OULUN YLIOPISTO
UNIVERSITY of OULU

OULU BUSINESS SCHOOL

Simo-Pekka Kiihamäki

A LATENT CLASS APPROACH: CHARACTERIZING THE WILLINGNESS TO SHARE PERSONAL HEALTH INFORMATION IN FINLAND

Master's Thesis
Marketing
August 2019

UNIVERSITY OF OULU
Oulu Business School

ABSTRACT OF THE MASTER'S THESIS

Unit
Department of Marketing

Author Simo-Pekka Kiihamäki	Supervisor Juntunen, J.

Title
A Latent Class Approach: Characterizing the Willingness to Share Personal Health Information in Finland

Subject	Type of the degree	Time of publication	Number of pages
Marketing	Master’s thesis	2019	50 + 2

Abstract

BACKGROUND: With the fast advances in technology, the aging populations, and the climate change, the amount of data in our hands has become enormous, and the ways of handling it has become better. There has been large amount of privacy concerns as well due to the fast-growing data that are spread everywhere. This study focuses on health data to find out whether personal characteristics can be associated with the willingness to consent it for secondary purposes.

METHODS: A sample data (n=2338) concerning the Finnish populations attitudes towards secondary uses of health data was acquired and analyzed. The questionnaire included 14 questions regarding the willingness to consent data for different purposes. The dimensionality of this issue was reduced with a latent class analysis, and the information was condensed into one latent variable with 5 classes. After that a latent class regression was performed to find out whether the willingness could be explained with the help of other background information.

RESULTS: A statistically significant association between the willingness to consent health data and the following characteristics; Gender, Age, Education, Perception of health, Number of visits to health or social care, and Financial situation. Political orientation had a high value of estimate, but no significance.

CONCLUSIONS: Secondary uses of health data can achieve improvements in public health and welfare and health equality. Therefore, it is important that we make sure that the privacy concerns of using and sharing health data are taken care of. Methods for increasing the citizens willingness to consent their health data could be done through education and by building mutual trust between the health care system and the patients.

Keywords
health data, consent, latent class analysis, public health, marketing

Additional information

CONTENTS

TABLES

FIGURES

1 INTRODUCTION

This study aims to characterize what, and how, certain personal characteristics affect the willingness to share personal health information for secondary purposes. Personal health information can mean, for example, a) data collected by medical personnel at healthcare centers, or b) data collected by the patient by using health tracking devices, for example. This study focuses mainly on the aforementioned; data collected by authorities and medical personnel.

The European Data Protection Supervisor (2018) is an independent authority, that has the power to oversee the General Data Protection Regulation (GDPR), which was introduced on May 25th 2018. In *Recital 35* they provide the following definition for health data:

> "Personal data concerning health should include all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the data subject. This includes information about the natural person collected in the course of the registration for, or the provision of, health care services as referred to in Directive 2011/24/EU of the European Parliament and of the Council to that natural person; a number, symbol or particular assigned to a natural person to uniquely identify the natural person for health purposes; information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples; and any information on, for example, a disease, disability, disease risk, medical history, clinical treatment or the physiological or biomedical state of the data subject independent of its source, for example from a physician or other health professional, a hospital, a medical device or an in vitro diagnostic test."

Secondary uses of health data cover a wide range of applications, which means that some purposes are more likely to be accepted by societies than others are. Therefore, it is both interesting and useful for policy makers to understand where the societies set their boundaries regarding their personal data.

Due to the improvements in computer and data sciences, it has become increasingly efficient to crunch large amounts of data. Maybe the most important applications for health data are addressing conditions causing high morbidity and mortality in populations (Ballantyne & Schaefer 2018), and evaluating how well a regional or national healthcare system is working. Evidence based science can lead to improvements in healthcare and the quality of life, nationally and globally. We will take a more in depth look on the different use purposes later in the chapter 2.

The problems regarding the use of health data are many, however. The data is scattered across multiple entities and is costly to gather and transfer (Kruse et al. 2016). By nature, health data is also very personal and hence it needs to be handled accordingly. This leads to a requirement of regulations on who can use it and for what purposes. Inefficiently protected data can cause a large variety of issues for those whom the data applies to. For example, personal health information can be used maliciously if the subjects can be identified from it. With modern machine learning concepts, it can also be possible to identify subjects even from unidentified data sets given that there is a large enough amount of data linked together.

The main research question that this study aims to answer is; *"How personal characteristics affect the willingness to share personal health information for secondary purposes?"*

It has been shown, that certain factors affect the willingness to participate in health data sharing and consenting to record linkage (Huang et al. 2007, Kim et al. 2017). However, even the direction of the factor estimates can vary between different populations. With methods, like regression analysis, these estimates can be assessed on a population level given a sufficient amount of data. Understanding the phenomenon can help, for example, to create better national and international regulations regarding the use of health data.

The empirical analysis presented later in the thesis (chapter 3) provides insight to the main research question in a rather similar manner to what Kim et al. (2017) have done in their study. However, due to the nature of the data, the approach in this study differs slightly from the logistic regression used there. In this study, the response variable is generated by Latent Class Analysis (LCA), which can be used to combine multiple questions into a latent variable.

LCA is a clustering method that allows classifying the observations in mutually exclusive classes based on responses to multiple categorical questions (Collins & Lanza 2010). The latent class type variable can be used as a response in Latent Class Regression (LCR) by adding the predictor variables into the model. The LCR method is rather similar to logistic regression, but it should provide estimates that are more robust when the response variable is generated via LCA (Linzer, D., Lewis,J. 2011). These methods will be discussed more in depth in chapter 3.1.2.

2 THEORETICAL BACKGROUND

This chapter presents a literature review regarding the benefits and challenges related to the secondary use of personal health information. Firstly, it is necessary to understand what the secondary use applications of health data are, and what benefits it may help to achieve. The data itself can be applied to many issues, but not all of them are ethical or increase the public wellbeing.

The second part discusses the challenges related to the secondary use of the information. The main issues are related to privacy of the data in hand. Issues may arise if the data is identifiable, or if the methods used to transfer and store the data are insufficient (Scott et al. 2017). Also, it needs to be noted, that some restrictions on data linkage are required, so that one user cannot collect and link all available data of a population (Holman 2001, Xafis 2015).

Another interesting question, which will be covered in the third part, is whether a consent is required from the patients the data applies to. Asking consent is considered ethical, and the public often feels, that they should be in control on who has access to their personal information. However, getting consent from each individual can be very difficult and costly. This can also lead to situations, where certain groups do not provide enough information, which staggers the medical care development in these conditions (Ballantyne & Schaefer 2018). For example, requiring a consent from patients suffering from a rare illness, that makes them unable to provide a consent, can lead to a situation where progress in research cannot be made, thus denying the future generation a potential cure.

The fourth part addresses the ethical side of using health data. Certain use purposes might not be as ethical as others, or not ethical at all, for example. Another interesting point of view is the ethics related to sharing the data. It has been argued, that there is an ethical duty for the public to share their health information in order to contribute to the general wellbeing (Ballantyne & Schaefer 2018).

The fifth part covers the current regulations regarding the use of health data in Finland. To assess whether something should be changed, or reworked, we first have to know what the current status quo is. The content of this chapter is based on a) Ministry of Social Affairs and Health of Finland (2019) regarding the national level regulations, and b) the GDPR (European Data Protection Supervisor 2018) regarding the international regulations that affect all countries in the European Union.

2.1 Secondary Applications of Health Data

As stated earlier, there are wide range of applications where health data could be used for the public benefit (Canaway et al. 2019). The public considers some of the purposes more acceptable than others. Likewise, they consider some actors more trustworthy to handle the data. Other things to consider could be, whether certain types of data are necessary to a specific study. Access to data could be restricted to different subsets depending on the use purpose.

A study by Bietz et al. (2016) surveyed researchers, to find out whether they would find use for high quality personal health data. The researchers seemed to agree quite heavily, that a rather unique and accurate data could be gathered from the patients themselves using some sort of health tracking equipment. However, this type of data can currently be very difficult and expensive to obtain. This further highlights, that at least the researchers themselves believe that they could potentially make scientific breakthroughs if access to such data was provided.

2.1.1 Assessment of conditions causing mortality and morbidity on a population level

One of the most prevalent secondary use purposes for health data is assessing conditions that cause high mortality and morbidity on populations. Epidemiological studies have been a major part in building welfare globally (Cutler & Miller 2004). Already in the ancient Rome, sanitary removal of human waste, and the access to clean water became public health concerns. The issues were tackled by some of the most sophisticated feats of engineering of that time; the aqueducts were built to deliver clean water from sources located far from the cities, and the sewers were built to drain the dirty excess water from the streets.

The use of data and more sophisticated statistical methods for the benefit of public health started decades later, however. John Snow, who is considered as the founding father of epidemiology, used his prowess in statistics to find the source of a raging cholera epidemic in 1854 in London. Effectively, due to his contribution, the epidemic could be contained (Rothman 2012).

Research is still needed in our era to find and explore effect-exposure relationships that are harmful for the public health. However, in these days the conditions can be very subtle and therefore difficult to notice and explain, thus requiring vast amounts data. In addition, the constantly developing methods and computers can help us assess phenomenon that previously could not be analyzed.

For example, the research of temperature related excess mortality has been under a lot of interest lately (Gasparrini et al. 2015, Guo et al. 2018, Ryti et al. 2016). While the phenomenon and the results are interesting, it is a rather crude way of assessing the health risks. Other health outcomes could be measured in a similar fashion if the researchers had access to other types of health measures. These could be, for example, daily numbers of hospitalizations due to strokes, or asthma.

Understanding how our actions, and the changes in the world affects our health and lives can lead to better foundations in policy making. Not only can we find out ways to reduce harmful actions, but we also gain more tools to fight against climate change and to preserve the diversity of life on the planet. Often, exposures that are harmful to us, humans, are also harmful to other forms of life.

2.1.2 Evaluating and enhancing the healthcare system

A healthcare system is never perfect. Some flaws, and processes to improve on, can always be found. Also, as the methods and equipment in healthcare develops, and the population distribution evolves, the whole healthcare systems requires constant evaluation and enhancement. The population base that requires the health services is growing due to aging populations in developed countries, and therefore the services need to be constantly developed so that the capacity and resources can sustain the growing number of customers.

One of the biggest issues in terms of evaluating the healthcare system is, that it is often done by the providers themselves. Thus, the evaluation can often have a high bias towards financial indicators, that do not tell much about the actual outcome; how well have they managed to treat the patients (Naranjo-Gil et al. 2016). While the financial indicators are obviously important in order to finance the system in the first place, it is also highly important to find out the best practices for reaching the desired health outcomes (Scott et al. 2017). Especially from the point of view of the public sector this reduces other costs caused by illnesses and improves wellbeing of the citizens.

