Healthcare data-sharing from the perspective of a scientist
Ciprian Mihai Cornea

In partial fulfilment of the requirements for the degree of Master of Science in Business Information Systems

Supervisor:
dr. ir. J.J.M. (Jos) Trienekens

Examination committee:
dr. ir. J.J.M. (Jos) Trienekens, TU/e, IE&IS
dr. M. (Mykola) Pechenizkiy, TU/e, WSK&I
ir. M. (Marcin) Klecha, Philips Research

Eindhoven, February 2015
MANAGEMENT SUMMARY

This thesis provides an analysis of the data-sharing practices within a research and development organisation, from the point of view of a researcher that needs healthcare data from outside his organisation in order to validate a research hypothesis.

The thesis starts by setting a scientific foundation through structured literature review of various aspects related to data-sharing. The thesis describes the type of shared data, the stakeholders involved, their requirements, the data sharing process, what data quality is and what the bottlenecks are regarding the data-sharing process and regarding the data quality.

During this orientation phase a number of questions arose on how things are done in practice. Each of the resulted questions is based on investigated literature. These formed a questions-based model that was used for interviewing a number of researchers working in Philips Research and in a hospital regarding their experience with data-sharing. The interviewees were selected among those scientists working in projects using healthcare data. All interviews followed the same discussion structure based on the questions-based model.

All questions were answered by the interviewees, a fact that demonstrates the validity of the questions-based model and veridical of the topic's importance. This questions-based model is seen as valuable from an academic point of view. All interviews can be found in the appendix.

The interviews reveal that the legal aspects are the most frequent bottlenecks for partnership projects, while trust in partners and willingness of partners to share data, are determining the speed of reaching to the data, when working in a partnership.

For some Philips-only projects, data quality is the most prominent bottleneck, specifically data-format and data-consistency. For all projects, relevant data ranks as the first data quality.

Based on the analysis of these findings, the author proposes a data-sharing model to serve the mentioned use case. The model is a succession of steps and points to take into consideration at each step. Based on the consideration points a scientist might select one of the steps for accessing data. The model promotes internal data-sharing together with other means of accessing the data, while it indicates for each the correlated project type. As applicable for all, the specifics of the internal data sharing process are presented.

This model is validated through dedicated sessions with experts from the organisation. As such, the applicability of the model is recognized, its fitness with the organisation's current processes and its understandability. Making reuse of the internal data sharing, is seen by the experts as a relevant contribution. The model may be used by a scientist requiring any sort of data type, since the model is not data specific.

The limitations of the data-sharing model refer to the fact this model is the first proposal and there are improvements that can be made, also the graphical representation of the data access types characteristics has to be validated. Finally, the model contains parts specific to the Philips Research organisation but also generic parts that may be encountered also in other research organisations.

From the overall discussions with the experts, data-sharing is recognized as an interesting subject for Philips Research. The data-sharing model helps in familiarising a researcher with the organisation, increases the contextual information around the known facts and challenges around data-sharing. Given the rising costs of running experiments, using this data-sharing model is an appealing workaround for supporting the research needs of the organisation.
ACKNOWLEDGMENTS

I thank my family for their support throughout the master studies.
I thank my supervisors for their guidance, patience and understanding.
### Table of Contents

1. Introduction .................................................................................................................. 6
2. The research problem ................................................................................................... 8
   2.1. Problem statement .................................................................................................. 8
   2.2. Research approach ............................................................................................... 8
3. Orientation .................................................................................................................... 10
   3.1. Data sharing ......................................................................................................... 10
       3.1.1. The patient's position within the data sharing initiative .................................. 12
       3.1.2. The data ......................................................................................................... 13
       3.1.3. Costs of data sharing ..................................................................................... 13
   3.2. The data sharing process ...................................................................................... 14
   3.3. Data sharing stakeholders ..................................................................................... 15
   3.4. Bottlenecks in the data sharing process ................................................................. 16
       3.4.1. Problems faced by data-producers .................................................................. 16
       3.4.2. Problems faced by data-consumers ............................................................... 17
   3.5. Data quality in data sharing ................................................................................... 17
   3.6. Bottlenecks in data quality ................................................................................... 18
   3.7. The questions-based model ................................................................................... 18
4. Analysis ......................................................................................................................... 21
   4.1. Answers summary matrix ...................................................................................... 21
   4.2. Observations based on the interviews ................................................................... 24
5. Design ........................................................................................................................... 25
   5.1. The data-sharing model ......................................................................................... 25
   5.2. Consideration points .............................................................................................. 26
       5.2.1. Encourage internal data sharing ..................................................................... 26
       5.2.2. Data access types .......................................................................................... 26
       5.2.3. Project types .................................................................................................. 32
       5.2.4. Internal approval process ............................................................................. 35
6. Validation ....................................................................................................................... 36
   6.1. The overall model .................................................................................................. 37
       6.1.1. Internal data sharing ....................................................................................... 37
       6.1.2. Data access types .......................................................................................... 38
       6.1.3. Project types .................................................................................................. 38
       6.1.4. The internal approval process ........................................................................ 38
   6.2. Validation conclusions ............................................................................................ 38
7. Conclusions ................................................................................................................... 40
   7.1. Limitations and future work ................................................................................... 40
8. References ..................................................................................................................... 42
1. Introduction

In the healthcare world there are high volumes of data available with unused potential that could help improve the quality of care, ease the burden on the healthcare systems, empower patients and provide industry with much needed input for services, devices and drug improvement. In reality, a large portion of data remains unused being obstructed by security concerns, privacy aspects and legislation. And as such, the “Research world” has unmet input needs regarding data generated within the “Clinical world”.

The thesis starts by presenting the research problem and the motivation for this problem. Next, the thesis presents the phases of the research: orientation, analysis, design and validation, as indicated by Illustration 1.

In the orientation phase, a review is made of the available literature around data-sharing. Within this theme, the following aspects describing it are discussed: what the data-sharing is, what the data-sharing process is, who the stakeholders are and what the bottlenecks are. In addition, it presents the data quality attributes and bottlenecks mentioned in the literature.

Literature review was done on the topic of data-sharing in the healthcare domain to understand what the current situation is, what the scientific challenges are and what kind of work was already done.

Philips Research, a research and development organisation, is interested in the topic of data-sharing. As a result, the thesis investigates for this organisation the specific use case of how does a scientist get access to healthcare data external to the organisations through data-sharing.

All of the aspects mentioned above, were translated into a questions-based model. This model is used in understanding, characterizing how data-sharing works in a real-life project. The orientation phase is concluded by interviewing scientists, interviews during which the questions-based model is followed.

The interview results are analysed during the Analysis phase. In this phase, similarities and differences are sought in the interview responses in order to identify similarities and differences regarding the current way-of-working for data-sharing, but also the bottlenecks around the process and quality aspects that such a model should help to overcome. In other words, this phase it is a search for the requirements of a data-sharing model.

The interview answers are presented side-by-side for a facile response comparison.

During the following phase, the Design, a data-sharing model is proposed using the requirements gathered during the Analysis phase. The model is composed by a succession of steps and per step, the points to take into consideration. The steps represent choices a scientist can make for accessing data. Each choice is characterised on four coordinates in order to help a scientist make a decision. The model is meant to address the quest for any kind of data-type of a scientists regardless the fact it is live data, offline data, patient data or hospital information data.

Afterwards, the model is validated during the Validation phase, in which discussions took place with scientists in the organisation on aspects as usability, accessibility, understandability, relevance and fit with the organisation's way of working.

The thesis concludes by presenting the findings of the thesis and the limitations of research.
Illustration 1: The thesis development phases
2. The research problem

2.1. Problem statement

The healthcare world has unused, high potential, high volumes of data that could be put to good use if shared with interested parties.

In the healthcare domain, the healthcare organizations and patients aim for the improvement of healthcare quality and cost reduction by using information technology (Leitheiser, 2001), data sharing (Gøtzsche, 2011), (Wilhelm, Oster & Shoulson, 2014), (Ross & Krumholz, 2013) and by asking for data quality (Leitheiser, 2001), (Parker, Stofberg, De la Harpe, Venter & Wills, 2006).

The healthcare industry is interested to provide solutions as services or products to the domain's important topics. For best performance, these solutions must make use of reliable data input from the field.

This is a subject interesting for Philips Research, a research and development organisation that often runs projects where researchers are in need of data for the testing and validation of various research hypotheses.

