological studies and is not generalizable to larger population. Table 4.3 shows another extreme example where quasi-identifiers are suppressed yet the sensitive attribute is presented in original value. Although it is person-specific, this table can not be used for causal relationship as the values of independent values are completely concealed, preventing definite determination of confounding factors.

Table 4.4 shows a more realistic example of a 2-anonymous dataset where quasi-identifiers are generalized and sensitive attributes are in non-aggregated form. This dataset is suitable for causal relationship studies with less information content about the independent variables. Information content , in this context, refers to the quasi-identifiers and sensitive attribute without leaking identity information.

Table4.2: Dataset with suppressed quasi-dientifiers and aggregated sensitive attributes

age	zip	gender	condition
*	*	*	14.5
*	*	*	14.5
*	*	*	14.5
*	*	*	14.5
*	*	*	14.5

Table4.3: Suppressed dataset with non-aggregated dependent variable

age	zip	gender	condition
*	*	*	12
*	*	*	18
*	*	*	14
*	*	*	15
*	*	*	16

Table4.4: 2-anonymous dataset with non-aggregated sensitive attributes

age	zip	gender	condition
10-50	2**	***	12
10-50	2**	***	18
10-50	2**	***	14
20-40	34*	F	15
20-40	34*	F	16

Table 4.5 shows another example where the sensitive attribute has been aggregated for some of equivalent clusters. Such clusters can only be used for ecological conclusions. The problem of generalization of the quasi-identifiers for non-aggregated records and its effect on causal inferences remains the same as in Table 4.4. Table 4.1 represents the original data gathered from individuals. Table 4.2 at the other hand is not representable of any particular individual despite the fact that the value of sensitive attribute belong to a real person. We defined the extend of resemblance of a record in anonymized data to that in sample as 'representability'. It provides information as to how much an anonymized record, or collectively a dataset, represents the original sample.

The aim of the 'representability' concept is to quantify the amount of bias caused by anonymizing a dataset in order to determine the generalizability of the results to wider population. The representability refers to one or more numerical values for information content of a record or dataset. It also contains other information about the dataset and dependent and independent variables without revealing the identity of individuals. As a result, we call this concept as Representability Vector as it contains more than one numerical and non-numerical values. This can be used by public health researchers, specifically epidemiologists, to ascertain the amount of bias in the anonymous dataset.

Table4.5: 2-anonymous dataset with aggregated and non-aggregated per different clusters

age	zip	gender	condition
10-50	2**	***	12
10-50	2**	***	18
10-50	2**	***	14
20-40	34*	F	15.5
20-40	34*	F	15.5

4.4 Application of cluster validity indices to calculate informatio loss as an essential part of the Representability Vector(RV)

The most important part of the Representability Vector, or RV, is the overall information content of the dataset. Many anonymization algorithms based on k-anonymity, calculate the inevitable information loss in a different way. The primary goal of registries in calculating the information loss is to ensure that the dataset has not been distorted more than necessary. The prominent example for the application of information loss for such purpose is MinGen algorithm [7, 48]. It is possible to have several datasets conforming to k-anonymity in high dimensional datasets. In such cases, information content measurement is used for choosing the disclosable datasets based on the policy of the organization. Another theoretical purpose of information content metrics is to report it to researcher to provide an objective measurement of the bias and to facilitate the incorporation of information loss in epidemiological models. This particular application of information loss is not practiced in real data disclosure activities. However, it is an essential part of $PARV$ model. Another use of information content in $PARV$ model is to compare the datasets anonymized with different anonymity methods and to disclose the preferred dataset by registry or the researcher based on evidence. Different methods based on k-anonymity have dissimilar metrics for information content. While normalization can be used to compare the information content of two different methods, it would become difficult over time when new algorithms emerge according to the needs. This problem arises from the fact that in dynamic disclosure of data, the constant review and evaluation of policies lead to changes in methods and the procedures to use them. All of these updates, ultimately affect the choice of information content metric.

In order to tackle the problem of changing methods over time and comparing the results of different algorithms we used clustery validity indices as a uniform method to measure information content. Figure 4.3 shows the comparison between conventional methods and levels 1 and 2 of *PARV* model. In this model, clustering analysis is applied on anonymization results, as describes in the first layer of *PARV* in chapter 3. The clustering validity indices are applied in the second level of *PARV* model as a level on top of the results of the first level of *PARV* to derive a generic and unified value to be used in *RV*.

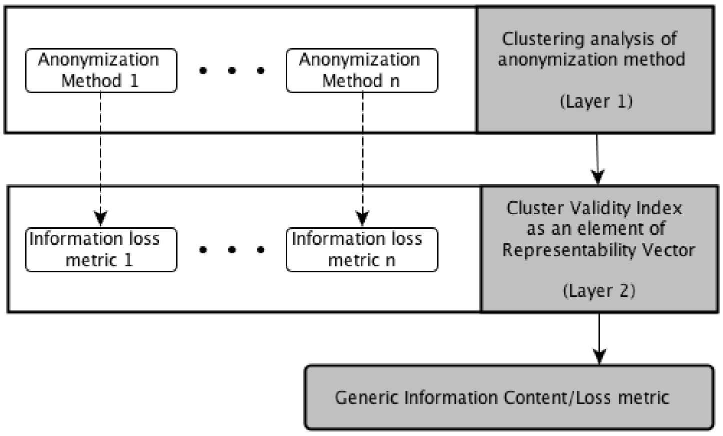
Anonymization Method 1
Anonymization Method n
Clustering analysis of anonymization method
(Layer 1)
Information loss metric 1
Information loss metric n
Cluster Validity Index as an element of Representability Vector
(Layer 2)
Generic Information Content/Loss metric

Figure 4.3: Cluster Validity Index as a generic method to calculate information content/loss of anonymized datasets

The advantages of applying clustering based anonymization on k-anonymous datasets and and using a uniform information content measurement based on cluster validity indices are:

1. Anonymization based on k-anonymity is actually a clustering problem.
2. The concept of Domain Generalization Hierarchy is similar to agglomerative hierarchical clustering.
3. There are few categories of clustering algorithms and at the same time they cover all the issues of k-anonymization.
4. A combination of different clustering algorithms facilitate anonymization in high dimensional data.
5. The cluster validity index facilitates the inclusion of semantics of data in information content metrics.
6. The same clustering rules applied on quasi-identifiers can be applied to non-quasi identifiers.
7. Depending on the context, semantic of data may require a fuzzy approach to data specification, generalization, and anonymization. Fuzzy anonymization can be realized though fuzzy clustering techniques.

We used the following metrics as the main numerical elements of *RV*:

1. Precision metric. Which is the original metric proposed by Latanya Sweeney as the basic information content metric for k-anonymity [7].

2. Semantic precision metric. Application of hyper-plane distance measurement to incorporate semantic of data in generalization hierarchy and precision metric

3. Internal criteria indices: There are indices used to validate the clustering of partitioning based algorithms. We chose Davies-Bouldin validity index as a candidate as it is independent of the number of clusters and clustering algorithm.

4. External criteria indices: such as Rand statistics, Jaccard coefficient, Folkes and Mallow index, and Hubert's statistics. We chose this family of indices because their fundamental idea is to compare the resulting clustering with an external pattern. In the context of this study, the non-anonymized sample dataset is considered as the the original pattern that can be used in external indices as the reference pattern.

4.4.1 Representability and information content

Representability aims at measuring the resemblance of an anonymized dataset and the original sample dataset held by a registry. From an information theory point of view, sample dataset contains the maximum information content about the dependent and independent variables. It should be noted that the datasets are k-anonymized therefore 'information content' in RV calculation does not refer to the identity of persons.

Given a dataset D^0 as the original de-identified dataset and assuming the application of cluster analysis of $PARV$ model on D^0, let $f_{msc} = f_s(f_k(D^0))$ denote a multi-step composite function manipulating the dependent and independent variables where f_s is the function to manipulate sensitive attribute and f_k is the k-anonymization function. The manipulation to quasi-identifiers and sensitive attributes are applied in several steps as shown in figure 4.2 . Let L_i denote the i^{th} level of alteration done on variables, with L_0 showing original values and L_m showing the maximum manipulation.

For simplicity, assume that disclosed dataset is as the same size of the P^{rs}, all records are identical in both datasets, and only generalization approach is applied. Let $P(\omega^{L_i})$ be the probability that the associated values $X_i^{L_n}$ of $q_i \in Q^{D^0}$ at the generalization level $L \leq L_{max+1}$ is identical to the original value in registry sample population P^{rs}. $X_i^{L_n}$ denotes the associated values of quasi-identifier q_i at n^{th} level of generalization and L_{max+1} is the highest level of manipulation, where in case of generalization approach is the suppression level. In this scenario, he entropy of $P(\omega^{L_n})$ is $H(\omega^{L_n}) = -\, \omega^{L_n} \log(\omega^{L_n})$.

Let I^o denote the overall information content of disclosed dataset after de-identification, k-anonymization, and sensitive attribute manipulation where 'information' refers to the resemblance of the disclosed dataset to the original dataset. Consider the following alteration extremes applied on independent attributes or quasi-identifiers.

- No manipulation is applied on Q^D. As a result, the associated values of each $q_i \in Q^D$ represents the original values belonging to a person. In terms of information content, $P(\omega^{L_0})=1 \Rightarrow H(\omega^{L_0})=-1\ \log(1)=0 \Rightarrow I^o(D)=1-H(\omega^{L_0})=1$. An example of this case is shown in table 4.1 where all records are de-identified and the level of manipulation is identical to the original sample dataset.

- Associated values of all quasi-identifiers are suppressed. Since no information is contained in quasi-identifiers, they practically fail to represent the original records in terms of information content of independent variables. In other words $I^o(D)=1-H(\omega^{L_{max+1}})=1-H(\omega_i^{L_{max+1}})=1-0\ \log(0)=0$ Therefore the probability of $X_i^{L_{max+1}}$ representing the original values of sample dataset is zero. An example of this case is shown in table 4.3.

The aim of the information content measurement, as a part of RV, is to find the quantitative representability of records in more realistic situations such as example tables 4.4 and 4.5. The main methods to capture the information content, that are analyzed here, as tools for measuring representability are original precision metric used in k-anonymity and cluster validity indices.

4.4.2 Precision metric analysed as the benchmark metric in *PARV* model

Precision metric is the original metric introduced with k-anonymity algorithm. It is "an information theoretic metric that reports the amount of distortion of a table caused by generalization and suppression" [7, p. 92]. Precision metric is different from classical measurement of entropy in terms of adding the semantics of anonymization, namely generalization and suppression, to the information content measurement [7, p. 92]. The main use of precision metric is in MinGen algorithm [7, 8] where its aim is to find the least distorted k-anonymized table. Precision metric measures the information content at record level by incorporating the characteristics of generalization applied on the associated values of quasi-identifiers and the related maximum level of generalization.

Let D^0 be the de-identified dataset with m records and Q^{D^0} be the set of n quasi-identifiers. Also let $h_{i.j} = L_{i.j}$ denote the actual generalization at level i of quasi-identifier belonging to record j, and $|DGH_j| = L_{max_j+1}$ denote the size of the generalization hierarchy, or the length of the maximum generalization level including suppression, of i^{th} quasi-identifier. Precision metric is shown in equation 4.1.

$$Prec(D^0)=1-\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{h_{i.j}}{|DGH_j|}}{m\cdot n} \tag{4.1}$$

In the context of multi-step clustering analysis of *PARV* model, precision metric can be calculated for each equivalent cluster. In this case, precision metric is calculated per cluster at record level. Given $\Xi^*=\{\xi_1^*,\xi_2^*,\ldots,\xi_{\lambda'}^*\}$ as the set of all equivalent clusters, and $\Delta_\xi=\emptyset$ indicating that all $\xi_p^* \in \Xi^*$ are conforming clusters, the cluster based precision metric is as follows. Figure 4.4 shows the height matrix of all conforming equivalent clusters needed for calculation of precision metric. The reason for simplification of precision metric per conforming equivalent cluster ξ_i^* is that associated values of all $Q^{\xi_i^*}$ are identical for all records.

A shortcoming of original precision metric is that the only semantic that is incorporated are generalization level. Other semantics related to the nature of quasi-identifier are not included in this metric. Another shortcoming is that it is only applicable to generalization approach for k-outlier handling. It does not capture the information loss due to changes in sample size or other kinds of alterations of record or sensitive attributes. The aim of applying cluster validity indices in *PARV* model is to calculate information content is to address these shortcomings. The following section describes the measurements of semantics in the context of public health beyond the characteristics of generalization of quasi-identifiers.

$$Prec(\xi_p^*)=1-\frac{\sum_{q=1}^{|\xi_p^*|}\sum_{j=1}^{n}\frac{h_{q.j}}{|L_{max_j}+1|}}{|\xi_p^*|.n}=1-\frac{\sum_{j=1}^{n}\frac{h_{q.j}}{|L_{max_j}+1|}}{n} \tag{4.2}$$

$$\text{Height matrix of } \Xi^* = \begin{array}{c c} & \begin{array}{cccc} q_1 & q_2 & \cdots & q_n \end{array} \\ \begin{array}{c} \xi_1^* \\ \xi_2^* \\ \vdots \\ \xi_{\lambda'}^* \end{array} & \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{\lambda'1} & h_{\lambda'2} & \cdots & h_{\lambda'n} \end{pmatrix} \end{array}$$

Figure 4.4: Height matrix of conforming equivalent clusters

4.4.3 Context-oriented and evidence-based cluster distance measurement in *PARV* model

The basis for any clustering based analysis method, including cluster validity index, is measuring the distance between different data points and clusters. The ideal clustering technique partitions data points into groups with low intra-cluster and high inter-cluster distance. The prominent problem with measuring distance is that quasi sub-records are comprised of several attributes with different data types. This multi-dimensionality, which is one of the compounding issues of k-anonymity [70], also complicates the distance measurement in clustering analysis.

$$M_{dist}(D^0)= \begin{array}{c c} & \begin{array}{cccc} r_1 & r_2 & \cdots & r_m \end{array} \\ \begin{array}{c} r_1 \\ r_2 \\ \vdots \\ r_m \end{array} & \begin{pmatrix} 0 & dist_{1,2} & \cdots & dist_{1,m} \\ dist_{2,1} & 0 & \cdots & dist_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ dist_{n,1} & dist_{n,2} & \cdots & 0 \end{pmatrix} \end{array}$$

Figure 4.5: Dissimilarity matrix for records of D^0 in classical clustering

There is a fundamental difference between classical clustering and using it as an analysis method in *PARV* model. In classical clustering the aim is "to create groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct" [125]. However, in *PARV* model the aim of using clustering is to compare the anonymized dataset D^* to the original de-identified dataset D^0 and express their similarity or dissimilarity in terms of cluster validity index as an essential element of RV.

The starting point for classical clustering is to find and represent the similarity and dissimilarity matrices of data points or objects. One way to do so is to use proximity matrix which represents a pairwise proximity distance measurement, either in terms of similarity or dissimilarity [125]. Figure 4.5 shows a dissimilarity matrix used in classical clustering for de-identified dataset D^0=$\{r_1, r_2, \ldots, r_m\}$ with m records. Figure 4.6, on the other hand, shows the dissimilarity matrix for quasi sub-records used in *PARV* model. In this matrix, the dissimilarity is calculated not between the records of Ξ^0 but between equivalent clusters of Ξ^* and equivalent classes of D^0. In this matrix, $d^{\xi}_{i,j}$ denotes the distance between the conforming equivalent cluster $\xi^*_i \in \Xi*$ and equivalent class $\xi^0_j \in \Xi^0$.

$$M_{dist}(\Xi^0, \Xi^*)= \begin{array}{c} \\ \xi^*_1 \\ \xi^*_2 \\ \vdots \\ \xi^*_{\lambda'} \end{array} \begin{array}{c} \begin{array}{cccc} \xi^0_1 & \xi^0_2 & \cdots & \xi^0_\lambda \end{array} \\ \begin{pmatrix} d^{\xi}_{1,1} & d^{\xi}_{1,2} & \cdots & d^{\xi}_{1,\lambda} \\ d^{\xi}_{2,1} & d^{\xi}_{2,2} & \cdots & d^{\xi}_{2,\lambda} \\ \vdots & \vdots & \ddots & \vdots \\ d^{\xi}_{\lambda',1} & d^{\xi}_{\lambda',2} & \cdots & d^{\xi}_{\lambda',\lambda} \end{pmatrix} \end{array}$$

Figure 4.6: Dissimilarity matrix of Ξ^0 and Ξ^* in *PARV* model

4.4.3.1 Multi-dimensional distance measurement and evidence-based parameters in *PARV* model

Datasets used in epidemiological research are usually multi-dimensional comprising of several variables of different types [129]. This requires multi-dimensional and mixed type distance measurements for clustering analysis. The common types of variables used in public health research and epidemiology are as follows:

1. Numerical or quantitative. Representing numerical values.
   - (a) Discreet, which are numbers belonging to a finite set, such as number of households.
   - (b) Continuous, which are numbers of infinite sets such as weight and blood pressure.
2. Categorial or qualitative
   - (a) Ordinal, which are finite ordered values, such as rank.
   - (b) Nominal, which are finite unordered values, such as gender, blood type, or city of birth.
   - (c) Binary, which have two possible values, such as "yes" or "no".

One of the widely used measurement of similarity for mixed-type data is the Gower's general distance coefficient [125, 130]. Our motivation is not to recommend the best measurement, instead, we aim at describing the concept of applicability of cluster analysis on k-anonymized dataset. As a result, while other mixed-type measurements, such as General Minkowski distance [131, 132](as cited in [125]) are available, we chose the Gower's distance due to its simplicity in explaining the distance measurement of mixed-type data. In order to incorporate the semantics of data and k-anonymization in multi-dimensional cluster analysis, we used the general distance measurements with the inclusion of semantics of k-anonymization using a method we call hyper-plane distance measurement. Our method also includes other parameters related to epidemiological studies in order to adjust distance measurement according to the context and the requirements of researcher. The rest of this section describes the application of Gower's general distance coefficient in multi-step clustering analysis of *PARV* model without the inclusion of semantics of k-anonymization or evidence-based parameters. We modified this measurement to reflect such parameters and epidemiological context which will be explained in section 4.4.4.

Let $\xi_i^0 \in \Xi^0$ be an equivalent class and $\xi_j^* = f^*(\xi_i^0)$ be a conforming equivalent cluster in Ξ^*. It is assumed that both ξ_i^0 and ξ_j^* have the same set of quasi-identifiers $Q^D = \{q_1, q_2, \ldots, q_l\}$. As a result, both clusters are l dimensional with probably different types of attributes comprising Q^D. Both clusters are represented by a unique combination of the associated values x_n of quasi-identifiers which at the same time are the centroid of the equivalent class or cluster. The Gower's general distance coefficient is shown in equation 4.3.

$$d_{gower}(\xi_i^0, \xi_j^*) = \left(\frac{1}{\sum_{n=1}^{l} \omega(x_n^{\xi_i^0}, x_n^{\xi_j^*})} \sum_{n=1}^{l} \omega(x_n^{\xi_i^0}, x_n^{\xi_j^*}) d^2(x_n^{\xi_i^0}, x_n^{\xi_j^*}) \right)^{\frac{1}{2}} \tag{4.3}$$

Where $d(x_n^{\xi_i^0}, x_n^{\xi_j^*})$ is the distance between the associated values of centroids of equivalent cluster ξ_j^* and equivalent class ξ_i^0. For numerical q_n, the distance measurements is based on Manhattan distance. For categorial q_n, the measurement is based on simple matching distance. $\omega(x_n^{\xi_i^0}, x_n^{\xi_j^*})$ specifies whether the comparison between centroids are valid. It equals 0 if any of the associated values of x_n are missing values, otherwise, $\omega(x_n^{\xi_i^0}, x_n^{\xi_j^*})$=1.

Classical clustering aims at finding groups of similar objects to put them in the same cluster and to ensure that different clusters are quite far from each other. In *PARV* model, however, distance measurement between clusters is not necessary because all the clusters of Ξ^0 and Ξ^* are maximally packed. However, taking the current values of records in Ξ^0 and Ξ^* for cluster analysis is biased due to the possible replacement of actual values in Ξ^*. The cluster analysis must be applied taking the original values of records contained in each equivalent class mapped to Ξ^*.

Distance measurement, regardless of the metric used, depends on the approach for k-outlier handling. The reason is the assigned value of the mapped equivalent class, and its effect on the distance, is determined by the applied k-outlier handling approach. The following list summarized the characteristics of distance measurement based on the approaches for k-outlier handling presented in chapter 3.6.

- *Generalization approach.* In this approach, each $\xi_j^* \in \Xi^*$ contains generalized records belonging to one or more $\xi_i^0 \in \Xi^0$. While all the elements in ξ_j^* have the same value for quasi sub-records, their corresponding values in non-generalized records in Ξ^0 might be different. The challenge is to include the semantics of data and k-anonymization into the distance measurement. The hyper-plane distance measurement concept of *PARV* model aims at incorporating the semantics of data and k-anonymization, both affected by evidence and context, into clustering distance metrics.
- *Deletion approach.* In this approach the sample size is reduced. This can be reflected in the metric by reporting the sample size, incorporating the ratio of actual and disclosed sample size, or considering the deleted records as maximally suppressed records with maximal distance from the original cluster.
- *Merging approach.* Is the simplest approach in which the values in Ξ^* is at the same level of generalization as in Ξ^0. Therefore, the classical distance measurements suffice in reflecting the semantics of data.
- *Fabrication approach.* Fabricated records do not exist in any equivalent class of Ξ^0. One approach to handle this situation is to consider them as a separate cluster in Ξ^0 with suppressed values or maximal distance from all other clusters in order to reflect the bias they introduce. They can also be considered as missing values , hence, be setting $\omega(x_n^{\xi_i^0}, x_n^{\xi_j^*})$ to zero. In this case, the sample size difference must be reported in RV to reflect the introduced bias.

Besides the approach used for k-outlier handling, other requirements of epidemiology methods, context, and the existing evidence also affect the interpretation of the clustering. These parameters are not included in k-anonymity or classical distance measurements. In *PARV* model, however, these parameters reflect the context based evidence and shape the basis for reporting the bias to the researcher. Figure 4.7 shows the interaction between evidence, RV, and *PARV* model. Ideally , the evidence of interest to epidemiologists must be included in anonymization process. Therefore, the distance measurement needs to be modified by including the quantitative parameters reflecting the context-oriented evidence.

The following list shows the evidence-based parameters that need to be included in cluster analysis and distance measurement in epidemiological and public health studies. It should be noted that, in many of weight parameters, where the range is between 0 and 1, the relationship between the value of weight parameter and its interpretation depend on whether the entropy or the information content is measured.

- *Semantics of generalization of quasi-identifier q_n as the centroid of either ξ_i^0 or ξ_j^*.* They refer to the semantics presented in precision metric(section 4.4.2), namely the height of the current generalization level of the centroid $h_{q.i}$, and the length of the generalization hierarchy $\mid L_{max_i} \mid$. For example in the dataset shown in table 3.4, the 'zip' quasi-identifier is generalized as $34232 \xrightarrow{1} 3423* \xrightarrow{2}$

$342**\xrightarrow{3} 34***\xrightarrow{4} 3****\xrightarrow{s=5} *****$. If a particular record have 'zip' value generalized as $3423**$ then the current generalization level is 2 and the maximum level is 5.

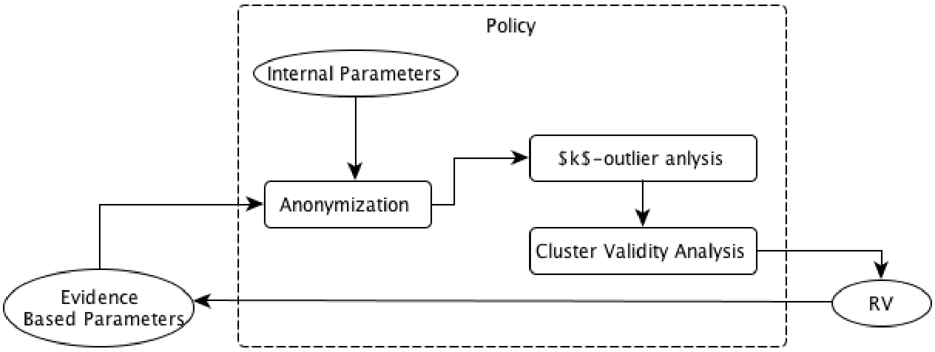

Figure 4.7: Interaction between evidence-based parameters, RV, and $PARV$ model

- *Normalization of generalization hierarchy.* This parameter specifies the normalization applied on domain generalization hierarchy for all quasi-identifiers or quasi sub-records. When maximum length of the generalization hierarchy are different between quasi-identifiers, bias occurs by the quasi-identifier with higher levels of generalization. We used min-max normalization to counter the bias introduced by different levels of generalization.

- *Hyper-plane distance.* Denoted by d^h, it measures the distance between ξ_i^0 and ξ_j^* in terms of semantics of actual data affected by generalization characteristic. This is used in generalization and merging approaches mentioned in sections 3.6.2 and 3.6.3 respectively. In order to calculate this distance, I proposed the hyper-plane distance, we conceptualized the hyper-plane distance measurement which will be explained in section 4.4.3.2.

- *Evidence-based generalization level weight parameter.* Denoted by w_i^e, it shows the semantic weight given to each generalization level to reflect the effects of generalization on actual epidemiological variables. This weight is different from weighted precision metric [7] as it is not applied to rank the quasi-identifiers based on their importance. Instead, it is applied on different generalization levels of the same quasi-identifier. For example, consider the 'zip' attribute in dataset shown in table 3.4. The requirement of the researcher might indicate that any generalization higher than level 3 would negatively affect the accuracy of the research. In this case, a smaller weight is assigned to lower levels of generalization in order to reflect the acceptable level of introduced entropy or distance. In this example, a hypothetical weight assigned to the chain of 'zip' attribute generalization would look like $w_0^e{=}0.1 \xrightarrow{1} w_1^e{=}0.2 \xrightarrow{2} w_2^e{=}0.4 \xrightarrow{3} w_3^e{=}0.5 \xrightarrow{4} w_4^e{=}0.9 \xrightarrow{s=5} w_s^e{=}1$ where more importance is given to generalization levels higher than 3 in terms of the entropy or distance they introduce.

- *Evidence-based quasi-identifier weight parameter.* Denoted by w_n^q, it is a weight assigned to each quasi-identifier in order to adjust their contribution to distance measurement. This is similar to the weight measurement in weighted precision metric [7] . For example, assume that the researcher needs all quasi-identifiers in table 3.4 , however is more interested in bias caused by manipulation applied on 'age' and 'gender' attributes. In this case, a hypothetical example would look like w^q_{age}=0.45, w^q_{zip}=0.1, and w^q_{gender}=0.45.

- *Required quasi-identifiers.* This parameter reflects the set of quasi-identifiers that researcher actually needs. In many of public health data disclosures, the registry does not have prior information about possible uses of disclosed data which leads to high dimensionality of data and higher information loss. This parameter, therefore, is only applicable to research based disclosures. Exclusion of un-required quasi-identifiers reduces the bias and complexity of anonymization considerably as it reduces the number of clustering features. This exclusion either does not change the generalization of other quasi-identifiers or leads to improved information content by having the remaining quasi-identifiers generalized at lower levels.