Demartini and Trucco (2017) discussed the development of the healthcare *Performance Measurement Systems* (PMSs). According to the study, the most effective PMS, currently, seems to be a rather complex pathway model, which requires longitudinal data from the customers journey. This, however, should not be much of an issue to handle with the technologies today. The main issue is gathering and storing the data in a unified form that an automated system can crunch into an evaluation.

The most important question, regarding the subject of this study, is whether it is enough that the healthcare provider alone has the access and opportunity to evaluate the performance. By having an option for third parties, for example academic researchers, to access the data, the healthcare systems could be benchmarked and compared against each other to find out even more information. This way, the best practices could be identified more easily and be adopted elsewhere. Also, third party evaluations could give the customers more knowledge about where they should seek help for their illnesses, which in turn might create more pressure for both public and private healthcare centers to improve on the quality of their services.

2.1.3 Development of treatments and medicine

The main function of the healthcare system is to improve and sustain health of the citizens, and to provide cures for illnesses that they might face. This not only directly improves the overall wellbeing, but when done effectively, it also cuts other costs in the society (Bhattarai et al. 2016, Jones et al. 2019). Like reduces the amount of sick leaves and early retirements, thus improving the collective longevity in employment, increasing productivity of the society as whole (Rashidian et al. 2017).

Development of the treatments and medicine is highly reliant on accurate data. For certain types of morbidities more than for others, but a certain amount of testing and confirmation on whether, and how, the treatment is working is always necessary. Having some data readily available, can further improve the effectiveness of the medicine, and cut down development time and costs (Costeloe et al. 2018). This can benefit all the stakeholders, by reducing both the costs of development and the prices for the customers, and thus improve the overall wellbeing in the society (Warren 2016).

2.1.4 Personal exposure and treatment

Personal exposure can be a rather tricky subject to study. One example of such assessment could be to estimate, to how much air pollution a person has been exposed to. For an accurate prediction, the researcher would need location information of the patient and very high-resolution data of the air pollution concentrations. Neither of these can be acquired from the patient's health records per se. However, many electronic devices such as phones, or health trackers have the potential to collect this data, which could then be linked with the health records (Banerjee et al. 2018). And with the help of analytics a) to find out whether certain exposures have any health effects on a population basis, or b) to warn the patients, that due to their condition they should avoid certain places at certain times.

Smartphone collected health data could also be used in the primary healthcare. This could lead into better and more efficient diagnoses (N. Chen et al. 2010). The main issue still remains, it can be difficult for the healthcare centers to collect such data. Currently the health tech companies, that manufacture the devices and design the software hold the data. And due to rising privacy concerns of private companies holding sensitive information about their customers, the regulations regarding the use and distribution of such data has been increased. The GDRP, for example, was published and established between 2016 and 2018 (Yuan & Li 2019).

To collect this type of data, the primary healthcare could possibly develop and publish their own application, that the consumers could use if they want to. However, this would require incentivizing the customers to actually download and use the applications. This could provide to be a difficult task when considering the primary health care's core competencies and that the competition on the market is already miles ahead.

2.2 Challenges of Sharing and Storing Health Data

Sharing and storing is likely the one of the biggest concerns for health data. Due to its nature, it is considered very sensitive and private. Thus, we would not want anyone to be able to access it without permission to see what is happening within our lives and bodies. Therefore, when handling such sensitive information, it is rather important to be able to protect the data at every stage of the process. This can provide to be a difficult challenge, because the more the data is distributed the more the likelihood of a leakage increases. Even if the original storage unit is secured well enough that no one can break into it, who knows how the secondary user treats and protects it. Thus, it is important to acknowledge the challenges and build up certain rules and regulations on how the data must be stored and used if a use permission is granted.

2.2.1 Collecting and transferring data

Data collection per se is already a rather trivial task. It is already done at health care, academies, and by all kinds of other organizations. There are also sufficient ways of ensuring the privacy of the collection. The main issue, however, rises when the data needs to be accessible by multiple parties. Transferring data does not only cause additional privacy concerns, but it can also be expensive (Kruse et al. 2016). Also, it should be always evaluated whether a certain set of data should be anonymized, and if so, at what stage (Canaway et al. 2019). Related to this, Lee & Gostin (2009) has made an important notion by pointing out, that the data collected, or transferred, for the secondary users should be kept at *a minimum that is required for the task* in order to provide more security for those whom the data may concern. Malin et al. (2018), for example, suggest that for some cases it could be enough to provide only aggregates of the original data.

A couple different methods can be used to transfer the data between the parties. It can be done via cloud computing so that the parties that have gained an access to the data, could retrieve at any time. The other options are various methods of manually sending the data to a party that has made an inquiry and gotten the permission for using the data. For example, transferring them via internet tools, like e-mail, or handing a physical drive with the data stored in it. The latter methods, however, may require more work and are often slower. They also contain risks of the data being stolen or lost along the way (Evans 2016).

If the users are many, and if the dataset is constantly updated, the cloud strategy becomes increasingly more attractive. A cloud-based system gives the users the freedom and flexibility to update and access the data whenever it suits them (Kruse et al. 2016). However, the main issue of the cloud is that the data is basically stored online, which can make it easier for outsiders to access the data as well (D. Chen & Zhao 2012). Thus, the main issue would be to secure the cloud well enough by using encryption or any other suitable method.

Due to the nature of the issue, the cloud computing system would likely be preferable. Access to the data is often already needed in multiple places, and thus already stored and transferred via cloud. For example, different health care centers sharing the patient data between each other. Also, since the secondary use purposes are many, as are the secondary users, it would be highly beneficial for all the parties involved to cut down extra work and costs related to manually transferring the data.

2.2.2 Storing the data

Data can be stored in various different ways, depending on the type and format of the data. For example, questionnaires or medical records might be physically stored on pieces of paper. Usually, the most reliable piece of data is the original one, since any errors can happen when, for example, a physical record is saved into an electronic format. (Meingast et al. 2006)

When storing the data, its physical state must be taken into account. Paper files need to be stored in a way, that outsiders are not able to access them, and that they remain as safe from accidents, such as fires or water leaks, as possible. Nowadays, however, most of the files are either stored electronically from the get-go, or they are copied into an electronic format afterwards (Evans 2016). Electronic files also need a physical storage, such as a hard drive or a server. A good practice is to also make backups so that the data cannot be lost due to an accident or theft.

Due to the large amounts of data that a health care sector can collect and produce, and due to the need to access the data in multiple locations, combining big data and cloud solutions seems enticing. Big data solutions require, however, a good foundation of data preprocessing, since these databases need to be as simple and user-friendly as possible (Scott et al. 2017). And raw data from health records can often be highly unstructured, such as from natural language processing (Kruse et al. 2016). The big data solutions often go hand in hand with the cloud technology, which is used to access the data anywhere in real-time (Jee & Kim 2013).

One method to increase the security and give the subjects more control over their own data could include using the blockchain method. Its efficiency has been proved with the use of cryptocurrencies, and other business solutions (Zyskind & Nathan 2015). The technicalities of blockchain technologies, as well as other encryption methods, and their use in the health care sector are out of the scope of this study. However, due to blockchain technology's recent popularity and buzz it deserves a mention as a potential solution.

In addition to privately held data, there are also open access databases comprised of health data. These services require both the subjects and the data owner's consent. Due to the fact, that consent can be difficult to obtain, the datasets may have varying levels of reliability to represent the population. Also, the datasets may not be homogenous between countries. (Riso et al. 2017)

2.2.3 Access and limitations

There is one main reason to limit the access to health data for some, while allowing access for others. If the data consists of sensitive information, and the subjects can be identified from it, it is a privacy hazard, and thus the access should be limited (Souhami 2006). This type of data is often required, when there is a need to link multiple databases together (Anderson 2015). In many cases, however, this issue can be circumvented by removing any identification tags, aggregating the data, and/or providing only such information that a person or an algorithm cannot identify anyone from it (Canaway et al. 2019, Lee & Gostin 2009).

When assessing the worth of secondary use of personal information, we must think about both the risks and rewards. Just like when investing on the stock markets, we would like to maximize the value and minimize the risks. Value, in this case, could be seen as the positive social contributions or financial gains for the party using the data. The risks, however, are directed towards those who are the subjects of the data, as in, the citizens (Riso et al. 2017). This means that the risks and rewards are in an imbalance in a way where one party might reap the rewards while the citizens bear the risks. To alleviate the risks and even the playing field we need a rulebook that obligates the users of the data to treat it with respect and make sure that a certain standard of privacy is met (Anderson 2015, Porsdam Mann et al. 2016).

To maximize the gains, we would provide access to the data for everyone. To minimize the risks, however, we would limit the access from everyone. Anything in between is a compromise between these two options. With smart regulation and good judgment we can likely find a middle road, where the access to the data is given only to those, who can justify their cause and make good out of the data with only a small risk (Anderson 2015).

2.3 Consent Requirements for the Use of Health Data

The pros and cons of a consent requirement provide us an interesting confrontation. The pros mainly benefit the subjects of the data, by providing them protection and control. The cons, however, are directed towards the society and the parties that may gain benefits from using the data. It is to be noted, though, that the detriments faced by the society indirectly affect the subjects as well.

A consent requirement for using personal information of citizens can provide protection for them. If all the data in the world could be used freely, then we would all be completely exposed. If no one could use the data, we would have way less understanding about the world and human behavior. Therefore, a middle road between these is a great compromise, where we can achieve the benefits from both while sacrificing very little (Porsdam Mann et al. 2016). However, it is important to decide what types of data require a consent, and from whom. The citizens, for example, might not have a good understanding of the risks or the use purpose of the data. On the other side, institutions might not understand the feelings of the subjects of whom the data concerns.

2.3.1 Benefits of a consent requirement

A consent requirement offers protection for the citizens against abusive use of the data. It also protects their privacy, since some information might be so sensitive, that they would not want anyone, or someone, to know about it. Additionally, a consent requirement gives the citizens some control over who they are willing to concede the data for.

Asking the subjects consent improves their trust on the project and those who are in charge of it. This trust, however, is not far-reaching without additional transparency (Kaplan 2016). A lack of transparency may also lead to a withdrawal of the consent afterwards. Singleton & Wadsworth (2006) mention that giving the participants a real choice is more important than just obtaining a consent.

For the subjects it is important to understand what their data is used for and how. This is further highlighted by some studies (Bietz et al. 2016, Kim et al. 2015) that show differences in the willingness to consent data for different purposes. A consent requirement makes sure, that the subjects have control over their own data.