For this reason the thesis conducts an investigation regarding the options and challenges that a researcher has when healthcare data is needed for verifying a research hypothesis or developing services or products. It is assumed that the researcher works for a research organisation active in the healthcare domain.

Thus, the thesis analyses the options the researchers of Philips Research have and the challenges they face for obtaining healthcare data from the field.

2.2. Research approach

The thesis starts with an orientation on the topic of sharing healthcare data through literature research. At the end of the orientation phase, a research question is formulated. This is answered through a number of interviews with experts in the field. Based on the interviews, a model is created. This model is validated with other experts. This is a summary of the thesis' development phases as depicted in Illustration 1.

The author defines through literature research what data sharing is, the patient's point of view, what kind of data is sharing about, the costs of data, but it also explains what the data sharing process is, it explains who the data sharing stakeholders are and what the bottlenecks in data sharing are.

The thesis continues by defining based on literature research what the data quality in data sharing is and what are the bottlenecks in data quality.

Out of the literature research, areas of unclarity were identified. In order to clarify them, a number of questions were formulated. Each of the questions cover a data-sharing characteristic discussed in literature. The collection of questions it is named the questions-based model.

This model is used to interview experts involved in projects in the healthcare domain. The purpose of the interviews is to learn what the practical issues are with respect to data-sharing.

The orientation phase concludes with the interview sessions.

The analysis phase starts by presenting the outcomes of the interviews.
The question-based model is used to landscape the current situation, with respect to the data sharing process and quality, in various research projects. The outcome of these interviews present the situation on various projects that are in need of data, and the challenges they are facing. For the readability of the thesis, the answers are presented side by side in an overview matrix, and for each interview the answers are presented based on the questions model.

Next, the answers are analysed for similarities and differences among the projects, and a number of observations are presented. And this concludes the analysis phase.

The observations were the starting point for defining a model that would help address the research question. The design phase starts by presenting the proposed model.

The model is a mix of a succession of steps and points to take into consideration. The steps lead a scientist down the path of accessing the data, first by taking advantage of the internal data sharing and afterwards by following already practised means of data access. At each step, its characteristics are listed and the most likely to encounter corresponding project type. The model also presents the internal approval process, mandatory for using external healthcare data.

This thesis is meant to serve the needs of a new researcher which searches healthcare data for testing a research hypothesis. The data can be of any type: live or offline data, patient data or hospital information data.

The model provided by the thesis is meant to guide the researcher on the steps to take to reach his goal. Along the road the researcher is made aware of certain points to take into consideration with respect to the specificity of data sharing within Philips Research. By specificity it is understood the types of data access and the types of projects running within the organisation.

The design phase is followed by the validation phase, that also concludes the thesis, phase that presents the validation of the model by a panel of experts of the same organisation. The model was validated for usability, applicability, understandability, relevance and fit with the organisation.

For the sake of clarity, the author defines as an expert, a researcher with experience in data-sharing projects in the healthcare domain.
3. Orientation

This section presents the theoretical foundation of the thesis. This foundation is based on literature research.

The literature review is realized by searching publications available via PubMed, Google Scholar, Thomson Reuters’ Web of Science and the EBSCO Discovery Service using the keywords: healthcare data quality, data quality aspects healthcare, {reliability, integrity, security} aspects, data quality problems healthcare, healthcare data quality important for research, data-sharing quality. Similarly, the following journals were inquired: the New England Journal of Medicine, Lancet, the Journal of the American Medical Association, the British Medical Journal and PLoS Medicine. These journals were selected by the author due to their presence in top 6 highest impact factor journals in the General Medicine section as published by Thomson Reuters (2015).

In this section, the thesis describes what data sharing is, what the patient’s position is related to sharing data, what kind of data is shared and what the costs of sharing data are. It continues with what the sharing process is, who the stakeholders of data-sharing are and if there are any bottlenecks during the data sharing process. In addition, this section enumerates the data quality attributes in data sharing and lists the known literature bottlenecks in data quality.

The section concludes with a structured list of questions, derived from literature, that is used in interviewing experts in order to discover the current situation in the organisation regarding the sharing of data.

The Orientation phase is followed by the Analysis phase that discusses the interviews' answers.

3.1. Data sharing

This section describes what data-sharing is, what it means for the patient – the one that data speaks about -, what kind of data is shared and the costs surrounding the data-sharing activities.

Data sharing in the healthcare domain means sharing healthcare data among different parties.

This thesis only considers the data sharing between Philips Research and external partners. Data sharing within the same organization is out-of-scope for this thesis. The reason is that once that data arrives in an organization, this can be shared everywhere within the organization.

Data sharing means that the data owned by a party is shared with a different party. Each party has its own reasons and benefits for participating in the sharing process.

According to Fear (2013), shared data opens the door to further research that builds on an original set of findings or supports an innovative re-purposing of data. If researchers beyond the original project can squeeze more value out of data, they can produce more science and more knowledge from that initial investment than would otherwise have been possible.

The same author, also describes the following benefits of data sharing for scientists: increased efficiency in the research cycle, new research capabilities, effectiveness for a wider scrutiny of research results, knowledge exchange and impact.

For funders, the data sharing benefits are considered as savings, if scientists are able to reuse data instead of re-collecting them (Fear, 2013).

The sharing of the patient data can have multiple uses as:
• Population health monitoring – done by lifesciences companies, by the biomedical studies/companies, by health authorities and by researchers;

• Measurement of quality of care;

• Used in clinical trials;

• Other uses, when data-sharing happens at the patient’s initiative.

There are three types of parties or stakeholders involved. These are: data-producers, data-consumers and data-regulators.

The regulating stakeholders also ask for data sharing to increase for various reasons.

As an example, the European Medicines Agency (EMA), a regulating stakeholder of the European healthcare landscape, provides the following reasoning for data sharing in the healthcare world, specifically for the data originating from the clinical trials (Secretary's Advisory Committee on Human Research Protections, 2013). These are:

• Reducing the selective reporting, as confirmed by the research world (Ross, & Krumholz, 2013), (Gøtzsche, 2011), (European Science Foundation, 2009);

• Allowing for study replication, as suggested by Gøtzsche (2011), Leo (2009) and The Economist (2013a);

• Giving clinical trials participants greater confidence that their contribution will be used to further medical knowledge;

• Increasing efficiency of research by allowing secondary analyses of data sets, also confirmed by Gøtzsche (2011);

• Providing patients and their advocates a greater ability to analyse relevant data. Largely described by Gøtzsche (2011).

From a researcher perspective, Gøtzsche (2011) adds other aspects, such as:

• The society being much better informed over the true benefits and harms of certain interventions or products;

• Accessing raw data makes meta-analysis of trials reliable (more than only based on published summary data);

• Helping exploratory analyses identify groups of patients where a treatment would be beneficial or harmful, resulting in cost-effective and evidence-base interventions.

Regarding clinical trials, the European Science Foundation (2009), a non-governmental organisation promoting scientific research collaboration and science policy at European level, suggests:

• Obliging sponsors, funders and all responsible organisations to register and to publish all clinical trial data regardless of the type of trial or phase; there exists the open access registry
in Europe\(^1\) and also in the US\(^2\).

According to Riveros et al. (2013), “serious adverse events are more likely to be published on the website than in the published article”. However, almost half of the clinical trials are not published despite the FDA requirement (Ross, & Krumholz, 2013). And worldwide there exists a registry\(^3\) composed of national registries of various countries.

- Facilitating the transfer of results into clinical practice.

Within The Netherlands, patient support organizations lobby the Dutch government for putting healthcare providers under pressure of delivering higher quality standards of healthcare.

The European Science Foundation (2009), when discussing about data-sharing, suggests taking into consideration the topics of:

- Curation and preservation of data;
- The ethical use of shared data;
- Consent to use the data;
- Regulatory mechanisms to ensure data is used appropriately.

### 3.1.1. The patient’s position within the data sharing initiative

This section describes what data sharing is, from a patient’s point of view. Once the patient signed the informed consent, his/her data is owned by the data-producer. The majority of literature does not speak about the patient’s position after the informed consent moment. Recent research (Reti, Feldman, Ross & Safran, 2009) argues that the concept of healthcare data co-creation empowers the patients in the form of the PHR (Patient Health Record). The patients perceive as positive to be provided with more access to their healthcare data, as such being helped to touch upon the healthcare aspects important to them (see Appendix B).