- *Required clusters.* This parameter reflects the values of quasi-identifiers that the researcher needs. Applying this parameter leads to deletion of unnecessary records, which in turn leads to to less variability. For example, the researcher might be interested in the relationship between smoking and pregnancy. In this case, only the records belonging to females are needed. Deletion of clusters including male individuals leads to less dimensions and clustering features and reduces the level of generalization applied on other quasi-identifiers.

- *Evidence-based initial generalization level.* This parameter is specified by researcher and shows the required initial generalization level of a quasi-identifier. This reduces the complexity of k-anonymization and affects the calculation of the bias as it changes the domain generalization hierarchy. For example, the initial level for geographical area might be the county code, instead of zip code, which naturally increases the number of identical values per cluster leading to shorter generalization hierarchy and less variability.

- Sample size ratio. Denoted by r^s, it is useful in calculating the overall distance measurement between Ξ^0 and Ξ^*. This parameter is used in deletion approach mentioned in section 3.6.1.

- Aggregation of sensitive attributes. Specifies the manipulation applied on sensitive attributes according to the requirements of the researcher.

4.4.3.2 Conceptualization of hyperplane distance measurement in *PARV* model for inclusion of semantics in distance measurement

The primary manipulations applied on data in order to achieve of k-anonymity are generalization and suppression. These techniques are presented as one of the possible k-outlier handling in generalization approach in section 3.6.2. The semantics of generalization that are included in information content metric or original k-anonymity, or precision metric, are height of the current generalization level and the maximum generalization height for each quasi-identifiers, as mentioned in section 4.4.2. While precision metric captures the characteristics of generalization hierarchy, it does not include the semantics of data or other factors affecting epidemiological inferences. For example, zip codes 34232, 34233, and 34500 may all generalize to 34*** while usually zip codes do not represent actual geographical distance. Capturing the real semantic of data becomes complicated when generalization approach is applied. The reason is that the values in generalized clusters of Ξ^* are different from the values of equivalent classes of Ξ^0.

In order to address this problem, we presented the concept of inclusion of semantics of actual data in the distance measurement in a method we call 'hyperplane distance measurement'. In this approach, each cluster $\xi_j^* \in \Xi^*$ is assigned with a hypothetical center which corresponds to a real point that represents all equivalent classes in Ξ^0 that have been mapped to ξ_j^*. Assume the following given sets:

- $D^0 = \{r_1^0, r_2^0, \ldots, r_d^0\}$ as the dataset.
- $Q^0 = \{q_1, q_2, \ldots, q_l\}$ as the set of quasi-identifiers.
- $\Xi^{0i} = \{\xi_1^{0.i}, \xi_2^{0.i}, \ldots, \xi_{\lambda^0}^{0.i}\}$ as the set of equivalent-classes with only i^{th} quasi-identifier considered at level 0 of generalization.
- $F_{k_i} = \{f_i^1, f_i^2, \ldots, f_i^*\}$ as the anonymization function applied on q_i. f_i^j is the j^{th} level in domain generalization hierarchy for q_i. f_i^* denotes the last application of anonymization function whose results comprise the the disclosed dataset.

Table4.6: Interplane Contingency Table $ICT_i(HP_i^0, HP_i^1)$

$ICT_i(HP_i^0, HP_i^1)$	$\xi_1^{1_i}$	$\xi_2^{1_i}$	$\cdots$	$\xi_{\lambda_1-1}^{1_i}$	$\xi_{\lambda_1}^{1_i}$
$\xi_1^{0_i}$	0	$d^h(\xi_1^{0_i}, \xi_2^{1_i})$	$\cdots$	$d^h(\xi_1^{0_i}, \xi_{\lambda_1-1}^{1_i})$	$d^h(\xi_1^{0_i}, \xi_{\lambda_1}^{1_i})$
$\xi_2^{0_i}$	$d^h(\xi_2^{0_i}, \xi_1^{1_i})$	0	$\cdots$	$d^h(\xi_2^{0_i}, \xi_{\lambda_1-1}^{1_i})$	$d^h(\xi_2^{0_i}, \xi_{\lambda_1}^{1_i})$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\xi_{\lambda_0-1}^{0_i}$	$d^h(\xi_{\lambda_0-1}^{0_i}, \xi_1^{1_i})$	$d^h(\xi_{\lambda_0-1}^{0_i}, \xi_2^{1_i})$	$\cdots$	0	$d^h(\xi_{\lambda_0-1}^{0_i}, \xi_{\lambda_1}^{1_i})$
$\xi_{\lambda_0}^{0_i}$	$d^h(\xi_{\lambda_0}^{0_i}, \xi_1^{1_i})$	$d^h(\xi_{\lambda_0}^{0_i}, \xi_2^{1_i})$	$\cdots$	$d^h(\xi_{\lambda_0}^{0_i}, \xi_{\lambda_1-1}^{1_i})$	0

We call each level $\Xi^{n.i}$ as a hyper-plane denoted by $HP^{n.i}$. Beginning from the first hyper-plane $HP^{0.i}$, assume that $\xi_1^{1.i} \in \Xi^{1.i}$ is an equivalent-cluster and $\xi_1^{1.i} = f_i^1(b)$ where $b = \{\xi_m^{0.i}, \ldots, \xi_{m'}^{0.i}\}$. Consider a real, or unreal, data point $C_{\xi_1^{1.i}}$ in $HP^{0.i}$ which is equivalent to a centroid of a hypothetical cluster containing b. Hyperplane distance measurement is defined in equation 4.4. It measures the semantic distance between associated values of i^{th} quasi-identifier for all corresponding records in $HP^{*.i}$ and b regardless of generalization or other approaches applied.

$$d^h(b, \xi_j^{*.i}) = \frac{\sum_{p=m}^{m'} d(\xi_p^{0.i}, C_{\xi_j^{*.i}})}{m' - m} \tag{4.4}$$

Where, $C_{\xi_j^{*.i}}$ is a hypothetical semantic center corresponding to a real or unreal value in clusters in original data, which are included in $\xi_j^{*.i}$, and $(m' - m)$ indicates number of equivalent classes in $\Xi^{0.i}$ that are mapped to $\xi_j^{*.i}$. Figure 4.8 shows an example for hyperplanes 0 and 1 for a hypothetical dataset with only one quasi-identifier. Table 4.6 shows the Inter-plane Contingency Table (ICT) for HP_i^0 and HP_i^1 of i^{th} quasi-identifier as shown in figure 4.8.

The concept of hyperplane distance measurement, especially for categorial data, is similar to the semantic relationship between the values in a taxonomy tree [74]. The difference here is that the semantic distance, and the hypothetical center of the first hyperplane, in the context of *PARV* model is defined by the epidemiological requirements. In other words, the distance between points in the same dataset might vary from one research to another based on the context.

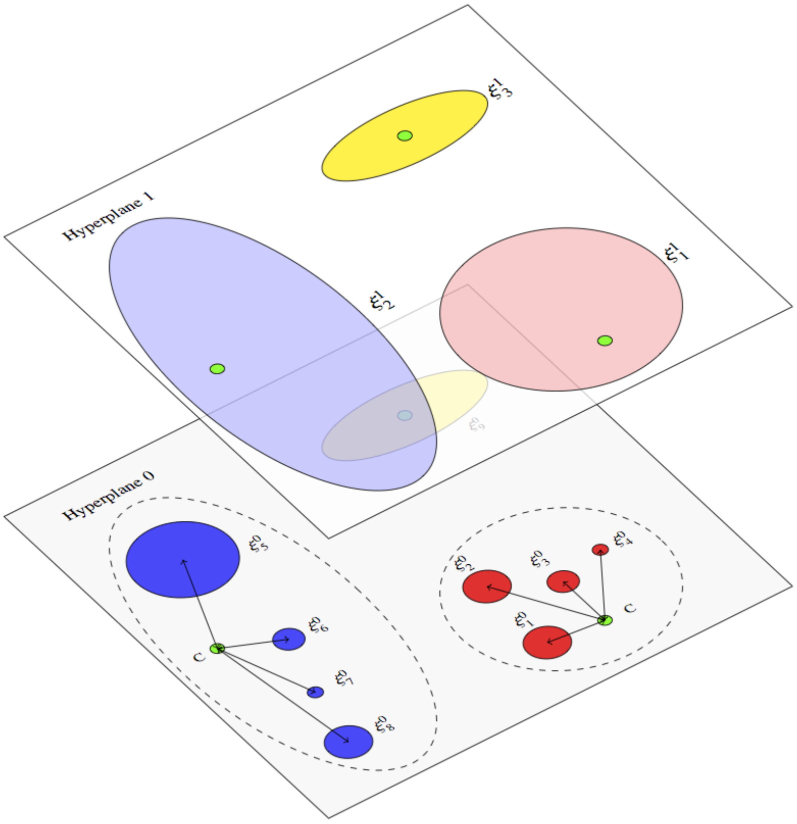

Figure 4.8: Hypothetical example for hyperplane distance measurement from HP_i^0 to HP_i^1

4.4.4 Modified evidence-based and multi-dimensional distance measurment for *PARV* model

In order to include the context-oriented evidence, researcher requirements, and semantics of data in cluster distance measurement, we modified the Gower's general distance coefficient. The modified distance used in *PARV* model, denoted by d_M is shown in equation 4.5, where the distance is shown between an equivalent cluster ξ_j^* and one of the equivalent classes ξ_i^0 that have been mapped to ξ_j^*. For simplicity, the distance is shown only for valid comparisons without a need for $\omega(x_n^{\xi_i^0}, x_n^{\xi_j^*})$. Let b=$\{\xi_m^0, \dots, \xi_{m'}^0\}$

be the set of all equivalent classes in Ξ^0 that have been mapped to ξ^*_j. Equation 4.6 shows the accumulative distance between ξ^*_j and b.

$$d_M(\xi^0_i,\xi^*_j)=\sum_{n=1}^{l} d^h(\xi^{0.n}_i,\xi^{*.n}_j)w^q_n G_n \tag{4.5}$$

$$d_M(b,\xi^*_j)=\frac{1}{m-m'}\sum_{c=m}^{m'} R_{\xi^0_c}\sum_{n=1}^{l} d^h(\xi^{0.n}_c,\xi^{*.n}_j)w^q_n G_n \tag{4.6}$$

Where:

- $R_{\xi^0_c}$ is the sample size ratio. Let $\xi^{0'}_i$ denote the records from ξ^0_i that are included in ξ^*_j. The size of $\xi^{0'}_i$ is equal to ξ^0_i unless the deletion approach is applied or some records have been removed due to the characteristics of sensitive attribute. Therefore, $R_{\xi^0_c}=\frac{|\xi^{0'}_i|}{|\xi^0_i|}$.
- w^q_n is the evidence-based quasi-identifier weight parameter for n^{th} quasi-identifier. Given $Q^0=\{q_1,q_2,\ldots,q_l\}$, $\forall q_i \in Q^0, 0 < w^q_i \leq 1$ and $w^q_1+w^q_2+\cdots+w^q_l=1$. Functionally, this parameter is similar to scaling the quasi-identifier in terms of actual distance [77].
- G_n is a parameter reflecting the semantics of generalization. It can take different forms and parameters based on policy and context including the following:
  - $\frac{h^i_d}{L_{max_i}+1}$ is the semantics of generalization used in precision metric (4.4.2), where h^i_d is the d^{th} level of domain generalization hierarchy of i^{th} quasi-identifier, and L_{max_i} is the last height of the generalization prior to suppression.
  - $\frac{h^i_d w^g_{i.d}}{L_{max_i}+1}$ where $w^g_{i.d}$ is the weight assigned to the d^{th} generalization level of i^{th} quasi-identifier. Given h^i_x as the current generalization at x$^{\text{th}}$ level, and $\forall 0 \leq x < (L_{max_i}+1)$, we have $0 \leq w^g_{i.d} \leq 1$ and $\sum_{d=0}^{L_{max_i}+1} w^g_{i.d}=1$. If $w^g_{i.d}$ is used as the semantics of hyperplane distance measurement, this form of G_n should not be used.
  - H^n_i is the normalized height measurement for i^{th} quasi-identifier. It is a min-max normalization of heights of all quasi-identifiers. Let $|DGH_i|$ denote the size of the domain generalization hierarchy including the suppression level for i^{th} quasi-identifier. Let m_g denote the minimum and M_g denote the maximum domain length among all quasi-identifiers. With min-max normalization, for all datasets where $m_g \neq M_g$ $H^n_i=\frac{m_g-h^i_d}{M_g-m_g}$.

4.4.5 Cluster validity indices with internal criteria

Internal clustering validity indices aim at evaluating the clustering using the characteristics of dataset [125]. Cluster evaluation in the context of *PARV* model falls in the hard clustering and partitioning category where a record belongs to one and only one cluster. The aim of validity index is to re-cluster the dataset to achieve desirable clustering results, including optimized number of clusters. In *PARV* model, however, the aim is to use validity index as a measurement and comparison method for bias caused by k-outlier handling. As a result, a more general, and data independent index suits better for aforementioned goal. W chose the Davies-Bouldin (DB) index as it is not dependent on the number of clusters or the clustering algorithms [125]. The advantage of using this index is that the evaluator is not need to know the underlying algorithm used for k-outlier handling. Although k-anonymization is not a clustering solution, its results resemble the outcomes of clustering. DB index in classical clustering is calculated by specifying dispersion and cluster similarity measures. Given the following given sets, the dispersion measure for cluster ξ_i^*, which indicates the average distance between the data points inside a cluster and its centroid, is shown in equation 4.7.

- Q^D=$\{q_1, q_2, \ldots, q_l\}$
- Ξ^*=$\{\xi_1^*, \xi_2^*, \ldots, \xi_{\lambda'}^*\}$
- b_i=$\{\xi_{i.m}^0, \ldots, \xi_{i.m'}^0,\}$ as the set of clusters of Ξ^0 that are mapped onto ξ_i^*.

$$S_{\xi_i^*} = \left(\frac{1}{\mid \xi_i^* \mid} \sum_{x=m}^{m'} \left(d_M(\xi_{i.x}^0, \xi_i^*) \right)^2 \right)^{\frac{1}{2}} \tag{4.7}$$

The next step is to measure the distance between clusters of Ξ^*, or dissimilarity measure,for each ξ_i^*. Dissimilarity measure is shown in equation 4.8, where $c_{\xi_i^*}$ is the centroid of ξ_i^*.

$$D_{\xi_i^*, \xi_j^*} = \left(\sum_{n=1}^{l} \mid c_{\xi_i^*} - c_{\xi_j^*} \mid^2 \right)^{\frac{1}{2}} \tag{4.8}$$

The cluster similarity measure is defined in equation 4.9.

$$R_{\xi_i^*, \xi_j^*} = \frac{S_{\xi_i^*} + \xi_j^*}{D_{\xi_i^*, \xi_j^*}} \tag{4.9}$$

Davies-Bouldin index is defied in equation 4.10, where $R_i = max(R_{\xi_i^*, \xi_j^*})$ for all $i \neq j$.

$$V_{DB} = \frac{1}{\lambda'} \sum_{i=1}^{\lambda'} R_i \tag{4.10}$$

Smaller values of V_{DB} indicate better clustering with high intra-cluster similarity and high inter-cluster dissimilarity. This indicates that any equivalent cluster Ξ^* with lower

DB index represents the Ξ^0 better than clusters with higher values for DB. V_{DB} can be used for comparison of different clustering results. For bias reporting, the value must be normalized using the indices of best and worst datasets in terms of their resemblance to the original dataset.

The advantage of using DB index over precision metric is that sensitive attributes can also be included in the analysis. Another advantage is that it can also be used for non-generalization approaches where precision metric fails to capture the bias caused by merging, deletion, or fabrication. With modified distance d_M (equation 4.6), the semantics of generalization are also included in DB index. The disadvantage of this index is that the value of DB index is not necessarily falls between 0 and 1. Min-max normalization can be applied to tackle this problem by taking the DB index of the original dataset as the minimum and that of the most distorted dataset as the maximum values.

4.4.6 Cluster validity indices with external criteria

Cluster validity indices with external criteria aim at comparing the resulted clustering to a pre-specified structure on the dataset [125]. In the context of *PARV* model, Ξ^0 is considered as the basis for comparison as it presents the actual patterns in dataset. The resulted anonymized dataset Ξ^* then can be compared to Ξ^0 using cluster validity indices using external criteria. Higher similarities indicate that the anonymized dataset highly represents the original data and its generalizability to the population can be performed with lower bias.

Given the $D^0{=}\{r^0_1, r^0_2, \ldots, r^0_d\}$, $\Xi^0{=}\{\xi^0_1, \xi^0_2, \ldots \xi^0_\lambda\}$, and $\Xi^*{=}\{\xi^*_1, \xi^*_2, \ldots \xi^*_{\lambda'}\}$,the following contingency table shows the number of records that are mutual between the equivalent classes of Ξ^0 and equivalent clusters of Ξ^*.

	ξ^*_1	ξ^*_2	.	.	$\xi^*_{\lambda'}$
ξ^0_1	n_{11}	n_{12}	.	.	$n_{1\lambda'}$
ξ^0_2	n_{21}	n_{22}	.	.	$n_{2\lambda'}$
.	.	.	.	.	.
.	.	.	.	.	.
ξ^0_λ	$n_{\lambda 1}$	$n_{\lambda 2}$	.	.	$n_{\lambda\lambda'}$

Four different values, are calculated in external criteria in order to derive the final index. The first value, a, denotes the number of pair of records that are in the same clusters in Ξ^0 and Ξ^* and is expressed as $a = \sum_{i,j} \binom{n_{i,j}}{2}$. The second value, b, denotes the number of pair of records that are in the same clusters in Ξ^0 but map to different clusters in Ξ^* and is expressed as $b = \sum_i \binom{n_i}{2} - \sum_{i,j} \binom{n_{i,j}}{2}$. The third value, c, denotes the number of pair of records that are in different clusters in Ξ^0 but map to the same cluster in Ξ^* and is expressed as $c = \sum_j \binom{n_j}{2} - \sum_{i,j} \binom{n_{i,j}}{2}$.Total number of records is $M = \binom{n}{2} = \frac{n(n-1)}{2}$. The fourth value, d, denotes the number of pair of records that are in different clusters in Ξ^0 and map to different clusters in Ξ^* and is calculated as $d = \binom{n}{2} - a - b - c$.

The following indices are calculated based on a, b, c, d, M.

- Rand statistics, $R = \frac{a+d}{M}$
- Jaccard coefficient, $J = \frac{a}{a+b+c}$
- Folkes and Mallows index, $FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$

The range of all these indices are $[0, 1]$. High values of these indices indicate that two compared clusters are similar. In other words, higher values indicate that Ξ^* represents Ξ^0 better than clusters with lower indices.

4.4.7 Advantages of applying cluster validity indices on anonymized datasets

k-anonymity has some inherent shortcomings that prevents it from being used directly in real situations without enhancements. Some of the shortcomings of k-anonymity are

- Basic k-anonymity is not directly applicable to real disclosure needs.
- The semantic of generalization is not considered beyond a user specific weight assigned to levels of generalization hierarchy.
- It reduces information content in temporal disclosures and joining data from different registries.
- It does not take into account the privacy risks posed by non quasi-identifier attributes.
- It becomes complicated in high dimensional and heterogeneous datasets.

The advantages of applying clustering based anonymization on k-anonymous datasets are:

1. It is a general method of which k-anonymity is a specific case.
2. A combination of different clustering algorithms facilitate anonymization in high dimensional data.
3. The cluster validity index facilitates the inclusion of semantics of data in information content metrics.
4. The same clustering rules applied on quasi-identifiers can be applied to non-quasi identifiers
5. It is a generic information content measurement independent of the underlying k-anonymization algorithm.

4.5 Representability Vector (RV) of $PARV$ model

Representability Vector, or RV, is a set of elements containing details of anonymization induced information loss and pieces of information reflecting the context-oriented, especially epidemiological, parameters. It is a means to report the bias affecting epidemiological studies and also a tool to collect the context-oriented requirements of the researcher that reflect the related evidence. RV set is denoted as $RV=\{I^0, I^c, A^e, A^s, A^q, K, M, S, D\}$ as depicted in figure 4.9. In case that disclosing data is performed in several steps based on the researcher feedbacks, one problem arises from the possibility of linking the different anonymized version of the same dataset. This can be avoided by disclosing the RV only in preliminary steps instead of the whole dataset. Therefore, only the meta-data about anonymized dataset is disclosed in RV. This prevents the chance of linking of datasets and any possible privacy risks.

4.5.1 Overal information content I^o

I^o represents the overall information content of Ξ^*. It is a set containing at least one member representing the value associated with an information content metric mentioned in section 4.4. Each $i^o_j \in I^o$ is a numerical value representing the information content obtained by methods based on precision metric, cluster validity index, or information theory. Ideally, all values are normalized, preferably as $[0, 1]$, in order to make them comparable to each other. Elements of I^o are calculated for the whole dataset. The decision about incorporating the original sample size in calculation of the elements of I^o depends on the expert decisions and researcher requirements.

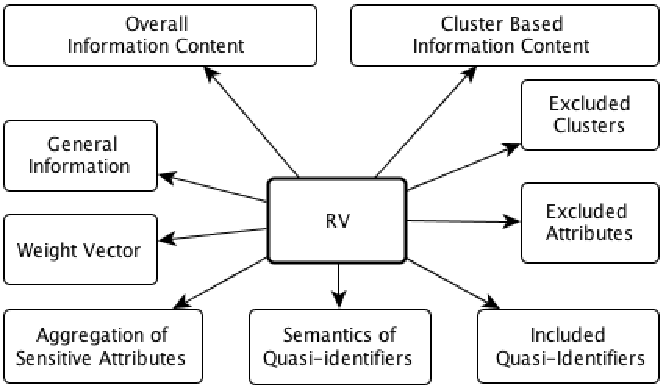
Overall Information Content
Cluster Based Information Content
Excluded Clusters
General Information
RV
Excluded Attributes
Weight Vector
Aggregation of Sensitive Attributes
Semantics of Quasi-identifiers
Included Quasi-Identifiers

Figure 4.9: Elements of RV

The primary aim of calculating information content in popular anonymization techniques is to find the least distorted dataset or to analyze the quality of the algorithm in terms of the balance between utility and information loss. In $PARV$ model, however, the aim of calculating the elements of I^o is to provide the researcher with the information loss of the dataset due to anonymization. High values of i^o_j indicate that the dataset is more suitable for ecological studies. Specifying the threshold value for determining whether the dataset is suitable for micro-data or macro-data analysis depends on policies and the evidence in epidemiological context. While the practicality of using the I^o in ascertaining the accuracy of generalizability of results is debatable, I^o can be used for comparison of two different datasets.

4.5.2 Cluster based information content I_i^c

I_i^c represents the information content per cluster $\xi_i^*$of Ξ^*. For $\Xi^*=\{\xi_1^*, \xi_2^*, \ldots, \xi_{\lambda'}^*\}$, the cluster based information content is the set $I_i^c=\{i_1^c, i_2^c, \ldots, i_{\lambda'}^c\}$, where i_n^c is the cluster based information content of the n^{th} equivalent cluster $\xi_n^* \in \Xi^*$. i_n^c is a set of values each calculated by an information content measurement method based on precision metric, cluster validity index, or information theory. To make the values comparable, they must be normalized to the same range. Comparing to I^o, I_i^c provides more information for analyzing the generalizability of the results . For example, assume that high levels of manipulation is applied on 'age' attributes where the values are higher than 70. In this example, the records with $age \geq 70$ are more biased comparing to those with $age < 70$. Therefore, the records with $age < 70$ are more accurately generalizable to the population. I_i^c provides an objective measurement of bias per equivalent cluster. The exact method to employ this measurement on epidemiological models is beyond the scope of this study.

4.5.3 Excluded attributes A^e

Although de-identified dataset D^0 does not include any directly identifiable attributes, the k-outlier handling approach may result in complete removal or suppression of quasi-identifiers. Further analysis might also result in the removal of sensitive attributes. The routine practice in anonymization of datasets, is to remove such attributes without giving information about their removal. Disclosing the meta-data about excluded attributes does not violate the k-anonymity constraints. Some of the independent variables, either as quasi-identifier or sensitive attribute, might be a probable confounding factor. Exclusion of attributes hinders the identification and analysis of confounding factor leading to selection and confounding bias.

A^e is a set of meta-data about excluded attributes without giving any information about their associated values. It should be noted that identifiable attributes are not included in A^e. If any of the attributes included in A^e is a probable confounding factor according to evidence, the researcher would have information whether the anonymized dataset D^* is suitable for causal relationship analysis for attributes in $(A_D - A^e)$. A^e can also be used as a means to collect researcher feedback about the required attributes in evidence-based anonymization.

4.5.4 Included quasi-identifiers A^q

In large datasets, especially in census data, there are many attributes gathered and included in sample and anonymized datasets. While the determination and specifying processes for of quasi-identifiers are known to the registry, the researcher is not usually provided with information about them. $A^q=(Q^{D^0} - Q^{A^e})$ is the set of meta-data about the quasi-identifiers included in D^*. The aim of A^q is to report the meta-data of quasi-identifiers in the disclosed data, which with combination of other RV elements would aid the researcher in determining the independent variables and better analysis of confounding factors.

A^q can also be used to collect the required dependent or possible confounding factors specified by the researcher. Some of these attributes might be labeled as quasi-identifiers while others might be labeled as sensitive attributes. If $A^q \subset Q^D$ but $A^q \neq Q^D$, it means that the number of required quasi-identifiers are less than those labeled by experts in original dataset. This situation leads to less dimensionality resulting in decreased complexity of k-outlier handling and possible improvement of information content.

4.5.5 Excluded clusters C^e

Some k-outlier handling approaches, such as deletion approach, change the sample size as a result of deleting some records. Other approaches might incorporates record deletion based on policy and the algorithm used. For example, some of the algorithms based on generalization approach delete the records whose the values of sensitivity attributes are less than a certain threshold, such as p-sensitive k-anonymity [52]. Such deletions are usually applied on clusters that have small number of records and are distant from other clusters. The inclusion of such clusters in non-delete based k-outlier handling approaches leads to high levels of information loss. However, deleting the records affect the sample size. If the size of the original dataset is important to the researcher, the size and meta-data of deleted non-conforming clusters, or records, must be reported. If record exclusion happens because of certain values of sensitive attribute, reporting C^e is of tremendous value as it provides information about possible extreme values that are present in the sample data and are excluded from disclosed dataset. In such cases, reporting the corresponding quasi-identifiers would probably leak privacy. While reporting the sensitive attributes without corresponding quasi-identifiers does not contribute to the objective calculation of bias, it gives an idea about the actual extreme or rare values of sensitive attributes.

C^e can also be used by researcher to indicate the clusters that are not needed for their research. For example, if the research is about pregnancy, the researcher's specified C^e would be the meta-data indicating the records belonging to male individuals. This leads to less dimensionality and variability of associated values of 'gender' attribute in each cluster leading to less k-outlier and higher information content. The sample size must be adjusted according to the new dataset which is $\mid P^{rs} - C^e \mid$.

4.5.6 Semantics of quasi-identifiers S^q

S^q contains the details of semantics of distance related to each quasi-identifier. It specifies the semantic relationship between the associated value of a quasi-identifier and the real context the attribute reflects. Take 'zip' attribute as an example with '20098', '21786' and '31990' as codes assigned to three different areas. In non-semantic coding, there is no relationship between the assigned code and the real geographical area. In this case, the '20098' area is not necessarily closer to '21786' than to '31990' area. In semantic coding the actual values reflect the real concept. In this case, '20098' is closer to '21786'. Besides the semantic of coding, S^q provides information about distance measurement used in cluster validity indices and in domain generalization hierarchy as mentioned in 4.4.3.2.