When talking about the secondary uses of health data, it is important to understand to what purpose the consent was given. Anderson (2015) argues, that it must be taken into consideration that any future use cases for the data need to be clarified as well in the original consent request, or otherwise a new consent must be sought.

In certain studies, where the data needs to be identifiable, the subject’s privacy risk rises. This is heavily related to so-called record linkage studies, where multiple different datasets need to be linked in order to study the phenomenon (Huang et al. 2007). When a subject can be identified from the data, a person with an access to it can easily retrieve sensitive information about any specific subject who is participating the study. For many, this can feel uncomfortable even if there are no suspects of malicious intents.

2.3.2 Detriments of a consent requirement

In a sense, the public suffers from providing consent requirements for the protection of citizens. Getting a consent from subjects can be a very difficult and expensive task. Especially, when the requirement for data is large, and/or the condition in question is rare. Effectively, the costs of seeking consent limits the possibilities for research (Singleton & Wadsworth 2006).

Another issue that arises from the need to get a consent is selection bias (Tu et al. 2004). Certain subgroups in a population can be less willing to provide a consent, or in a state where they are not capable of doing so. This may lead to health inequality between the sub populations (Ballantyne & Schaefer 2018). This means, that it can be nigh impossible to obtain enough of relevant data about certain types of conditions, for example. Which in turn means, that the development of healthcare and treatment towards those conditions may become stagnant.

In the benefits it was mentioned, that the secondary use purposes should, for the sake of the subjects, be foreseen when requesting the consent in the first place, or otherwise a new consent should be sought when starting a spin-off using the same data (Anderson 2015). This can be exceedingly difficult in many cases. Hence, it can cause serious issues for the user; the subjects may be unavailable, or difficult to contact, or they could even have passed away by the time. These issues severely hinder the opportunities of new studies.

2.4 Ethical Considerations

The word e*thics* is derived from a Greek word *ethos* which means “character”. In common speak, when we talk about something being ethical or unethical, we refer to it being right or wrong, or good or bad (Thiroux & Krasemann 1980). Most of our laws and regulations are based on philosophy and judgment calls – is an action ethical or not.

In the context of this study, we need to consider both the ethics regarding the users of the data, and the ethics regarding the subjects of the data. Only by understanding the moral limits, duties and obligations of all the parties can we find an optimal solution. Optimal solution being the equilibrium point where the regulations and freedom of action satisfies, if not everyone, at least the majority of parties involved.

From the point of view of the users, it is important to understand what are the ethical boundaries when using the data. What are the responsibilities of an organization that holds the data, or access to it? How far must they go to ensure the privacy of the subjects? (Porsdam Mann et al. 2016) On the other side are the responsibilities of the citizens. Can they use services, with a pure conscience, that have been made possible through science and the generosity of the previous generations, if they do not participate in contributing? (Ballantyne & Schaefer 2018) In modern welfare states the public services are provided through taxes. However, the importance and value of data can be difficult to measure.

2.4.1 Ethics concerning the user of the data

Organizations can have many different motivations for working in responsible or ethical ways (Schaltegger & Burritt 2018). These motivations can range from a genuine altruism and the want to make the world better, to a want to polish the image of the organization or a fear of facing sanctions. Regardless of the underlying motivations, being responsible can usually be considered a virtue.

What does ethics mean in the context of using health data? Ballantyne (2018) approaches the issue by dismantling the ethicality in to seven dimensions; referred to as ethical values. Social value, Harm minimization, Control, Justice, Trustworthiness, Transparency, and Accountability. These values represent both the ambitions that an organization should aim to achieve as well as the foundation for the legislation regarding the use of health data.

Social value defines the reason why the health data should be used, and why its use for beneficial purposes should be encouraged. Harm minimization refers to the responsibility of the controller to make sure that no harm will come to the subjects due to leaks or misuse of the data. The control aspects hold the rights of the subjects. For the subjects it can be important that they can decide whether to consent, or to withdraw the consent in case they dislike the way their data is being treated. Justice refers to the fair distribution of the benefits gained from the data. Trustworthiness and transparency define the controller's obligations towards the subjects. The controller of the data has to provide reasonable justification and be transparent about how the data is being used. Accountability means, that the user of the data needs to be accountable for their actions. If they make a mistake or a malicious act, they need to take responsibility or be brought to responsibility by the law. (Ballantyne 2018)

Another ethical issue arises when considering the analysis of the health data. It is a common fallacy to assume that outcomes produced via machine learning would be unbiased. In fact, it is rather easy to get biased results, either by mistake or on purpose. After all, the algorithms are made by humans and are thus suspect to errors and biases. Also, the results reflect the data that the scientist feeds the computer with. It is very important, that the scientists are equipped well enough to detect any bias issues within the data, or else we are teaching the machines to do the same mistakes as we, humans, do. (Char et al. 2018)

2.4.2 Ethical duty of citizens

In certain countries there are rules that obligate the citizens to participate in *easy rescue*. This means, that if a person is walking by someone who is, for example, about to drown in a pond they must help the person in immediate danger if it does not cause serious harm, trouble, or danger to the one helping. In a similar manner, contributing to public welfare by donating data can even save lives while causing minimal trouble for the citizen. (Porsdam Mann et al. 2016)

The concept of an ethical duty of citizens, in this context, is derived from the type of logic presented above. The citizens are enjoying the fruits of science by using the public health care and other government provided services and infrastructures. Therefore, contributing to further and speed up the progress of science can be seen as a moral duty of the citizens (Ballantyne & Schaefer 2018). In this case, though, it is assumed that there are no, or very little, drawbacks from it for the citizens. Which comes back to the regulatory organs and oversight to make sure that the parties in control of the data treat it with respect.

As mentioned in the chapter discussing the detriments of a consent requirement, the insufficient amount and imbalances in the data can lead to health inequality. Ballantyne & Schaefer (2018) claim, that it is the ethical duty of the citizens to provide this data, in order to make sure that all populations are represented. This way the science and the society can provide equally for everyone.

2.5 Regulations for the Use of Patient Health Information in Finland

The last chapter of the theory section goes through how the use of patient health information is regulated in Finland. It is seemingly the most important part of it, at least when considering the political or managerial implications of the results. How could we assess, improve, or make conclusion without understanding the current?

2.5.1 National regulations

The first part of this section is mostly comprised of information collected from the official website of Ministry of Social Affairs and Health of Finland (2019) and the Act on the Status and Rights of Patients (785/1992).

> “Health care professionals shall record in patient documents the information necessary for the arranging, planning, providing and monitoring of care and treatment for a patient.”

The first part declares, that the health care professionals have a right to write and store personal health information.

> For independent practitioners: “Patient documents, samples and models shall be disposed of immediately after there are no grounds as referred to above for keeping them.”

In this part it is made clear, that privately-operated health care units are not allowed to store the data indefinitely. Likely due to the lack of tools for the government to supervise it. However, this should be enough for the independent practitioners to manage with the needs of the patients.

> “Health care professionals or other persons working in a health care unit or carrying out its tasks shall not give information contained by patient documents to outsiders without a written consent by the patient. If a patient is not capable of assessing the significance of the consent, information may be given by his/her legal representative’s written consent.”

The law also forbids giving any data for outsiders. This means, that legally the doctors and the health care units are held responsible for the privacy of patient health records.

The previous part is, however, notwithstanding in the following situations:

1) The information is required for another purpose on the basis of some law
2) If the information is required in another healthcare unit and they can obtain a consent from the patient
3) If the patient is deemed unable to evaluate the implications, or completely unable to give a consent due to, for example, being unconscious, the patients documents can be given to another health care unit without the patients consent
4) In the previously mentioned conditions, the health information can be also given to family or to another person close to the patient if deemed necessary
5) The health information of a deceased person may be given to anyone with a well justified written application

These laws are meant to help the patients. There may come times when the patient has to use a different health care unit, and thus the regulation allows sharing the data in these situations.

The most interesting part for the context of this study is in the last part. For the sake of public good, there is also a section, that allows giving out health records for research. It says:

> “The National Institute for Health and Welfare may, in individual cases, grant permission to obtain information that is needed for purposes of scientific research from patient documents of more than one municipality or joint municipal board providing health and medical care services.”
>
> “The permission may be granted if it is obvious that the supplying of the information does not violate the interests for the protection of which the secrecy obligation has been prescribed.”

As we can see, there is an entity in Finland, The National Institute for Health and Welfare, that is overseeing the use of medical records for research. The law itself, however, does not define accurately to what type of research the data can be given. Thus, it seems that the National Institute for Health and Welfare has a great deal of power in making the decisions to grant access to the data, or to deny it. For the application, the researcher must provide a research plan, solid justifications on why the data is required and why is the research important (Terveyden ja Hyvinvoinnin Laitos (THL) 2019). More accurate descriptions of the application process and the forms required, as well as the available datasets, can be found from the institution’s webpage.

2.5.2 International regulations

The GDPR regulates and defines the rights to use and process personal health data even further (European Data Protection Supervisor 2018). GDPR contains three recitals that directly address health data, and numerous mentions in other recitals. First of them, *Recital 35*, contains the definition for health data, which can be found from the introduction of this study. *Recital 53* is about *Processing of sensitive data in health and social sector*, which refers to the primary use of health data. *Recital 54 - Processing of sensitive data in public health sector*, however, is of a high importance in the context of this study:

> "The processing of special categories of personal data may be necessary for reasons of public interest in the areas of public health without consent of the data subject. Such processing should be subject to suitable and specific measures so as to protect the rights and freedoms of natural persons. In that context, 'public health' should be interpreted as defined in Regulation (EC) No 1338/2008 of the European Parliament and of the Council (11), namely all elements related to health, namely health status, including morbidity and disability, the determinants having an effect on that health status, health care needs, resources allocated to health care, the provision of, and universal access to, health care as well as health care expenditure and financing, and the causes of mortality. Such processing of data concerning health for reasons of public interest should not result in personal data being processed for other purposes by third parties such as employers or insurance and banking companies."

Additionally, the GDPR (2018) also defines general ground rules that apply to all types of data, including health data. For example, they define consent and apply certain ground rules about how a consent agreement should be presented and when is it required. The main points being, that:

1) The request of contest shall be presented in a clearly distinguishable manner and,
2) It should clearly state what the subject is consenting to
3) The subject has the right to withdraw the consent at any time
4) The controller has to be able to prove the consent at any time

3 EMPIRICAL ANALYSIS

This chapter presents the empirical analysis and findings related to the topic. This empirical study explores what factors affect the willingness to share personal information for secondary purposes. The first part discusses the methods used in the study. The second part covers the results and findings from the analysis. In the third part, we discuss the findings and their implications.