By using PHR, the patient is empowered and can by himself/herself agree to share his/her own sensitive healthcare data that otherwise is the subject of restrictive information due to privacy policies. Empowering the patient means to allow the patient to be responsible for their own data and data exchange and consequently of protecting his/her own data confidentiality, to be a data steward. As such, this technique would simplify many of the privacy and consent issues data-producers have, it would simplify the consent protocols between healthcare data-producers and consumers. However, policies regarding privacy, security, personal control and data stewardship should be revised (Halamka, Mandl & Tang, 2008).

At patient level, the choices for consent (i.e. of opt-in and opt-out) vary. In some countries, the patient can do an opt-out as by default is opted-in, while in other countries is the opposite situation. For UK, the authors Hill, Turner, Martin and Donovan (2013) describe the views of patients about consent for sharing their details for research purposes towards entities, including industry, outside the NHS (National Health System). The authors conclude that consent rates could be improved or acceptability of research without informed consent for the greater good to be improved, if public education about benefits of research, safeguards and legislation would be increased such that public trust is increased and misconceptions reduced.

---

1. See https://www.clinicaltrialsregister.eu.
3. See http://apps.who.int/trialsearch/.
3.1.2. **The data**

This section describes what kind of data the thesis is discussing about. The author is interested in the data belonging to the healthcare domain. According to Lane and Schur (2010), a data-consumer is interested in:

- Survey data;
- Administrative data;
- Linked administrative and survey data;
- Clinical data;
- Social-spatial data.

Regarding the data format, in the healthcare provider settings, the patient data resides in paper form and electronic formats. Regarding the electronic formats, the thesis uses the definitions proposed by Conover (2014), though these are representative for the U.S.:

- **Electronic Medical Record (EMR):** An electronic record managed by staff within one healthcare organization;
- **Electronic Health Record (EHR):** An electronic record managed by staff from multiple healthcare organizations;
- **Personal Health Record (PHR):** An electronic record managed by the patient. This is a subset of an EHR.

3.1.3. **Costs of data sharing**

This section presents what the costs are with respect to data sharing. There are many research papers around the data-sharing principles, but few of them discuss what the costs are for having data sharing since it comes with a price. The author considers costs a dimension that the reader has to take into account when thinking of data-sharing.

Wilhelm, Oster and Shoulson (2014) identify 4 categories of data-sharing costs:

- **Infrastructure and administration** – costs for maintaining the data repository (i.e. storage, security and quality-control processes), data exchange interfaces, monitoring and maintenance;
- **Standardization** – costs for preparing the raw data fit for multiple uses, notify users of changes and quality control;
- **Human resources** – costs for the personnel involved in recording the data, in building and maintaining the database and its functionality and in reporting; the largest cost of all 4;
- **Opportunity costs** – financial losses by personnel (i.e. investigators) doing low level data-sharing tasks or support instead of focusing on research activities; it raises to 10-15% of overall costs, covering ~15% of an investigators time.

These data-sharing costs are insufficiently taken into account by research sponsors when budgeting.
a clinical trial, only a few research grants capturing them explicitly. As such, argues Wilhelm, Oster and Shoulson (2014), data-sharing costs are laying on the data-producers, while data-consumers make use of data at no costs. This situation could be different in the future when potential business models would assume sharing the cost with the data-consumers.

In order to have high quality research, the input of data is expected to come from different sources and as such, it is expected to hold different types of data that need first to be combined before being analysed.

### 3.2. The data sharing process

This section briefly explains what the data sharing process is.

The data sharing process implies a data-consumer (e.g. a research project) that consumes data provided by a data-provider or data-producer (e.g. a hospital). First, the legal aspects have to be agreed between the partners and afterwards data can be transferred from the data-producer to the data-consumer according to the agreement. This contract bears different names: data use agreement, master research agreement, etc.

The process means that data owned by a data-provider is shared with a data-consumer respecting the rules set by the data-regulators.

There are data-regulators at international level (e.g. European Union – European Medicines Agency), at country level (e.g. Netherlands – IGZ\(^4\)) and at the healthcare institution level (e.g. in Dutch hospitals, the METC\(^5\)).

This is a simplistic view of a more complex legal process derived by the author from the works of Gøtzsche (2011) and Wilhelm, Oster and Shoulson (2014), since no single paper was identified that describes the data sharing process.

In practice, there are various modalities of accessing data in a shared way. For instance, the literature identifies the following channels for obtaining data (Wolf, 2005):

- Collaborative reanalysis – for re-analysing published results. An academic practice;
- Data exchange – a two-way data exchange agreement;
- Unilateral sharing – one way sharing;
- Sharing only project – a project created solely for gathering data that is later shared;
- Public archives;
- Restricted data archives and research data centres.

Similarly, Lane and Schur (2010) describe data to be accessible and available:

- As public use files, part of government sponsored survey datasets;
- Within research data centres;
- Through licensing arrangements or data use agreements between a data-consumer and the data-provider.

For the thesis, it is important to know the expected means of accessing data in order to determine

\(^4\) IGZ = Inspectie voor de Gezondheidszorg
\(^5\) METC = Medisch Ethische Toetsing Commissie
whether or not the same rules apply for Philips.

### 3.3. Data sharing stakeholders

This section presents who the data sharing stakeholders are.

In the data sharing process there are multiple partners involved e.g. patients, healthcare providers, insurers, government and research entities. Their numbers might differ per project, but their types do not. There are three types of partners involved. These are data-producers, data-consumers and data-regulators. These are also named stakeholders.

In Illustration 2 there is a schematic representation of the stakeholders involved in the sharing of healthcare data.

The goal of the regulating stakeholders is to mitigate, facilitate or prohibit data exchange from producers towards consumers, but also to manage, support and request exchange of processed data from consumers back into the ecosystem, meaning back towards the clinical world or other entities of the research world.

The data-regulators impose rules on the data sharing process, rules that are meant to enforce the rights of patients, thus to protect the patients and their data, e.g. ZonMw, IGZ.

The data-producers are those entities that capture and retain the patient data. e.g. a hospital.

The data-consumers are those entities that need patient data in their business development.

The stakeholders active in “producing data” are the clinical world, namely the healthcare providers and the clinical trials. The stakeholders active in “consuming data” are represented by the research world of university hospitals, academia and industry. The border between “producing data” and “consuming data” is governed by the regulating stakeholders (the decision-makers) that are represented by government, insurers and patient organizations. Their mission is to mitigate, facilitate or prohibit data exchange from producers towards consumers.

This diagram is confirmed by Hanney, Kuruvilla, Soper, and Mays (2010) and the approach is confirmed by Hanney, Kuruvilla, Soper, and Mays (2010) and the approach is

---

6 ZonMW = De Nederlandse organisatie voor gezondheidsonderzoek en zorginnovatie ([http://www.zonmw.nl/](http://www.zonmw.nl/))

The “feedback of data” represents the activity of feeding back the obtained results after consuming data. In recent years, the “feedback of data” is requested to happen actively by the regulating stakeholders who demand transparency and sharing back, for instance, of both input and output data for public or private sponsored clinical trials regardless of the trials’ outcome or papers' submittal. This is the case for North America and Europe.

For the Netherlands, as an example, the author identified the following stakeholders:

- The patients represented by Consumentenbond and Nederlandse Patienten Consumenten Federatie;
- The healthcare providers represented by Nederlandse Federatie van Universitair medische centra, Nederlandse Vereniging van Ziekenhuizen, Orde van Medisch Specialisten, Verpleegkundigen & Verzorgenden Nederland;
- Insurers represented by Zorgverzekerings Nederland;
- The government represented by IGZ.

The stakeholders' interests as published by Hanney et al. (2010), are presented in Appendix 8.3.

Wolf (2005) describes that those organizations running biomedical and behavioural studies involving humans, have institutional review boards (IRBs) that are assessing, reviewing, approving and monitoring the experiment protocols from a privacy, a legal and an ethical point of view. Similarly, within the responsibility of the IRBs fall also all data sharing or data use agreements that an organization has in place with external parties.

### 3.4. Bottlenecks in the data sharing process

For intellectual property or competitive reasons private organisations are reluctant to share data, though different opinions are shared by the actual employed scientists (Fear, 2013).

The data is not shared as regulators do not require it to happen (Gøtzsche, 2011).