S^q can also be used by researcher to specify the context-oriented and evidence-based semantics pertaining the research. This will affect the domain generalization hierarchy and distance measurement which eventually will , positively or negatively, change the information loss.

4.5.7 Aggregation of sensitive attributes M^s

Public health data are usually disclosed with sensitive attributes aggregated for the whole dataset, or per equivalent cluster. The primary aim of *PARV* model is to disclose sensitive attribute in person-specific form per equivalent cluster as far as possible. In cases where person-specific values of sensitive attribute cause high number of record deletion or high information loss due to excess generalization, it might be preferable to apply some degree of aggregation on sensitive attributes per cluster. M^s is a set of meta-data about the characteristics of aggregation applied on $\xi^* \in \Xi^*$. This aggregation can be applied based on any central tendency methods, such as mean, mode, median, and standard deviation per cluster. M^s provides the researcher with a means to objectively analyze the causal relationship based on the value of the sensitive attribute. It also can be used by researcher to specify whether they want person-specific or aggregated values for sensitive attributes per cluster. However, it unlikely to be requested by researchers who pursue non-ecological studies.

4.5.8 Weight vector W

The aim of *PARV* model is to analyze anonymized datasets, focusing on the context of anonymization rather than its detailed technical aspects. The context in *PARV* model refers to issues of privacy protection in secondary uses of health data and actualizing the evidence-based approach to data disclosure for epidemiological research. In order to capture the contextual factors affecting the anonymization and bias, we specified several weight parameters that affect the distance measurement for cluster validity indices. Weight vector W is a set of weight parameters, all normalized to the same range. Reporting the weights to researcher is useful in determining the effect of quasi-identifiers, different clusters, generalization levels, and other elements of anonymization on causal relationship analysis. W can also be used to obtain the evidence-based weight parameters specified by researcher. The following list shows the weight parameters included in W:

- Evidence-based quasi-identifier weight parameter, w_i^q which specifies the weights assigned to i^{th} quasi-identifier. This weight can be used by researcher to determine the effect of including the i^{th} quasi-identifier on bias and the interpretation of the results. w_i^q can also be used by researcher to specify the required weight of a quasi-identifier based on evidence. It specifies the relative importance of i^{th} quasi-identifier comparing to the rest of the quasi-identifiers. Given Q^D=$\{q_1, q_2, \ldots, q_l\}$ with each quasi-identifier in Q^D assigned with the weight w_i^q, the sum of all weights equals 1. In other words $\sum_{i=1}^{l} w_i^q$=1. If w^q is not in effect, for all $i \leq l$, $w_i^q = 1$.

- Evidence-based generalization level weight parameter, w_n^e which specifies the weight assigned to n^{th} level of generalization along the generalization hierarchy per quasi-identifier. This weight can be used by researcher to determine the maximum level of generalization at which the value of a quasi-identifier can be used for non-ecological studies. The sum of all the weights assigned to the levels of generalization equals 1. If w^e is not in effect, for all $i \leq L_{max}$, $w_i^e = 1$.

4.5.9 General information G

G is a set of general information about anonymization that might help researcher in determining bias and comparing different datasets. The following list shows different elements of G:

- Minimum k for k-outlier handling
- k-outlier handling approach
- Characteristics of generalization hierarchy
- General characteristics of included, or excluded, sensitive attributes
- Sample size ratio $\frac{D^*}{D^0}$

CHAPTER 5

PARV MODEL, LEVEL 3 OF 3: A FRAMEWORK FOR EVIDENCE-BASED POLICY MAKING BASED ON K-ANONYMITY AND *RV*

5.1 Outline and contribution of the third layer of *PARV* model

This chapter covers the third layer of *PARV* model as depicted in figure 1.3. It presents a high level conceptual framework for privacy preserving policy making in context-oriented and evidence-based disclosure of health data. It provides a roadmap for designing, implementing, and evaluating of policies for privacy protection in public health domain without forcing the organization to adopt a certain policy model or technical solution. I provides pseudo-policies with high-level semi-technical and procedural guidelines for protecting the privacy, incorporating epidemiological evidence, and reporting the inevitable semantic information loss to the researcher.

Policy making assists registries in choosing, implementing, and dynamic reviewing of anonymity methods and procedures. This facilitates upgrading and changing the methods based on inevitable temporal and incremental dynamics of datasets, especially in multi-organizational data sharing and joining. A benefit of this framework is that lack of publicly accessible health-related individual data in Turkey. The value of this situation arises from the fact that the outcome of anonymization is considerably affected by publicly accessible data disclosed in the past. With none, or very few, person-specific data accessible, it is possible to regulate data disclosure initiations from scratch which results in higher information content and less privacy leaks. Another contribution of this framework is its emphasis on non-technical contextual aspects such as regulatory, policy, ethics, and expert intervention. The main focus of medical informatics studies in Turkey is on technical issues or clinical procedures. There is a gap in the areas of ethics, social sciences, and patient rights in that the social aspects surrounding public health informatics are usually overlooked. Existing regulations on patient privacy are limited to clinical uses in Turkey which are mandated several years ago (section refsec:regulationBackground). They do not cover the emerging issues of public health informatics and requirements of epidemiological studies. A dynamic model for policy making with flexibility in technical choices and their implementation, facilitates the adaptation of future regulations with minimum convergence time and effort.

5.2 Triangle and MSS approaches for policy process in *PARV* model

The policy is a broad term meaning different things based on context and goal. Policy is defined as "a high-level overall plan embracing the general goals and acceptable procedures especially of a governmental body" in the Merriam-Webster dictionary. It is also defined as "a course or principle of action adopted or proposed by an organization or individual" in Oxford dictionary. The general definition of policy is that it is a framework of activity based on a broad statement of goals, objectives, and means which may or may not be explicitly written [133]. In traditional models, policy making is a step-by-step process including stages such as formulation, implementation, and evaluation whereas modern policy making is seen as an iterative process [134]. One example of stagist approaches] is the ROAMEF cycle for policy making which includes the following [135, 134, 136]:

- A Rationale for policy
- Identification of objectives
- Appraisal of options
- Implementation and monitoring
- Evaluation
- Feedback

In privacy protection, most of the techniques are focused on technical aspects such as the effectivity of privacy protection, data quality, and complexity of methods. Only a few studies have considered a broader view to consider non-technical aspects including actors, people, experts, and policies (refer to section 2.5). However, in evidence-based and context-oriented approaches, focusing on technical issues leads to ineffectiveness of technically oriented policies. The technical approach fits well with rational approaches to policy making. Although it seems straightforward, such step by step and clear cut processes do not reflect the real policy making process as it involves different actors, conflicting interests, and non-compatible goals and values [134].

Since *PARV* model deals with secondary health data, it falls into the public health policy domain. One of the frameworks used in field of health policy is the policy analysis triangle [137, 138] that is adopted in *PARV* model. The health policy triangle is helpful in thinking systematically about policy making in healthcare and it includes four inter-dependent components namely context, content, process[137, 138]. Figure 5.1 shows the policy triangle tailored for *PARV* model.

The context of policy in *PARV* model refers to any setting or condition where privacy of secondary health data is to be protected. It also covers the population research, public health, and epidemiological uses of secondary health data where person-specific datasets are preferred. Additionally, it covers evidence-based research conducted on secondary health data. The involvement of researcher in anonymization process and policy making is a key feature in *PARV* model. This involvement is considered both as an input and output of *PARV* model. As an input, the requirements of researcher and the evidence surrounding the content and usage of dataset is considered in the policy process and the anonymization. As an output, *RV* provides researchers with information about bias which improves the interpretation of the generalizability of the results and also as a feedback mechanism to the registry. The actors in *PARV* model are data holders and registries, public health organizations, health ministry, epidemiologists, public health informaticists, legislators, and ethicists, to count a few. Although many of these actors do not directly involve in policy making process, their interests, roles, and responsibilities affect the context , content, and the processes of the policy. This is true for policy triangle in general where it is a simplified framework representing complex inter-relationship between these components.

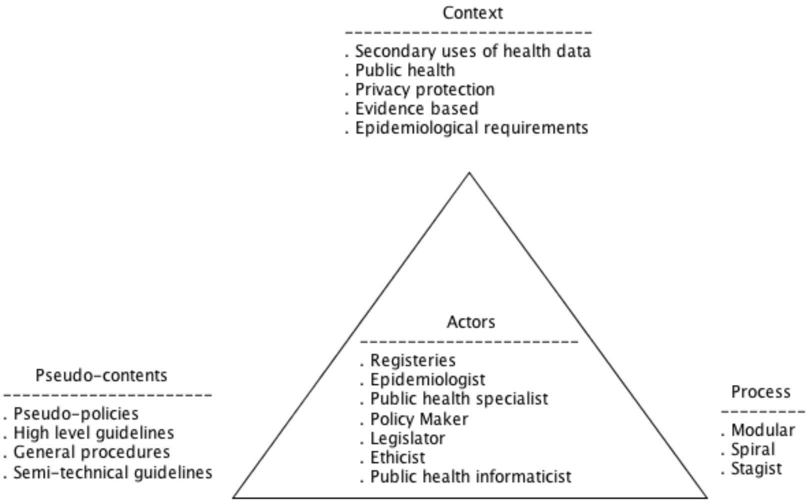

Figure 5.1: Policy Triangle tailored for *PARV* model

The content in *PARV* model is presented as pseudo-contents instead of detailed statements suitable for a specific context, dataset, organization, or requirement. The primary goal of *PARV* model is to provide a generic, dynamic, adaptable, and sustainable model for policy making and anonymization independent of underlying technical methods or policy priorities. Instead of presenting the policy as a product, *PARV* model discusses the policy as a process. As a results, pseudo-content are presented in our model as abstract and high level statements. These pseudo-contents include high level policies, guidelines, best practices, and procedures that affect the evidence-based anonymization, analysis, and *RV* reporting activities. Figure 5.2 depicts the elements of policy process based on [139]. Table 5.1 shows these basic elements along with external factors, actors, and corresponding functionalities and characteristics.

The process in *PARV* model is similar to stagist approaches. We call this approach as MSS which stands for Modular Spiral Stagist approach. Stagist approach follows the classical policy making cycle. It has an spiral characteristic where the policy is not considered as a product, rather, it is an ongoing loop of stages as shown in figure 5.3 and explained in the following list based on [140, 141].

1. *Strategic thinking.* The first step in policy making is strategic thinking which

refers to understanding the problem, context, actors, and desirable outcomes.

2. *Policy development.* The second step is to develop solutions for realizing the strategic points specified in the first step. This step includes collecting evidence, considering alternative options, consultation, expert review, risk assessment, and incorporating recommendations and best practices.

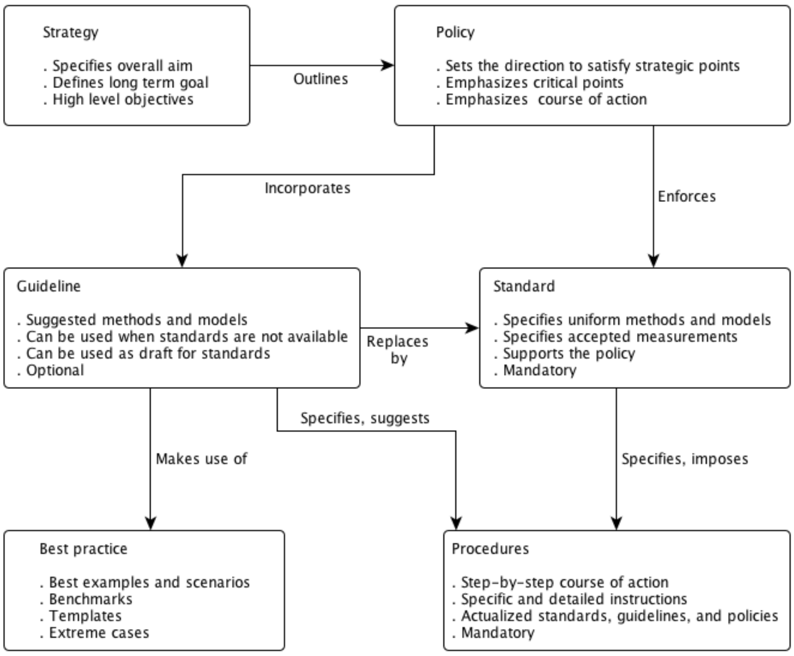
Strategy
. Specifies overall aim
. Defines long term goal
. High level objectives
Outlines
Policy
. Sets the direction to satisfy strategic points
. Emphasizes critical points
. Emphasizes course of action
Incorporates
Enforces
Guideline
. Suggested methods and models
. Can be used when standards are not available
. Can be used as draft for standards
. Optional
Replaces by
Standard
. Specifies uniform methods and models
. Specifies accepted measurements
. Supports the policy
. Mandatory
Specifies, suggests
Makes use of
Specifies, imposes
Best practice
. Best examples and scenarios
. Benchmarks
. Templates
. Extreme cases
Procedures
. Step-by-step course of action
. Specific and detailed instructions
. Actualized standards, guidelines, and policies
. Mandatory

Figure 5.2: Basic elements of policy process

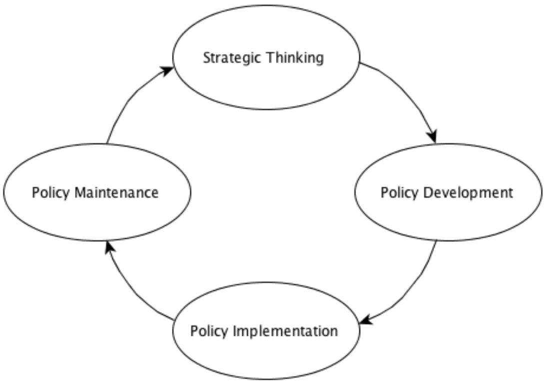

Figure 5.3: Policy making cycle

3. *Policy implementation.* The next step is to put the policy designed in the previous step into practice. In includes testing the options, decision and consensus about an option, supporting the actors, and implementing the policy.

4. *Policy maintenance.* The last step is to maintain the policy, to monitor its implementation, and to periodically evaluate and adjust it accordingly. This step brings the policy making process to the first step and initiates the next round of policy making.

Table5.1: Elements, actors, and factors of policy making in the context of *PARV* model

Element	**Function**	**Characteristics**
Actors	• Determining goals and priories • Performing risk assessment • Policy process	• Managers • Executives
Strategy	• Specifies the overall aim • Defines long term goals • High level objectives	•Abstract and high level
Policy	• Emphasized critical aspects • Highlights the course of action	•General statements •No details
Standards	• Specifies uniform models • Specifies common metrics • Supports the policy	•Mandatory
Guidelines	• Suggested methods and models • Can be used as draft for standards	•Optional
Procedures	• Course of action • Detailed instructions • Actualizes standards and policies	•Detailed •Step-by-step •Mandatory
Outside factors	• Guides strategic objectives • Affects policy cycle • Enforces standards	•Regulatory and legal aspects •Ethical issues •Evidence and context
Best practices factors	• Provides examples	•Scenarios •Benchmarks •Templates
Recipient	• Affected by policy • Receives information • Provides feedback	•Employee •End user •Researcher

5.3 A Modular framework of *PARV* model for evidence-based policy making and anlysis

The policy making layer of *PARV* model presents a modular framework dividing the analysis and policy making tasks into different, and yet, interrelated modules. MSS, which stands for Modular Spiral Stagist, refers to spiral approach for handling different stages of policy process divided among modules. Each module represents the major characteristics of policy process and analysis. These points are not necessarily exclusive as their functionalities are dependent on those of other modules.

Figure 5.4 depicts *PARV* policy making framework in its highest level of abstraction. The framework consists of several modules each following the stages of policy process, as shown in figure 5.3. The outcome of MSS approach is a set of pseudo-contents per module including policies, guidelines, and procedures that are guided by high level strategic thinking. Pseudo-contents are also provided with hypothetical scenarios (refer to chapter 6). The relationship between policy elements, namely strategy, policy, standards, guidelines, and procedures is shown and explained in figure 5.2 and table 5.1.

Figure 5.5 shows the levels and modules of *PARV* policy making framework. It has two levels with the first level covering the high level strategic points and the second level containing the modular policy making elements. These two levels affect each other by means, and as a result, of expert reviews and and researcher involvement.

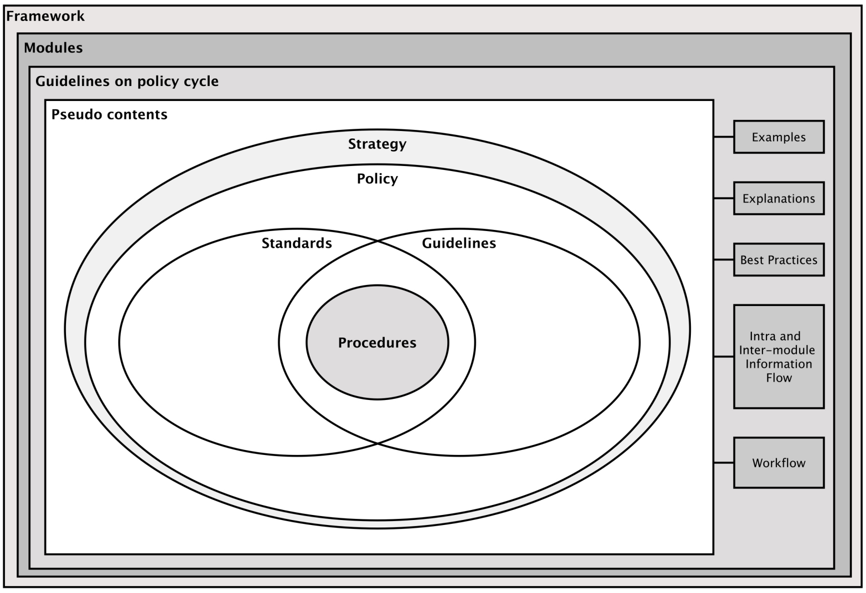

Figure 5.4: High level abstraction of the framework

5.4 First level of *PARV* policy making framework

The first level, as shown in figure 5.4, is the strategic level. It is the first step in policy stages of MSS shown in figure 5.3. In *PARV* model, with its emphasis on evidence-based and context-oriented approaches, it is possible to have a policy outcome which does not fully satisfy the strategic goals or the requirements of the researcher. The evaluation of the policy or the feedbacks from researcher leads to revising the strategies. Unlike local organizational policies that mostly interact with internal actors, in *PARV* model, the involvement of outside actors is emphasized. The reason for higher degree of researcher involvement is the evidence-based and context-oriented nature of *PARV* model. *RV* might seem to be a one way information channel to the researcher. However, it can also be used as a feedback mechanism in order to enhance the anonymization process and improving the generalizability of the results.I also facilitates the evaluation and revising of the policies beginning from new strategic goals. As a result, in *PARV* model, the involvement of actors in shaping the strategic point is not a one way top-down approach. It follows a spiral two-way approach where the decisions not only affect the actors but also shapes around their emerging requirements. Figures 1.3 and 4.7 show how *RV* act both as a reporting tool and a means to collect researcher requirements.

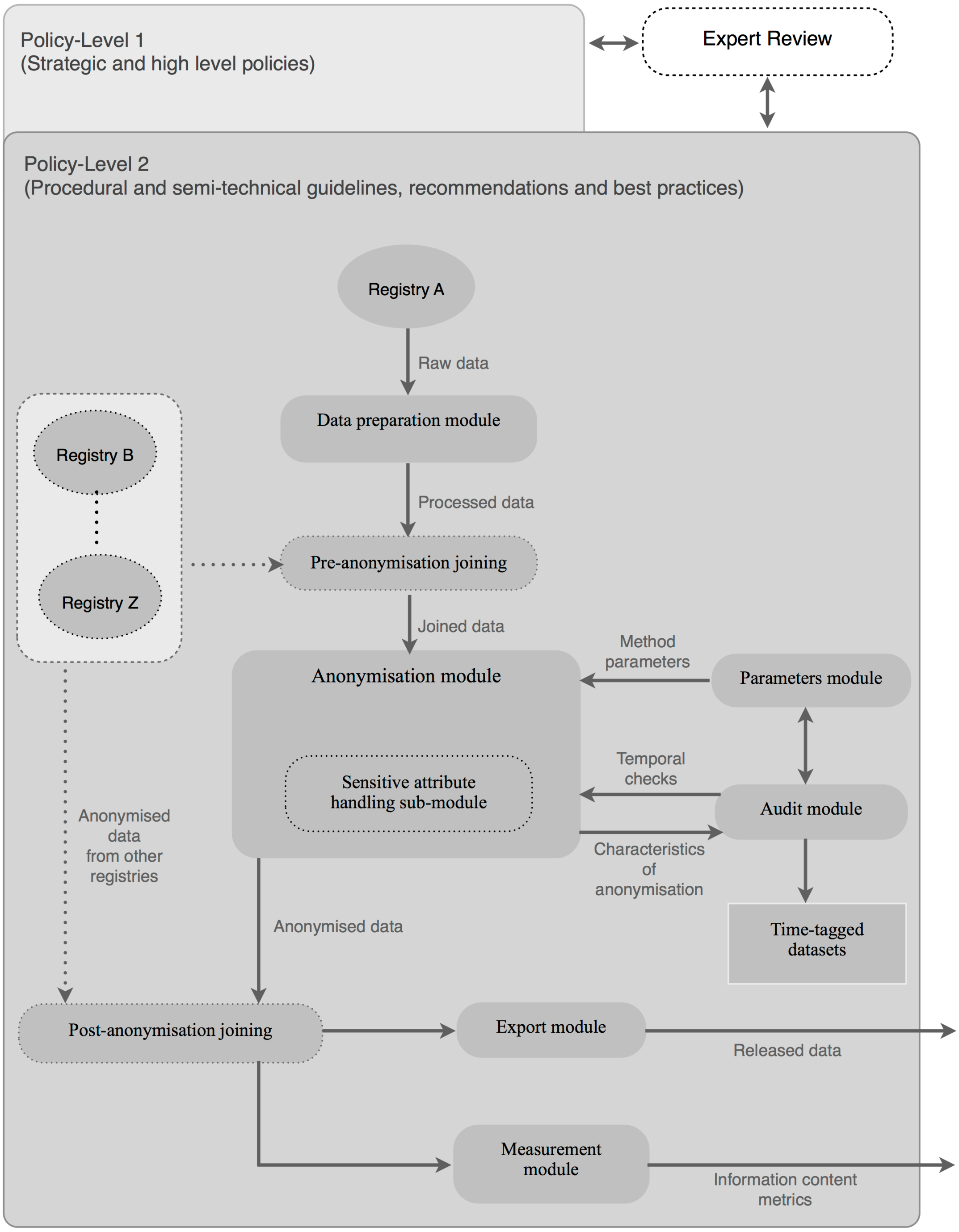
Policy-Level 1
(Strategic and high level policies)
Expert Review
Policy-Level 2
(Procedural and semi-technical guidelines, recommendations and best practices)
Registry A
Raw data
Data preparation module
Registry B
Registry Z
Processed data
Pre-anonymisation joining
Joined data
Anonymisation module
Method parameters
Parameters module
Sensitive attribute handling sub-module
Temporal checks
Audit module
Characteristics of anonymisation
Anonymised data from other registries
Anonymised data
Time-tagged datasets
Post-anonymisation joining
Export module
Released data
Measurement module
Information content metrics

Figure 5.5: Levels and modules of the framework

5.4.1 Prominent and proposed strategies in disclisure of secondary health data

With dynamic nature of data, requirements, and priorities, and also the availability of numerous privacy protection method,it is extremely difficult to choose the most suitable method applicable to all contexts. With predefined generic goals and strategic objectives, it is easier to guide the policy making processes, which includes the choice of the solution. In the context of *PARV* model, we have derived several high level strategic objectives that guide the overall direction of subsequent decisions on strategies and policy making.These high level strategies are specified based on general decision about the following key points:

1. **Data oriented vs Research oriented**
   This strategic point specifies the starting point for the analysis of data for anonymization and policy making. In data oriented approach, a registry initiates the policy making process and anonymization without knowing the recipient or taking into account the requirements of known recipients. Decisions about the parameters, policy choices, and evaluation are made solely based on internal priorities and strategic points and are usually not following the evidence. In research oriented approach, the policy making and anonymization processes are performed based on the requirements of researcher and the available evidence. In this approach, a registry makes the policies while incorporating at least one of the following:

   - Prior information about the recipient.
   - Performing the disclosure based on the requirements or characteristics of recipient
   - Making policies with recipient requirement as a critical determinant.
   - Repeating disclosure upon receiving feedback from researcher.
   - Considering the evidence and context.
   - Informing the researcher about the details of information loss.

2. **Aggregation vs person-specific**
   Data sensitivity, and the context in which it is to be disclosed, guide another important strategic objective which ultimately determines the specificity of data. When data is sensitive in terms of privacy, and the overall goal is to prioritize privacy protection to information content, aggregation strategy is preferable . In this approach, the policy mandates transforming datasets into statistical summaries for groups of people or the whole dataset. Aggregated datasets hamper the generalizability of the results to the whole population by introducing ecological fallacy. Person-specific strategy, on the other hand, emphasizes on disclosing the dataset as tables with each row containing the data belonging to an individual. The advantage of this strategy is that it enables epidemiologist to conduct case-control and cohort studies with high generalizability of the results. The disadvantage, however, is higher risk to privacy. This strategy is suitable for datasets that are less sensitive in terms of privacy, either due to the nature of chosen attributes or the effect of sample size on the outcome of k-anonymity. It

is also suitable when there is a lack of related publicly accessible datasets leading to less known information about an individuals. Additionally, it is preferable where information content and accuracy are more important than strict privacy protection.

3. **Restriction vs openness of information loss details**
   The restricted strategy refers to approaches where the details of information loss are not provided to the researcher. The recipient of the anonymized dataset would not have any information about the bias introduced as a result of anonymization. In an open strategy, the details of the information loss, assumptions prior to anonymization, and manipulations to the dataset are reported to the recipient. This can be used to objectively quantify, or have a general idea about, the introduced bias. Additionally, it can be used as a feedback mechanism for re-anonymizing the dataset or improving the policies that govern the anonymization.

Table 5.2 shows different combination of the key strategic points and their outcomes. The first row shows the strategic points and the outcome of the popular and prominent approaches taken by public health organizations in Turkey. While other combinations in other rows are possible, many of them are not practiced. The last row represents the characteristic of our model as a whole. It encompasses the other cases as special cases.

5.5 Second level of *PARV* policy making framework

The second level of *PARV* policy making framework divides the concepts and tasks of policy making into inter-related modules and elements. These modules cover the stages of policy making taking into account the public health elements such as data sources, registries, or data flows. This section covers the details of these modules and elements. Modules provide concepts, generic guidelines for policy process, pseudo-policies, actors, information flows, input, output, and abstract procedures.

5.5.1 Data sources and datasets

Several data sources are used and maintained in *PARV* policy framework. Data sources in *PARV* model refers to actual datasets containing person-specific or aggregated data. Some of these data sources include sensitive and health-related attributes while others are general datasets containing census or publicly accessible data. Primary data sources in *PARV* policy framework are shown in the following list:

- **Publicly accessible data sources.** Refers to datasets that are maintained about individuals by private or national organizations. Most of these data sources contain census and demographic information that might be used to link with other de-identified datasets to re-identify the individuals.