3.1 Methods

3.1.1 Data

The study is based on a dataset: “Terveys- ja hyvinvointitietojen toissijainen hyödyntäminen 2016 (Secondary use of health and well-being data 2016)” (Sitra & Hyry 2016) which can be acquired from the Aila services by Yhteiskuntatieteellinen tietoarkisto. The data contains a questionnaire regarding the subjects views on the secondary use of health and well-being data, and their sentiments on different types of actors potentially using the data, as well as what types of data they feel are the most private. The data also includes a variety of background variables that explain the characteristics of the population. The data collection itself was outsourced to TNS Gallup Oy, and due to their methods of randomizing and contacting the participants, it can be argued that the data represents the whole population of Finland relatively well. It should be noted, that the questionnaire was originally in Finnish, and all the translations to English has been done by the author of this study.

Table 1 presents the characteristics of the study population and the distributions for the most relevant background variables in this study. Some deviations from the actual Finnish population can be seen, like the small number of participants in the age group of 15-19 and the amount of women being quite high compared to men. However, this small deviation should not have much impact on the results.

Table 1. Characteristics of the Study Population

		N	%
Total		2338	100.0
Gender			
	Female	1315	56.2
	Male	1023	43.8
Age			
	15-19	72	3.1
	20-29	309	13.2
	30-39	293	12.5
	40-49	465	19.9
	50-59	418	17.9
	60-69	498	21.3
	70-79	281	12.0
Education			
	Elementary school	215	9.2
	Vocational school	483	20.7
	High school	282	12.1
	College level vocational	448	19.2
	Undergraduate degree	506	21.6
	Graduate degree	392	16.8
	Other	12	0.5
Perception of current health			
	Good	584	25
	Quite good	905	38.7
	Moderate	606	25.9
	Quite bad	181	7.7
	Bad	54	2.3
	Cannot tell	8	0.3
Have you used social or health care services within last 12 months?			
	No	291	12.4
	1-2 times	908	38.8
	3-6 times	699	29.9
	More than 6 times	408	17.5
	Cannot tell	32	1.4
Financial situation			
	Very good	200	8.6
	Quite good	825	35.3
	Get along	799	34.2
	Quite bad	288	12.3
	Bad	188	8.0
	Cannot tell	38	1.6
Political orientation			
	Left	212	9.1
	Somewhat left	521	22.3
	Cannot tell	661	28.3
	Somewhat right	638	27.3
	Right	306	13.1

Some of the questions had answer options like ‘Other’, ‘I don’t know’, or ‘Cannot tell’. These answer options did not fit the ordinal scale in many circumstances; those observations were handled as missing information to keep the ordinal scale intact. Fortunately, the amount of these types of answers was low and thus the number of missing observations in the analysis is almost non-existent. Hence, the exclusion method should not have much impact on the analysis. The number of observations (N) used in the final model is 2293 out of the 2338 participants.

3.1.2 Statistical Methods

The main methods applied in this study are Latent Class Analysis and Latent Class Regression. These methods allow the clustering of categorical data to create a latent variable, which can have one to n (number of observations in the data) number of classes. For the clustering step, unsupervised machine learning algorithm like K-modes and hierarchical clustering could have been used. These methods do not, however, save the uncertainty in a similar manner as LCA. Due to this, the LCA was deemed as the best fit for this study.

An interesting question when performing any type of clustering is; how many classes should I choose? Many types of methods have been developed to estimate the optimal number of latent classes. However, a more subjective approach was chosen for this study, since it is necessary to be able to categorize and explain the characteristics of the latent classes; otherwise, the results cannot be interpreted. After all, the main objective of this study is not finding the best fitting or most parsimonious model, but to assess what factors affect the willingness to share data (Linzer, D., Lewis,J. 2011). The Goodness-of-fit test-scores are shown in the appendix.

After the clustering process, the following predictors were added to the model one-by-one: gender (nominal categorical), age (continuous numerical), education (ordinal categorical), perception of current health (ordinal categorical), use of health services (ordinal categorical), financial situation (ordinal categorical), and political orientation (ordinal categorical). After examining the unadjusted results, the covariates were added simultaneously into the model. This forms the final model, which is referred to as the “Main Effects Model” later on.

All data analyses were conducted using R statistical software (version 3.5.0). The R package poLCA was used for the LCA and LCR analysis.

3.2 Results

The questions of main interest in the data were “Would you allow the use of your anonymous health and social care data for the following purposes?” addressing eight different institution, and “Would you allow the use of your anonymous genetics data for the following purposes?” which addresses six different institutions. **Figure 1** and **Figure 2** show the distributions for all the individual questions, and it can be seen that the distributions are rather similar. However, insurance companies were considered less trustworthy than the other institutions. The most likely reason is that the insurance companies are perceived as for-profit companies, thus their interests might contradict with the interest of the patients that the data applies to. This finding falls in line with, for example, a study by Bietz et al. (2016) where it was also noted that the respondents had more aversion towards commercial use of their data when compared to research use.

The LCA clustering was applied on all the 14 questions. The idea behind using LCA in this study is to get an overall view of each person’s willingness to concede their health data for secondary purposes. Another option could have been to apply logistic regression on each of the different items, which would give an idea how the participants views changed regarding each item. However, in this study the focus was more on the general attitudes towards conceding data, thus the LCA was applied. This not only reduces the dimensions of the data, but also nets a more robust outcome variable, since a single question might suffer from more random variance than a large set of them. Also, the LCA clustering seemed to provide rather good and interpretable 5-class variable, which has the potential to convey more accurate information than the original 4-class scale of a single question.

Figure 1. Distributions for Questions 17.1 – 17.8. Would You Allow the Use of Your Anonymous Health and Social Care Data for the Following Purposes?

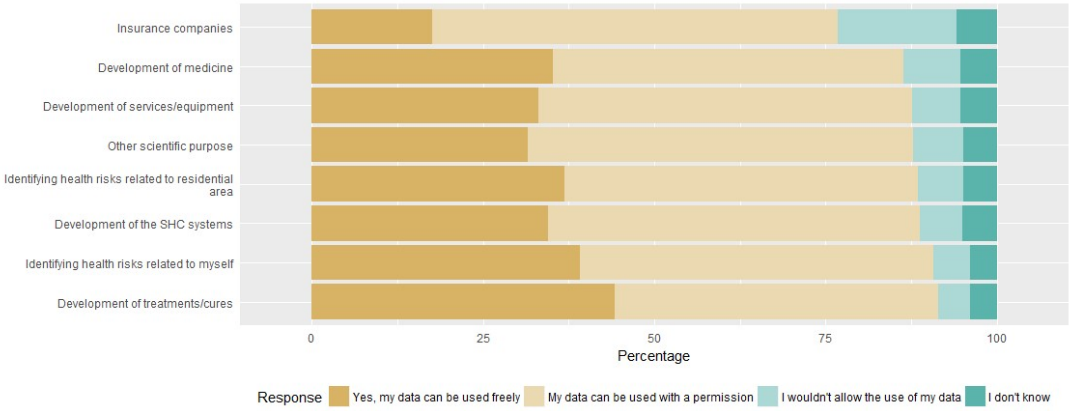
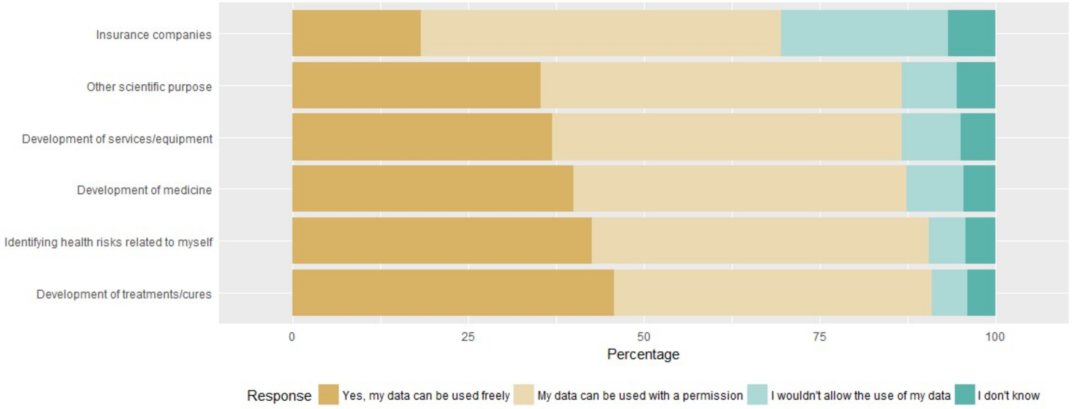


Figure 2. Distributions for Questions 18.1 – 18.6. Would You Allow the Use of Your Anonymous Genetics Data for the Following Purposes?



3.2.1 The Classification

Below, in **Figure 3**, is a visual presentation of the *Posterior Item Response Probabilities* (PIRP), where the classes have been reordered into a logical order. This visualization provides a good idea of how the classes are distributed, and what answers affected the membership of each class. Additional A.Table 1 can be found in the appendix. It tells the same information as **Figure 3**, but it has an edge when performing deductions that are more exact regarding single items and response.