In a data-sharing scenario, shared data is used as part of a study and the study results are published under the form of a scientific paper. The published papers are later used by peer scientists as foundation for further studies. However, it may happen that the papers' findings are biased towards positive reporting and thus unreliable to be considered as corner stones, as reported by The Economist (2013b). The author of the thesis considers the scientific paper publishing based on healthcare data as part of the feedback data loop.

There are other instances when clinical trial results are selectively reported, influencing acceptance of certain solutions on the market leading to lives losses or monetary losses, as it happened in the drug industry and reported by Gøtzsche (2011).

For example, it is difficult to obtain data from drug manufactures about licensed drugs if a researcher intends to scientifically research the product or the drug's effects on diseases not originally licensed for (European Science Foundation, 2009).

### 3.4.1. Problems faced by data-producers

Some clinical trial data-producers are reluctant to share data due to the financial costs, inappropriate data uses, deceptive secondary analyses or for fear that it would generate unwanted competition
Indeed, Wilhelm, Oster and Shoulson (2014) argues the financial costs of data-sharing are entirely on the data-producers who should capture the data, prepare it and format it in the right format so that it is usable off-the-shelf by the prospective data-consumers that mostly incur no costs. The signing of the data use agreement should help prevent inappropriate data use. Judging based on real-life situations, the secondary analyses of clinical trial data brought useful insights to the community on overestimating benefits and presented biased positive results of initial analysis. It may be beneficial for patients that more parties are active and competing in developing remedies for the good health of the patients (Gøtzsche, 2011).

The existing regulations governing data sharing apply to the healthcare professionals and not to the patients. The latter may decide to share their details on an online health community with the purpose of achieving a better health outcome (Wicks et al., 2010). Wicks also surveyed the users of this community for their reasons on sharing health data and benefits in doing so and discovering that those users with the most serious illnesses are also the most comfortable with sharing data. Among the benefits are mentioned: more informed treatment decisions around side effects, improved medication adherence and better managing their symptoms.

3.4.2. Problems faced by data-consumers

This section describes what the problems that data-consumers have based on the literature research. Scientists withhold data, like it happens in Genomics where a study has shown that between 3% and 15% of studies do not comply with the domain's requirement of publishing results in a domain's specialized database (Fear, 2013).

The willingness of scientist to share data is different to the actual sharing of data, according to Wallis, Rolando and Borgman (2013). As such saying yes, is a pro-social answer, while concerns for first right of publishing may hamper them from actual sharing of data. However, if data sharing occurs, it is conditioned to be attributed properly to the originating researchers. Sharing happens only when requesters are known and trusted individuals, as data sharing happens mostly upon individual request than via a data repository. The person sharing wants to ensure data will not be misused, while the requester needs to trust the accuracy and validity of shared data.

The negative aspects of data sharing, as perceived by scientists, are: (i) fear of losing privilege of publishing subsequent studies and (ii) the fear of using the dataset only once, because others will reuse the dataset faster than the data creator would be able to. One survey found that 35% of geneticists had seen others' work based on shared information being published, before publishing their own work (Fear, 2013).

3.5. Data quality in data sharing

Data is thought of having a high quality if it fits the consumer's needs of use, according to Strong, Lee and Wang (1997), and Parker, Stofberg, De la Harpe, Venter and Wills (2006), while Wilhelm, Oster and Shoulson (2014) argues that the value of sharing data depends on the prerequisites of data accessibility and data formatting to be satisfied.

Other research presents that low quality of data and missing information led to human lives losses (Fisher & Kingma, 2001). Also, Leitheiser (2001) argues that poor data quality leads to poor decision making, laborious data storage structures, raised re-engineering effort and pressure on the organization managing the data structures.
An example of data quality is fit-for-use, i.e. if data is meaningful for the business needs (Strong, Lee & Wang, 1997), thus if data is representative and accurate.

In the healthcare domain, most of the data is not captured primarily for research purposes but for treatment and operational purposes. However, when data is used for research purposes, the literature refers to it as secondary use of data.

In order to suffice the secondary role of data use for research, data quality should be first increased, as explained in (Ghosh, 2009). To ensure a high data quality, the entire chain of data collection, storage and sharing should be investigated, as each step in the workflow has an influence on the data quality, as presented by Ghosh (2009). For example, Coorevits et al. (2013) proposes a number of changes in the EHR design in order to make them more research friendly. However, the thesis will not advice on this aspect, since it does not depend on Philips Research, a data-consumer, but rather it depends on the data-producer.

Another challenge is the format in which the data is stored and shared that makes it inappropriate for use as confirmed by Coorevits et al. (2013). Yet another challenge is the fact that requests for data sharing are made by competitors.

The European Science Foundation recommends making available funding for supporting data-sharing with an appropriate storage and installation of relevant architectures and also harmonize the data management systems by creating a European standard e.g. ELIXIR (a distributed infrastructure for life-science information). In addition, the European Science Foundation (2009) recommends the quality of deposited data in clinical trials registries to be improved.

### 3.6. Bottlenecks in data quality

The literature speaks about several issues with respect to data quality problems:

- Relevance – data to be representative for the project (Strong, Lee & Wang, 1997);
- Data format (Fear, 2013), (Coorevits et al., 2013), (Wilhelm, Oster & Shoulson, 2014);
- Time-to-data – getting access to data within a projected time period (Ghosh, 2009), (Hill et al., 2013);
- Accurate (Coorevits et al., 2013);
- Contextual information (Fear, 2013);
- Data adaptation – budget implications (i.e. costs) (Fear, 2013);
- Data ownership (Fear, 2013);
- Confidentiality (Coorevits et al., 2013).

### 3.7. The questions-based model

Based on literature research, the research question of this thesis was formulated. This research question has a number of aspects that, when answered, create a concrete picture of how things are happening in reality. The targeted aspects were formulated as questions and were asked to professionals active in the healthcare domain. These questions are referred to as the questions-based model.

---

7 ELIXIR = http://www.elixir-europe.org/about
These questions are the following:

Q1. Do you use data from outside the company?

This thesis did not investigate data sharing within the same company. The assumption is that once data is in a company, it can be shared internally without limits based on the Binding Corporate Rules (BCRs) – as is the case for Philips. The BCR are company internal regulations that guarantee data privacy for any data that is transferred anywhere within the company. The BCR is a European policy, similar to the U.S. Safe Harbour policy. Therefore, the thesis focuses only on the data sharing between a company and an external party.

Q2. If yes, how do you get access to this data (e.g. partnership, purchase)? What is the source of data? (i.e. data access type)

The literature does not treat the data's source of origin directly, but indirectly by discussing the problems appeared in cases of industry sponsorships (Gøtzsche, 2011), of the insufficient use of existing data (Secretary’s Advisory Committee on Human Research Protections, 2013), (European Science Foundation, 2009) and of the non-complying with the mandatory publishing of clinical trial data in public registries (Gøtzsche, 2011), (European Science Foundation, 2009), (Ross, & Krumholz, 2013). Therefore, it is important for the thesis to take note of the means of obtaining the data and of its provenance. It is expected to have similarities with Lane and Schur's (2010) findings.

Q3. What is the data sharing process in your project?

The author intends to understand how the process works in the interviewed project. It is expected to be different depending on the source of data. This creates a base for the thesis to identify similarities and differences among different projects.

Q4. What are the parties involved in your project?

This questions helps in clarifying the conditions, meaning the partners, that a project is running as these might influence the problems identified.

Q5. Are there problems in the process?

The intention of this question is to understand whether in Philips Research the same problems apply as those identified in literature by Wallis, Rolando and Borgman (2013), and Fear (2013). Philips Research is seen as a data-consumer only, since due to intellectual properties rights the company feeds data back as new products or services.

Q6. What are the most important data quality aspects for you?

The literature provides a number of aspects to consider with respect to data quality (Strong, Lee & Wang, 1997), (Wilhelm, Oster & Shoulson, 2014). It is interesting to know what are the important aspects per project, such that it can be determined whether it differs and how it differs. Concretely, what are the guiding criteria the project looks for related to the data. e.g. whether data is representative, where it is clean, whether is accurate.

Q7. Are there problems related to data quality? What is the data quality you expected and what data quality did you receive from the field?