Table5.2: Oucome of the general strategies for privacy protection in epidemiological context

no.	Obj 1		Obj 2		Obj 3		Outcome
	Data oriented	Research oriented	Aggregation	Person specific	Internal bias	Reported bias	
1	✓	–	✓	–	✓	–	• Unknown recipient • Ecological studies only • Highest information loss • Inclusion of unnecessary attributes
2	✓	–	–	✓	✓	–	• Unknown recipient • Suitable for inference studies • High information loss • Unknown bias • Inclusion of unnecessary attributes
3	✓	–	*	*	–	✓	• Unknown recipient • Ecological for aggregated groups (in any) • Inference studies for person specific groups (if any) • Bias is known • Inclusion of necessary attributes only • Information loss is high or maximum based on Obj2
4	–	✓	✓	–	✓	–	• Demand and feedback based • Only for ecological studies • Bias is unknown • Necessary attributes only • Information loss drops
5	–	✓	–	✓	✓	–	• Demand and feedback based • Suitable for inference based • Bias is unknown • Necessary attributes only • Less overall information loss • Uncertainty about generalization is no measurable
6	–	✓	*	*	–	✓	• Demand and feedback based • Ecological for aggregated groups (in any) • Inference studies for person specific groups (if any) • Bias is known • RV is known (if applied) • Necessary attributes only • Less overall information loss • Satisfies the criteria for a comprehensive policy making

- **Population sample dataset.** Refers to health-related data gathered about individuals with privacy and security protection methods in place. These data sources are gathered directly from a sample population or from clinical datasets. The primary aim of the collection of these datasets is public health and epidemiological research.

- **Registry sample dataset.** Refers to a subset of population sample dataset that is collected and maintained by a registry.

- **De-identified dataset.** Refers to a processed registry sample dataset with its directly identifiable attributes removed.

- **Anonymized dataset.** Refers to k-anonymized, or aggregated, dataset that are finalized based on the privacy risk analysis and mitigation results of the registry.

- **Audit dataset.** Refers to anonymized datasets disclosed in the past. These datasets will be used in the future disclosed to spot any privacy leaks due to temporal disclosures.

- **Linking dataset.** Refers to anonymized or disclosed dataset with additional mechanisms for spotting duplicate records and linking datasets from different registries.

- **Exported dataset.** Refers to the post anonymization dataset that is to be disclosed to the researcher or the public.

5.5.2 Data Preparation Module (DPM)

Data Preparation Module (DPM) focuses on preparing data for anonymization. In *PARV* policy framework, DPM only contains concepts affecting anonymization and *RV*. It is assumed that the general pre-processing of data has already been applied. Table 5.3 shows the factors affecting the context, content, and actors of DPM.

Table5.3: Factors affecting the context, content, and actors of DPM

Component	**Explanation**
Aim	Preparing dataset for anonymization or RV analysis by choosing attributes and specifying their characteristics
Assumptions	Data pre-processing not related to anonymization or RV calculation has already performed. This pre-processing includes handling of missing values, normalization, coding , and classification.
Elements	Person p, dataset D, attributes A_D=$I_D \cup Q_D \cup S_D$, record r, Unique Health Identifier
Actions	Pre-anonymization processing, attribute tagging, de-identification f_d, unique record identification
Input	Dataset D belonging to registry sample population P^{rs}, previous datasets audited in Audit Module (adm)
Output	De-identified dataset D^0

The following list shows the key points in defining the policy process related to DPM.

Strategic thinking

The goal is to conceal personal identification of individuals revealed by inherent characteristics of the dataset.

Points for policy development and implementation

This section provides the high level policy points and pseudo-procedures of DPM. Assume that D=$\{r_1, r_2, \ldots, r_d\}$ is the dataset of d records representing the sample population P^{rs} held by registry reg_a, and A_D=$I_D \cup Q_D \cup S_D$ is the set of attributes . Let A_D=$\{a_1, a_2, \ldots, a_n\}$ be the set of n attributes of dataset D.

- *dpm-1.* **Identifying explicit identifiers.** All directly identifiable attributes must be tagged as identifiable. An explicit identifier reveals the identity of an individual without any further information or linkage to other attributes [7].
  - *dpm-1-1.* For all $1 \leq i \leq n$, if a_i is explicitly identifier then include a_i in I_D
  - Further tagging based on expert review and the analysis of characteristics of census data
  - Example: name, last name, social security number
- *dpm-2.* **Identifying implicit identifiers.** All attributes that might be linked with other personal information to infer the person identity must be tagged as identifiable .
  - *dpm-2-1.* For all $1 \leq i \leq n$, if a_i is implicitly identifier then include a_i in I_D
  - Further tagging based on expert review and the analysis of characteristics of census data
  - Examples: email, phone number, address
- *dpm-3.* **Identifying quasi-identifiers.** All attributes that might link with attributes in external sources to infer about person identity, or to increase the risk of re-identification, must be tagged as quasi-identifier.
  - *dpm-3-1.* Make $A^{'}_D$=$\{a_m, \ldots, a_{m'}\}$ as the set of attributes that are not directly identifiable. In other words, $A^{'}_D$=$(A_D - I_D)$.
  - *dpm-3-2.* Exclude non health-related attributes from $A^{'}_D$ that are also gathered by other organizations and are included in publicly accessible datasets, or in controlled datasets that might be disclosed in the future.
  - *dpm-3-3.* Find attributes in $A^{'}_D$=$\{a_m, \ldots, a_{m'}\}$ that when combined together and linked with external data, might uniquely identify a person. Include such attributes in Q_D.
  - *dpm-3-4.* Tag attributes in Q_D as quasi-identifiers.
  - Further tagging based on expert review and the analysis of characteristics of census data, population health data, data analysts, and risk analysis.

  - Examples: combination of zip code, gender, birthdate, ethnicity, occupation, and marital status.

- *dpm-4.* **Sensitive attributes.** The attributes that are exclusively gathered for the research purposes managed by the registry, are collected for the specific purpose of a study, are health-related, and are not gathered by any other registry must be tagged as sensitive attributes. These are the most sensitive attributes in terms of privacy protection in health domain. Any relation between a person and the value of sensitive attribute must be kept secret.
  - *dpm-4-1.* For all $1 \leq i \leq n$, if a_i is health-related, has not been gathered by the registry, contains confidential data, might be a dependent variable, and is considered confidential by other registries must be tagged as sensitive attribute and included in S_D.
  - Further tagging based on expert review and the analysis of characteristics of census data, clinical and population health data, epidemiology, legal and regulations, and ethics.
  - Examples: HIV status, presence of a disease, value of a test with a known cut-off point, psychiatry results, and drug abuse.

- *dpm-5.* **Specifying a unique record identifier.** Each record must be tagged with a unique identifier for future references and for joining tasks of JNM. Any relation between the assigned identifier and other personal identification attributes should not be included in the disclosed dataset.
  - *dpm-5-1.* Identify the set of attributes $I_D^{'}$ that are used, or can be used, as Unique Health Identifier UHI .
  - *dpm-5-2.* For all $1 \leq i \leq d$, assign a unique record identifier URI_i. This identifier is only known and assigned by registry. URI_i can be independent from UHI_i by linking it to an internally assigned identifier. URI_i can be dependent on UHI_i by applying a mechanism to transfer the value of UHI_i to URI_i. If the transfer mechanism is a two-way and reversible , the mechanism for transfer should not be disclosed.
  - Examples of UHI are social security number, combinations of demographic data, and national unique health identifier. Encryption can be used to transfer UHI to URI.
  - *dpm-5-3.* Make an index between URI_i and the corresponding value of UHI_i.
  - The inclusion of URI in final disclosed dataset depends on the risk analysis, expert decision, and possible future joining requirements.

- *dpm-6.* **De-identification.** All implicit, explicit, and other identity revealing attributes must be removed.
  - *dpm-6-1.* Make D^0=$f_d(D)$=$(D - I_D)$.

5.5.3 Parameters Module (PRM)

Parameters Module (PRM) is the module providing the means for incorporating the evidence in anonymization and analysis of *PARV* model. The policies and elements in this module specify the parameters used in anonymization or in analysis processes specified in *PARV* model parts 1 and 2. The parameters affect the choice of anonymization method, choice of the cluster validity index used to calculate RV, and incorporating evidence in distance measurement. Table 5.4 shows the factors affecting the context, content, and actors of PRM. Table 5.5 shown the main parameters used in *PARV* model.

Table5.4: Factors affecting the context, content, and actors of PRM

Component	Explanation
Aim	Specifying parameters based on evidence and researcher requirement and incorporating them in anonymization module (ANM) and measurement module (MRM)
Assumptions	The researcher needs are known and are included in risk analysis and decisions made by experts.
Actors	epidemiologist, public health specialists, public health informaticists, data mining expert, risk analyst, expert on census data
Elements	Q_D, S^D, f^G, f^s
Actions	Choosing quasi-identifiers and clusters, Assigning weights, Incorporating evidence and researcher needs
Input	Evidence, Researcher requirement, Expert decisions, D^0
Output	Parameters of anonymization, Parameters of RV calculation in Measurement Module MRM

The following list shows the key points in defining the policy process related to PRM.

Strategic thinking

To incorporate the evidence and research based parameters in anonymization, analysis of anonymized datasets, and in reporting RV.

Points for policy development and implementation

This section provides the high level policy points and pseudo-procedures in PRM.

- *prm-1.* **Specify k.** Specify the minimum k for k-outlier analysis.
  - option 1. based on a pre-defined k.
  - option 2. based on an optimized k based on dataset and constraints.
  - option 3. based on expert decision and risk analysis results.
- *prm-2.* **Set of quasi-identifiers.** Specifying the quasi-identifiers that are needed for the research context. The aim is to reduce the multi-dimensionality caused be un-necessary quasi-identifiers.
  - *prm-2-1.* For all $q_i \in Q_D$, specify whether q_i is required for the research context based on expert decision, risk analysis, researcher requirement, or known evidence.

Table5.5: Parameters of PRM

Parameter	**Explanation**
k	Minimum k for k-anonymization and k-outlier analysis
k-outlier handling	Preferred approach for k-outlier analysis based on expert decision, risk analysis, and research context
Q_D	Set of quasi-identifiers to be included in analysis based on the context of the research and the requirements of the researcher
Semantics of generalization	In generalization based k-outlier handling method, this parameter specifies the semantics of domain generalization hierarchy (DGH) or value generalization hierarchy (VGH) for each $q_i \in Q_D$
Normalization of DGH	Specifies whether the DGH of each $q_i \in Q_D$ is normalized based on the minimum and maximum generalization levels of all quasi-identifier
Hyper-plane distance measurement	Specifies whether hyper-plane semantic distance measurement is meaningful and prefereble to non-semantic distance measurement for each $q_i \in Q_D$
w_i^q	Evidence based quasi-identifier weight parameter, which indicates a weight between 0 and 1 given to each quasi-identifier based on its importance to the research context. All quasi-identifiers are assigned with a weight and the total weight of all quasi-identifiers is 1
w_n^e	Evidence based generalization level weight parameter, which indicates a weight between 0 and 1 given to all the levels of DGH for each $q_i \in Q_D$. This weight is assigned based on the researcher requirement indicating the negative affect of generalization on their models
Initial generalization level	Evidence based initial generalization level, which indicates the starting point for generalization based on research context. This parameter makes the DGH length smaller and prevents bias of unnecessary generalization levels in DGH
Required clusters	Specifies the equivalent clusters that are needed by researcher
Sample size details	Specifies the dynamics of sample size such as the ration of original sample size to the post-anonymization sample size
Aggregation	Specifies the dynamics of aggregation applied on sensitive attribute for the whole dataset or per equivalent cluster

- *prm-2-2.* Check whether the seemingly un-necessary q_i is a possible confounding factor based on evidence.
- *prm-2-3.* Exclude un-necessary quasi-identifier from Q_D.
- Example: if the aim of research is to find a causal relationship between age and drug abuse regardless of gender, the gender quasi-identifier can be excluded. If there is a chance that gender is a confounding factor it must be included in Q_D.

- *prm-3.* **Semantics of generalization.** Specifying the details of generalization for each $q_i \in Q_D$.
  - *prm-3-1.* For each $q_i \in Q_D$, specify the generalization hierarchy by applying $f^G_{q_i}$, the maximum level of generalization before suppression, and whether to include the suppression on top of the DGH.
  - *prm-3-2.* If at least one of the quasi-identifiers have the suppression included on top of the DGH, apply the same on other quasi-identifiers.

- *prm-4.* **Normalization of DGH.** Apply normalization on DGH of all included quasi-identifiers if varying levels introduce bias.
  - *prm-4-1.* Decide whether varying maximum level of generalization L_{max_i} introduces bias. If expect decision calls for normalization, use the following pseudo-procedures in related modules.
  - *prm-4-2.* Specify the minimum and maximum L_{max_i} of Q_D. Let G_{max} and G_{min} denote the maximum and minimum values found respectively. If $G_{max} \neq G_{min}$ t the next pseudo-procedure is applicable.
  - *prm-4-3.* Let h^d_i denote the generalization at d^{th} level for q_i. Normalize h^d_i as $\frac{|h^d_i - G_{min}|}{|G_{max} - G_{min}|}$.

- *prm-5.* **Hyper-plane distance measurement.** Specifies the parameters for semantic distance measurement for each quasi-identifier.
  - *prm-5-1.* Decide whether to incorporate semantics of the distance for each $q_i \in Q_D$ based on researcher needs, well known semantics of data, expert decisions, and evidence.
  - Example: decide whether to incorporate the actual geographical distance for zip code and whether the value of zip code contains information about actual geographical point or distance.
  - *prm-5-2.* Specify the minimum and maximum amounts for semantics values for any normalization.
  - *prm-5-3.* For each h^d_i of each q_i, specify the semantic centroid of each equivalent cluster ξ^* at level d in real data at Ξ^0.

- *prm-6.* **Quasi-identifier weight parameters.** Specify the wights to be assigned to each quasi-identifier. This balances the effect caused by each quasi-identifier or put a higher weight on the ones that the research context specifies.
  - Let Q_D=$\{q_1, q_2, \ldots, q_l\}$.
  - *prm-6-1.* Determine the importance of the each $q_i \in Q_D$ in the distance measurement based on its contribution to the causal relationship.

- *prm-6-2.* Assign a weight w_i^q to q_i where $0 < w_i^q \leq 1$ and $\forall 0 \leq i \leq l$: $\sum_{i=1}^{l} w_i^q$=1.
- This weight is specified by epidemiologist based on the context of the research, known dependent and independent variable and the degree of their importance, the possibility of an attribute to be a confounding variable, and the available evidence regarding the casual relationship.

- *prm-7.* **Generalization level weight parameters.** Specify the wights to be assigned to each generalization level of each quasi-identifier. This specifies the effect of generalization on the results based on researcher needs and the evidence.

  - *prm-7-1.* Determine the importance of the each level of generalization h_i^d of q_i.
  - *prm-7-2.* Assign a weight w_n^e ranged from 0 to 1 to each level h_i^d where $0 < w_n^e \leq 1$ and $\forall 0 \leq n \leq L_{max_i}$: $\sum_{n=1}^{L_{max_i}} w_n^e$=1.
  - This weight is specified by epidemiologist based on the effect of generalization of a quasi-identifier on the analysis of causal relationship.

- *prm-8.* **Initial generalization level.** Specify the zero level for the generalization hierarchy of a quasi-identifier based on the researcher need. This leads to less nonconforming equivalent classes leading to less required generalization for handling k-outliers.

  - *prm-8-1.* Provide the researcher with DGH for each q_i from a hypothetical dataset and ask for the minimum level of generalization to begin the DGH from.
  - *prm-8-2.* Assign to each q_i an initial generalization level based on researcher feedback.
  - Example: assume that the researcher is interested in finding the relationship between a health condition and people living in areas separated as NUTS (Nomenclature of Territorial Units for Statistics) [142]. In this case, the initial level for geographical location can be NUTS codes, eliminating lower levels. This reduces the number of non-conforming classes caused by geographical location and eventually leads to less manipulation required to maintain k-anonymity.

- *prm-9.* **Required clusters.** For each quasi-identifier, specify the groups and clusters which researcher requires.

  - *prm-9-1.* Ask for required groups of people that researcher requires based on evidence and context.
  - *prm-9-2.* Tag records of non-required clusters as un-necessary records.
  - Example: assume that researcher is interested in finding the relationship between a health condition and the age of people younger than 20. In this case, the records of people older than 20 are not required and their inclusion would introduce more variability to age quasi-identifier.

- *prm-10.* **Aggregation** Specify the aggregation type that must be applied on each cluster or the whole dataset.

5.5.4 Anonymization Module (ANM)

Anonymization Module (ANM) is the core module for analyzing the anonymization of the dataset based on k-anonymity. It contains the layer 1 of *PARV* model focusing on the general approach for k-outlier handling. As a result, the actual anonymization algorithm is considered as a black box in this module. The aim is to analyze the k-outlier handling method without the need to know the details of the anonymization algorithm. In order to shed light on the dynamics of anonymization, the general high level procedures for basic k-anonymization will also be provided in this section. Table 5.6 shows the factors affecting the context, content, and actors of ANM.

Table5.6: Factors affecting the context, content, and actors of ANM

Component	**Explanation**
Aim	Determining and enforcing k-outlier handling approach in order to render dataset k-anonymous based on the evidence reflected in the parameters determined in PRM
Assumptions	Anonymization is based on hard clustering, Anonymization results resembles k-anonymity, The un-anonymized de-identified dataset D^0 is accessible, Parameters and k-outlier handling approaches are known, The registry is aware of all the publicly accessible related datasets, Regardless of the public or controlled access to data, the dataset will be person specific
Actors	Privacy analyst, data analyst, data mining expert
Elements	D^0, k-outlier handing approach, Q^D, S^D, k. Ξ^0, Ξ^*
Actions	k-outlier detection, k-anonymization, Clustering, k-outlier handling approach, Multi-step cluster analysis
Input	De-identified dataset D^0, Parameters of anonymity including k, Q^D, and k-outlier handling approach, Expert decisions
Output	k-anonymized dataset D^*, Parameters for RV calculation in Measurement Module MRM

The following list shows the stages of the policy process of ANM. It is assumed that the dataset has already been anonymized. In this case, the anonymization is considered as a black box. The aim of ANM, therefore, is to find the k-outlier handling approach and determine the parameters that would be used by other modules including MRM.

Strategic thinking

To analyze the approach to handle k-outliers based on clustering viewpoint.

Points for policy development and implementation.

This section provides the high level policy points and pseudo-procedures in ANM.

- *anm-1.* **Specify k.** Determine the minimum k from *prm-1.*
    - option 1. based on a pre-defined k.
    - option 2. based on an optimized k based on dataset and constraints.

  - option 3. based on expert decision and risk analysis results.

- *anm-2.* **Performing first step clustering .** Find k-outliers of the original dataset.

  - *anm-2-1.* Find the unique combination of associated values of quasi-identifiers in D^0 and map them onto $\Xi^0=\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$.
  - *anm-2-2.* Find all $\xi^i \in \Xi^0$ with cardinality less than k. These clusters are called k-outliers. Add them to the set Δ_λ.

- *anm-3.* **Multi-step cluster analysis for anonymization.** Application of clustering in several steps based on k-outlier handling approaches until k-anonymity holds. *anm-3* is for anonymization purposes. It is assumed in *PARV* model that the details of the anonymization method is not known. The contents in this part are high level with an aim at applying the k-outlier approaches.

  - *anm-3-1.* Find the unique combination of the associated values of quasi-identifiers in D^0 and map them onto $\Xi^0=\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$.
  - *anm-3-2.* Choose the k-outlier handling method. Table 5.7 shows the points for expert decision making about the approach selection.
  - *anm-3-3.* Apply the k-outlier handling method.
  - *anm-3-4.* Repeat *anm-3-2* and *anm-3-3* if the expert review calls for more privacy protection. The final cluster is called $\Xi^*=\{\xi_1^*, \xi_2^*, \ldots, \xi_{\lambda'}^*\}$.
  - *anm-3-5.* Apply dataset-wide or cluster based aggregation according to *prm-3-4*

Table5.7: Points for expert decision making about the choice for k-outlier handling approach

Approach	**Points for decision making**
Delete	Low number of small sized k-outliers with larger inter-cluster distances larger than the maximum distance between conforming clusters. For algorithms that delete records based on patterns in sensitive attributes. If some clusters are not required according to policy point *prm-9*.
Generalization	The original sample size should be kept intact. Number of k-outliers are high k-outliers have small inter-cluster distance.
Merging	The original sample size should be kept intact. Number of k-outliers are low k-outliers have small inter-cluster distance.
Fabrication	Number of k-outliers are low The difference between the cardinality of non-conforming clusters and k is low. The original values of the attributes and having access to all sample records are more important than duplicate or actual outlier records.

- *anm-4.* **Multi-step cluster analysis for analysis of anonymizaiton.** Analysis of the k-outlier handling to make the necessary inputs for *RV* calculation.

  - *anm-4-1.* Find the unique combination of the associated values of quasi-identifiers in D^0 and map them onto $\Xi^0=\{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$.

- *anm-4-2* Specify the applied k-outlier handling methods on Ξ^0 by registry. The details of the k-anonymization is not required. Use points for expert decision making shown in *anm-3-2*, *prm-10*, and table 5.7 for specifying k-outlier handling method.
- *anm-4-3.* Specify the last result of the multi-step clustering as Ξ^*, which is the set of all equivalent clusters conforming to k-anonymity.

5.5.5 Measurement module (MRM)

Measurement Module (MRM) covers the second layer of *PARV* model, namely the clustering based analysis of the information content and calculating the elements of RV. It takes the characteristics of anonymization from ANM, along with parameters from PRM, and provides policy points regarding RV calculation. Table 5.8 shows the factors affecting the context, content, and actors of MRM. The following list shows the stages of the policy process for MRM.

Table5.8: Factors affecting the context, content, and actors of MRM

Component	**Explanation**
Aim	Specifying the RV based on the dynamics of anonymization in ANM and parameters of PRM without increasing the risk of re-identification beyond the point actualized by anonymization
Assumptions	k-outlier handling and the parameters determined by anonymization which might affect RV calculation are known without a need to know the details of the underlying algorithm, RV calculating team have access to the original data and are governed by the same policies
Actors	Privacy analyst, data analyst, data mining expert, risk analyst
Elements	Ξ^0, Ξ^*, k, Q_D, S^d, I^o, I^c, A^e, A^q, C^e, S^q, M^s, W
Actions	Calculating cluster validity index by comparing clusters of Ξ^0 and Ξ^*, Translating parameters in PRM into elements of RV
Input	De-identified dataset Ξ^0, Disclosed dataset Ξ^*, Parameters of anonymity including k, Q^D, and k-outlier handling approach, Expert decisions
Output	RV

Strategic thinking

To calculate the information loss, along with other details of anonymization, in order to report them to researcher in the form of elements of RV. For more detail refer to 4.5.

Points for policy development and implementation.

This section provides the high level policy points and pseudo-procedures in MRM.

- *mrm-1.* **Initializing the second step analysis of *PARV* model.** Preparing the input for second step clustering analysis.

  - *mrm-1-1.* Specify k from ANM and PRM.
  - *mrm-1-2.* Specify k-outlier handling approach from ANM and PRM.
  - *mrm-1-3.* Specify Ξ^0 and Ξ^* from ANM.
  - *mrm-1-4.* Specify the goal of RV calculation based on the strategies presented in section 5.4.1. Based on strategic priorities, RV might be used internally by the registry or reported to the researcher.

- *mrm-2.* **Caclulate information content.** Perform information content measurement per cluster and for the whole dataset.

  - *mrm-2-1.* Choose the information content measurement to use based on expert decision or PRM. For clustering based measurements of $PARV$ model go to *mrm-2-3.*
  - *mrm-2-2.* For non-clustering based measurements follow the procedures of each measurement. Add each measurement to I^o for overall measurement and I_i^c for per cluster measurement.
  - *mrm-2-3.* This step and the following steps are for clustering based measurement of information content. Let $\Xi^0 = \{\xi_1^0, \xi_2^0, \ldots, \xi_\lambda^0\}$ be the sample data and $\Xi^* = \{\xi_1^*, \xi_2^*, \ldots, \xi_{\lambda'}^*\}$ be the anonymized dataset.
  - *mrm-2-4.* For each $\xi_j^* \in \Xi^*$ find the set of equivalent classes in Ξ^0 that have been mapped to ξ_j^*. In other words find the set $b_j = \{\xi_m^0, \ldots, \xi_{m'}^0\}$ such that $f^*(b) = \xi_j^*$ where f^* is the multi-step clustering function for anonymization.
  - *mrm-2-5.* For each quasi-identifier of ξ_j^*, if generalization approach is applied, calculate the semantics of generalization using domain generalization hierarchy from ANM and PRM.
  - *mrm-2-6.* If semantics of data is to be applied according to PRM, incorporate the hyperplane distance measurement.
  - *mrm-2-7.* Apply the initial generalization level from PRM.
  - *mrm-2-8.* Apply evidence-based generalization level weight w_n^e, if specified in PRM.
  - *mrm-2-9.* Apply evidence-based quasi-identifier level weight w_n^e, if specified in PRM.
  - *mrm-2-10.* Repeat *mrm-2-5* to *mrm-2-9* for all quasi-identifiers of ξ_j^*.
  - *mrm-2-11.* For all quasi-identifier of ξ_j^*, find the distance between quasi sub-records of ξ_j^* and b.
  - *mrm-2-12.* Combine the distances found in step *mrm-2-11* in modified general distance coefficient shown in equation 4.6.
  - *mrm-2-13.* Repeat *mrm-2-4* to mrm-2-12 for all equivalent clusters in Ξ^*.
  - *mrm-2-14.* Use the distances found in previous procedures in cluster validity index.
  - *mrm-2-15.* Normalize the index based on a common normalization standard for I^o and I^c.
  - *mrm-2-16.* Add the results of *mrm-2-15* for each equivalent cluster to I^c if required. Add the results of *mrm-2-15* for the entire Ξ^* to I^o if required.

- *mrm-3.* **Excluded attributtes.**Make the set of excluded attributes from the de-identified dataset.
  - *mrm-3-1.* Take de-identified dataset D^0 as the starting point. For all attributes of Q^{D^0} and S^{D^0} perform the following steps.
  - *mrm-3-2.* Specify quasi-identifier that must be completely concealed or suppressed based on PRM, risk analysis, or researcher requirement.
  - *mrm-3-3.* Specify sensitive attribute that must be completely removed based on PRM or risk analysis.
  - *mrm-3-4.* Make the meta-data of excluded attributes by specifying their name, type, range, aim of collection, and reason for exclusion.
  - *mrm-3-5.* Add the meta-data obtained in step *mrm-2-4* to A^e.
  - *mrm-3-6.* Repeat steps *mrm-3-2* to *mrm-3-5.* for the next quasi-identifier or sensitive attribute.
- *mrm-4.* **Included quasi-identifiers.** Make meta-data for quasi-identifiers that must be, or requested to be, included in disclosed dataset.
  - *mrm-4-1.* Take de-identified dataset D^0 as the starting point. For all attributes of Q^{D^0} perform the following steps.
  - *mrm-4-2.* Specify quasi-identifier that must be included in anonymization based on PRM, risk analysis, or researcher requirement.
  - *mrm-4-3.* Make the meta-data of the included quasi-identifiers including their name, type, range, aim of collection, and reason for inclusion. If the inclusion is based on researcher requirement or evidence, also include the information about the evidence in meta-data.
  - *mrm-4-4.* Add the meta-data obtained in step *mrm-4-3* to A^q.
  - *mrm-4-5.* Repeat steps *mrm-4-2* to *mrm-4-4.* for next quasi-identifier.
- *mrm-5.* **Excluded clusters.** Make meta-data for equivalent clusters, or classes, that or requested to be, excluded from dataset.
  - *mrm-5-1.* Take de-identified dataset Ξ^0 as the starting point. For all equivalent classes $\xi_i^0 \in \Xi^0$ perform the following steps.
  - *mrm-5-2.* Specify equivalent classes that must be deleted based on the associated values of quasi-identifiers. Make this decision based on PRM, risk analysis, or researcher requirement.
  - *mrm-5-3.* Specify records that must be deleted based on the associated values of sensitive attributes. Make this decision based on the algorithm, PRM, risk analysis, or researcher requirement.
  - *mrm-5-4.* Perform another round of multi-step clustering to achieve anonymization.
  - *mrm-5-5.* Repeat *mrm-5-2* to *mrm-5-4* until Ξ^* is finalized.
  - *mrm-5-6.* Include the meta-data about the deleted records in C^e without giving any information about the relation between quasi-identifiers and sensitive attributes of the deleted records.
  - *mrm-5-7.* Calculate and report the new sample size.