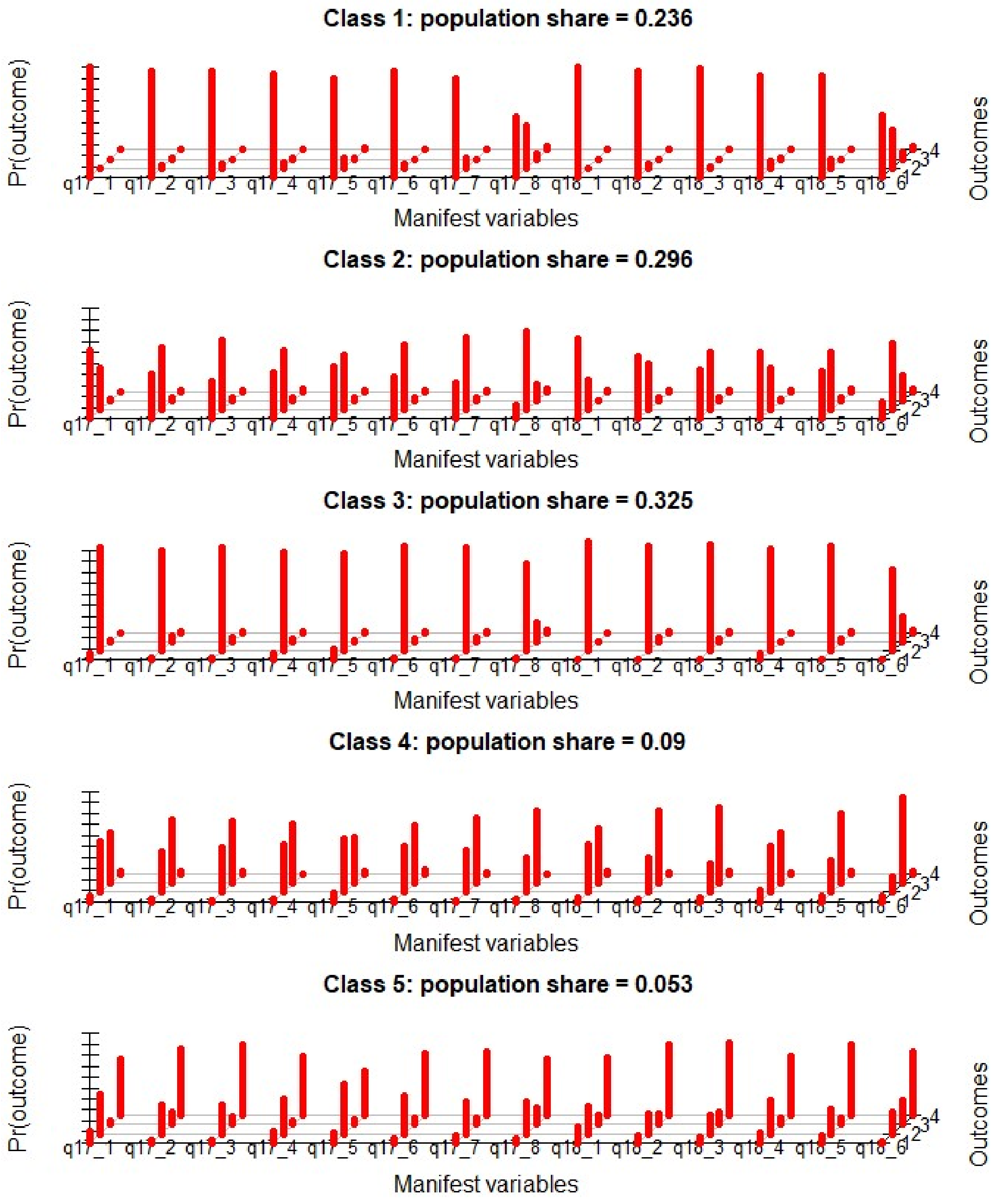
How should the PIRP's be interpreted? The A.Table 1 in the appendix is filled with probabilities (ranging 0-1), that tell us how the clustering decided to allocate the observations into the latent classes. If we take a look at the second column and first question (17_1), it has values of 1.00, 0.00, 0.00, and 0.00. This means, that for an observation to be classified into the Liberal class, the response for the question had to be 1 (Yes, my data can be used freely), since the probability for option 1 is 1.00 and the rest are 0.00. However, even if the respondent answered 1, it does not automatically mean that they were placed in that class. In fact, no confirmation can be made from a single item, the allocation to a class is a combination of all the answers, and each answer has its own probability for tilting the final allocation to a particular class.

The classes have been labeled as (where the percentage is indicating the population share of the class):

1. Liberal (23.6%)
2. Somewhat Liberal (29.6%)
3. With Permission (32.5%)
4. Conservative (9.0%)
5. Uncertain (5.3%)

So, how were the classifications defined? The visualization of the PIRP made assessing the content of the classification rather simple. The membership of the Liberal class requires that the respondent has answered mainly the option number one (Yes, my data can be used freely).

Figure 3. Ordered Classes from Latent Class Analysis



Classes: 1 - Liberal, 2 - Somewhat Liberal, 3 - With Permission, 4 - Conservative, 5 - Uncertain.

The Somewhat Liberal class consists of a mixture of response options one and two (My data can be used with permission). Therefore, this class consists of respondents who believe that for some purposes it is fine to have no consent requirements for using their health data, while some purposes it should be required. The third class, called as With Permission, consists of respondents who felt that when using health information for almost any purpose, a consent should be sought first. The conservative class is also a mixture, this time of response options two and three (I would not allow the use of my data). Finally, the Uncertain class consists mostly of responses to option four (I don't know). In this case, it is safe to say that this class consists of respondents that have not formed an opinion about the subject.

3.2.2 Assessing the Factors Affecting the Willingness to Share Health Data

Next step in the analysis was to assess the relationship between the latent classes describing the willingness to share health data, and the respondents' characteristics. A linear LCR was performed to calculate estimates for the influence of these factors on the outcome measure. The predictor variables used for the regression analysis are the same ones that were presented in **Table 1**, and below in **Table 2**.

Table 2. Spearman rank correlation matrix for predictor variables

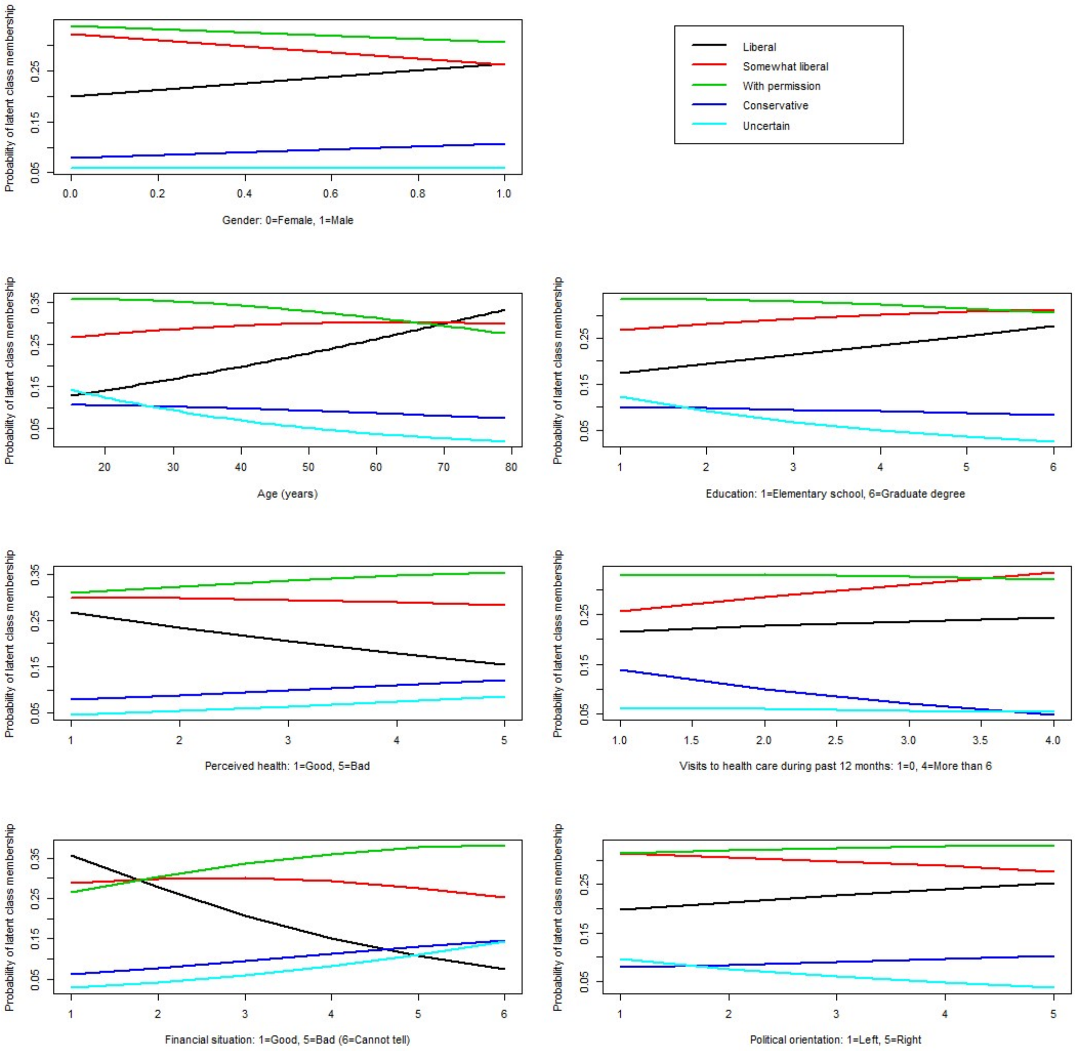
Variables	Gender	Age	Education	Perceived health	Visits to healthcare	Political orientation	Financial situation
Gender	1.00						
Age	0.22	1.00					
Education	0.00	0.07	1.00				
Perceived Health	0.07	0.14	-0.17	1.00			
Visits to Healthcare	-0.14	0.07	-0.01	0.24	1.00		
Political orientation	0.14	0.11	0.12	-0.09	-0.07	1.00	
Financial Situation	0.16	0.21	0.34	0.17	-0.04	0.19	1.00

Table 2 presents the Spearman correlation coefficient for the predictor variables. This is done to avoid collinearity in the final model, since it can cause errors with the accuracy of the estimates (Mason & Perreault Jr 1991). In this data the highest correlation coefficient was found between Education and Financial situation, which should not be too surprising. None of the coefficients are very high, and thus we accept all the variables in the analysis.

The first step in the actual regression analysis was to assess the crude effects (also known as unadjusted), where each variable was added as a lone predictor in the model. **Figure 4** displays the results of the regression model. Since these figures do not present the final outcome, a decision was made to leave out the odds ratio estimates and instead present them graphically as probabilities of latent class membership.

The first graph in the figure describes how gender affects the willingness to share personal health data. In the unadjusted model it looks like the female population is more likely to require a consent to allow the use of their health records for secondary purposes, while the males have a slightly higher chance of belonging to the liberal class.

Figure 4. Probabilities of Latent Class Membership. Crude Models.



In the second graph we can see how age affects the outcome variable. Since here we have a continuous variable, instead of dichotomous, as the predictor, the curves can bend into non-linear shapes, making it slightly more difficult to interpret. The most notable thing is, that the older people are a lot more likely to belong in the Liberal class than the younger ones. Another interesting thing to note is, that the younger ones are a lot more likely to belong in the Uncertain class. It seems that the younger population could be considered more conservative regarding this topic, since they are most likely to belong in the With permission class, while for the oldest part of the population it is only the third most likely class to belong to.

The probabilities for education look highly similar to those of age. Those with a higher level of education seem to be more likely to share their personal health data, than those with less education. Also, the uncertainty of the subject matter seems to fall when the level of education rises.

Possibly the most interesting results can be found in the fourth graph, where the outcome was predicted with the perceived health of the subject. Those that felt less healthy are more likely to require permission for the secondary use of their health data. In a way it is surprising and counter-intuitive for the less healthy ones to be more strict about sharing their data, since it could be used to help out finding cures and treatments for their conditions. However, it is also understandable that they feel that the data is more sensitive or personal due to a) likely having provided larger amount of data due to visiting healthcare more often, and b) feeling that the data reveals more about them.