Gøtzsche (2011) argues that in some cases of pharma clinical trials, only the beneficial results are presented and later products are launched on the market, while the negative effects are on purpose withheld. This leads to situations when drugs are pulled out of the market after being released, but meanwhile derivated drugs from the original drug are tested and brought to the market. The old drug is pulled back from the market due to its doubtful clinical trial test results. The questions is whether the clinical trial results data used was relevant, thus meaningful, at the time of deciding to launch the drug to the market. This means that the data quality has to be thoroughly checked. This idea is also sustained by the studies of Fisher and Kingma (2001).

For the thesis, it is important to know what kind of data quality problems are in practice. The intention is to summarize them and establish whether they differ among different projects.

The questions-based model can be used for characterizing the data-sharing activities in a project for all types of data and not limited to healthcare data.

Due to the focus of the thesis, the questions-based model is demonstrated to be able to characterize the activities of sharing of healthcare data. The questions-based model is derived from the reviewed literature. Based on the genericity of the data-sharing aspects, the author believes this model may be used for describing the data-sharing activities in projects that work with healthcare and/or non-healthcare data. Thus, from an academic point of view, this questions-based model is seen as valuable.

The answers given by the interviewees to the above questions are summarized in the next section.
4. Analysis

This section presents an analysis of the answers provided by the interviewees to the questions-based model described in previous section. The purpose of the interviews is to gather the requirements the researchers have regarding the data-sharing process. The requirements are extracted by analysing the answers for patterns of similarity.

The interviewees answered all of the addressed questions. This demonstrates the fact the questions made sense and that these were rightly focusing on aspects of their domain of knowledge. This is a small validation of the questions-based model.

The reason to request clinical data is to develop new techniques, procedures, drugs and devices or to improve existing ones based on existing clinical data. This data can be used for instance in the identification of new usage patterns of a medical device, but also for testing of a new drug or assessment of a new back-office patient administration technique.

But, there are also other types of data that are still interesting to be used. For instance according to one interviewed expert, Philips collects regularly usage logs of its devices in the field, in agreement with the customer and respecting the data sharing regulations. The data is used for maintenance purposes, but also for analysing usage patterns with the purpose of improving the product or as part of defining new services. This approach is also in use for non-medical devices that the company produces and similarly, an agreement with the customer is in place and regulations are respected.

Another interviewed expert presented details about a research project that aggregates administrative hospital data. It means there is no patient data, but it contains information on the number of occupied beds during a certain period of time, average hospital stay length per sickness type, number of patients treated per sickness, etc. By having access to such data, a researcher is able to verify the applicability and performance of data analytics techniques as part of the effort of developing new services targeting healthcare institutions.

In the next section, the interviews are summarized based on the questions-based model and displayed in a matrix, side by side, in order to highlight their similarities and differences.

Next, some observations are made about the interviews.

The Analysis phase is followed by the Design phase that proposes a model that is constructed to address the requirements discovered in the Analysis phase.

4.1. Answers summary matrix

The purpose of the interviews is to find out how a researcher gets access to data and what challenges are encountered in this process and with respect to data quality.

The interviewees were Philips Research staff and staff from a Dutch hospital. Their roles range from software designer, to project manager, to researcher and to roles supporting the scientists.

In Table 1 are presented the summary of the interview answers. For each interview are presented the characteristics of the project using the questions-based model. For confidentiality reasons, all interviewees received a code and are addressed as being men, though in reality they can be men or women.
<table>
<thead>
<tr>
<th>Projects\Questions</th>
<th>Do you use data from outside the company? (Q1)</th>
<th>How do you get access to data? (i.e. data access type) (Q2)</th>
<th>What is the data sharing process in your project? (Q3)</th>
<th>What are the parties involved in your project? (Q4)</th>
<th>Are there problems in the process? (Q5)</th>
<th>What are the most important data quality aspects for you? (Q6)</th>
<th>Are there problems related to data quality? (Q7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewee 4</td>
<td>Yes</td>
<td>Philips-only. Buy or access free of charge data: US, UK, DE, SE, NL.</td>
<td>Scout for data. Process defined by sharing party.</td>
<td>Governmental bodies and insurance companies.</td>
<td>Legal aspects.</td>
<td>Data relevance.</td>
<td>Data format. Inconsistent data.</td>
</tr>
<tr>
<td>Interviewee 5</td>
<td>No, see Q5.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>Ideally, partnership</td>
<td>Lack of data due to missing partnership with Data privacy.</td>
<td>Data ownership and data responsibility.</td>
<td></td>
</tr>
<tr>
<td>Projects\Questions</td>
<td>Do you use data from outside the company? (Q1)</td>
<td>How do you get access to data? (i.e. data access type) (Q2)</td>
<td>What is the data sharing process in your project? (Q3)</td>
<td>What are the parties involved in your project? (Q4)</td>
<td>Are there problems in the process? (Q5)</td>
<td>What are the most important data quality aspects for you? (Q6)</td>
<td>Are there problems related to data quality? (Q7)</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------------------</td>
<td>---------------------------------------------------------</td>
<td>---------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>--------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Interviewee 7</td>
<td>Yes</td>
<td>Consortium.</td>
<td>Hospital shares data with industry.</td>
<td>Hospital and academia.</td>
<td>Legal aspects.</td>
<td>Data relevance, Access rights to data, Data storage.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Interviewee 8</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>Philips scientists.</td>
<td>Laborious internal approval process.</td>
<td>Data security, Data storage.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Interviewee 9</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>Philips scientists.</td>
<td>Challenging to work with other organisations.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

*Table 1: Interview summaries*
4.2. Observations based on the interviews

This section presents some observation made around the interview matrix presented in the previous section. These observations are seen by the author as requirements the data-sharing process.

All interviewed projects have to solve the following common challenges:

i. Data access type – meaning where to take relevant data from;

ii. Legal aspects – always the case when using external data, meaning when legal aspects have to be organised for a partnership, a data sharing agreement or a data use agreement.

All the projects are looking for relevant data. This is the highest priority with respect to data quality. By analysing Table 1, it can be noticed that projects have similar ways of accessing data. It can also be observed that a data access type is used only by certain project types. Based on this reasoning the data-sharing model in section 5 is proposed by the author.

Still related to data quality, in literature, data format is seen as a bottleneck, since proprietary format of shared data requires extra processing time for the researcher (Fear, 2013). However, for one interviewee the data format heterogeneity poses an advantage since this researcher works with a team of engineers able to handle any format. This is different than the average researcher that lacks access to such a team and thus encounters delays analysing same types of data.

During discussions with scientists, internal data sharing came up as not being sufficiently promoted within the company. For certain projects, if the needed data is available internally, there would be no need to use external data and thus process of accessing data is shorter.

The approval body, the IRB, of Philips Research is the ICBE$^9$.

The interviews reveal that the legal aspects are the most frequent process bottlenecks for consortium projects. The interviews also reveal that for EU partnership projects, trust among partners and willingness of partners are determining the speed of reaching the data.

For the Philips-only projects that are accessing external data off-the-shelf the challenges refer to inconsistent data through obfuscation or by truncation, on purpose.

An example of differences among projects, actually among roles is how open innovation is perceived to be. For instance, for an interviewee in a supporting role for the scientists of Philips Research, open innovation cannot be possible within a partnership since it acts as a closed cooperation project. This is confirmed by an interviewee in a healthcare institution who says that there cannot be an open innovation with the industry, but only with not-for-profit organizations and academia.

---

$^9$ ICBE = Internal Committee Biomedical Experiments
5. Design

Currently, each project (i.e. a team of scientists) is free to look anywhere for data that is relevant for the project's requirements.

In this section a data-sharing model is presented, an instrument through which the reader is able to understand what a new scientist should do for getting access to data. This model is based on the findings, the requirements identified in the observations section.

The model is a succession of steps the scientist should follow in order to discover how to find data. Along this road, the scientist learns more information about the types of data accesses, learns about the projects Philips runs and learns about their characteristics.

This model is developed by the author based on the similarities and differences observed in the interviews and discussions around the interviews. Parts of the model are confirmed by literature. This is pointed out during the in detail description of the model.

The Design phase is followed by the Validation phase that discusses the findings of the model validation by a panel of experts.

5.1. The data-sharing model

For sake of clarify, the author reiterates that the research question is how does a scientist or a project he is working in, be able to access data from the healthcare domain for using the data for the validation of a hypothesis. By data it is meant information of different types, see section 3.1.2.

The model starts by proposing to **first search relevant data internally** by verifying the internal Philips Research catalogue that holds information regarding existing datasets.