- *mrm-6.* **Semantics.** Specify the overall information about the semantic of data and generalization hierarchy.
  - *mrm-6-1.* Take de-identified dataset D^0 as the starting point. For all attributes of Q^{D^0} perform the following steps.
  - *mrm-6-2.* Specify the inherent semantics of the coding applied on quasi-identifier. This semantics reveals the relationship between the value assigned to a quasi-identifier and the real situation that the quasi-identifier is representing.
  - *mrm-6-3.* Specify the domain generalization hierarchy following PRM and ANM.
  - *mrm-6-4.* Specify the semantics of associated values at each level of generalization and its relationship with higher and lower levels.
  - *mrm-6-5.* Specify which parts of this semantic must be included in the meta-data of the quasi-identifier based on risk analysis, PRM, or researcher requirement. Add this meta-data to S^q.
  - *mrm-6-6.* Repeat steps *mrm-6-2* to *mrm-6-5.* for the next quasi-identifier.
- *mrm-7.* **Sensitive attribute.** Specify the aggregation details of sensitive attribute.
  - *mrm-7-1.* Take de-identified dataset D^0 as the starting point. For all sensitive attributes of S^{D^0} perform the following steps.
  - *mrm-7-2.* Specify the type of aggregation that is to be applied on each cluster $\xi_j^* \in \Xi^*$ based on risk analysis, PRM, or researcher requirement.
  - *mrm-7-3.* Add this meta-data to M^s.
  - *mrm-7-4.* Repeat steps *mrm-7-2* to *mrm-7-3.* for the next sensitive attribute.
- *mrm-8.* **Weight measurements.** Specify the weight parameters.
  - *mrm-8-1.* Take de-identified dataset D^0 as the starting point. For each $q_i \in Q^{D^0}$ perform the following steps.
  - *mrm-8-2.* Specify the k-outlier handling approaches applied on q_i from PRM and ANM. If generalization approach has not been applied in any steps of multi-step clustering, then go to step *mrm-8-6.*
  - *mrm-8-3.* Specify the domain generalization hierarchy of q_i from PRM and ANM.
  - *mrm-8-4.* Determine w^e which is the weight to be assigned to each levels of domain generalization hierarchy based on expert decision, evidence, predefined values, or researcher requirement. Each weight is $0 \leq w_n^e \leq 1$ and the sum of all weights along domain generalization hierarchy is 1.
  - *mrm-8-5.* Assign the set of generalization level weights to q_i.
  - *mrm-8-6.* Determine the evidence-based quasi-identifier weight parameter, w_i^q which specifies the contextual importance of i^{th} quasi-identifier in the research.
- *mrm-9.* **General information.** Specify aggregation details of sensitive attribute.

- *mrm-9-1.* Take de-identified dataset D^0 as the starting point. For all sensitive attributes of S^{D^0} perform the following steps.
- *mrm-9-2.* comparing to D^0 specify the type of aggregation that is to be applied on each cluster $\xi_j^* \in \Xi^*$ based on risk analysis, PRM, or researcher requirement.
- *mrm-9-3.* Add this meta-data to M^s.
- *mrm-9-4.* Repeat steps *mrm-9-2* to *mrm-9-3.* for the next sensitive attribute.

- *mrm-10.* **Make *RV*.** Make Representability Vector, RV, by following any of the following point that are applicable.
  - *mrm-10-1.* Add I^o, and I_j^c for each equivalent cluster $\xi_i^* \in \Xi^8$ from *mrm-2* to RV.
  - *mrm-10-2.* Add A^e from *mrm-3* to RV.
  - *mrm-10-3.* Add A^q from *mrm-4* to RV.
  - *mrm-10-4.* Add C^e from *mrm-5* to RV.
  - *mrm-10-5.* Add S^q from *mrm-6* to RV.
  - *mrm-10-6.* Add M^s from *mrm-7* to RV.
  - *mrm-10-7.* Add W from *mrm-8* to RV.
  - *mrm-10-8.* Add G from *mrm-9* to RV.

5.5.6 Export module (EXM)

Export Module (EXM) contains policy points for defining the format of anonymized dataset Ξ^* and RV from ANM and MRM respectively to researcher or public. The aim is to make the internally used codes meaningful to the end user. In multi-organizational models, EXM transforms data using agreed upon standards and data interfaces. Table 5.9 shows the factors affecting the context, content, and actors of EXM.

Table5.9: Factors affecting the context, content, and actors of EXM

Component	Explanation
Aim	Specifying the dynamics of exporting of the anonymized dataset and RV from ANM and MRM respectively
Assumptions	Previous disclosed datasets are known and have taken into consideration from ADM
Actors	privacy analyst, data analyst, risk analyst
Elements	Ξ^*, Ω, $D_s^{k^*}$
Actions	Specifying the coding for disclosure, Specifying the format of attributes, Hiding internal characteristics
Input	De-identified dataset Ξ^*, RV, Expert decisions
Output	k-anonymized dataset $D_s^{k^*}$, RV

The following list shows the stages of the policy process.

Strategic thinking

The final format and coding of the dataset must be different from internal characteristics.

Points for policy development and implementation.

This section provides the high level policy points and pseudo-procedures in EXM.

- *exm-1.* **Coding and formating** Preparing the coding and formatting of the attributes.
  - *exm-1-1.* Change the name of attributes.
  - *exm-1-2.* For each quasi-identifier in Ξ^* specify the coding for all records.
  - *exm-1-3.* For each sensitive in Ξ^* specify the coding for all records.
  - *exm-1-4.* If policy calls for it in **prm-10**, add the number of records per cluster for records with all sensitive attributes aggregated.
- *exm-2.* **Hiding internal characteristics** Hide all characteristics of the dataset that is used by procedures of DPM, PRM, ANM, and MRM.
  - *exm-2-1.* Hide or pseudonymize key attribute.
  - *exm-2-2.* Change the order of records.

CHAPTER 6

HYPOTHETICAL APPLICATION OF *PARV* MODEL ON 2008 GATS DATA OF TURKEY: EXAMPLE SCENARIOS

6.1 Application of scenarios on real dataset

This chapter presents three scenarios for privacy protection of public health data, k-anonymization, RV, and policy issues discussed in three layers of *PARV* model in chapters 3, 4, and 5. The aim of the scenarios is to provide hypothetical examples on actual data rather than presenting best practices and optimized solutions.

The scenarios are applied on a sample portion of data from Global Adult Tobacco Survey (GATS) of Turkey conducted in 2008. GATS is used as a standard to monitor tobacco use in adults of 15 age or older with an aim to provide comparable data in and across participating countries [143]. It's a part of Global Tobacco Surveillance System Data (GTSSData) which hosts data from four surveys related to tobacco use around the world [144]. The four surveys include Global Youth Tobacco Survey (GYTS), Global School Personnel Survey (GSPS), Global Health Professions Student Survey (GHPSS), and Global Adult Tobacco Survey (GATS) which have been used in several studies related to tobacco use [145, 146, 147]. GATS data and related results become publicly available after the ministry of health finalizes the country's report [148]. The GATS survey was conducted in Turkey in 2008 [149]. The dataset for this survey is available to the public by the Center of Control Disease [143]. This dataset is in de-identified person-specific form containing the answers of participants to questions regarding tobacco use including personal characteristics, tobacco use and cessation, second hand smoke, media, knowledge, attitude and perception [150]. The reason for choosing this dataset is that it is person-specific, non-aggregated, and relates to real situation.

The following list shows the categories of personal characteristics fields of the survey according to the codebook of GATS [151]. The rest of the variables used in GATS dataset, are related to tobacco use gathered in different categories as mentioned above.

- Age and data of birth including year and month
- Gender
- Education and literacy

- Occupation
- Household items including electricity, fixed phone, television, mobile phones
- Transportation including car
- Residence
- Language
- Household including number of age

Table6.1: Characteristics of sample datasets used in *PARV* scenarios based on GATS data of Turkey for 2008

Characteristic	Explanation
Key attribute	Case ID of GATS data is replaced by new key values
Identifiable attribute	Not applicable. GATS data is de-identified.
Quasi-identifiers	• Age • Education • Occupation • Gender • Number of people living in a household • Being smoker might be considered as quasi-identifier based on sample policies
Sensitive attributes	• Daily smoking habits • Doctor visits • Paying attention to anti-smoking messages on cigarette packs • Knowledge about health hazards of smoking • A hypothetical numerical value to demonstrate aggregation

The attributes used in the scenarios of this chapter, are subsets of GATS dataset. Table C.1 shows a sample dataset extracted from the 2008 GATS dataset for Turkey. In order to show the sample datasets in this manuscript for easier comparison, a subset of 40 records were selected from GATS dataset. The 'value' attribute is not included in the original dataset. The reason for its addition to these examples is to demonstrate the effect of aggregation on numerical sensitive attributes.

Table 6.1 shows the characteristics of sample datasets used in the example scenarios in this chapter. The choice of dependent and independent variables is determined by research context and does not necessarily follow the pre-determined attribute tagging of a registry.

6.2 Scenario 1: One way data anonymization and basic policy making

This first scenario presents the basic policy making procedures of *PARV* model along with general anonymization procedures. It demonstrates the most popular approaches for disclosures of public health data where internal parameters are determined according to the nature of data or the policies without considering the research context . As a result, this scenario is based on 'data oriented' strategy with 'restricted reporting of information loss' as explained in section 5.4.1 and table 5.2. This scenario provides examples of both person-specific and aggregated datasets. It should be noted that the aim of *PARV* model is not to provide new methods for anonymization. Instead, it focuses on determining the general k-outlier handling method and the application of cluster validity indices to calculate information loss elements of RV. Therefore, the procedures for anonymization in this scenario are based on original k-anonymity and presented with the most basic concepts of anonymization.

Strategic points

The aim of this scenario is to anonymize a subset of GATS dataset for public access. No details about information loss is to be disclosed. Public access is maintained for aggregated dataset while person-specific dataset is accessible in controlled environment.

Context and actors

The context of this scenario relates to a registry who has access to the whole GATS dataset and aims at disclosing dataset in two different forms. The end user is likely to be a researcher in public health or epidemiological domain, yet other legitimate users might have gain access to data. It is probable that the dataset would be used for a research about tobacco cessation. The first form of disclosure is person-specific accessible only to known recipients or in controlled environments. In case that it is not disclosed in a controlled environment, it is possible that the dataset would be accessed by recipients not known by the registry. The second from is aggregated dataset accessible to the public. The main actors of this scenario are organizations collecting and disseminating GATS data, the registry obtaining GATS data, risk analyst, and experts in the registry, informaticist, and public health policy analyst.

Basics

Considering the context in which GATS survey is performed [149], the sample population P^s is the survey data gathered from one person p per 11,200 households in Turkey except for villages with less than 200 residents. The person p is of age or older.

Registry sample population P^{rs}, in this scenario, is a set of 40 records $\{r_{101}, \ldots, r_{140}\}$ selected from P^s. The dataset D representing P^{rs} is shown in table C.1 In this case, D=$\{r_{101}, r_{102}, \ldots, r_{140}\}$.

DPM

Since GATS dataset is de-identified, it is assumed that *dpm-1*, *dpm-2*, and *dpm-6* have already been applied.

The decision about the quasi-identifiers in *dpm-3* depends on risk analysis and expert decisions. The basis for decision making about quasi-identifiers is the existence of same attribute in other datasets outside the registry. In this scenario, all attributes related to the characteristics of individuals are considered as quasi-identifier. Although being a smoker or non-smoker is not included in the individual characteristics part of the survey, it might be considered as a known personal trait by outsiders. In this scenario, however, it is considered as a dependent variable. Therefore Q^D={ number of people in a household, age, gender, education, job}.

According to *dpm-4*, assume that the policy of the registry of this scenario is to tag dependent variables as sensitive attributes. Since GATS data is related to tobacco use and cessation, it can be concluded that the possible sensitive attributes are S^D={smoking, noticing health-related PSAs, perception about health hazards of smoking, and the value for a medical condition }.

And finally, according to *dpm-5*, the key attribute is added by replacing the values of CaseID attribute in GATS dataset with new values. The aim of key attribute is for internal usage and comparison. Since identification of persons participating in GATS survey is not included in dataset, no index is made between key attribute and the unique identification of the participants.

PRM

The following list shows the elements of PRM for scenario 1.

- $k = 3$ based on option 1 of *prm-1*.
- Case 1, D_1: If all quasi-identifiers are specified by policy or expert review to be included, then only *prm-2-1* is applied. In this case, all attributes tagged as quasi-identifier in *dpm-3* will be included in Q_1^D. Case 1 is shown in table C.1.
- Case 2, D_2: Assume that 'number of people in a household' and 'job' attributes are determined as independent variables. In this case, other quasi-identifiers specified in *dpm-3* are removed. Following *prm-2-2* and *prm-2-3*, the set of quasi-identifiers are specified as Q_2^D={ Hh, job }. It should be noted that there is a semantic relationship between 'job' and 'gender' attributes as all the records with 'housewife' for job attribute belong to individuals with 'female' value for 'gender' attributes. Case 2 is shown in table C.2.
- Domain generalization hierarchies based on *prm-3-1*, including suppression for all quasi-identifiers according to *prm-3-2* are shown in the following list. In this scenario, hierarchies are arbitrary and based on expert review without considering researcher requirements.
  - Hh: $2 \xrightarrow{f^G_{Hh_1}} \text{1-5} \xrightarrow{f^G_{Hh_2}} \text{1-10} \xrightarrow{f^G_{Hh_3}} 10+ \xrightarrow{f^s_{Hh}} ***$
  - age: $23 \xrightarrow{f^G_{age_1}} \text{21-30} \xrightarrow{f^G_{age_2}} \text{0-25} \xrightarrow{f^G_{age_3}} \text{0-50} \xrightarrow{f^s_{age}} ***$
  - gender: M $\xrightarrow{f^s_{gender_1}}$ ***
  - education: high school $\xrightarrow{f^G_{education_1}}$ (none, pre high school, post high school) $\xrightarrow{f^G_{education_2}}$ none, some $\xrightarrow{f^s_{education}}$ ***

– job: employer $\xrightarrow{f^G_{job1}}$ works, doesn't work $\xrightarrow{f^s_{job}}$ ***

- *prm-4* to *prm-9* are not applicable in this scenario.
- The sensitive attribute 'value' will be presented in original and aggregated forms in this scenario based on *prm-10.* For case 2, S^D={ smoking, value }.

ANM

The minimum k in this scenario is 3 following *prm-1* and *anm-1.*

The next course of action is the first step cluster analysis as stated in *anm-2* and to specify Ξ^0={$\xi^0_1, \xi^0_2, \ldots, \xi^0_\lambda$} and Δ_λ. All records in case 1, shown in table C.1, are non-conforming. In other words, all equivalent classes of Ξ^0_1 have less than 3 members. Case 2, shown in table C.2, contains mostly conforming equivalent clusters. Table 6.2 shows the members of Ξ^0_2 along with the cardinality of each equivalent cluster. The cardinality of Δ_λ is 40, 5 cases 1 and 2.

Table6.2: Ξ^0_2 for Case 2 shown in table C.2

ξ^0_i	$\mid \xi^0_i \mid$	Member of Δ_λ
{5, Self}	3	x
{4, Self}	3	x
{5, House}	4	x
{4, House}	6	x
{2, House}	4	x
{6, House}	1	✓
{7, House}	1	✓
{4, Employee}	3	x
{3, Employee}	5	x
{2, Employee}	3	x
{6, Employee}	1	✓
{4, Retired}	2	✓
{3, Retired}	4	x

The next course of action is *anm-3* which is the second step clustering analysis. *anm-3* is applied in this scenario to provide examples of the application of k-outlier handling methods. The justification or optimization of chosen methods are beyond the scope of this study. Other scenarios will use k-outlier handling methods implemented in this step in *anm-4* which is the analysis of the multi-step clustering analysis without actual anonymization procedures.

Table C.3 shows an example of generalization approach applied on case 2 dataset. The 'household' attribute is generalized one level. The 'job' attribute has not been generalized except for records r_{109}, r_{112}, and r_{116} where their associated values are suppressed. Table C.4 shows another example of generalization approach where 'household' attribute generalized one level only for records r_{109}, r_{112}, and r_{116}. The 'job' attribute is generalized one level for all records and suppressed for records r_{109}, r_{112}, and r_{116}.

Table C.5 shows a dataset with deletion approach applied. All the members of Δ_λ shown in table 6.2 are deleted from dataset. Table C.6 shows the merging approach applied where non-conforming clusters of Δ_λ are merged into the closest conforming clusters. In all these cases, the shown dataset are the final dataset Ξ^* following *anm-3-4* or *anm-4-3*.

To provide examples for aggregation of sensitive attributes according to *anm-3-5*, two different approaches are applied on . The first is shown in table6.10 where the dataset is aggregated for all records. The second example, shown in table C.7, follows a exemplary policy of aggregating the 'value' sensitive attributes for clusters with less than 10 members.

MRM

Scenario 1 is focused only on disclosing the anonymized dataset without reporting information content. Therefore, MRM is not applicable here in terms of reporting RV. However, information content measurement can be used internally in order to choose the final dataset from the pool of Ξ^*s. Each of these Ξ^*s are anonymized by different k-outlier handling approach or the same method with different characteristics.

Assume that the preferred k- outlier handling approach in scenario 1 is generalization. Also assume that the priority of registry is on minimum complexity of the methods and least amount of processing time. As a result, it is not important to find the minimally distorted dataset among all possible outcomes.In this example, one of the generalized tables of case 2, shown in tables C.3 and C.4, will be used for disclosure. According to *mrm-1-4*, the information content measurement is used internally based on priorities of scenario 1. Therefore, the aim is to find the I^o for both generalized datasets according to *mrm-2-2*.

Applying precision metric, as shown in equation 4.1 for the whole dataset produces the following values:

- $Prec(Example1)$=0.8375
- $Prec(Example2)$=0.7989

According to the policies of scenario 1, the information content of the example 1 (table C.3) is higher than that of example 2 (C.4) while both datasets satisfy 3-anonymity and preserving the number of records. As a result, dataset presented in table C.3 is chosen for disclosure. Also assume that policy calls for aggregation of sensitive attributes with clusters with less than 10 members. Therefore table C.7 is sent to EXM for final procedures. Since no information regarding bias and RV is to be reported in scenario 1, other procedures of MRM are skipped.

EXM

The outcome of EXM is highly dependent on internal policies in single organizational and on agreed upon data standards and interfaces in multi-organizational settings models. The following list shows an example of *exm-1* on table C.7.

- *exm-1-1.* Changing the name of the attributes
  - Hh → Household size
  - PSA → Noticed PSAs?
  - Illness → Smoking causes illness?
  - Value → Test result
- *exm-1-2.* and *exm-1-3.* Review coding of the values.
  - For 'Job' Quasi-identifier replace 'House' with 'Housewife', 'Self' with 'Self employed', and '***' with 'Unknown'.
- *exm-1-4.* is not applicable because final dataset is in person-specific form.
- *exm-2.* The key attribute is used internally by the registry. As a result, it must be removed. Order of the records are also changed randomly.

Table 6.3 shows the first five records of the final dataset D^{k^*} to be disclosed.

Table6.3: First five records of the output of scenario 1

Household size	**Job**	**Noticed PSA?**	**Smoking causes illness?**	**Test result**
1–5	Retired	No	Yes	40.16
6–10	***	Yes	Yes	44
1–5	Employee	Yes	Yes	34
1–5	House	Yes	Yes	51
1–5	Retired	Yes	Yes	40.16

6.3 Scenario 2: One way data disclosure with *RV*

The second scenario provides further examples applied on datasets presented in scenario 1. These examples focus mainly on the second layer of *PARV* model, namely *RV* calculation. *RV* in this scenario is only used to report the information content and other aspects of anonymization to researcher. This scenario is based on 'the openness of information loss details' as explained in section 5.4.1 and table 5.2.

Strategic points

The aim is to calculate *RV* using the second layer of *PARV* model without a need to know the technical aspects of underlying anonymization algorithm.

Context and actors

The context of this scenario is related to disclosing to controlled environments where the end user is known. Although it is possible to disclose *RV* to the public, its aim and components more suitable for controlled disclosures where the research context is known. The main actors of this scenario are organizations collecting and disseminating GATS data, the registry obtaining GATS data, risk analyst, and experts in the registry, informaticist, public health policy analyst, epidemiologists, public health professionals, and population health researchers.

Basics

The focus of this scenario is on measurement module (MRM). The application of cluster validity indices will be examined on the same anonymized datasets presented in scenario 1. It is also assumed that DPM procedures are similar to those of scenario 1.

PRM

PRM is similar to scenario 1 except for the following procedure.

MRM

Assume that Ξ^0=D_2 explained in page 110. Also assume that Ξ^*=D_2^* shown in table C.3.

- *mrm-1.* Preparing the input for second step clustering analysis.
  - *mrm-1-1.* k=3.
  - *mrm-1-2.* k-outlier handling approach is generalization.
  - *mrm-1-3.* Ξ^0=D_2 (table C.2) and Ξ^*=D_2^* (table C.3).
  - *mrm-1-4.* The goal of *RV* is to report the information content characteristics to the researcher. The components of *RV* are determined based on internal policy rather than the requests of researcher.
- *mrm-2.* Based on policy, clustering based analysis is used for information content measurement. The rest of this section presents information content measurement without referring to sub-procedures of *mrm-2.*

The clusters of D_2^* and D_2^0 are shown in table 6.4. The first index presented in this scenario is Davien-Bouldin presented in section 4.4.5. First, the dispersion measure, S, will be calculated. In order to calculate $S_{\langle 1\breve{}5,S\rangle}$ first the distance between each points in the receiving clusters must be calculated.

- $d_M(\langle 5,S\rangle,\langle 1\breve{}5,S\rangle)$. Parameters of d_M are based on policy and evidence. In this scenario, the w_n^q is not included and G_n represents semantics of generalization similar to precision metric.

Table6.4: Ξ^0 and Ξ^* for D_2

Clusters in D^0	**Cluster size in D^0**	**Receiving cluster in D^***
$\langle 5, Self\rangle$	3	$\langle 1\breve{}5, Self\rangle$
$\langle 4, Self\rangle$	3	$\langle 1\breve{}5, Self\rangle$
$\langle 5, House\rangle$	4	$\langle 1\breve{}5, House\rangle$
$\langle 4, House\rangle$	6	$\langle 1\breve{}5, House\rangle$
$\langle 2, House\rangle$	4	$\langle 1\breve{}5, House\rangle$
$\langle 6, House\rangle$	1	$\langle 6\breve{}10, *\,*\,*\rangle$
$\langle 7, House\rangle$	1	$\langle 6\breve{}10, *\,*\,*\rangle$
$\langle 6, Employee\rangle$	1	$\langle 6\breve{}10, *\,*\,*\rangle$
$\langle 4, Employee\rangle$	3	$\langle 1\breve{}5, Employee\rangle$
$\langle 3, Employee\rangle$	5	$\langle 1\breve{}5, Employee\rangle$
$\langle 2, Employee\rangle$	3	$\langle 1\breve{}5, Employee\rangle$
$\langle 4, Retired\rangle$	2	$\langle 1\breve{}5, Retired\rangle$
$\langle 3, Retired\rangle$	4	$\langle 1\breve{}5, Retired\rangle$

- Hyperplane distance measurement, presented in section 4.4.3.2,is used for $q_1 ='Hh'$ which is the 'size of the household'. Values of ξ^0_{Hh} are considered in HP^0_{Hh}. For cluster $\langle 1\breve{}5\rangle$ in Hh dimension , which is in HP^1_{Hh}, the hypothetical center is considered as the mean of the values of Hh in HP^0_{Hh}. Therefore $C^{*,Hh}_{\langle 1\breve{}5\rangle}=\frac{5+4}{2}=4.5$.

- Hyperplane distance is also specified for $q_2 ='Job'$. The 'evidence-based generalization level weight measurement'w^e_2, is used for this purpose. The distance between different clusters are specified by experts based on researcher requirements or evidence. A hypothetical example is shown in figure 6.1.

- For 'household attribute' $G_{\langle 1\breve{}5\rangle}=\frac{h^i_d}{L_{max_i}+1}=\frac{1}{4}$ $= 0.25$. G_n for job attribute is zero for all clusters except for $G_{\langle ***\rangle}=\frac{2}{2}=1$. (For details for G_n refer to 4.4.4)

- The following equations show the rest of the calculations.

$$d_M(\langle 5,S\rangle,\langle 1\breve{}5,S\rangle))=\Big((d^h(\langle 5\rangle,\langle 1\breve{}5\rangle))^2+(d^h(\langle S\rangle,\langle S\rangle)^2)\Big)^{\frac{1}{2}}=0.125$$

$$S_{\langle 1\breve{}5,S\rangle}=\left(\frac{1}{6}\left(0.125^2+0.125^2\right)\right)^{\frac{1}{2}}=0.3061862178$$

$$S_{\langle 1\breve{}5,H\rangle}=\left(\frac{1}{14}\left(0.3214285714^2+0.4285714286^2+0.4285714286^2\right)\right)^{\frac{1}{2}}=1.4890269192$$

$$D_{\langle 1\breve{}5,S\rangle,\langle 1\breve{}5,H\rangle}=\left(\mid c_{\langle 1\breve{}5\rangle}-c_{\langle 1\breve{}5\rangle}\mid^2+\mid c_{\langle S\rangle}-c_{\langle H\rangle}\mid^2\right)^{\frac{1}{2}}=0.7857142857$$

$$R_{\langle 1\breve{}5,S\rangle,\langle 1\breve{}5,H\rangle}=\frac{S_{\langle 1\breve{}5,S\rangle}+S_{\langle 1\breve{}5,H\rangle}}{D_{\langle 1\breve{}5,S\rangle,\langle 1\breve{}5,H\rangle}}=\frac{0.3061862178+1.4890269192}{0.7857142857}=2.2848167199$$

Table 6.5 shows the cluster similarity measures for all clusters of Ξ^*_D. Davien-Bouldin index is calculated as shown in the following equation:

$$V_{DB}(Example1) = \frac{1}{5}\,(2.2848167199 + 4.6231301713 + 1.7030510724 + 2.1732976315 + 4.6231301713) = 3.0814851533$$

Table6.5: Cluster similarity measures for table 6.4

ξ^*_i, ξ^*_j	$R_{\xi^*_i, \xi^*_j}$
$\langle 1\breve{}5, Self\rangle, \langle 1\breve{}5, House\rangle$	2.2848167199
$\langle 1\breve{}5, Self\rangle, \langle 6\breve{}10, ***\rangle$	1.5926131832
$\langle 1\breve{}5, Self\rangle, \langle 1\breve{}5, Employee\rangle$	0.5056354898
$\langle 1\breve{}5, Self\rangle, \langle 1\breve{}5, Retired\rangle$	0.495730067
$\langle 1\breve{}5, House\rangle, \langle 6\breve{}10, ***\rangle$	1.7030510724
$\langle 1\breve{}5, House\rangle, \langle 1\breve{}5, Employee\rangle$	2.0846376869
$\langle 1\breve{}5, House\rangle, \langle 1\breve{}5, Retired\rangle$	4.6231301713
$\langle 6\breve{}10, ***\rangle, \langle 1\breve{}5, Employee\rangle$	1.281167683
$\langle 6\breve{}10, ***\rangle, \langle 1\breve{}5, Retired\rangle$	1.311765094
$\langle 1\breve{}5, Retired\rangle$, $\langle 1\breve{}5, Retired\rangle$	2.1732976315

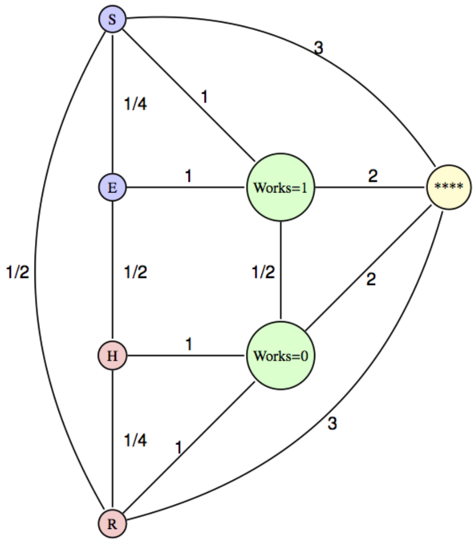

Figure 6.1: Hypothetical example for hyperplane distance measurement from $HP^0_{'Job'}$ to $HP^1_{'Job'}$

DB index for the second example of generalization approach shown in table C.4 is 3.8372883535. Lower values of DB index indicate a better clustering scheme. As a result, Example 1 is a better clustering in terms on information content and representability. This accords with the results of the precision metric presented at page 112.