The fifth one of the graphs displays visits to healthcare during past 12 months as the predictor variable. The direction of the results is surprisingly different to those of perceived health. Those subjects that have visited healthcare more seem to also be more likely share their health information. Not only is the likelihood of belonging to the liberal class rising when visiting healthcare more, but basically none of those who have visited healthcare more than 6 times during past 12 months belong to the conservative class.

Financial situation, in the sixth graph, seems to have a large effect on the outcome. Those respondents who felt that their financial situation is good were also most likely to belong in the liberal, while those who had it bad or could not tell were very unlikely to belong the liberal class. A good financial situation seems to also decrease the likelihood of belonging to the Uncertain class.

The very last graph in the figure shows political orientation as the predictor variable. All the curves seem to remain rather constant throughout the spectrum. The right-hand side being slightly more tilted towards the liberal class, and the left-hand side being more uncertain.

Table 3. Latent Class Regression: Adjusted Main Effects Model. Demographic Factors Predicting Class Membership. Class "Liberal" as the Reference.

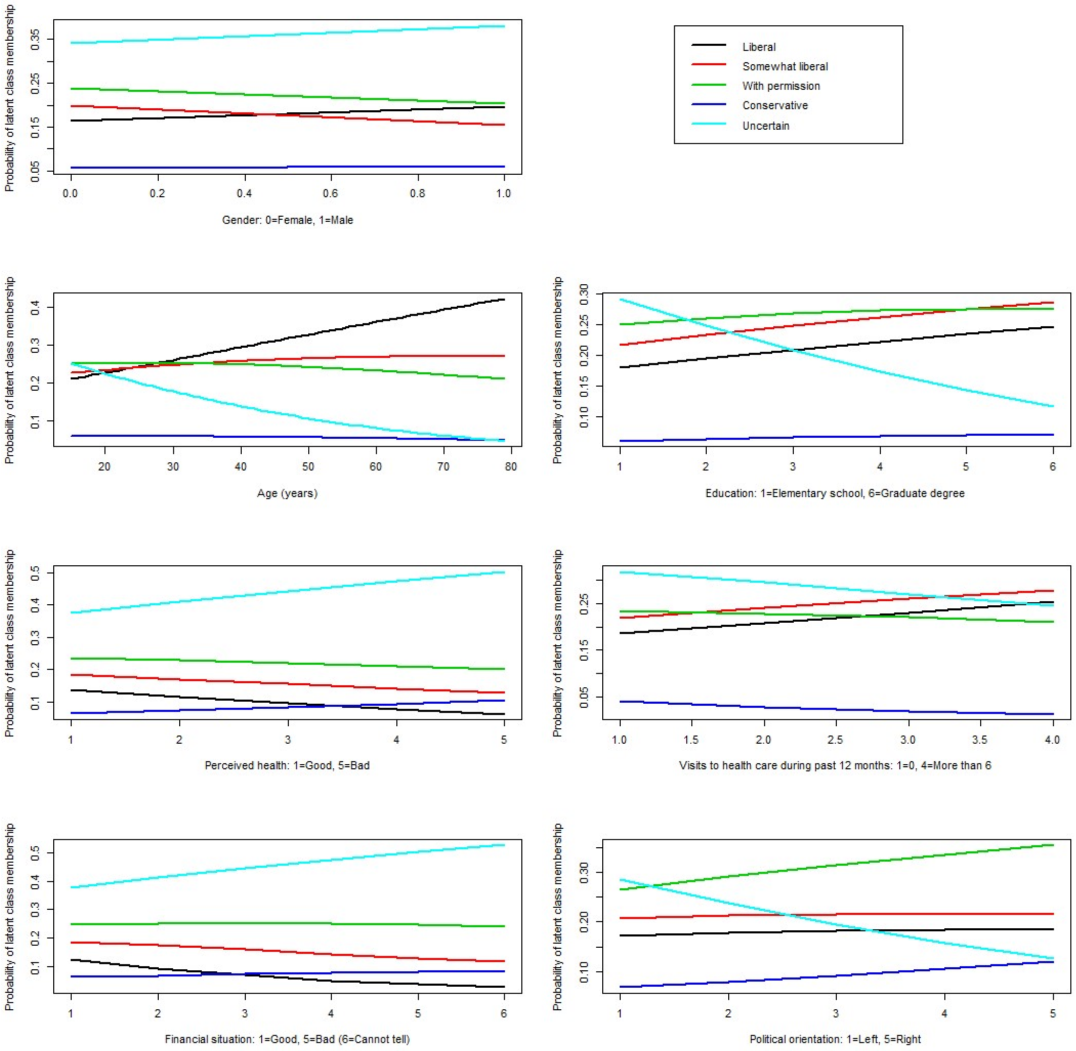
	Somewhat liberal Odds ratio (95% CI)	With permission Odds ratio (95% CI)	Conservative Odds ratio (95% CI)	Uncertain Odds ratio (95% CI)
Intercept	1.21 (0.53 – 2.77)	1.46 (0.67 – 3.16)	0.34 (0.09 – 1.30)	2.08 (0.52 – 8.34)
Gender (ref=female)	0.66*** (0.50 – 0.85)	0.71*** (0.56 – 0.92)	0.92 (0.63 – 1.36)	0.93 (0.56 – 1.57)
Age	0.99** (0.98 – 1.00)	0.99*** (0.98 – 0.99)	0.99** (0.97 – 1.00)	0.96*** (0.95 – 0.98)
Education	0.99 (0.92 – 1.08)	0.96 (0.89 – 1.04)	0.97 (0.86 – 1.10)	0.78*** (0.67 – 0.91)
Perceived health	1.11 (0.96 – 1.27)	1.17** (1.02 – 1.34)	1.37*** (1.12 – 1.67)	1.31* (0.99 – 1.73)
Visits to health or social care	0.98 (0.84 – 1.13)	0.87** (0.75 – 1.00)	0.62*** (0.50 – 0.76)	0.83 (0.64 – 1.06)
Financial situation	1.24*** (1.08 – 1.41)	1.35*** (1.20 – 1.53)	1.45*** (1.20 – 1.74)	1.45*** (1.14 – 1.85)
Political orientation (1=Left wing, 5= Right wing)	0.99 (0.89 – 1.10)	1.06 (0.95 – 1.17)	1.13 (0.96 – 1.34)	0.80* (0.63 – 1.02)

The asterisks in the table denote P-values: *** < 0.01, ** < 0.05, * < 0.1

Table 2 presents the effect estimates for the adjusted Main Effects Model. One important thing to note in the table is, that many of the odds ratio estimates are statistically significant (level of significance denoted by the asterisks). The effects are also shown in **Figure 5** in a similar manner as the crude models were presented earlier. The graphical representation in this case, however, is not ideal since the probability curves shown are affected by the model holding each of the covariate's constant at a value of 0. Therefore, interpreting the results from the graphs can be misleading. After that being said, once understanding the dynamic of the representation, the graphs can help to understand the direction and the magnitude of the effect, since the multinomial regression estimates can be difficult to interpret. Also, the graph can show relationships between all of the classes, while the table only displays a comparison with the reference class, which in this case is the Liberal class.

The LCR produces a similar output as logistic regression and should be interpreted in similar fashion. In these analyses the Liberal class was treated as the reference. The columns of **Table 2** represent the different classes, and the rows represent the predictor variables. First, we will take a look at the estimate for Gender in Somewhat liberal class (OR=0.66). Since the estimate is lower than 1 and highly significant (P<0.01) it suggests that men are less likely to belong to the Somewhat liberal class or the With permission class than the Liberal class when compared to women. From **Figure 5** we can also see, that men also seem more likely to belong in the Uncertain class.

Figure 5. Visualization of the Odds Ratios from Table 3. Displayed as Probabilities of Latent Class Membership.



Both age and education, according to **Figure 5**, reduce the likelihood of belonging to the Uncertain class. **Table 2** shows, that age is statistically significant determinant when comparing the Liberal class to any of the other classes.

Just like in the unadjusted model, the perception of one's health has interesting results. The respondents, who felt that their health is good, are more likely to belong in the Liberal class. When the Perceived Health gets worse, the likelihood of belonging to any other class increases. For With Permission and Conservative groups the P-value goes below 0.05, which is considered as statistically significant. However, the confidence intervals for the Somewhat Liberal and Uncertain classes are just barely over the 1.0, so a real effect is likely to be present in there as well. Interestingly enough, the Visits to Healthcare variable seems to provide completely opposite results than the Perceived Health.

Judging from the amount of asterisk's in the next row of the **Table 4**, Financial Situation seems have good predictive power for the outcome. It is to be noted, that income is somewhat correlated with at least two other predictors; age and education. The results suggest, that the respondents with a better financial situation are more likely to belong in the Liberal class, than those with a worse financial situation.

The last predictor variable is Political Orientation. In the table, the only estimate with a P-value close to statistical significance is the estimate for the odds of belonging to the Uncertain class when compared to Liberal class. This suggests, that the respondents whose political orientation is more towards the Left wing are more likely to be uncertain about the subject. When looking at the corresponding graph, it seems that the likelihood of belonging to Liberal or Somewhat liberal class stays almost the same while moving through the political spectrum. However, when moving from Left wing to Right wing, the likelihood of belonging to the Uncertain class changes into an increased likelihood of belonging to the Conservative and With permission classes.

It seems, that the estimates for individual predictor variables stay rather similar when adding other predictors in the models. This is a good sign when assessing the robustness of the model. Also, the unadjusted models can provide some help for us in interpreting the individual results. Overall the final model seems to fit well since all the variables have some statistically significant power behind them.

3.3 Discussion

At this point we can conclude, that certain personal characteristics undisputedly affect the willingness to consent personal health data. For the sake of public health, welfare, and health equality, we should try to engage as many citizens as possible for sharing data. Especially those subjects that represent minorities or other underrepresented sub populations in our current data. With the help of these results we might be able to identify those populations that are less likely to consent their data and focus our actions of promoting the cause towards them.

However, before promoting the consent of data to anyone, we should make sure that the privacy concerns are dealt with. If a health data outbreak were to happen, it would be a catastrophe. The citizens would lose much of their privacy, and the health care system would lose its trust. Therefore, the government needs to take responsibility in regulating the use and collection of the data, as well as give guidelines and provide information on how to. The health care units and other actors who are authorized with the use of the data must also be willing to oblige the rules and appropriate sanctions need to be in place in case of any misuse.