If relevant data is found internally for the project, the search stops.

If no relevant data was found internally, explore next the following options:

- a) Search if relevant data can be obtained **off-the-shelf** from data-providers. If yes, the project reached its goal and looks no further. It means the project is a **Philips-only project**. See section 5.2.2.3 for considerations for this type of project.

- b) If off-the-shelf data is irrelevant or data is not found, a project looks if it can create its own data. It can be done within Philips or it can be delegated to a healthcare institution to collect the targeted signals from patients, in which case the institution acts as a supplier. This project is a **Philips-only project**. See section 5.2.2.3 for considerations for this type of project.

- c) If it is not possible due to technical limitations, legislation, etc., the last chance of a project is to look if it can partner to obtain data. Thus, it can be a **European partnership or consortium project**. See the next sections on consideration points regarding these type of projects.

It holds for all projects, that once a project uses external data, **follow the internal approval process**, as it is already done.

Once an option was found for the project, it helps discussing matters with an experienced colleague, as it completes the overall picture offered by this thesis.

The graphical representation of the data-sharing model is presented in Illustration 3.
In the next section detailed information is presented about each of the steps of the model i.e. the bold text above.

It is the conviction of the author that the model can handle all types of healthcare data since the requirements identified in the observations section are based on interviews with experts using different types of healthcare data, depending on the specifics of their projects.

5.2. Consideration points

This section presents information to clarify the model presented in the previous section.

5.2.1. Encourage internal data sharing

During discussions with scientists on the workfloor, the topic of internal data sharing came up as not being sufficiently promoted within the company. If the needed data is available internally, there would be no need to use external data and thus, there is no need of solving legal and such similar aspects.

Philips Research has an internal catalogue of studies. When data is available from the internal catalogue, there would be no need to use external data and thus, it means no legal aspects have to be solved to reuse the data. It is the fastest and cheapest way of obtaining data as there are no delays due to the legal aspects.

It means data from the catalogue might have a secondary use. Before reusing it, it should be verified that it is allowed to be reused for other purposes than those initially used for collection. This information should be displayed within the catalogue as an attribute of the data described.

If however, reusing internally available data is not possible, then external data has to be found and accessed. The types of data access are presented in detail in the next section.

5.2.2. Data access types

This section presents in more detail what the data access types are. The data access types are correlated to Q3 of the interview.

This section presents the channels that an institution is able to use for obtaining data from the field for the improvement or the creation of services or products. Each of the projects presented above, uses one of the channels described in this section.

As described in section 3.2, the research of Wolf (2005) identifies the following channels for

Illustration 3: The graphical representation of the model
obtaining data:

a) Collaborative reanalysis;
b) Data exchange;
c) Unilateral sharing;
d) Sharing only project;
e) Public archives;
f) Restricted data archives and research data centres.

Through the interviews, the channels used by Philips Research to access data are the following:

1) Access off-the-shelf data;
2) Create own data;
3) Partner to have access to data;
4) Trade data for other data – mentioned in the interviews, but not used by the company.

When putting the two sources side by side for comparison, the mapping of data sharing in literature versus practice is the following:

- The literature “Public archives”, “restricted data archives, research data centres”, “sharing only project” can be mapped to practice’s “access of-the-shelf data”;
- The literature “Data exchange” mapped to practice's “trade data for other data”;
- The literature “Collaborative reanalysis”, “unilateral sharing”, “sharing only project” mapped to practice's “partner to have access to data”;
- There is no mapping for “create own data”.

Each data access type has its own characteristics. These can influence the selection of a certain data access type. The characteristics are:

i. Relevance: how relevant is data for the project's purpose;
ii. Costs of data: the costs needed to procure data. The costs associated for preparing and processing the data once it is received within the project, are not taken into account.

iii. Data format: the format the data has when it is delivered to the project;
iv. Time-to-data: how fast does the project get access to data.

The author proposed to use these characteristics as these were found as representative in literature and discussed about in the interviews. When reading the devil's quadrangle, it is best to consider the textual interpretation since it depends per dimension what it means to be close to 0.

The details of each data access type are presented in the subsequent sections. For each it is indicated whether data is collected for the first time for a given purpose or it is a secondary use of data. The reason to specify this, is to make the reader aware of the fact that the literature observations related to the secondary use of data are applicable.

The order of presenting the data access types is based on the lowest relevance and cost to the highest relevance and cost.
Access off-the-shelf data

Accessing off-the-shelf data means that data is already available off-the-shelf from external institutions. The data was captured specifically for the purpose of research and development. In the interviews, there are 2 projects that make use of this data access type, as visible in the matrix.

This data access type can be further split into:

(a) Buying healthcare data

This option is specific for U.S. For example, for research purposes it is possible to buy patient identifiable data\(^{10}\) or even more focused, patient discharge records\(^{11}\).

(b) Accessing free data

This refers to population health indicators or statistical hospital numbers.

This option is available in many countries. For example in the Netherlands, the healthcare performance indicators are freely available from a governmental agency\(^{12}\).

Also in The Netherlands, data it is not centralized, but scattered among multiple governmental institutions e.g. Dutch Hospital Data, IGZ, Ministerie van Volksgezondheid, Welzijn en Sport.

In Germany, this is centralized by the government and it is freely available.

In Sweden, such data is available from the health insurance agencies and as such it is less focused onto the hospital's occupancy but more onto the treatment that took place, their length, start date, end date, treatment's conclusion, etc.

The characteristics of this data access type are represented in Illustration 4 following the devil's quadrangle for quality (Limam Mansar & Reijers, 2007). The interpretation of these dimensions is the following:

i. Relevance: the data has a relatively low relevance (negative aspect i.e. -);

ii. Cost: It has the lowest cost possible, it costs only to purchase the data and to clarify the legal aspects of the purchased data (positive aspect i.e. +);

iii. Data format: It is expected to have a proprietary format of the organisation providing the data (negative aspect i.e. -);

iv. Time-to-data: it has the shortest time possible for accessing data since data is available off-the-shelf. There is no need to wait for data to be collected or to prepare it for sharing (positive aspect i.e. +).

---


\(^{11}\) See [http://www.hcup-us.ahrq.gov/sidoverview.jsp](http://www.hcup-us.ahrq.gov/sidoverview.jsp).

\(^{12}\) See [http://www.ziekenhuizentransparant.nl/](http://www.ziekenhuizentransparant.nl/).
The project choosing this channel is likely to be a Philips-only project.

Create own data

Creating own data it means that the data is collected for the first time for a given purpose. An interviewee acknowledged this case as being representative for his project.

The characteristics of the data access type when Philips Research collects the data are represented in Illustration 5 following the devil's quadrangle for quality (Limam Mansar & Reijers, 2007). The interpretation of these dimensions is the following:

i. Relevance: It has the highest relevance since data captures and delivers what the researcher needs (positive aspect i.e. +);

ii. Cost: On average it has high costs, since it depends on the type of data to collect. For some signals it is sufficient to invite the organisation's employees to take part in the study, while for others signals an external party is involved in collecting the data (negative aspect i.e. -);

iii. Data format: It has the expected format, since the organisation itself decides the format (positive aspect i.e. +);

iv. Time-to-data: it has a low waiting time before accessing data, as it depends solely on Philips Research. However, if delays occur in organising the project, accessing the data is also delayed (positive aspect i.e. +).
The project choosing this channel is likely to be a Philips-only project. However, sometimes given the specificity of the signals, Philips cannot record its own data of wants to share the costs and as such, it partners to have access to data.

Partner to have access to data

When partnering for having access to data, the data is collected for the first time for an intended use.

There are 2 projects that are using this data access type, as presented in the matrix.

This data access type can be further decomposed into:

(a) European project, e.g. ACT\(^{13}\), HeartCycle\(^{14}\);
(b) Consortium project, e.g. e-Vita (Emerce, 2013);
(c) not-for-profit sponsorship e.g. the Cochrane Collaboration\(^{15}\);
(d) public-private healthcare partnership, i.e. mostly in developing countries.

The characteristics of this data access type are represented in Illustration 6 following the devil’s quadrangle for quality (Limam Mansar & Reijers, 2007). The interpretation of these dimensions is the following:

i. Relevance: It has a high relevance (positive aspect i.e. +);

ii. Cost: It has some costs for organising the partnership and for making the partnership operational (medium negative aspect i.e. +/-);

iii. Data format: The format is decided by the sharing organisation (negative aspect i.e. -);

\(^{13}\) ACT = Advancing Care Coordination and Telehealth Deployment (http://www.act-programme.eu)

\(^{14}\) See http://www.heartcycle.eu/.