The next example shows the application of cluster validity indices with external criteria in order to demonstrate the action of comparing different datasets and approaches. Table 6.6 shows the contingency table for generalization table C.3.

Table6.6: Contingency table for generalization approach shown in table C.3

	$(1-5,E)$	$(1-5,H)$	$(1-5,R)$	$(1-5,S)$	$(6-10,*)$
$(5,S)$	0	0	0	3	0
$(4,S)$	0	0	0	3	0
$(5,H)$	0	4	0	0	0
$(4,H)$	0	6	0	0	0
$(2,H)$	0	4	0	0	0
$(6,H)$	0	0	0	0	1
$(7,H)$	0	0	0	0	1
$(4,E)$	3	0	0	0	0
$(3,E)$	5	0	0	0	0
$(2,E)$	3	0	0	0	0
$(6,E)$	0	0	0	0	1
$(4,R)$	0	0	0	0	2
$(3,R)$	0	0	4	0	0

If these indices are used for comparison, it can be concluded that Generalization approach shown in example 2 and table C.4 are preferable to example 1 shown in table C.3.

Table6.7: Contingency table for generalization approach shown in table C.4

	$(5,W)$	$(4,W)$	$(2,W)$	$(3,W)$	$(6-10,*)$	$(5,U)$	$(4,U)$	$(2,U)$	$(3,U)$
$(5,S)$	3	0	0	0	0	0	0	0	0
$(4,S)$	0	3	0	0	0	0	0	0	0
$(5,H)$	0	0	0	0	0	4	0	0	0
$(4,H)$	0	0	0	0	0	0	6	0	0
$(2,H)$	0	0	0	0	0	0	0	4	0
$(6,H)$	0	0	0	0	1	0	0	0	0
$(7,H)$	0	0	0	0	1	0	0	0	0
$(4,E)$	0	3	0	0	0	0	0	0	0
$(3,E)$	0	0	0	5	0	0	0	0	0
$(2,E)$	0	0	3	0	0	0	0	0	0
$(6,E)$	0	0	0	0	1	0	0	0	0
$(4,R)$	0	0	0	0	0	0	2	0	0
$(3,R)$	0	0	0	0	0	0	0	0	4

Table6.8: Comparison of Rand statistics, Jaccard coefficient, and Folkes and Mallows index for tables 6.6 and 6.7

	R	J	FM
Generalization, Ex1, Tables 6.6 and C.3	0.8449	0.3164	0.5625
Generalization, Ex2, Tables 6.7 and C.4	0.9692	0.7000	0.8367

RV_1 for generalization approach example 1 (table C.7)

- Overall Information content I^o={ *Precision metric* : 0.8375, V_{DB} : 3.08148, *Rand index* : 0.8449}

- Excluded Attributes A^e={ *age*, *gender*, *job*, *smoking*}. A hypothetical example is a case where there is an evidence of the *job* attribute being a confounding factor. The epidemiologist having access to A^e would know that a highly probable confounding factor is excluded from C.4).

- Excluded cluster C^e=$\emptyset$. This means that sample size has not been changed.

- Semantics of quasi-identifiers S^q

  - 'Hh' represents the number of people ing a household. For grouped values, such as '6-10' lower and upper values are included in the possible values.
  - 'Job' attribute includes self employers, housewives, employees, and retired people.

- Aggregation of sensitive attributes M^s. Values for PSA, $Illness$ are not aggregated. $Value$ attribute is aggregated for cluster with less than 10 records. Table 6.9 shows the central tendency values for aggregated clusters.

- Weight vector W. No weight difference is in effect. Therefore for all quasi-identifier w^q= 1 and for all generalization levels w^e=1.

- General Information G. Minimum k is 3 and in order to maintain that the values of 'Household' and 'Job' are generalized. Sample size ratio is 1. Clusters with less than 10 records are presented with aggregated values for $Value$ attribute.

Table6.9: M^s element of RV for table C.7

Cluster	Size	Mean	Standard Deviation
$\langle 1\breve{}5, Retired\rangle$	6	40.16	5.67
$\langle 1\breve{}5, Self\rangle$	6	39	3.09
$\langle 6\breve{}10, ***\rangle$	3	44	6.55

RV_2 for generalization approach example 2 (table C.4)

- Overall Information content I^o={ $Precision\ metric$: 0.7989, V_{DB} : 3.8372, $Rand\ index$: 0.9692}

- Excluded Attributes A^e={ *age*, *gender*, *job*, *smoking*}. A hypothetical example is a case where there is an evidence of the *job* attribute being a confounding factor. The epidemiologist having access to A^e would know that a highly probable confounding factor is excluded from C.4).

- Excluded cluster C^e=$\emptyset$. This means that sample size has not been changed.

- Semantics of quasi-identifiers S^q

  - 'Hh' represents the number of people ing a household. For grouped values, such as '6-10' lower and upper values are included in the possible values.

  - 'Job' attribute includes self employers, housewives, employees, and retired people.

- Aggregation of sensitive attributes M^s. Values for PSA, $Illness$, and $Value$ are not aggregated for any cluster.

- Weight vector W. No weight difference is in effect. Therefore for all quasi-identifier w^q= 1 and for all generalization levels w^e=1.

- General Information G. Minimum k is 3 and in order to maintain that the values of 'Household' and 'Job' are generalized. Sample size ratio is 1.

6.4 Scenario 3: Research oriented data disclosure with RV reporting and feedback

The third scenario presents the role of researcher in the analysis of anonymization results using RV. The policy points are similar to of scenarios 1 and 2. Therefore, only the issues relating to layers 1 and 2 of $PARV$ model is presented in this scenario.

Case 1.
Assume that the researcher obtains the generalized table C.4 with RV_2 shown on page 118. The following shows the hypothetical context of the research:

- The aim is to find the association between $Illness$ and $Value$ attributes. Although $Illness$ is tagged by registry as a sensitive attribute, it does not prevent it to be considered as a dependent variable by researcher.

- A study in the literature shows a possible association between the job and rate of smoking. This study shows that self-employed people have a lower chance of smoking.

- The researcher considers the Job attribute as a possible confounding factor.

- Comparing to previous studies performed by researcher, the useful minimum precision metric is 0.8.

In this case, table C.4 might not be the suitable dataset for research context. One reason is that the values of Job attribute are generalized as 'works', and 'unemployed' and it is impossible to know which records belong to 'self employed' individuals in the group generalized as 'works'. This can lead to confounding bias considering the available evidence applicable to the sample dataset. Another reason is that the precision metric of dataset shown in table C.4 is less than 0.8.

In this case, the researcher provides feedback, in the form of RV, to the registry about the dataset that is more suitable for their research. The following shows the Representability Vector of case 1 (RV_{case1}) indicating the researcher requirements based on context and evidence

- Minimum overall information content $I^o = \{\ Precision\ metric : 0.8\}$.

- Included quasi-identifiers $A^q\{\ Household,\ Job\}$.
- General information G.
    - Generalization characteristics: No generalization of Job attributes for clusters with values generalized to 'works'.
    - General characteristics of sensitive attributes. $Illness$ attribute is required as independent variable. $Value$ attribute is required as dependent variable.

Case 2.
In the second case, the registry provides the researcher with dataset shown in table C.7. The associated RV_1 shown on page 117. Assume that the research context is same as the context in case 1. This table provides the conditions that researcher stated in RV of case 1, which is the minimum precision metric of 0.8. The registry provides the aggregated version of C.7 along with RV_1. This aggregated dataset is shown in table 6.10. It acts to provide more information, along with RV_1 as meta-data, about the final dataset. The dataset provided to the researcher might seem more suitable for the context of the research. The most important reason is that Job attribute is not generalized. This facilitates the analysis of Job attribute as a possible confounding factor. However, for clusters with less than 10 records, the associated values of $Value$ attribute are aggregated. Although M^s element of RV_2 contains information about aggregation characteristics shown in table 6.9, it still gives rise to ecological fallacy for such clusters. Take the cluster $\langle 1\breve{}5, Self\rangle$ as an example. This cluster has 6 records and the mean of the values for $Value$ attribute is 39. As mentioned in case 1, there is an evidence of being a self employed and high rates of smoking. Since $Value$ attribute s related to smoking, and Job attribute is a possible confounding factor, using the mean value leads to ecological bias for this cluster. This bias becomes unavoidable if the registry do not include M^s in RV_2. This leads to a wrong assumption that all people in $\langle 1\breve{}5, Self\rangle$ cluster have 39 as the value of $Value$ attribute.

Table6.10: Database-wide aggregation of generalized D_2, example 1 (C.3)

Count	Hh	Job	PSA	Illness	Value
11	1-5	Employee	Yes=100%, No=0%	Yes=90.90%, No=9.09%	40.18
14	1-5	House	Yes=100%, No=0%	Yes=100%, No=0%	38.78
6	1-5	Retired	Yes=100%, No=0%	Yes=100%, No=0%	40.16
6	1-5	Self	Yes=100%, No=0%	Yes=100%, No=0%	39
3	6-10	***	Yes=66.66%, No=33.33%	Yes=100%, No=0%	44

Therefore, the RV provided by researcher to the registry contains the following element about sensitive attribute:

- Aggregation of sensitive attributes M^s. No aggregation for $Value$ attribute, especially for records with values for Job attribute that are generalized to 'works'.
- General information G.

- General characteristics of sensitive attributes. *Illness* attribute is required as independent variable. *Value* attribute is required as dependent variable.

Table6.11: Contingency table for merge approach shown in table C.6

	$(5,S)$	$(4,S)$	$(5,H)$	$(4,H)$	$(2,H)$	$(4,E)$	$(2,E)$	$(3,E)$	$(3,R)$
$(5,S)$	3	0	0	0	0	0	0	0	0
$(4,S)$	0	3	0	0	0	0	0	0	0
$(5,H)$	0	0	4	0	0	0	0	0	0
$(4,H)$	0	0	0	6	0	0	0	0	0
$(2,H)$	0	0	0	0	4	0	0	0	0
$(6,H)$	0	0	1	0	0	0	0	0	0
$(7,H)$	0	0	1	0	0	0	0	0	0
$(4,E)$	0	0	0	0	0	3	0	0	0
$(3,E)$	0	0	0	0	0	0	0	5	0
$(2,E)$	0	0	0	0	0	0	3	0	0
$(6,E)$	0	0	0	0	0	1	0	0	0
$(4,R)$	0	0	0	0	0	0	0	0	2
$(3,R)$	0	0	0	0	0	0	0	0	4

Case 3. In this case, the decision of the registry is to disclose the data with merging approach applied. This dataset is shown in table C.6. Since no generalization is applied, the cluster validity indices with external criteria are chosen for this case. Table 6.11 shown the contingency table for the merge approach of dataset C.6.

The associated cluster validity indices are shown bellow[1].

- Rand index. R=0.9744
- Jaccard coefficient. J=0.7368
- Folkes and Mallows index. FM=0.8584

It should be noted that these indices do not capture the semantics of generalization. Instead, they are sensitive to changes in cluster size and moving of records from one cluster to the other. Since the least changes in terms of moving the records have been applied in merge approach, shown in table C.6, the validity indices have higher values comparing to generalization approach applied on table C.2 with RV_2 . Therefore, merging approach is the most similar clustering to the sample dataset shown in table C.2. This contradicts with V_{DB} value calculated in page 116. The reason is, the semantics of generalization and hyper-plane distance measurement are not included in cluster validity indices based on external criteria chosen here. The choice of the metric for comparison, and the interpretation of those values by researcher, depend on context, evidence, and expert review.

The following shows the Representability Vector of case 3 (RV_{case3}) provided by the registry.

- Overall Information content I^o={ *Rand index* : 0.9744}
- Excluded Attributes A^e={ *age*, *gender*, *job*, *smoking*}.

[1] Refer to chapter 4.4.6

- Excluded cluster C^e=$\emptyset$.
- Semantics of quasi-identifiers S^q
  - 'Hh' represents the number of people ing a household.
  - 'Job' attribute includes self employers, housewives, employees, and retired people.
- Aggregation of sensitive attributes M^s. No aggregation is applied.
- General Information G. Minimum k is 3. Clusters with less than 3 members are moved to the semantically closest cluster. Sample size ratio is 1.

CHAPTER 7

CONCLUSION AND DISCUSSION

Protecting the right to privacy in secondary uses of health data, such as public health, is of utmost importance. However, there is an unavoidable tension between privacy protection and usefulness of data for research. This leads to either unavailability of data due to privacy concerns, or biased research results due to highly distorted datasets. Both of these situations affect the outcome of public health research which will eventually affect the health of individual determined by clinical processes, public health policies, and regulations. Although there are several technical methods to counter the problem of balancing the privacy and utility, these measures fall short to address the issue in real situation. With changing nature of researches, datasets, regulations, and policies, purely technical methods become ineffective over time or in certain contexts. We believe that privacy protection goes beyond choosing certain technical methods. Instead, it requires a broader approach, based on flexible policies, new demands, adoption of innovative technical methods, and improvements in order to adapt to new regulations. We also believe that all privacy protection measures, be it technical or procedural, must be based on evidence and researcher requirements. Also the researcher must be informed about the bias introduced by privacy protection methods, or policies, in order to be able to assess the accuracy of research and its generalizability to the whole population. Our study, therefore, provides a three level model (*PARV*) for policy making with a focus on protecting the privacy of person-specific data based on researcher requirements and reporting the bias to researcher.

7.1 Aim and methods to achieve it

The overall aim of our study is privacy protection of secondary data, especially in public health and epidemiology domains. An eminent problem in public health is unavailability of data due to privacy concerns. Health organizations are unwilling to disclose their datasets due to uncertainties about the possible risks to privacy. This impedes with the need to perform research and to improve the health of population. Where data is available, they are distorted without considering the epidemiology or public health requirements and evidence. This leads to highly biased research which will eventually reflect in clinical practices or public services.

secondary health data in our study primarily refer to person-specific data which are necessary for accurate and effective public health research. The core method we studied for anonymizing person-specific data is k-anonymity. Many algorithms are available based on k-anonymity, each suitable for a certain situation or type of data. In order to present a generic approach for privacy analysis, that would cover the person-specific and aggregated datasets, we used clustering concept. The reason is anonymous datasets, either in aggregated or k-anonymous forms, resemble the results of clustering. Our aim of applying clustering is to provide an analysis method independent of underlying algorithms. Our goal of clustering analysis is to construct a vector that we named Representability Vector (RV), to provide information about the resemblance of disclosed dataset to original sample. In order to calculate a uniform metric for information content, we applied cluster validity indices on datasets anonymized using non-clustering methods. RV acts as a medium that carries several parameters regarding the anonymization and information content from an epidemiological and public health perspectives.

The first goal of RV is to inform the researcher about the amount and details of bias introduced in dataset due to anonymization. In this sense, RV contains information content, characteristics of anonymization, and details of dependent and independent variables. RV helps epidemiologists in assessing causal relationship analysis and having more information about potential confounding factors. The second use of RV is for epidemiologists and public health researchers to specify their research based requirements. This enables the registries to incorporate the evidence in privacy protection and to reduce the bias accordingly. RV, in both of its uses, reflects the context of data usage and privacy protection.

The context-oriented approach of our study is based on the fact that privacy protection is not a merely technical issue and depends on ever-changing data, demands, priorities, legislations, and ethical viewpoints. In order to combine technical and policy aspects of privacy protection, we presented a three level model ($PARV$) which is comprised of clustering based analysis of anonymization, RV, and policy making. The policy making level of $PARV$ model is a framework emphasizing expert review and high level strategies for privacy protection and evidence-based approaches. The framework has a modular structor with two major modules covering the first and second levels of $PARV$ model, namely clustering based analysis of anonymity and RV, respectively. The policy making level of $PARV$ model is structured as pseudo-contents which are high level statements of strategies, policies, and procedures, for modules of the framework.

7.2 Results

The first level of $PARV$ model is focused on k-anonymization and aggregation approaches. We showed that the records which pose privacy risks can be considered as outliers in terms of their conformity to k-anonymity. We demonstrated that outlier handling methods based on clustering analysis can be applied on datasets to handle outlier records, which we named as k-outliers. We also presented popular, and other hypothetical, approaches for k-outlier handling. The main value of this method is that it can be used to analyze the outcomes of any anonymized dataset without knowing the details of underlying algorithms. This level of $PARV$ model is presented in chapter 3.

The second level of $PARV$ model presents the concept of Representability Vector (RV).

We showed that access to anonymized datasets without knowing the information content can lead to biased results in epidemiological and public health research. The reason is that datasets are not anonymized with the requirements of epidemiology, or the context of its possible usage, in mind. *RV* comprises of several informational elements presenting the researcher with details of anonymized dataset. We also demonstrated how these elements can be used by researcher to ascertain the accuracy of their research and, in case of epidemiological research, the generalizability of the results to the population and spotting potential confounding factors. In short, *RV* shows how an anonymized dataset represents the original dataset. The main element of *RV* is information content measurement. In order to keep this measurement independent of underlying algorithm, we applied cluster validity indices on the results of the first level of our model. We demonstrated that this concept is more flexible as cluster validity indices can be used for different k-outlier handling methods. We also argued that this method applies for uniformly comparing the results of different algorithms. We also demonstrated that *RV* can be used in evidence-based approaches where anonymization is performed based on researcher requirements, facilitating the inclusion of evidence and epidemiological parameters in the dataset. The second level of *PARV* model is presented in chapter 4.

The third level of *PARV* model presents a framework for privacy preserving policy making to address the apparent lack of flexible policies pertaining secondary uses of health data in Turkey. We argued that merely technical methods, without robust and dynamic policies, will fall short to address the balance between privacy protection and data utility. Our framework has two levels with the first level stressing the issues of expert intervention and strategic decision making. This facilitates the adaptation of policy to changing contexts, legislations, and ethical viewpoints. The second level of our framework presented a modular approach to policy making with modules covering the first and second levels of *PARV* model. This framework emphasizes a holistic approach for policy making taking technical, procedural and social aspects of the problem as inter-connected issues. The third level of *PARV* model is presented in chapter 5.

We also conducted an e-survey in order to elicit the demand for and difficulties of gaining access to person-specific secondary health data, including its technical, procedural, policy-based, awareness, and legal aspects. The e-survey was sent to the members of the most active mailing lists of medical informaticist and public health specialist in Turkey. Due to the nature of e-surveys and technical difficulties in ascertaining view rate, the participation and completion rates were unknown. As a result, we used the responses to conduct a descriptive, rather than conclusive, analysis. The results suggest that the demand for person-specific data is higher among public health specialists comparing to medical informaticists. The main difficulty in accessing person-specific data is the existence of strict organizational policies limiting data disclosure to only aggregated forms. Where individual data were available, partitioned data and lack of standards and coordination between organizations were mentioned as main problems in effective data usage. There is an apparent misconception in participants in public health group about the inadequacy of removal of directly identifiable attributes to protect privacy. While medical informaticists are generally neutral toward policy issues, public health specialists are more aware of policy needs and the effect of legal aspects. The majority of public health specialists suggest the need for a centralized and multi-disciplinary effort to design policy and a central multi-disciplinary team to act as the liaison between organizations. While the generalizability of our e-survey is limited to the members of the chosen mailing lists, the results are valuable suggesting that our model is in-line with the expectations of the participants. Details of the e-survey are presented in Chapters 2.5.2, A, and B.

7.3 Limitations and recommendations for future research

Our model can be applied in approaches as strict as restrictive access and as open as publicly accessible disclosure. However, it is only suitable for situations where disclosure of RV is not prohibited by law and the involvement of researcher in anonymization process is legally possible. RV in publicly accessible disclosure approaches tend to be highly general and not informational enough for researcher in terms of assessing the introduced bias. More study is needed in eliciting the required parameters in subdomains of epidemiological studies in order to guide the calculation of RV accordingly.

The framework we presented in the third level of our model has several modules. The joining and temporal modules are presented conceptually as their details were beyond the scope of this study. They are indeed necessary for a comprehensive policy making framework and require elaborate study.

We presented the application of cluster validity indices, including Davies-Bouldin, Rand statistics, Jaccard coefficient, and Folkes and Mallows index. These are presented as example methods of indices with internal and external criteria. These methods are limited to hard clustering approaches. As a result, our study is not suitable for fuzzy clustering analysis approach. Further study is required for soft clustering and comparing different indices to elicit the suitability of each to certain contexts.

7.4 Contribution and significance to the field of epidemiology and public health informatics

The general contributions of our study is its emphasis on public health informatics, epidemiology, and policy making as an essential part of medical informatics. Incorporating epidemiological viewpoint is indispensable as popular privacy protection methods of secondary data simply overlook the effects on epidemiological research. In Turkey, most of the studies in medical informatics field are focused on clinical aspects of informatics. Our study incorporates other aspects of privacy protection of the secondary health data, such as policy issues and evidence-based approach. We believe that this approach would help to fill the gap in medical informatics in Turkey which is undercoverage of some aspects of health informatics such as public health, epidemiological viewpoint, policy issues, social factors, and expert involvement aspects.

To actualize this goal, our study provides a generic policy making framework that covers the technical and procedural issues of handling evidence-based anonymization and reporting information loss. This framework stresses that the effectiveness of technical measures depends on human factor, legislations, and procedures, all of which can be bound to technical aspect using dynamic policies. Hence, our framework provides high level and pseudo-contents underlying the major steps required for policy protection of secondary health data. It also emphasizes the expert intervention and the importance of strategic decision making. This facilitates incorporation of context, evidence, legislations, and priorities, all of which are too complex to actualize thorough pure technical measures. We believe that our generic approach is of value for public health organizations, policy makers, and legislators by providing a means for joining technical and policy based aspects.

Technically, the contribution of our study is the incorporation of clustering based analysis of anonymization and associated information loss. This enables public health data holders to assess the information content of their datasets independent of the details of underlying anonymization methods. We also employed cluster validity indices as a metric to measure the information loss of anonymized datasets. This makes it possible to use a unified metric on datasets that are anonymized using different algorithms. This information loss, along with other details of anonymization, constitute our proposed Representability Vector (RV). We believe that information about induced distortions leads to more accurate public health research. RV can also be used as an input for anonymization process or bias analysis. In this case , RV acts as a medium to ascertain the requirements of the researcher and the available evidence that affects the parameters of anonymization, bias calculation, and policy review. Therefore, our framework encourages the involvement of epidemiologists and researcher and an evidence-based approach in determining the parameters of privacy protection process.

REFERENCES

[1] P. A. Bath, "Health informatics: current issues and challenges," *Journal of Information Science*, vol. 34, no. 4, pp. 501–518, 2008.

[2] L. O. Gostin, J. G. Hodge Jr., and R. O. Valdiserri, "Informational privacy and the public's health.," *American Journal of Public Health*, vol. 91, no. 9, p. 1388, 2001.

[3] L. O. Gostin, "Future of public health law, the," *Am. JL & Med.*, vol. 12, p. 461, 1986.

[4] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[5] S. Fortin and B. M. Knoppers, "Secondary uses of personal data for population research.," *Genomics, Society and Policy*, vol. 5, no. 1, pp. 80 – 99, 2009.

[6] L. Gordis, *Epidemiology.* W.B. Saunders, 2 ed., 2000.

[7] L. A. Sweeney, "Computational disclosure control: a primer on data privacy protection," 2001.

[8] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression.," *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems*, vol. 10, no. 5, p. 571, 2002.

[9] H. WR, "Medical informatics: Improving health care through information," *JAMA*, vol. 288, no. 16, pp. 1955–1958, 2002.

[10] C. Safran, M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, and D. E. Detmer, "Toward a national framework for the secondary use of health data: An american medical informatics association white paper," *Journal of the American Medical Informatics Association*, vol. 14, no. 1, pp. 1 – 9, 2007.

[11] H. A. Taylor and S. Johnson, "Ethics of population-based research," *The Journal of Law, Medicine & Ethics*, vol. 35, no. 2, pp. 295–299, 2007.

[12] P. Huston and C. D. Naylor, "Health services research: reporting on studies using secondary data sources.," *CMAJ: Canadian Medical Association Journal*, vol. 155, no. 12, p. 1697, 1996.

[13] T. G. Savel, S. Foldy, *et al.*, "The role of public health informatics in enhancing public health surveillance," *CDC's Vision for Public Health Surveillance in the 21st Century*, vol. 61, p. 20, 2012.

[14] J. S. Lombardo and D. L. Buckeridge, *Disease surveillance: a public health informatics approach.* Wiley-Interscience, 2007.

[15] J. Childress and R. Gaare Bernheim, "Public health ethics," *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, vol. 51, no. 2, pp. 158–163, 2008.

[16] J. F. Childress, R. R. Faden, R. D. Gaare, L. O. Gostin, J. Kahn, R. J. Bonnie, N. E. Kass, A. C. Mastroianni, J. D. Moreno, and P. Nieburg, "Public health ethics: Mapping the terrain," *The Journal of Law, Medicine & Ethics*, vol. 30, no. 2, pp. 170–178, 2002.

[17] H. T. Sorensen, S. Sabroe, and J. Olsen, "A framework for evaluation of secondary data sources for epidemiological research," *International Journal of Epidemiology*, vol. 25, no. 2, pp. 435–442, 1996.

[18] D. Capurro Nario, *Secondary Use of Electronic Clinical Data: Barriers, Facilitators and a Proposed Solution.* PhD thesis, University of Washington, 2012.