Not many studies have explored the characterization of willingness to consent health data. In fact, only two such studies were found in the literature review (Huang et al. 2007, Kim et al. 2017), and none related to the Finnish population. Multiple studies, however, have been conducted on population level about the willingness to consent without examining the characteristics of the respondents (Kim et al. 2015, Page et al. 2016, Patel et al. 2015, Riordan et al. 2015, Whiddett et al. 2016). Thus, comparisons between existing research knowledge are limited.

The research studying the Californian population by Kim et al. (2017) found a similar associations between both education, age, and the willingness to consent that was also present in this study. Their study, however, did not find any association between health status, and the association between financial situation (income) was reversed when compared to our results. Other interesting finding from the Californian study, that we did not measure, was the effects of ethnicity. They report that the ethnic minorities are less likely to consent their health data.

The Taiwanese study by Huang et al. (2007) found similar associations with financial situation (household income) and education. However, according to their results, a higher level of education had no effect on willingness to consent. Only the illiterate group had a meaningful estimate. Interestingly, in the Taiwanese population age had an opposite effect to those found from the Californian population and this study, and no difference between the genders. Also, they report that in Taiwan the minorities are more likely to consent their health data, which is the opposite of what was observed in California.

From the regression models we can see an interesting phenomenon. Those who use health care services more are also more likely to consent their data. It can seem surprising for two reasons. First, these respondents leave a larger trail of data in the system due to visiting the health care more often. It would be understandable if those patients that are the subjects of larger amounts of data would be more worried of it being leaked. On the other hand, it is likely that due to using the service more they have managed to build trust in the system. This trust might also affect the willingness to consent data; the patients believe that the health care professionals can also make good judgment on who they share the data with.

Second, this might seem to be in contradiction with the estimate for the perception of health; those respondents who felt that they are not healthy were also less likely to consent their data. However, there is not necessarily a contradiction between those two. It might simply imply that those who visit health care more often are actually healthier, because they are getting treatment on their current conditions and their yet undetected health issues can be diagnosed at an earlier stage.

Based on these two observations, it might be beneficial for all parties if the citizens were encouraged to seek medical care more often. Not only after they get sick, but also to visit for standard check-ins. This could prevent many health issues before they even become prevalent. For example, a nurse might notice a lift in the patient’s blood pressure and thus prevent a stroke that would have happened couple years later. Not only would it improve public health and introduce cost savings on the health sector (Rose et al. 2019), but it could also help to build trust between the health care system and the citizens.

Interestingly, almost all the factors used in the model had high estimates for uncertainty. In science, data analytics, and our everyday lives we try to minimize the uncertainty with a varying success. Uncertainty often springs from a lack of knowledge. Age and education both seem like the most obvious determinants for explaining uncertainty. Along with both age and education we gain more knowledge and are likely more willing to express it as a formed opinion. Financial situation also had, expectedly, a rather high correlation with education, and thus it is not surprising to see similar results from it. However, the other associations do not seem as obvious.

4 CONCLUSIONS

The development of medicine and technology has made data even more vital. It is no secret, that appropriate use of health data can benefit the public health and welfare. Also, due to recent developments the governments and citizens everywhere in the world have raised concerns about privacy, and the ethical use of health data. This has led to both national and international agreements being made about how data should be used and how do we ensure that the risks do not outweigh the benefits. While advances in science can be considered a priority in modern societies, the social costs need to be minimized in order to protect the citizens, and to achieve mutual trust between the parties.

This study has approached the issue by examining the willingness to consent health data for secondary uses in Finland. Majority of the population are willing to give a consent, and many of them would not be opposed if their data was used without a separate consent agreement. At least for specific purposes. From a scientific point of view, this can be considered good news, that the citizens are willing to contribute as well.

The main contribution of this study was the characterization of the respondents, which had not been done before on Finnish population. The results shown provide some insight on which people are most likely to consent their health data, and which are less likely. We could also see a distinct group which consisted of those who had not formed an opinion about the subject. By recognizing and identifying target groups, in this case those who are less likely to consent or who have not formed an opinion, we can target the promotion and education about the subject to them specifically.

The results also imply that via education we could guide the citizens, and especially the younger generations, into becoming more willing to consent their health data. Education obviously has other benefits, and thus it should be encouraged and funded by government anyways, but possibly education about science and the scientific methods could be a way to teach the population about why, and how, their contribution is needed.

Another key finding of this study is the trust between health care system and the patients. Those patients that visited health care more often were also more willing to consent their health data, which leads to conclusion, that the citizens should likely be encouraged to visit health care for regular checks.

Because the field of research still leaves many things unknown, it also leaves a lot of room for future research. This study, for example, could be improved by examining the clustered variables one by one as dependent variables in a logistic regression. Also, other clustering methods could be applied for the data to explore how those compare, or to find out whether more dimensions could be recognized and defined.

Furthermore, the dataset includes many variables that were left completely out of this study. These could be used to explore other types of relationships regarding the willingness to consent health data or some other outcome. For example, machine learning could be applied to find patterns from the data with unsupervised learning, or other relationships could be analyzed through supervised learning methods.

REFERENCES

Anderson, R. J. (2015). The collection, linking and use of data in biomedical research and health care: Ethical issues.

Ballantyne, A. (2018). Where is the human in the data? A guide to ethical data use. *GigaScience* 7(7), giy076.

Ballantyne, A. & Schaefer, G. O. (2018). Consent and the ethical duty to participate in health data research. *Journal of medical ethics* 44(6), 392-396.

Banerjee, S. (., Hemphill, T. & Longstreet, P. (2018). Wearable devices and healthcare: Data sharing and privacy. *The Information Society* 34(1), 49-57.

Bhattarai, N., McMeekin, P., Price, C. & Vale, L. (2016). Economic evaluations on centralisation of specialised healthcare services: A systematic review of methods. *BMJ Open* 6(5), e011214.

Bietz, M. J., Bloss, C. S., Calvert, S., Godino, J. G., Gregory, J., Claffey, M. P., Sheehan, J. & Patrick, K. (2016). Opportunities and challenges in the use of personal health data for health research. *Journal of the American Medical Informatics Association : JAMIA* 23 e42-e48.

Canaway, R., Boyle, D. I. R., Manski-Nankervis, J., Bell, J., Hocking, J. S., Clarke, K., Clark, M., Gunn, J. M. & Emery, J. D. (2019). Gathering data for decisions: Best practice use of primary care electronic records for research. *Medical Journal of Australia* 210 S12-S16.

Char, D. S., Shah, N. H. & Magnus, D. (2018). Implementing machine learning in health care - addressing ethical challenges. *The New England journal of medicine* 378(11), 981-983.

Chen, D., & Zhao, H. (2012). Data security and privacy protection issues in cloud computing. *2012 International Conference on Computer Science and Electronics Engineering, , 1* 647-651.

Chen, N., Rabb, M., Lee, Y. Y. & Schatz, B. (2010). Feasibility of long-term monitoring of everyday health through smartphones.

Collins, L. M. & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences.*John Wiley & Sons.

Costeloe, K., Turner, M. A., Padula, M. A., Shah, P. S., Modi, N., Soll, R., Haumont, D., Kusuda, S., GÃ¶pel, W., Chang, Y. S., Smith, P. B., Lui, K., Davis, J. M. & Hudson, L. D. (2018). Sharing data to accelerate medicine development and improve neonatal care: Data standards and harmonized definitions. *The Journal of pediatrics* 203 437-441.e1.

Cutler, D. M. & Miller, G. (2004). The role of public health improvements in health advances: The 20th century united states. *National Bureau of Economic Research Working Paper Series* No. 10511

Demartini, C. & Trucco, S. (2017). Are performance measurement systems useful? perceptions from health care. *BMC health services research* 17(1), 96; 96-96.

European Data Protection Supervisor (2018). *GDPR guidelines and full documentation.* Retrieved 04/18, 2019, from https://gdpr-info.eu/

Evans, R. (2016). Electronic health records: Then, now, and in the future. *Yearbook of medical informatics* 25(S 01), S48-S61.

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., RocklÃ¶v, J., Forsberg, B., Leone, M., De Sario, M., Bell, M. L., Guo, Y. L., Wu, C., Kan, H., Yi, S., de Sousa Zanotti, S. C., Saldiva, P. H. N., Honda, Y., Kim, H. & Armstrong, B. (2015). Mortality risk attributable to high and low ambient temperature: A multicountry observational study. *Lancet (London, England)* 386(9991), 369-375.

Guo, Y., Gasparrini, A., Li, S., Sera, F., Vicedo-Cabrera, A., de Sousa Zanotti, S. C., Saldiva, P. H. N., Lavigne, E., Tawatsupa, B., Punnasiri, K., Overcenco, A., Correa, P. M., Ortega, N. V., Kan, H., Osorio, S., Jaakkola, J. J. K., Ryti, N. R. I., Goodman, P. G., Zeka, A., Michelozzi, P., Scortichini, M., Hashizume, M., Honda, Y., Seposo, X., Kim, H., Tobias, A., Ã   Ã±iguez, C., Forsberg, B., Ã…strÃ¶m, D. O., Guo, Y. L., Chen, B., Zanobetti, A., Schwartz, J., Dang, T. N., Van, D. D., Bell, M. L., Armstrong, B., Ebi, K. L. & Tong, S. (2018). Quantifying excess deaths related to heatwaves under climate change scenarios: A multicountry time series modelling study. *PLOS Medicine* 15(7), e1002629.

Holman, C. D. (2001). The impracticable nature of consent for research use of linked administrative health records. *Australian and New Zealand Journal of Public Health* 25(5), 421-422.

Huang, N., Shih, S., Chang, H. & Chou, Y. (2007). Record linkage research and informed consent: Who consents? *BMC Health Services Research* 7(1), 18.

Jee, K. & Kim, G. (2013). Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system. *Healthcare informatics research* 19(2), 79-85.

Jones, C., Verstappen, S. M. M. & Payne, K. (2019). A systematic review of productivity in economic evaluations of workplace interventions: A need for reporting criteria? *Applied Health Economics and Health Policy*

Kaplan, B. (2016). How should health data be used?: Privacy, secondary use, and big data sales. *Cambridge Quarterly of Healthcare Ethics* 25(2), 312-329.

Kim, K. K., Joseph, J. G. & Ohno-Machado, L. (2015). Comparison of consumers' views on electronic data sharing for healthcare and research. *Journal of the American Medical Informatics Association* 22(4), 821-830.

Kim, K. K., Sankar, P., Wilson, M. D. & Haynes, S. C. (2017). Factors affecting willingness to share electronic health data among california consumers. *BMC Medical Ethics* 18(1), 25.

Kruse, C. S., Goswamy, R., Raval, Y. & Marawi, S. (2016). Challenges and opportunities of big data in health care: A systematic review. *JMIR medical informatics* 4(4), e38; e38-e38.