\(^{15}\) See http://www.cochrane.org/.
iv. Time-to-data: there is some throughput time for accessing data, as it depends on the sharing partners (medium negative aspect i.e. +/-).

The project choosing this channel is likely to be a European or consortium project.

**Trade data for data**

For this data access type, the data might be used for secondary purposes.

When an organization has data, it can access an exchange platform where data is traded for other data. An example is the OpenfMRI\(^{16}\), exchange platform for MRI images.

This channel is not used in Philips Research since it might lead to intellectual property losses. However it is mentioned during the interviews as general information and not as used in daily research activities.

Despite not being used, the author still describes the characteristics of such a data access type in order to address all items discussed with the researchers.

Illustration 7 presents them using the devil's quadrangle for quality (Limam Mansar & Reijers, 2007). The interpretation of these dimensions is the following:

i. Relevance: It has a low relevance (negative aspect i.e. -);

ii. Cost: It has very low costs for accessing the data, since all data is available free of charge; the only costs are related to the clarifying of the legal aspects (positive aspect i.e. +);

iii. Data format: It has a proprietary unknown format, it has the format of the organisation sharing the data (negative aspect i.e. -);

iv. Time-to-data: there is some throughput time for accessing data depending on the exchange platform (medium negative i.e. +/-).

\(^{16}\) See [http://www.openfmri.org](http://www.openfmri.org).
Table 2 presents the summary of the characteristics of the data access types presented above.

<table>
<thead>
<tr>
<th>Data access type characteristics</th>
<th>Relevance</th>
<th>Cost</th>
<th>Data format</th>
<th>Time-to-data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access off-the-shelf data</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Create own data</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Partner for data</td>
<td>+</td>
<td>+/-</td>
<td>-</td>
<td>+/-</td>
</tr>
<tr>
<td>Trade data</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+/-</td>
</tr>
</tbody>
</table>

Table 2: Summary of the data access type

Depending on the data access type, there are different type of projects that are initiated. These are presented in detail in the next section.

5.2.3. Project types

It can be observed that a data access type is used only by projects with a certain funding. Thus, projects can be grouped in types given their type of funding. The funding party has also the role of a stakeholder that exercises its governance role.

This section presents the characteristics of each project type and the encountered challenges for the research problem. This section corresponds to Q2 of the interviews.
The project types are the following:

1. European project – European Union funding, multiple European partners;
2. Consortium project – country government funding or funding from a group of companies;
3. Philips-only project – solely Philips funding.

By looking at the interview summaries comparison in Table 1, there are different consideration points particular to each project types. These are presented in the next sections.

**European project**

It is a project accessing European Union (EU) funds together with partners from multiple European countries. This type of project runs in an EU framework e.g. FP7\(^\text{17}\), H2020\(^\text{18}\). The partners are industry, healthcare institutions (e.g. hospitals) and academia.

Challenge types for the research problem:

i. Who to partner with – partners join the project based on previous interactions in European projects.
   
   In a partnership it is important to have a good reputation and to maintain it. That is why, it is important for Philips not to loose its credibility, as this is an important criterion in a partnership.

ii. Ad-hoc data-sharing processes among partners – address it by making clear commitments on due dates if feasible;

iii. Willingness – the partner is not willing to share data, since the data is not ready for sharing. Meaning, the partner has no extra time to spend for recording contextual information essential for shared data. Currently, the costs for collecting data that are borne by the data-producer only.

iv. Doctor's opinion about his/her ownership of the data – though legislation sets boundaries regarding medical data ownership, in reality it happens that when a patient undergoes treatment in a healthcare institution (e.g. hospital), the specialist in charge of the treatment also acts as owner (in possession of and responsible for) of the patient's medical data

v. Dummy data – Sometimes temporary lack of data is solved in practice by creating dummy data based on the real data. This might solve problems on short term, but can create problems on the long run.

vi. Inconsistency of data – due to time pressure, data is shared in a raw form that might hold inconsistencies. By using contextual information, the real inconsistencies might be detected and eliminated, and not that information that is left out on purpose as part of a process.

   However, once an inconsistency is observed, it is good to feed the situation back to the data producer. This is done in academia.

Not a challenge:

i. Legal – it is not an issue.

---


Consortium project

It is a project funded by a group of companies. In addition this project type may also access governmental funds, e.g. of the Dutch government. It has several partners from the same country. Overall a consortium has less partners than a European project does. Similarly to the European project, the partners are coming from the industry, healthcare institutions and academia.

Challenge types for the research problem:

i. Legal – it is an issue;

Take advantage of a Master Research Agreement (MRA) if it is in place. Such an agreement allows projects to take place without worrying about the legal terms since an agreement is already in place. Such an MRA is usually signed between a hospital and Philips Research and it functions as an umbrella for any projects taking place between the two parties. See Lewin (2009) and Pardo Roques (2014).

Not a challenge:

i. Source of data – because it is provided by a partner;

ii. Representative data – to have data available from a high sample of population, population of different ages and from different sites. An interviewee places data quality as a top priority, surpassing costs and time-to-(access)data.

Philips-only project

It is a project funded solely by Philips. All data collected is ownership of Philips.

Challenge types for the research problem:

i. Source of data – It differs. It can be:

(a) Philips user trial

(b) provided by a partner;

(c) bought or free of charge from governments in US, UK, DE, SE, NL;

Depending on the country a researcher needs to access data from, there are different points of accessing such administrative hospital data. See section 5.2.2.1.

ii. Legal – it is an issue, as time is needed to sort out the legal aspects depending on the source of data. As such, these can be:

(a) Philips terms are leading;

(b) The terms of the sharing data party are leading;

(c) The terms of the sharing data party are leading.

iii. Ad-hoc process when trusted relationships – privacy regulations not followed due to unawareness.

(a) Currently, privacy awareness is increasing, but there is still room for improvement.

iv. Data quality- it depends on the data source.

i.e. Data format – the format in which the data is stored and shared that makes it inappropriate for use, the coding of data that is at times premeditated incorrect or too vague, as one of the interviewees stated and confirmed by Coorevits et al. (2013) or the fact that
requests are made by competitors.
The identified challenges are presented to an expert panel to determine whether these are recognizable in other projects as well. If they do, it means these challenges are considered representative for Philips Research.
In the next section solutions are proposed to address those challenges that fit within the thesis' scope.

5.2.4. **Internal approval process**

The internal approval process is an implementation of the generic IRB process.
6. Validation

This section discusses the findings during the validation sessions of the model proposed in the Design section.

The model was presented to 4 experts for validation, of which 2, participated also in the interviews held during the orientation phase. These experts are scientists working in Philips Research in projects making use of data for validation of various hypotheses. The model was discussed with each of the experts in individual meetings of 1 hour.

Before explaining the dimensions used in the validation, as a reminder, the use case that model has to serve is the following: when a new scientist needs healthcare data to validate his/her hypothesis. By following this model, it will become clear to the scientist what the steps to take are, what the characteristics of data access types are and finally, what the characteristics of the project types are.

During the validation sessions, the following constituent parts of the model and their characteristics, were discussed:

1. The internal data sharing;
2. The data access types;
3. The project types;
4. The internal approval process.

For the entire model, the following validation aspects were discussed:

i. Usability of the model;
ii. Applicability of the model;
iii. Understandability of the model;
iv. Relevance of the model;
v. The fit of the model within the organisation's culture.

The understanding of the author regarding the validation aspects presented above, is the following:

i. Usability means whether one can work easily with the model, that the model is uncomplicated to use.

ii. Applicability means that the model can be taken as-is and followed out-of-the-box without adaptations. Thus, it covers already the practical aspects of the business daily life.

iii. Understandability means the model is explicit in the steps the researcher has to take, and whether the notions used by the model are clear.

iv. Relevance means whether the model serves the purpose, if it will become clear to the scientist what the steps to reach healthcare data are, what the consideration points are, etc.

v. Fitness means whether the model fits with the organisation's way of working, whether is respecting the organisation's culture.

In literature there are multiple dimensions to consider when validating models (Sargent, 2013). From these, the author selected the aspects presented above as being representative for the proposed model proposed. Given the limited amount of time of the experts and deadline for completing the thesis, the validation of the model is limited and it is no full-fledged academic validation.
The views of the experts are presented in the following sections.