[19] J. G. Hodge, "Health information privacy and public health," *The Journal of Law, Medicine and Ethics*, vol. 31, no. 4, pp. 663–671, 2003.

[20] N. E. Kass, "An ethics framework for public health," *Journal Information*, vol. 91, no. 11, 2001.

[21] J. Pandiani, S. Banks, and L. Schacht, "Personal privacy versus public accountability: A technological solution to an ethical dilemma," *The Journal of Behavioral Health Services & Research*, vol. 25, no. 4, pp. 456–463, 1998.

[22] L. O. Gostin and J. G. Hodge Jr, "Personal privacy and common goods: a framework for balancing under the national health information privacy rule," *Minn. L. Rev.*, vol. 86, p. 1439, 2001.

[23] R. B. Ness, "Influence of the hipaa privacy rule on health research," *JAMA*, vol. 298, no. 18, pp. 2164–2170, 2007.

[24] D. Deapen, "Cancer surveillance and information: balancing public health with privacy and confidentiality concerns (united states)," *Cancer Causes & Control*, vol. 17, no. 5, pp. 633–637, 2006.

[25] G. Church, C. Heeney, N. Hawkins, J. de Vries, P. Boddington, J. Kaye, M. Bobrow, and B. Weir, "Public access to genome-wide data: five views on balancing research with privacy and protection," *PLoS genetics*, vol. 5, no. 10, p. e1000665, 2009.

[26] C. I. A. Emerson, *Private interests in the public domain: Privacy and confidentiality in observational health research.* PhD thesis, McMaster University, 2008.

[27] T. C. Rindfleisch, "Privacy, information technology, and health care," *Communications of the ACM*, vol. 40, no. 8, pp. 92–100, 1997.

[28] R. Gavison, "Privacy and the limits of law," *Yale LJ*, vol. 89, p. 421, 1979.

[29] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Toward privacy in public databases," in *Theory of Cryptography*, pp. 363–385, Springer, 2005.

[30] R. Lazarus, K. Yih, and R. Platt, "Distributed data processing for public health surveillance.," *BMC Public Health*, vol. 6, pp. 235 – 11, 2006.

[31] T. Walley, "Using personal health information in medical research: Overzealous interpretation of uk laws is stifling epidemiological research," *BMJ: British Medical Journal*, vol. 332, no. 7534, p. 130, 2006.

[32] M. A. Harris, A. Levy, and K. Teschke, "Personal privacy and public health: potential impacts of privacy legislation on health research in canada," *Can J Public Health*, vol. 99, no. 4, pp. 293–96, 2009.

[33] D. J. Willison, L. Schwartz, J. Abelson, C. Charles, M. Swinton, D. Northrup, and L. Thabane, "Alternatives to project-specific consent for access to personal information for health research: What is the opinion of the canadian public?," *Journal of the American Medical Informatics Association*, vol. 14, no. 6, pp. 706 – 712, 2007.

[34] M. R. Robling, K. Hood, H. Houston, R. Pill, J. Fay, and H. M. Evans, "Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study," *Journal of Medical Ethics*, vol. 30, no. 1, pp. 104–109, 2004.

[35] J. Strobl, E. Cave, and T. Walley, "Data protection legislation: interpretation and barriers to research," *BMJ: British Medical Journal*, vol. 321, no. 7265, p. 890, 2000.

[36] "*Turkish Statistical Institute. Information Request. Use of Micro Data.*." Retrieved August 25, 2013, from http://www.turkstat.gov.tr.

[37] L. O. Gostin, J. G. Hodge Jr, and R. O. Valdiserri, "Informational privacy and the public's health: the model state public health privacy act," *Journal Information*, vol. 91, no. 9, 2001.

[38] S. Greenland and H. Morgenstern, "Ecological bias, confounding, and effect modification," *International journal of epidemiology*, vol. 18, no. 1, pp. 269–274, 1989.

[39] D. A. Freedman, "Ecological inference and the ecological fallacy," *International Encyclopedia of the social & Behavioral sciences*, vol. 6, pp. 4027–4030, 1999.

[40] N. Pearce, "The ecological fallacy strikes back," *Journal of epidemiology and community health*, vol. 54, no. 5, pp. 326–327, 2000.

[41] D. Lambert, "Measures of disclosure risk and harm," *Journal of Official Statistics-Stockholm-*, vol. 9, pp. 313–313, 1993.

[42] "Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule," 05 2013.

[43] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the hipaa privacy rule," *J Am Med Inform Assoc*, vol. 17, no. 169e177, p. 169, 2010.

[44] J. J. Gardner, *Privacy Preserving Medical Data Publishing.* Phd, Emory University, Atlanta, 2007.

[45] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, p. 14, 2010.

[46] A. Friedman, R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," *The VLDB Journal*, vol. 17, no. 4, pp. 789–804, 2008.

[47] L. Sweeney, "Replacing personally-identifying information in medical records, the scrub system.," in *Proceedings: a conference of the American Medical Informatics Association. AMIA Fall Symposium*, p. 333, 1996.

[48] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, p. 188, ACM, 1998.

[49] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the datafly system.," in *Proceedings of the AMIA Annual Fall Symposium*, p. 51, American Medical Informatics Association, 1997.

[50] L. Sweeney, "Foundations of privacy protection from a computer science perspective," 2000.

[51] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. S. Nordholt, G. Seri, and P.-P. De Wolf, "Handbook on statistical disclosure control," 2007.

[52] T. Truta and B. Vinay, "Privacy protection: p-sensitive k-anonymity property," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pp. 94–94, 2006.

[53] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 3, 2007.

[54] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 106–115, April 2007.

[55] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *KDD 06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 754–759, ACM, 2006.

[56] J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang, "Maintaining k-anonymity against incremental updates," in *SSDBM '07: Proceedings of the 19th International Conference on Scientific and Statistical Database Management*, (Washington, DC, USA), p. 5, IEEE Computer Society, 2007.

[57] M. Lunacek, D. Whitley, and I. Ray, "A crossover operator for the k- anonymity problem," in *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, (New York, NY, USA), pp. 1713–1720, ACM, 2006.

[58] K. Chatzikokolakis, C. Palamidessi, and P. Panangaden, "Anonymity protocols as noisy channels," *Information and Computation*, vol. 206, no. 2–4, pp. 378 – 401, 2008.

[59] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," *The VLDB Journal*, vol. 15, no. 4, pp. 316–333, 2006.

[60] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, *Anonymizing Tables*, pp. 246–258. 2005.

[61] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain k-anonymity," in *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, (New York, NY, USA), pp. 49–60, ACM, 2005.

[62] H. Park and K. Shim, "Approximate algorithms for k-anonymity," in *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, (New York, NY, USA), pp. 67–78, ACM, 2007.

[63] C. Yao, X. S. Wang, and S. Jajodia, "Checking for k-anonymity violation by views," in *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pp. 910–921, VLDB Endowment, 2005.

[64] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on*, pp. 25–25, April 2006.

[65] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through microaggregation," *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 195–212, 2005.

[66] W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity," in *Data and Applications Security XIX* (S. Jajodia and D. Wijesekera, eds.), vol. 3654 of *Lecture Notes in Computer Science*, pp. 166–177, Springer Berlin / Heidelberg, 2005.

[67] T. Li and N. Li, "Towards optimal k-anonymization.," *Data & Knowledge Engineering*, vol. 65, no. 1, pp. 22 – 39, 2008.

[68] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, (New York, NY, USA), pp. 279–288, ACM, 2002.

[69] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira, "Efficient multi-dimensional suppression for k-anonymity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 334–347, 2010.

[70] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pp. 901–909, VLDB Endowment, 2005.

[71] B. Gedik and L. Liu, "Protecting location privacy with personalized k-anonymity: Architecture and algorithms," *Mobile Computing, IEEE Transactions on*, vol. 7, no. 1, pp. 1–18, 2008.

[72] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *Proceedings of the 33rd international conference on Very large data bases*, VLDB '07, pp. 758–769, VLDB Endowment, 2007.

[73] J. Domingo-Ferrer and V. Torra, "A Critique of k-Anonymity and Some of Its Enhancements," in *Proceedings of the 2008 Third International Conference on Availability, Reliability and Security*, pp. 990–993, IEEE Computer Society, 2008.

[74] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, *Efficient k -Anonymization Using Clustering Techniques*, pp. 188–200. 2008.

[75] Sheng-yizJiang and Q. bo An, "Clustering-based outlier detection method," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, vol. 2, pp. 429 –433, oct. 2008.

[76] K. Jajuga, A. Sokolowski, and H. H. Bock, *Classification, clustering and data analysis: recent advances and applications.* Springer Germany, 2002.

[77] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving anonymity via clustering," in *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 153–162, ACM, 2006.

[78] C. A. Cassa, S. J. Grannis, J. M. Overhage, and K. D. Mandl, "A context-sensitive approach to anonymizing spatial surveillance data impact on outbreak detection," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 160–165, 2006.

[79] B. Fung, K. Wang, L. Wang, and M. Debbabi, "A framework for privacy-preserving cluster analysis," in *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on*, pp. 46–51, IEEE, 2008.

[80] B. C. Fung, K. Wang, L. Wang, and P. C. Hung, "Privacy-preserving data publishing for cluster analysis," *Data Knowledge Engineering*, vol. 68, no. 6, pp. 552 – 575, 2009.

[81] J. Domingo-Ferrer, "Microaggregation for database and location privacy," *Lecture notes in computer science*, pp. 106–116, 2006.

[82] H. Zhu and X. Ye, "Achieving k-anonymity via a density-based clustering method," in *Advances in Data and Web Management*, pp. 745–752, Springer, 2007.

[83] C.-C. Chiu and C.-Y. Tsai, "A k-anonymity clustering method for effective data privacy preservation," in *Advanced Data Mining and Applications*, pp. 89–99, Springer, 2007.

[84] J.-L. Lin, M.-C. Wei, C.-W. Li, and K.-C. Hsieh, "A hybrid method for k-anonymization," in *Asia-Pacific Services Computing Conference, 2008. APSCC '08. IEEE*, pp. 385–390, 2008.

[85] J. Liu and K. Wang, "Enforcing vocabulary k-anonymity by semantic similarity based clustering," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 899–904, 2010.

[86] J. Li, K. Yi, and Q. Zhang, “Clustering with diversity,” in *Automata, Languages and Programming*, pp. 188–200, Springer, 2010.

[87] J.-L. Lin and M.-C. Wei, “Genetic algorithm-based clustering approach for k-anonymization,” *Expert Systems with Applications*, vol. 36, no. 6, pp. 9784–9792, 2009.

[88] J.-L. Lin and M.-C. Wei, “An efficient clustering method for k-anonymization,” in *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, PAIS ’08, (New York, NY, USA), pp. 46–50, ACM, 2008.

[89] G. Loukides and J.-H. Shao, “An efficient clustering algorithm for k-anonymisation,” *Journal of Computer Science and Technology*, vol. 23, no. 2, pp. 188–202, 2008.

[90] G. Loukides and J. Shao, “Greedy clustering with sample-based heuristics for k-anonymisation,” in *Data, Privacy, and E-Commerce, 2007. ISDPE 2007. The First International Symposium on*, pp. 191–196, 2007.

[91] G. Loukides and J. Shao, “Capturing data usefulness and privacy protection in k-anonymisation,” in *Proceedings of the 2007 ACM symposium on Applied computing*, SAC ’07, (New York, NY, USA), pp. 370–374, ACM, 2007.

[92] L. Lu and X. Ye, “An improved weighted-feature clustering algorithm for k-anonymity,” in *Information Assurance and Security, 2009. IAS’09. Fifth International Conference on*, vol. 1, pp. 415–418, IEEE, 2009.

[93] M. Nergiz, C. Clifton, and A. Nergiz, “Multirelational k-anonymity,” in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 1417–1421, 2007.

[94] T. S, *Epidemiyoloji Tıbbi Araştırmaların Yöntem Bilimi.* No. 92/1, Hacettepe Halk Sağlığı Vakfı Yayınları, 1992.

[95] E. Lynge, “European directive on confidential data: a threat to epidemiology.,” *BMJ: British Medical Journal*, vol. 308, no. 6927, p. 490, 1994.

[96] M. Rahu and M. McKee, “Epidemiological research labelled as a violation of privacy: the case of estonia,” *International Journal of Epidemiology*, vol. 37, no. 3, pp. 678–682, 2008.

[97] A. V. Diez-Roux, “Bringing context back into epidemiology: variables and fallacies in multilevel analysis.,” *American journal of public health*, vol. 88, no. 2, pp. 216–222, 1998.

[98] F. Department of Psychology California State University, “Confounding variables-psychology 144, research methods, department of psychology california state university, fresno.”

[99] S. Greenland and H. Morgenstern, “Confounding in health research,” *Annual Review of Public Health*, vol. 22, no. 1, pp. 189–212, 2001.

[100] S. Greenland, J. M. Robins, and J. Pearl, “Confounding and collapsibility in causal inference,” *Statistical Science*, vol. 14, no. 1, pp. pp. 29–46, 1999.

[101] "*Evidence Based Medicine Toolkit. Clinical Epidemiology Glossary.*" (n.d.). Retrieved June 01, 2013, from http://www.ebm.med.ualberta.ca/Glossary.html.

[102] M. Jenicek, "Epidemiology, evidenced-based medicine, and evidence-based public health.," *Journal of epidemiology/Japan Epidemiological Association*, vol. 7, no. 4, p. 187, 1997.

[103] M. J. Dobrow, V. Goel, R. Upshur, *et al.*, "Evidence-based health policy: context and utilisation," *Social Science and Medicine*, vol. 58, no. 1, pp. 207–218, 2004.

[104] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Clustering validity checking methods: part ii," *ACM Sigmod Record*, vol. 31, no. 3, pp. 19–27, 2002.

[105] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 187–194, IEEE, 2001.

[106] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster validity methods: part i," *SIGMOD Rec.*, vol. 31, pp. 40–45, June 2002.

[107] C. T. Di Iorio, F. Carinci, J. Azzopardi, V. Baglioni, P. Beck, S. Cunningham, A. Evripidou, G. Leese, K. F. Loevaas, G. Olympios, M. O. Federici, S. Pruna, P. Palladino, S. Skeie, P. Taverner, V. Traynor, and M. M. Benedetti, "Privacy impact assessment in the design of transnational public health information systems: the biro project," *Journal of Medical Ethics*, vol. 35, no. 12, pp. 753–761, 2009.

[108] D. Deapen, "Cancer surveillance and information: Balancing public health with privacy and confidentiality concerns (united states).," *Cancer Causes and Control*, vol. 17, no. 5, pp. 633 – 637, 2006.

[109] J. Myers, T. R. Frieden, K. M. Bherwani, and K. J. Henning, "Ethics in public health research," *Journal Information*, vol. 98, no. 5, 2008.

[110] "*European Commision. Protection of personal data.*" Retrieved June 01, 2013, from http://ec.europa.eu/justice/data-protection.

[111] "*Understanding Health Information Privacy.*" Retrieved June 01, 2013, form http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html.

[112] "*Data Protection Act 1998.*" Retrieved June 01, 2013, from http://www.legislation.gov.uk/ukpga/1998/29/contents.

[113] "*Summary of the HIPAA Privacy Rule.*" Retrieved June 01, 2013, from http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html.

[114] S. A. Redsell and F. M. Cheater, "The data protection act (1998): implications for health researchers," *Journal of Advanced Nursing*, vol. 35, no. 4, pp. 508–513, 2001.

[115] R. Al-Shahi and C. Warlow, "Using patient-identifiable data for observational research and audit: Overprotection could damage the public interest," *BMJ: British Medical Journal*, vol. 321, no. 7268, p. 1031, 2000.

[116] "*TUIK, Turkish Statistical Institute.*" Retrieved June 01, 2013, from http://www.turkstat.gov.tr.

[117] "*Resmi İstatistiklerde Veri Gizliliği Ve Gizli Veri Güvenliğine İlişkin Usul Ve Esaslar Hakkinda Yönetmelik.*" Retrieved June 01, 2013, from http://www.tuik.gov.tr/UstMenu/yonetmelikler/Veri_Gizliligi.pdf.

[118] "*Türkiye Halk Sağlığı Kurumu. Aile Hekimliği.*" Retrieved June 01, 2013, from http://www.ailehekimligi.gov.tr.

[119] E. Aydin, "Rights of patients in developing countries: the case of turkey," *Journal of Medical Ethics*, vol. 30, no. 6, pp. 555–557, 2004.

[120] "*Hasta Haklari Yönetmeliği (Statute of patient rights). Official Gazette 1998 Aug 1 (no 23420).*" Retrieved June 01, 2013, from http://www.saglik.gov.tr/TR/belge/1-555/eski2yeni.html.

[121] "*Turkish Institute of Public Health-Regulations.*" Retrieved June 01, 2013, from http://www.thsk.gov.tr/tr/index.php/mevzuat/yonetmelikler.

[122] "*General Directorate of Development of Legislation and Publication.*" Retrieved June 01, 2013, from http://www.mevzuat.gov.tr.

[123] "*Hekimler Ve Tabip Odası Yöneticileri için Mevzuat. Tıbbi Deontoloji Tüzüğü.*" Retrieved June 01, 2013, from http://www.ttb.org.tr.

[124] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, pp. 283–304, 1998.

[125] G. Gan, P. Ma, Chaoqun, and J. Wu, *Data clustering : theory, algorithms, and applications / Guojun Gan, Chaoqun Ma, Jianhong Wu.* ASA-SIAM series on statistics and applied probability: 20, Philadelphia, Pa. : Society for Industrial and Applied Mathematics, 2007., 2007.

[126] W. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of classification*, vol. 1, no. 1, pp. 7–24, 1984.

[127] B. Ange, J. Symons, M. Schwab, E. Howell, and A. Geyh, "Generalizability in epidemiology: an investigation within the context of heart failure studies," *Annals of Epidemiology*, vol. 14, no. 8, pp. 600 – 601, 2004.

[128] S. Schwartz, "The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences.," *American journal of public health*, vol. 84, no. 5, pp. 819–824, 1994.

[129] L. Meo-Evoli, E. Nardelli, D. M. Pisanelli, and F. L. Ricci, "Adams: an object-oriented system for epidemiological data manipulation," in *Proceedings of the 1993 ACM/SIGAPP symposium on Applied computing: states of the art and practice*, SAC '93, (New York, NY, USA), pp. 652–659, ACM, 1993.

[130] J. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.

[131] M. Ichino, “General metrics for mixed features the cartesian space theory for pattern recognition,” in *Systems, Man, and Cybernetics, 1988. Proceedings of the 1988 IEEE International Conference on*, vol. 1, pp. 494–497, IEEE, 1988.

[132] M. Ichino and H. Yaguchi, “Generalized minkowski metrics for mixed feature-type data analysis,” *IEEE Transactions On Systems, Man, and Cybernetics*, vol. 24, no. 4, 1994.

[133] K. Buse, N. Mays, and G. Walt, *Making health policy / Kent Buse, Nicholas Mays and Gill Walt.* Understanding public health, Maidenhead ; New York, NY : Open University Press, 2005, 2005.

[134] H. M. Bochel and S. Duncan, *Making policy in theory and practice / edited by Hugh Bochel and Sue Duncan.* Bristol : Policy, 2007., 2007.

[135] “*The Green Book. Appraisal And Evaluation in Central Government.*” Retrieved August 25, 2013, from https://www.gov.uk/government.

[136] A. Chorley, P. Edwards, A. Preece, and J. Farrington, “Tools for tracing evidence in social science,” in *Proceedings of the Third International Conference on eSocial Science*, Citeseer, 2007.

[137] K. Buse, N. Mays, and G. Walt, *Making health policy / Kent Buse, Nicholas Mays and Gill Walt.* Understanding public health, Maidenhead ; New York, NY : Open University Press, 2005, 2005.

[138] G. Walt and L. Gilson, “Reforming the health sector in developing countries: the central role of policy analysis,” *Health policy and planning*, vol. 9, no. 4, pp. 353–370, 1994.

[139] D. L. Cannon, *CISA [electronic resource] : certified information systems auditor study guide / David L. Cannon.* Indianapolis, Ind. : Wiley Pub., c2008., 2008.

[140] “Professional policy making for the twenty first century,” report, Great Britain. Cabinet Office. Strategic Policy Making Team, September 1999.

[141] Office of the First Minister and Deputy Frist Minister, *A Practical Guide to Policy Making in Northern Ireland*, 2003.

[142] A. Casellas and C. C. Galley, “Regional definitions in the european union: a question of disparities?,” *Regional Studies*, vol. 33, no. 6, pp. 551–558, 1999.

[143] “*Global Adult Tobacco Survey (GATS).*” Retrieved June 01, 2013, from http://www.cdc.gov/tobacco/global.

[144] “*Global Tobacco Surveillance System Data (GTSSData).*” Retrieved June 01, 2013, from http://nccd.cdc.gov/gtssdata/Default/Default.aspx.

[145] T. Baška, C. Warren, M. Bašková, and N. Jones, “Prevalence of youth cigarette smoking and selected social factors in 25 european countries: findings from the global youth tobacco survey,” *International Journal of Public Health*, vol. 54, no. 6, pp. 439–445, 2009.

[146] T. Erguder, T. Soydal, M. Ugurlu, B. Cakir, and C. Warren, “Tobacco use among youth and related characteristics, turkey,” *Sozial- und Praventivmedizin*, vol. 51, no. 2, pp. 91–98, 2006.

[147] G. A. Giovino, S. A. Mirza, J. M. Samet, P. C. Gupta, M. J. Jarvis, N. Bhala, R. Peto, W. Zatonski, J. Hsia, J. Morton, *et al.*, “Tobacco use in 3 billion individuals from 16 countries: an analysis of nationally representative cross-sectional household surveys,” *The Lancet*, vol. 380, no. 9842, pp. 668–679, 2012.

[148] “*Global Adult Tobacco Survey (GATS). Frequently Asked Questions.*” Retrieved June 01, 2013, from http://nccd.cdc.gov/gtssdata/Ancillary/DownloadAttachment.aspx?ID=948.

[149] “Global adult tobacco survey, turkey report,” tech. rep., The Ministry Of Healthof Turkey, Primary Health Core General Director, 2010.

[150] “Global adult tobacco survey (gats) fact sheet turkey 2008,” tech. rep., Ministry Of Health Turkey, 2010.

[151] “Global adult tobacco survey (gats) codebook, turkey 2008,” tech. rep., 2011.

APPENDIX A

WEB BASED QUESTIONNAIRE

Please read this section before filling the questionnaire. This e-survey has been prepared as a part of research conducted by Mehrdad Alizadeh Mizani with an aim to assess the secondary uses of health data in public health contexts and related privacy issues. the goal of this study is to elicit the accessibility and usability of secondary health data, concern about patient privacy, and ideas about privacy policies. Filling this questionnaire is voluntarily. No personal information will be asked throughout the questionnaire. The raw answers will not be disclosed to third parties. They will be only be used by those who conduct this study in order to share the results for scientific purposes.

The questionnaire is an e-survey and will be filled via web. None of the questions contains personal or sensitive issues. However, if at any stage you feel uncomfortable about filling the questionnaire, you can quit the questionnaire. Not clicking on the 'Submit' button at the end of the web page is equivalent to quitting the questionnaire. In such cases, the given answers to the point of quitting will not be submitted.

You can share your concerns and suggestions about this questionnaire at the end the page or you can send us an email to the address provided. To obtain more details about this study you can contact Informatics Institute of Middle East Technical University.

If you don't click on 'Submit' button no information will be submitted to the researcher. Clicking on 'Submit' button indicates that you agree with the following statement:

I attended this study and filled the questionaire voluntarily and at my own will. I was aware that I could quit the questionnaire at any time at my own will. I give permission for the analysis of my answers to be used in academic and scientific research and publications

Thank you for your contribution.
Mehrdad Alizadeh Mizani

Terms used in this questionnaire: Secondary uses of health data = Refers to non-clinical uses of health data which are not related to providing direct healthcare to patients. Examples of secondary uses are epidemiology, public health, quality analysis, disease surveillance, and market analysis. Individual data = Refers to person-specific data in which each records belongs to a specific person. Aggregate data = Refers to statistical representation of data in non-individual forms such as means and percentage.

Questions

1. **What is your main area of expertise/profession? (eg. MD, Epidemiologist, Medical Informaticist, Public health)**:
   ____

2. **If you've ever needed health data for secondary uses, what were the purposes of secondary uses? (You can choose more than one option.)**

   ○ Public health

   ○ Epidemiology

   ○ Disease surveillance

   ○ Market analysis

   ○ Other: ____________________

3. **Which data format do you USUALLY prefer for your research?**

   ○ Aggregated data

   ○ Individual data

   ○ Other: ____________________

4. **If you've ever obtained individual data for secondary uses, how easy or difficult was it for you to obtain it?**
   very easy ○—○—○—○—○ very difficult

5. **If you've ever obtained aggregated data for secondary uses, how easy or difficult was it for you to obtain it?**
   very easy ○—○—○—○—○ very difficult

6. **If you've ever used aggregate data, how was the overall validity of data for research in best case?**
   very valid ○—○—○—○—○ extremely invalid

7. **Rank the following problems with obtaining/using aggregate health data for secondary uses. (Scalled as not applicable-minor-high but managable-major-extremely complicated)**

a **Data were highly aggregated.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

b **No information provided to me about the missing data in the original individual data.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

c **No information provided to me about the invalid data in the original individual data.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

d **I had difficulty in generalizing the result to the population.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

e **Generally the amount of useful information in aggregated data was very low.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

f **Generally the amount of useful information in aggregated data was very low.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

g **Add any problems and their severity that are not included**

____

____

8. Which one of the following reasons is the most important barrier in accessing/using individual data

○ The data holder has a fixed policy of releasing data in aggregate form only

○ Lengthy formalities of the data holder makes it practically impossible to obtain individual data

○ The prohibitions imposed by data holder limited the use of individual data for my research purposes

○ The data holder has concerns over legal aspects of privacy leaks

○ Other: ____________________

9. To what extent the following problems have affected your research (Scalled as not applicable-minor-high but managable-major-extremely complicated)

a **The data I needed was scattered among different organizations.]**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

b **The subjects of my research had records in different organizations.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

c **I had difficulty with different data standards used by data holders.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

d **The amount of missing data in individual data affected my research.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

e **The amount of invalid data in individual data affected my research.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

f **Since I didn't know the amount of missing data in aggregate data I was unable to verify the bias of my research.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

g **Since I didn't know the amount of invalid data in aggregate data I was unable to verify the bias of my research.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

h **Since I used aggregate data I was unable to generalize the results.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

i **I couldn't use data due to wrong data collection methods used by data holder.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

j **I was unable to access the individual data.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

k **I had concerns about privacy of the subjects.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

l **I had concerns about legal issues over use of individual data.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

m **I had difficulty in obtaining data due to the lack of coordination between different data holders made.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

n **I had difficulty in obtaining data due to the lack of legislations and policies to govern secondary uses of data.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

o **I had difficulty in obtaining the whole data due to the lack of joining methods between different data holders.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

p **Data holders were reluctant to publish medical data for secondary uses.**
○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

q **Add any problems and their severity that are not included**

____

____

10. **How important it is to have access to individual data in most of your researches? (Leave unanswered in non applicable)**
no importance ○—○—○—○—○ very important

11. **How serious is the privacy problem in individual data? (Leave unanswered in non applicable)**
no problem ○—○—○—○—○ serious important

12. **Removing directly identifiable fields, like name, last name, TC kimlik numarasi, adres, telefon, etc. is enough to protect privacy**

    ○ Yes

    ○ No

    ○ Not sure

    ○ Other: ____________________

13. **Rank the following problems in accessing to available medical data for secondary uses.(Scalled as not applicable-minor-high but managable-major-extremely complicated)**

    a **Lack of effective technical solutions to protect privacy.**
    ○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

    b **Lack of coordination between organizations.**
    ○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

    c **Lack of effective detailed policies to govern the secondary uses of data.**
    ○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

    d **Availability of aggregate data instead of individual data.**
    ○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

    e **Complexity of privacy protection in releasing data.**
    ○ n/a ○ minor ○ high but managable ○ major ○ extremely complicated

    f **Add any problems and their severity that are not included**

    ____

    ____

    ____

14. **Rank the following barriers in maintaing uniform policies for publishing health data for secondary uses.(Scalled as not applicable-little effect-minor-major-extremely serious)**

    a **Cost of designing and applying policies.**
    ○ n/a ○ very little ○ minor ○ major ○ extremely serious

    b **Unawareness toward the importance of secondary uses of medical data.**
    ○ n/a ○ very little ○ minor ○ major ○ extremely serious

    c **The need for a centralized and governmental effort to design and implement a policy.**
    ○ n/a ○ very little ○ minor ○ major ○ extremely serious

    d **Organizations are reluctant to share their data.**
    ○ n/a ○ very little ○ minor ○ major ○ extremely serious

e **Complexity of privacy protection in secondary uses of data.**
○ n/a ○ very little ○ minor ○ major ○ extremely serious

f **Uncertainty about the effectiveness of technical solutions.**
○ n/a ○ very little ○ minor ○ major ○ extremely serious

g **Prohibiting state regulations.**
○ n/a ○ very little ○ minor ○ major ○ extremely serious

h **Add any barriers and their effects that are not included**

____

____

15. Please add any additional comments in the following area.

____

____

____

____

APPENDIX B

CHECKLIST FOR REPORTING RESULTS OF INTERNET E-SURVEYS (CHERRIES)

This appendix presents the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) filled for the questionnaire presented in appendix A.