Lee, L. M. & Gostin, L. O. (2009). Ethical collection, storage, and use of public health data: A proposal for a national privacy protection. *Jama* 302(1), 82-84.

Linzer, D., Lewis,J. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software* 42(10),

Malin, B., Goodman, K. & Section Editors for the IMIA Yearbook,Special Section. (2018). Between access and privacy: Challenges in sharing health data. *Yearbook of medical informatics* 27(1), 55-59.

Mason, C. H. & Perreault Jr, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research* 28(3), 268-280.

Meingast, M., Roosta, T., & Sastry, S. (2006). Security and privacy issues with health care information technology. *2006 International Conference of the IEEE Engineering in Medicine and Biology Society,* 5453-5458.

Act on the Status and Rights of Patients 785.

Ministry of Social Affairs and Health (2019). *Client and patient rights.* Retrieved 04/20, 2019, from https://stm.fi/en/client-and-patient-rights

Naranjo-Gil, D., Sánchez-Expósito, M. & Gómez-Ruiz, L. (2016). Traditional vs. contemporary management control practices for developing public health policies. *International journal of environmental research and public health* 13(7), 713.

Page, S. A., Manhas, K. P. & Muruve, D. A. (2016). A survey of patient perspectives on the research use of health information and biospecimens. *BMC Medical Ethics* 17(1), 48.

Patel, V., Beckjord, E., Moser, R. P., Hughes, P. & Hesse, B. W. (2015). The role of health care experience and consumer information efficacy in shaping privacy and security perceptions of medical records: National consumer survey results. *JMIR Med Inform* 3(2), e14.

Porsdam Mann, S., Savulescu, J. & Sahakian, B. J. (2016). Facilitating the ethical use of health data for the benefit of society: Electronic health records, consent and the duty of easy rescue. *Philosophical transactions.Series A, Mathematical, physical, and engineering sciences* 374(2083), 20160130.

Rashidian, A., Arab, M. & Souri, A. (2017). Comparison the effects of poor health and low income on early retirement: A systematic review and meta-analysis.

Riordan, F., Papoutsi, C., Reed, J. E., Marston, C., Bell, D. & Majeed, A. (2015). Patient and public attitudes towards informed consent models and levels of awareness of electronic health records in the UK. *International journal of medical informatics* 84(4), 237-247.

Riso, B., Tupasela, A., Vears, D. F., Felzmann, H., Cockbain, J., Loi, M., Kongsholm, N. C. H., Zullo, S. & Rakic, V. (2017). Ethical sharing of health data in online platforms - which values should be considered? *Life sciences, society and policy* 13(1), 12; 12-12.

Rose, A. J., Timbie, J. W., Setodji, C., Friedberg, M. W., Malsberger, R. & Kahn, K. L. (2019). Primary care visit regularity and patient outcomes: An observational study. *Journal of General Internal Medicine* 34(1), 82-89.

Ryti, N. R. I., Guo, Y. & Jaakkola, J. J. K. (2016). Global association of cold spells and adverse health effects: A systematic review and meta-analysis. *Environmental health perspectives* 124(1), 12-22.

Schaltegger, S. & Burritt, R. (2018). Business cases and corporate engagement with sustainability: Differentiating ethical motivations. *Journal of Business Ethics* 147(2), 241-259.

Scott, P. J., Rigby, M., Ammenwerth, E., McNair, J. B., Georgiou, A., Hyppönen, H., de Keizer, N., Magrabi, F., Nykänen, P. & Gude, W. T. (2017). Evaluation considerations for secondary uses of clinical data: Principles for an evidence-based approach to policy and implementation of secondary analysis. *Yearbook of medical informatics* 26(01), 59-67.

Singleton, P. & Wadsworth, M. (2006). Consent for the use of personal medical data in research. *BMJ (Clinical research ed.)* 333(7561), 255-258.

Souhami, R. (2006). Governance of research that uses identifiable personal data. *BMJ (Clinical research ed.)* 333(7563), 315-316.

Suomen itsenäisyyden juhlarahasto Sitra & Hyry, Jaakko (TNS Gallup Oy): Terveysja hyvinvointitietojen toissijainen hyödyntäminen 2016 [sähköinen tietoaineisto]. Versio 1.0 (2017-08-04). Yhteiskuntatieteellinen tietoarkisto [jakaja]. http://urn.fi/ urn:nbn:fi:fsd:T-FSD3132

Terveyden ja Hyvinvoinnin Laitos (THL) (2019). *The national institute of health and welfare biobank.* Retrieved 04/20, 2019, from https://thl.fi/en/web/thl-biobank/main-page

Thiroux, J. P. & Krasemann, K. W. (1980). *Ethics: Theory and practice.*Glencoe Publishing Company.

Tu, J. V., Willison, D. J., Silver, F. L., Fang, J., Richards, J. A., Laupacis, A. & Kapral, M. K. (2004). Impracticability of informed consent in the registry of the canadian stroke network. *N Engl J Med* 350(14), 1414-1421.

Warren, E. (2016). Strengthening research through data sharing. *New England Journal of Medicine* 375(5), 401-403.

Whiddett, D., Hunter, I., McDonald, B., Norris, T. & Waldon, J. (2016). Consent and widespread access to personal health information for the delivery of care: A large scale telephone survey of consumers' attitudes using vignettes in new zealand. *BMJ Open* 6(8), e011640.

Xafis, V. (2015). The acceptability of conducting data linkage research without obtaining consent: Lay people's views and justifications. *BMC medical ethics* 16(1), 79; 79-79.

Yuan, B. & Li, J. (2019). The policy effect of the general data protection regulation (GDPR) on the digital public health sector in the european union: An empirical investigation. *International Journal of Environmental Research and Public Health* 16(6), 1070.

Zyskind, G., & Nathan, O. (2015). Decentralizing privacy: Using blockchain to protect personal data. *2015 IEEE Security and Privacy Workshops,* 180-184.

APPENDICES

A.Table 1. Posterior Item Response Probabilites

		Liberal	Somewhat liberal	With permission	Conservative	Uncertain
Class membership probabilities		0.24	0.29	0.33	0.09	0.05
Item response probabilities						
17_1						
	1	**1.00**	**0.62**	0.05	0.05	0.11
	2	0.00	**0.37**	**0.94**	**0.46**	**0.35**
	3	0.00	0.01	0.01	**0.46**	0.03
	4	0.00	0.00	0.00	0.03	**0.51**
17_2						
	1	**0.97**	**0.40**	0.02	0.02	0.03
	2	0.03	**0.56**	**0.91**	**0.37**	0.26
	3	0.00	0.03	0.06	**0.58**	0.11
	4	0.00	0.01	0.02	0.03	**0.61**
17_3						
	1	**0.97**	**0.34**	0.01	0.01	0.02
	2	0.03	**0.63**	**0.94**	**0.40**	0.27
	3	0.00	0.02	0.03	**0.56**	0.07
	4	0.00	0.01	0.01	0.02	**0.65**
17_4						
	1	**0.89**	**0.47**	0.10	0.08	0.09
	2	0.09	**0.49**	**0.88**	**0.48**	**0.44**
	3	0.00	0.02	0.02	**0.42**	0.05
	4	0.01	0.01	0.00	0.02	**0.41**
17_5						
	1	**0.89**	**0.47**	0.10	0.08	0.09
	2	0.09	**0.49**	**0.88**	**0.48**	**0.44**
	3	0.00	0.02	0.2	**0.42**	0.05
	4	0.01	0.01	0.00	0.02	**0.41**
17_6						
	1	**0.96**	**0.38**	0.01	0.02	0.04
	2	0.04	**0.58**	**0.95**	**0.43**	**0.34**
	3	0.00	0.02	0.02	**0.52**	0.05
	4	0.00	0.01	0.01	0.03	**0.56**
17_7						
	1	**0.90**	**0.32**	0.02	0.01	0.07
	2	0.09	**0.65**	**0.94**	**0.38**	0.29
	3	0.00	0.02	0.04	**0.59**	0.06
	4	0.01	0.01	0.01	0.02	**0.58**
17_8						
	1	**0.55**	0.12	0.01	0.02	0.04
	2	**0.38**	**0.71**	**0.79**	**0.31**	0.29
	3	0.04	0.14	0.18	**0.66**	0.16
	4	0.03	0.03	0.02	0.01	**0.51**

Table Continued

		Liberal	Somewhat liberal	With permission	Conservative	Uncertain
Item response probabilities						
18_1						
	1	**1.00**	**0.72**	0.01	0.04	0.14
	2	0.00	0.27	**0.99**	**0.44**	0.25
	3	0.00	0.00	0.00	**0.50**	0.08
	4	0.00	0.00	0.00	0.03	**0.53**
18_2						
	1	**0.97**	**0.57**	0.01	0.02	0.06
	2	0.03	**0.41**	**0.95**	**0.32**	0.18
	3	0.00	0.02	0.03	**0.66**	0.11
	4	0.00	0.00	0.01	0.01	**0.65**
18_3						
	1	**0.99**	**0.57**	0.01	0.02	0.06
	2	0.01	**0.41**	**0.95**	**0.32**	0.18
	3	0.00	0.02	0.03	**0.66**	0.11
	4	0.00	0.00	0.01	0.01	**0.65**
18_4						
	1	**0.92**	**0.61**	0.06	0.11	0.10
	2	0.07	**0.37**	**0.93**	**0.43**	**0.31**
	3	0.01	0.01	0.01	**0.45**	0.06
	4	0.01	0.01	0.00	0.01	**0.54**
18_5						
	1	**0.92**	**0.43**	0.01	0.05	0.07
	2	0.07	**0.52**	**0.95**	0.29	0.19
	3	0.00	0.02	0.03	**0.63**	0.21
	4	0.01	0.02	0.01	0.03	**0.58**
18_6						
	1	**0.56**	0.15	0.00	0.05	0.02
	2	**0.35**	**0.60**	**0.73**	0.14	0.19
	3	0.06	0.22	0.23	**0.78**	0.21
	4	0.03	0.03	0.03	0.02	**0.58**

A.Table 2. LCA Fit-Statistics

Classes	AIC	BIC	G^2	X^2
5	43 309.59	44 541.59	16 340.00	6 743 056 751.06
6	42 500.45	43 980.01	15 444.86	7 320 346 417.46
7	41 845.46	43 572.58	14 703.88	7 192 797 801.07
8	41 254.14	43 228.81	14 026.55	7 682 175 006.19
9	40 693.70	**42 915.92**	13 380.11	28 889 116 254.52
10	40 453.95	42 923.73	13 054.36	**6 602 563 464.04**
11	**40 257.60**	42 974.93	**12 772.01**	9 853 182 521.77
diff-%	7.6	3.8	27.9	2.1