6.1. The overall model

In this section are presented the views of the experts regarding the overall model.

During the validation sessions, the discussions were triggered over the current processes within the organisation around data-sharing. This proves the model was clear, understandable in its steps and the characteristics of each step.

An expert found the model as a solution for some projects, as it depends on the kind of data one is searching for. For instance, for some projects one must follow the model a few times, as a project makes use of multiple source of data e.g. partially bought and partially in-house created. However, the same expert said the model is good to raise awareness among the researchers dealing with shared data.

The experts found the model, the steps and their characteristics, as describing facts and situations they were confronted with during their projects.

Concretely in the day to day work, there are a multitude of aspects that have to be solved with respect to data sharing. Some can be solved in technique, others are solved using processes and some after a while are no longer relevant, thus get solved by themselves.

Depending on the role and type of projects an expert is working in, an expert can find certain aspects more relevant to him/her, than other experts would. For those working in European projects, the model is more meaningful related to the European aspects. For those working in consortium, the Philips-only projects are not so meaningful.

Another dimension suggested to be investigated by the thesis is related to validity of data. An expert argued that data has a limited relevancy. Over 20 years it may be that the signals collected today are not representative any more e.g. some parameters related to blood samples.

An expert asked if this model can be applied in other companies. The model has generic components that are applicable for other organisations as well. The Philips specific parts refer to the internal studies catalogue and the internal approval process. However, it is expected an ICBE-like body to exist also in other organisations that are working with healthcare data.

The experts agreed the model increase the knowledge of a researcher related to characteristics of data accesses, project types, points to consider and challenges. In addition to it, it also helps to discuss with colleagues that have experience in the subject of using healthcare data.

In the next sections are presented the opinions of the experts regarding the constituent parts of the model.

6.1.1. Internal data sharing

In this section are presented the views of the experts regarding the internal data sharing.

The experts agreed that reusing the internal data sharing catalogue is a good place to start, since it can reduce considerably the effort spent in finding the external data and everything else that comes along, i.e. fulfilling the legal requirements, keeping in contact with external parties, time delays and hours of back-office activities, instead of research work.

They also agreed the catalogue should gain more content and increase its visibility among the researchers. Simply seen it is a chicken-egg problem. The catalogue is not visible since there is
insufficient content. No content is added since the catalogue is not sought actively by colleagues. An expert proposed to use creative methods to engage researchers in enriching the catalogue's data. As a future line, instead of a scientist searching manually the internal data base for relevant data, an automatic process of data mining should be put in place for providing to the scientist the most relevant study that has such data.

**6.1.2. Data access types**

In this section are presented the views of the experts regarding the data access types.

The experts agreed with the dimensions proposed by the author for the quadrangles. For one expert, the devil's quadrangles from the data access types were interesting to see, as it is a new point of view. One expert proposed to introduce data storage as another dimension in future versions of the model.

**6.1.3. Project types**

In this section are presented the views of the experts regarding the project types.

The experts recognized the project types.

Related to the partnership with partners, credibility is an important criterion in a partnership. A data quality aspect to consider is traceability i.e. who provided the data, who is the data-producer. It does not refer to identifying a patient but the entity that shared the data.

**6.1.4. The internal approval process**

In this section are presented the views of the experts regarding the internal sharing process.

The data-sharing model triggered discussions about how certain things related to the internal process for accessing and storing data can be improved, about the internal approval process and about the relationship with partners.

It is good to present the ICBE process as a part of using the external data, said one expert, since a researcher has to be aware of this approval process. The internal approval process has to be followed in all cases, except when reusing internally available data under certain conditions, and this is something the data-sharing model takes into consideration.

**6.2. Validation conclusions**

In conclusion, the experts confirmed the model is understandable since it triggered discussion on how to improve the existing processes and the tools for data sharing within the organisation.

Regarding the relevance of the model, an expert said it is needed to promote the internal data sharing, since it is in the advantage of the organisation. The reuse of internal data sharing was found as a relevant and of a positive impact for a scientist's workload.

Regarding the applicability of the model, an expert mentioned that some projects may combine different sources to reach relevant data.

The model makes use of principles the organisation uses, it makes use of processes the organisation uses, thus the model fits within the organisation's current practices.

Thus, the experts discussed the model on the validation aspects presented previously. They also
made suggestion for future work and improvements points for the model. Given the time limit to finish this master thesis, these suggestions will be placed on the list of future work.

The final conclusions of the master thesis are presented in the next section.
7. Conclusions

The research question of this thesis is how does a scientist get access to healthcare data from the outside of his organisation.

For using the solid academic base in addressing this topic, the thesis started with a literature research. During the literature research, numerous papers were consulted with respect to data-sharing. The papers describe various facets related to the stakeholders involved in data-sharing, the costs of the data-sharing, related data quality aspects and the challenges surrounding these topics. However, the author found no model in the literature describing how to reach data through data-sharing from a scientist's perspective.

Due to this fact, firstly a questions-based model is created and used through interviews with Philips Research scientist with the purpose of investigating how data-sharing works in practice. The interviews revealed that a common challenge for partnership projects are the legal aspects, followed by trust and willingness of partners. For Philips-only projects, it is the data format.

Secondly, from the similarity patterns, from these requirements, a model is proposed that promotes internal data sharing, that explains what the means of accessing data are, what kind of projects types -are in the organisation and for each some characteristics to take into consideration. This model may be used in projects using different healthcare data types.

The model was validated with experts from the same organisation during individual discussions. The experts recognized the applicability of such a model, that it is understandable and that it fits with the current practices of the organisation. They saw the relevance of the internal data sharing through data re-use.

They also provided suggestions for improvement points of the model in order to increase its relevance.

In the next section the limitations and future work are presented.

7.1. Limitations and future work

Based on the literature review a questions-based model is derived. This questions-based model can be used for characterising the data-sharing activities in a project. It may be applied out-of-the-box for projects active in the healthcare domain, but it may also be applied in non-healthcare projects by adjusting it slightly. This is one direction regarding the future work on indicated how this questions-based model may be developed into a stable and mature generic model.

The literature research did not reveal a model or method on how a researcher can access healthcare data through data sharing. The model proposed by the author is the first of his kind. This model was validated by a panel of scientists on a number of dimensions, nevertheless it is a limited validation and not an academic validation. Certainly, there is space of improvement for this model.

Another limitation is related to the devil's quadrangle used to describe the data access types. The quadrangle's dimensions are proposed by the author and did not go through a validation process before using it.

Another limitation is that the data-sharing model might be tailored to Philips Research. In its current form, the model contains generic parts, i.e. the data access types and the project types, but also organisation specific parts, i.e. the internal data sharing and the internal approval process. However, in any organisation working with healthcare data, such specific parts should exist as well, since
these are required by the legislation (i.e. the IRB) or business practices (i.e. internal data sharing).

The validation sessions may be extended to include discussions with other experts that those involved in the first round of interviews during the orientation phase. This would eliminate a potential over-fitting of the model and increase the confidence of the validation results. In addition, by using a different experts cohort, the validation findings may be different than the current ones. This is another line of future work.

Some of the improvement points were identified by the author, while others were pointed out during the validation sessions. All improvement ideas are presented in the following sections.

**Internal data sharing**

The fastest and cost effective mean for a researcher to reach and use data from the healthcare domain, is by reusing data already available within the company. As future work, it is interesting to research the reasons why data is not shared sufficiently internally and how can that be improved.

**The use of dummy data**

In some projects, due to the limited availability of real data, researchers create based on it, dummy data. This is used for testing of a limited scoped hypothesis. If successful, this grows towards an extended hypothesis. It is a question whether dummy data can still be used in this case. It is a question whether dummy data brings benefits when used on long term, as it might introduce bias into the hypothesis results.

**Data storage**

During the validation sessions, an expert suggested adding an extra dimension to the devil’s quadrangle when defining the data access types, that is the data storage. Within the organisation data storage is one aspect that can be analysed for potential improvement, in the expert’s opinion. However, the first step is to raise the awareness of the researcher that uses the model also with respect to this dimension.

**Analyse the Philips internal ICBE approval process**

The internal ICBE approval process allows a project to use external data. This process may be further analysed for potential improvements, since this was not done given the limited time available of this thesis.
8. References