1. **Design**

   The survey was prepared in Google Docs and the link was sent by email to the mailing lists of the public health and medical informatics professionals. The survey link was posted to the owners of the mailing lists. The most popular mailing lists with large number of members were selected. These include the mailing lists for public health specialists and epidemiologists in Turkey and the mailing lists of medical informatics in METU university and Turkey. Some mailing list members also forwarded the link to other related mailing lists they were a member of.

2. **IRB**

   (a) **IRB approval.** The survey was approved by UEAM , or Research Center for Applied Ethics of the METU university. UEAM is the official board that approves the ethical aspects of the studies conducted in METU.

   (b) **IRB informed consent** The participants were provided with the information about the survey in the first page of the survey, as required by UEAM. The information provided the following items:

   - Length of the survey
   - The fact that it is not mandatory to fill the survey
   - That the participants can quit filling the survey whenever they like
   - The purpose of the research: to assess the need for and difficulties of accessing person-specific health data and related policy issues
   - Name, affiliation, and contact address of the investigators
   - That the answers will be saved on Google Docs server
   - That no personal information will be asked
   - That no information about their IP or internet connection will be gathered

   (c) **IRB data protection.** No personal data was gathered from participants.

3. **Development and pre-testing**

(a) **Development and testing**

- The first step was to formulate the main questions which were
- What is the demand for individual health data
- What are the barriers in accessing individual data
- What are the related policy issues
- General awareness about privacy protection in secondary usage of health data
- Then we designed the questions which were checked by 4 experts in the fields of medical informatics, public health, and epidemiology
- The questionnaire were pre-tested by 7 experts and 6 students from medical informatics, public health and epidemiology fields
- Experts and students participated in reviews and pre-tests were asked to refuse filling the final survey
- The visibility and functionality of online survey forms were tested on most popular internet browsers.

4. **Recruitment process and description of the sample having access to the questionnaire**

(a) **Open survey versus closed survey.** The e-survey link was available without password. However, the link was sent to the mailing lists. The link was not searchable on search engines. As a result, only the professionals who received the link had access to the survey.

(b) **Contact mode.** No personal contact was made. The link of the web based survey was sent by email to selected mailing lists.

(c) **Open survey versus closed survey.** No special advertisement were used. The link accompanied with the information given at "IRB informed consent" was emailed to the mailing lists.

5. **Survey administration**

(a) **Web/E-mail.** Web based Google Docs e-survey. The results were entered into a spreadsheet automatically.

(b) **Context.** The survey was not published on a web site. The link to a web based Google Docs survey was emailed to selected mailing lists.

(c) **Mandatory/voluntary.** It was a voluntary survey.

(d) **Incentives.** No incentives offered.

(e) **Time/Date** 20 Dec 2011 - 30 Jan 2012

(f) **Randomization of items or questionnaires.** No randomization or alternate items were used.

(g) **Adaptive questioning.** Some questions were conditioned to the answers given to the previous ones.

(h) **Number of Items**

- 15 main questions some of which had several sub-items.
- 45 including multi-item questions and open questions for optional explanations about the issues not asked in the questionnaire.

(i) **Number of screens (pages).** Single webpage (Seven printed pages) including the consent, information about researchers, and boxes for open questions.

(j) **Completeness check.** Mandatory completeness check was not used. All questions had 'not applicable' or 'other' items with a box for optional explanation.

(k) **Review step.** The survey was a single web page with all the given answers visible by scrolling. In Google Docs it is not possible to review or change the answers after submitting them.

6. **Response rates**

(a) **Unique site visitor.** For privacy issues Google Docs does not provide mechanisms for unique identification of visitors. No IP or cookie information is stored on Google Docs. This issue was provided at the beginning of the survey asking the participants to refuse resubmitting the survey.

(b) **View rate (Ratio unique site visitors/unique survey visitors).** Google Docs does not provide information about the number of visitors.

(c) **Participation rate (Ratio unique survey page visitors/agreed to participate).** Since the link was mailed to mailing lists and the information about number of visitors was unavailable.

(d) **Completion rate (Ratio agreed to participate/finished survey).** Not possible to calculate on Google Docs.

7. **Preventing multiple entries from the same individual**

(a) **Cookies used.** Not available on Google Docs.

(b) **IP check.** Not available on Google Docs.

(c) **Registration.** No login used. Members of selected mailing lists had open access to the survey.

8. **Analysis**

(a) **Handling of incomplete questionnaires.** Questioners with large number of missing answers, or missing profession information, were not considered in the analysis. In two questions with large number of responses, the missing values were replaced by mode of the answers given by participants from the same profession group.

(b) **Questionnaires submitted with an atypical timestamp** The e-survey was a single page webpage. Google Docs does not provide functionalities for saving the survey for later or information about the time spent on filling the survey.

(c) **Statistical correction.** Due to the nature of e-surveys and the uncertainty about the number of mailing list members who had read the sent email, the results of our survey is not generalizable to whole professionals. It was conducted as supporting information and can only represent the members of mailing lists.

APPENDIX C

SAMPLE DATASETS BASED ON 2008 GATS DATA

This appendix presents the sample datasets for scenarios of Chapter 6.

TableC.1: Registry sample dataset D_1, Case 1

Key	**Hh**	**Age**	**Gender**	**Education**	**Job**	**Smoker**	**PSAs**	**Illness**	**Value**
101	5	42	M	Elementary	Self	No	Yes	Yes	35
102	4	24	F	Elementary	House	No	Yes	Yes	36
103	3	48	F	Vocational	Retired	Daily	No	Yes	46
104	2	27	F	Vocational	Employee	No	Yes	Yes	34
105	3	49	M	Elementary	Retired	Daily	Yes	Yes	45
106	2	16	F	Primary	House	No	Yes	Yes	42
107	3	48	M	High school	Employee	Daily	Yes	Yes	33
108	3	23	F	High school	Employee	No	Yes	Yes	32
109	6	38	M	Vocational	Employee	No	Yes	Yes	51
110	4	30	F	Elementary	House	No	Yes	Yes	38
111	5	34	M	Vocational	Self	No	Yes	Yes	42
112	7	46	F	Elementary	House	No	No	Yes	43
113	5	58	F	None	House	No	Yes	Yes	33
114	3	35	M	Elementary	Employee	Daily	Yes	Yes	32
115	4	27	F	High school	House	No	Yes	Yes	39
116	6	65	F	None	House	No	Yes	Yes	38
117	4	42	M	Vocational	Self	Daily	Yes	Yes	42
118	5	42	F	None	House	No	Yes	Yes	51
119	2	35	M	High school	Employee	No	Yes	Yes	51
120	3	22	F	High school	Employee	Daily	Yes	Yes	39
121	4	37	M	Vocational	Employee	No	Yes	Yes	39
122	2	24	F	High school	House	No	Yes	Yes	38
123	4	36	M	High school	Employee	No	Yes	Yes	41
124	4	34	F	Vocational	House	Daily	Yes	Yes	42
125	4	80	M	None	Retired	Daily	Yes	Yes	44
126	3	49	F	Elementary	Retired	Daily	Yes	Yes	36
127	3	54	M	High school	Retired	Daily	Yes	Yes	32
128	2	49	F	Elementary	House	No	Yes	Yes	48
129	4	65	M	Elementary	Retired	No	Yes	Yes	38
130	5	32	F	High school	House	No	Yes	Yes	40
131	4	32	M	High school	Self	No	Yes	Yes	41
132	5	24	M	High school	Self	No	Yes	Yes	36
133	4	31	F	High school	House	Daily	Yes	Yes	37
134	4	34	F	College	House	No	Yes	Yes	37
135	4	32	M	High school	Self	Daily	Yes	Yes	38
136	2	27	F	College	Employee	No	Yes	Yes	50
137	4	35	M	College	Employee	No	Yes	No	61

Continued on next page

Table C.1 continued.

Key	Hh	Age	Gender	Education	Job	Smoker	PSAs	Illness	Value
138	5	35	F	Elementary	House	No	Yes	Yes	34
139	3	34	M	College	Employee	No	Yes	Yes	30
140	2	23	F	High school	House	No	Yes	Yes	28

TableC.2: Registry sample dataset D_2, Case 2

Key	Hh	Job	PSA	Illness	Value
101	5	Self	Yes	Yes	35
111	5	Self	Yes	Yes	42
132	5	Self	Yes	Yes	36
117	4	Self	Yes	Yes	42
131	4	Self	Yes	Yes	41
135	4	Self	Yes	Yes	38
113	5	House	Yes	Yes	33
118	5	House	Yes	Yes	51
130	5	House	Yes	Yes	40
138	5	House	Yes	Yes	34
102	4	House	Yes	Yes	36
110	4	House	Yes	Yes	38
115	4	House	Yes	Yes	39
124	4	House	Yes	Yes	42
133	4	House	Yes	Yes	37
134	4	House	Yes	Yes	37
106	2	House	Yes	Yes	42
122	2	House	Yes	Yes	38
128	2	House	Yes	Yes	48
140	2	House	Yes	Yes	28
116	6	House	Yes	Yes	38
112	7	House	No	Yes	43
121	4	Employee	Yes	Yes	39
123	4	Employee	Yes	Yes	41
137	4	Employee	Yes	No	61
107	3	Employee	Yes	Yes	33
108	3	Employee	Yes	Yes	32
114	3	Employee	Yes	Yes	32
120	3	Employee	Yes	Yes	39
139	3	Employee	Yes	Yes	30
104	2	Employee	Yes	Yes	34
119	2	Employee	Yes	Yes	51
136	2	Employee	Yes	Yes	50
109	6	Employee	Yes	Yes	51
125	4	Retired	Yes	Yes	44
129	4	Retired	Yes	Yes	38
103	3	Retired	No	Yes	46
105	3	Retired	Yes	Yes	45
126	3	Retired	Yes	Yes	36
127	3	Retired	Yes	Yes	32

TableC.3: Generalization approach applied on D_2, example 1

Key	Hh	Job	PSA	Illness	Value
121	1-5	Employee	Yes	Yes	39
123	1-5	Employee	Yes	Yes	41
137	1-5	Employee	Yes	No	61
104	1-5	Employee	Yes	Yes	34
119	1-5	Employee	Yes	Yes	51

Continued on next page

Table C.3 continued.

Key	Hh	Job	PSA	Illness	Value
136	1-5	Employee	Yes	Yes	50
107	1-5	Employee	Yes	Yes	33
108	1-5	Employee	Yes	Yes	32
114	1-5	Employee	Yes	Yes	32
120	1-5	Employee	Yes	Yes	39
139	1-5	Employee	Yes	Yes	30
102	1-5	House	Yes	Yes	36
110	1-5	House	Yes	Yes	38
115	1-5	House	Yes	Yes	39
124	1-5	House	Yes	Yes	42
133	1-5	House	Yes	Yes	37
134	1-5	House	Yes	Yes	37
106	1-5	House	Yes	Yes	42
122	1-5	House	Yes	Yes	38
128	1-5	House	Yes	Yes	48
140	1-5	House	Yes	Yes	28
113	1-5	House	Yes	Yes	33
118	1-5	House	Yes	Yes	51
130	1-5	House	Yes	Yes	40
138	1-5	House	Yes	Yes	34
125	1-5	Retired	Yes	Yes	44
129	1-5	Retired	Yes	Yes	38
103	1-5	Retired	No	Yes	46
105	1-5	Retired	Yes	Yes	45
126	1-5	Retired	Yes	Yes	36
127	1-5	Retired	Yes	Yes	32
117	1-5	Self	Yes	Yes	42
131	1-5	Self	Yes	Yes	41
135	1-5	Self	Yes	Yes	38
101	1-5	Self	Yes	Yes	35
111	1-5	Self	Yes	Yes	42
132	1-5	Self	Yes	Yes	36
109	6-10	***	Yes	Yes	51
116	6-10	***	Yes	Yes	38
112	6-10	***	No	Yes	43

TableC.4: Generalization approach on D_2, example 2

Key	Hh	Job	PSA	Illness	Value
101	5	Works	Yes	Yes	35
111	5	Works	Yes	Yes	42
132	5	Works	Yes	Yes	36
117	4	Works	Yes	Yes	42
131	4	Works	Yes	Yes	41
135	4	Works	Yes	Yes	38
121	4	Works	Yes	Yes	39
123	4	Works	Yes	Yes	41
137	4	Works	Yes	No	61
104	2	Works	Yes	Yes	34
119	2	Works	Yes	Yes	51
136	2	Works	Yes	Yes	50
107	3	Works	Yes	Yes	33
108	3	Works	Yes	Yes	32
114	3	Works	Yes	Yes	32
120	3	Works	Yes	Yes	39
139	3	Works	Yes	Yes	30
109	6-10	***	Yes	Yes	51
116	6-10	***	Yes	Yes	38
112	6-10	***	No	Yes	43
113	5	Unemployed	Yes	Yes	33
118	5	Unemployed	Yes	Yes	51
130	5	Unemployed	Yes	Yes	40
138	5	Unemployed	Yes	Yes	34

Continued on next page

Table C.4 continued.

Key	Hh	Job	PSA	Illness	Value
102	4	Unemployed	Yes	Yes	36
110	4	Unemployed	Yes	Yes	38
115	4	Unemployed	Yes	Yes	39
124	4	Unemployed	Yes	Yes	42
133	4	Unemployed	Yes	Yes	37
134	4	Unemployed	Yes	Yes	37
125	4	Unemployed	Yes	Yes	44
129	4	Unemployed	Yes	Yes	38
106	2	Unemployed	Yes	Yes	42
122	2	Unemployed	Yes	Yes	38
128	2	Unemployed	Yes	Yes	48
140	2	Unemployed	Yes	Yes	28
103	3	Unemployed	No	Yes	46
105	3	Unemployed	Yes	Yes	45
126	3	Unemployed	Yes	Yes	36
127	3	Unemployed	Yes	Yes	32

TableC.5: Deletion approach applied on D_2

Key	Hh	Job	PSA	Illness	Value
101	5	Self	Yes	Yes	35
111	5	Self	Yes	Yes	42
132	5	Self	Yes	Yes	36
117	4	Self	Yes	Yes	42
131	4	Self	Yes	Yes	41
135	4	Self	Yes	Yes	38
113	5	House	Yes	Yes	33
118	5	House	Yes	Yes	51
130	5	House	Yes	Yes	40
138	5	House	Yes	Yes	34
102	4	House	Yes	Yes	36
110	4	House	Yes	Yes	38
115	4	House	Yes	Yes	39
124	4	House	Yes	Yes	42
133	4	House	Yes	Yes	37
134	4	House	Yes	Yes	37
106	2	House	Yes	Yes	42
122	2	House	Yes	Yes	38
128	2	House	Yes	Yes	48
140	2	House	Yes	Yes	28
121	4	Employee	Yes	Yes	39
123	4	Employee	Yes	Yes	41
137	4	Employee	Yes	No	61
104	2	Employee	Yes	Yes	34
119	2	Employee	Yes	Yes	51
136	2	Employee	Yes	Yes	50
107	3	Employee	Yes	Yes	33
108	3	Employee	Yes	Yes	32
114	3	Employee	Yes	Yes	32
120	3	Employee	Yes	Yes	39
139	3	Employee	Yes	Yes	30
103	3	Retired	No	Yes	46
105	3	Retired	Yes	Yes	45
126	3	Retired	Yes	Yes	36
127	3	Retired	Yes	Yes	32

TableC.6: Merge approach applied on D_2

Key	Hh	Job	PSA	Illness	Value
101	5	Self	Yes	Yes	35
111	5	Self	Yes	Yes	42
132	5	Self	Yes	Yes	36
117	4	Self	Yes	Yes	42
131	4	Self	Yes	Yes	41
135	4	Self	Yes	Yes	38
113	5	House	Yes	Yes	33
118	5	House	Yes	Yes	51
130	5	House	Yes	Yes	40
138	5	House	Yes	Yes	34
116	5	House	Yes	Yes	38
112	5	House	No	Yes	43
102	4	House	Yes	Yes	36
110	4	House	Yes	Yes	38
115	4	House	Yes	Yes	39
124	4	House	Yes	Yes	42
133	4	House	Yes	Yes	37
134	4	House	Yes	Yes	37
106	2	House	Yes	Yes	42
122	2	House	Yes	Yes	38
128	2	House	Yes	Yes	48
140	2	House	Yes	Yes	28
121	4	Employee	Yes	Yes	39
123	4	Employee	Yes	Yes	41
137	4	Employee	Yes	No	61
109	4	Employee	Yes	Yes	51
104	2	Employee	Yes	Yes	34
119	2	Employee	Yes	Yes	51
136	2	Employee	Yes	Yes	50
107	3	Employee	Yes	Yes	33
108	3	Employee	Yes	Yes	32
114	3	Employee	Yes	Yes	32
120	3	Employee	Yes	Yes	39
139	3	Employee	Yes	Yes	30
125	3	Retired	Yes	Yes	44
129	3	Retired	Yes	Yes	38
103	3	Retired	No	Yes	46
105	3	Retired	Yes	Yes	45
126	3	Retired	Yes	Yes	36
127	3	Retired	Yes	Yes	32

TableC.7: Aggregation of generalized D_2, example 1, aggregation of $\mid \xi^* \mid < 10$

Key	Hh	Job	PSA	Illness	Value
121	1-5	Employee	Yes	Yes	39
123	1-5	Employee	Yes	Yes	41
137	1-5	Employee	Yes	No	61
104	1-5	Employee	Yes	Yes	34
119	1-5	Employee	Yes	Yes	51
136	1-5	Employee	Yes	Yes	50
107	1-5	Employee	Yes	Yes	33
108	1-5	Employee	Yes	Yes	32
114	1-5	Employee	Yes	Yes	32
120	1-5	Employee	Yes	Yes	39
139	1-5	Employee	Yes	Yes	30
102	1-5	House	Yes	Yes	36
110	1-5	House	Yes	Yes	38
115	1-5	House	Yes	Yes	39
124	1-5	House	Yes	Yes	42
133	1-5	House	Yes	Yes	37
134	1-5	House	Yes	Yes	37

Continued on next page

Table C.7 continued.

Key	**Hh**	**Job**	**PSA**	**Illness**	**Value**
106	1-5	House	Yes	Yes	42
122	1-5	House	Yes	Yes	38
128	1-5	House	Yes	Yes	48
140	1-5	House	Yes	Yes	28
113	1-5	House	Yes	Yes	33
118	1-5	House	Yes	Yes	51
130	1-5	House	Yes	Yes	40
138	1-5	House	Yes	Yes	34
125	1-5	Retired	Yes	Yes	40.16
129	1-5	Retired	Yes	Yes	40.16
103	1-5	Retired	No	Yes	40.16
105	1-5	Retired	Yes	Yes	40.16
126	1-5	Retired	Yes	Yes	40.16
127	1-5	Retired	Yes	Yes	40.16
117	1-5	Self	Yes	Yes	39
131	1-5	Self	Yes	Yes	39
135	1-5	Self	Yes	Yes	39
101	1-5	Self	Yes	Yes	39
111	1-5	Self	Yes	Yes	39
132	1-5	Self	Yes	Yes	39
109	6-10	***	Yes	Yes	44
116	6-10	***	Yes	Yes	44
112	6-10	***	No	Yes	44

Glossary

aggregated data refer to datasets where records of individuals are grouped based on the central tendency and aggregation of a certain attribute. They are also reffered to as statistical datasets or macro-data. 1, 3, 4, 7, 8, 15, 20, 22, 27–30, 34, 51, 56–60, 79, 89, 108, 109, 124, 125

aggregation refers to applying statisticsl methods on individual data in order to transform them into aggregated data. . 15, 20, 22, 25, 29, 58, 79, 88, 97, 99, 103, 108, 112, 120, 122, 124

confounding bias refers to a bias caused by a wrong assumption that there is a meaningful relationship between a confounding factor, as a cause, and the effect. Confoudning factor refers to an independent variable that is not the cause of an effect, reflected by the dependent variable, and yet might be considered as the actual cause. The exclusion or genealization of quasi-identifiers might hinder the accurate determining the confounding factors. . 22, 57, 59, 77, 78, 96, 97, 118–120

ecological refers to the type of studies based on aggregated and statistical datasets. Although ccological studies are not generalizable to the population due to ecologicla bias, they provide useful and overall information. . 1, 4, 20, 56, 58, 59, 76, 79

ecological bias refers to a wrong assumption that the characteristics in a aggregated sample dataset is the same as that of the population. It is also refered to as "ecologial fallacy" and "aggregation bias". . 15, 120

generalizability refers to "generalization" concept in epidemiology domain, which indicates the extent to which the results of a study on a sample can be extended to the population. In order to prevent the confusion of this term with "generalization" in k-anonymity domain, we used this term as "generalizability" throught the manuscript.. 1, 6, 9, 11, 18, 55, 56, 59, 74, 77, 83, 86, 88, 123, 125

generalization refers to "generalization" concept in k-anonymity domain, which indicates representing the values of quasi-identifiers with less detail in order to make the corresponding records to count at least k times based on their quasi-identifiers. Reducing the amount of details of quasi-identiferis lead to an inevitable reduction of information content of the changed records.. 5, 6, 10, 16–24, 50–52, 55, 57, 59, 61–64, 67–70, 72, 74, 75, 78–80, 96, 97, 101, 103, 110–112, 115, 116

information loss , in the context of our study, refers to the reduction of attribute details, or complete exclusion of records and attributes, dur to the application of privacy methods. It is also refered to as the "reduction of information content" throgout this manuscript. . 2, 6–11, 21, 23, 27, 33, 43, 49, 55, 60, 64, 69, 76, 78, 79, 81, 88, 89, 100, 109, 126, 127

outlier , in the context of our study, refers to records in dataset that contain quasi-identifiers with cardinalities less than k, hence, break the k-anonymity constraint. In our study we refer to such records as k-outliers.. 8, 10, 19, 20, 33, 34, 43, 48, 124

person-specific data refer to datasets where each row or records belongs to a real or unreal person. They are also reffered to as individual data or micro-data. . 1, 3–5, 7, 8, 14–16, 18, 27–29, 33, 34, 56, 57, 59, 79, 81, 82, 88, 89, 107–109, 113, 123, 125

representability , as defined in our study, refers to the extend of resemblence of a record in anonymized data and sample data. High levels of privacy protection measures applied on datasets make them less representable of the original datasets. . 24, 59, 63, 116

Representability Vector ,abbreviated as RV, is a set of elements that reflect the resemblence of anonymized dataset to sample dataset.. 8, 9, 11, 40, 55, 60, 76, 104, 119, 121, 124, 127

secondary health data refers to health related data that are used for non-clinical purposes such as public health, epidemiology, disease surveillance, quality assessment, and marketing. . 2–4, 7, 8, 13, 14, 25, 28, 30, 35, 39, 82, 123, 125–127

secondary uses refers to uses of helath data for non-clinical purpose, such as public health and epidemiology, where the primary aim is not to provide direct healthcare to individuals. The results of secondary uses of health data may eventually affect the individuals by their influence on clinical methods and public health services.. 1–3, 7, 8, 13–15, 21, 26–28, 33, 79, 123, 125

selection bias refers to a bias caused by choosing the wrong individuals for a study. Examples of causes for selection bias in privacy protection methods are suppression at record level and aggregation of independent variables. . 22, 77

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Alizadeh Mizani, Mehrdad
Nationality: Iran (IR)
Date and Place of Birth: 21.03.1974, Urmia
Marital Status: Single

EDUCATION

Degree	**Institution**	**Year of Graduation**
M.S.	Medical Informatics, METU	2006
B.S.	Computer Engineering, Isfahan Uni.	1998
High School	Dehkhoda,Urmia	1993

PUBLICATIONS

Journal papers:

M. A. Mizani, N. Baykal, *A software platform to analyse the ethical issues of electronic patient privacy policy: the S3P example*, Journal of Medical Ethics, 2007 December; 33(12):695-698.

M. A. Mizani, N. Baykal, *Privacy preserving policy making for disclosure of public health data: a suggested framework*, Journal of Medical Ethics, (Under final revision for grammatical and language corrections)

Conferences:

T. Temizel Taskaya. M. A. Mizani, T. Inkaya, S. C. Yucebas, *The Effect of Data Set Characteristics on the Choices of Clustering Validity Index Type*, 22nd International Symposium on Computer and Information Sciences (ISCIS07), 2007, Ankara.

L. E. Akman, T. Beyan, O. D. Kocgil, M. A. Mizani, *Continuity of Care Records: A New Healthcare Standard*, Proceeding for the 2nd symposium of the Turkish Medical Informatics Association, 2005, Belek.

N. Baykal, M. A. Mizna, *Ethical Issues of Designing Security and Privacy Policies in Healthcare*, 2nd International Conference on Teaching Applied and Professional Ethics in Higher Education. Real World Real People: Ethics in a Virtual World. 2005, London.

Posters:

M. A. Mizani, N. Baykal, *An Assessment of the Implementation of Mobile Solutions at a State Hospital in Turkey*, American Medical Informatics Association Annual Symposium 2009, San Francisco.

Abstracts:

M. A. Mizani, M. O. Cinar, S. C. Yucebas, *Application of SOM based Clustering on Aggregated Epidemiological Data Based on Nomenclature of Units for Territorial Statistics (NUTS)*, Proceeding for the 8th symposium of the Turkish Medical Informatics Association, 2011, Antalya.

Books:

A. Arikan, M. A. Mizani, Y. A. Dilek, *Number One - Vocational Medical English*, 2010, ISBN: 9786055930943, Data Yayinlari - Ders Kitaplari, Ankara.

Languages:

English, Persian, Turkish, Azerbaijani