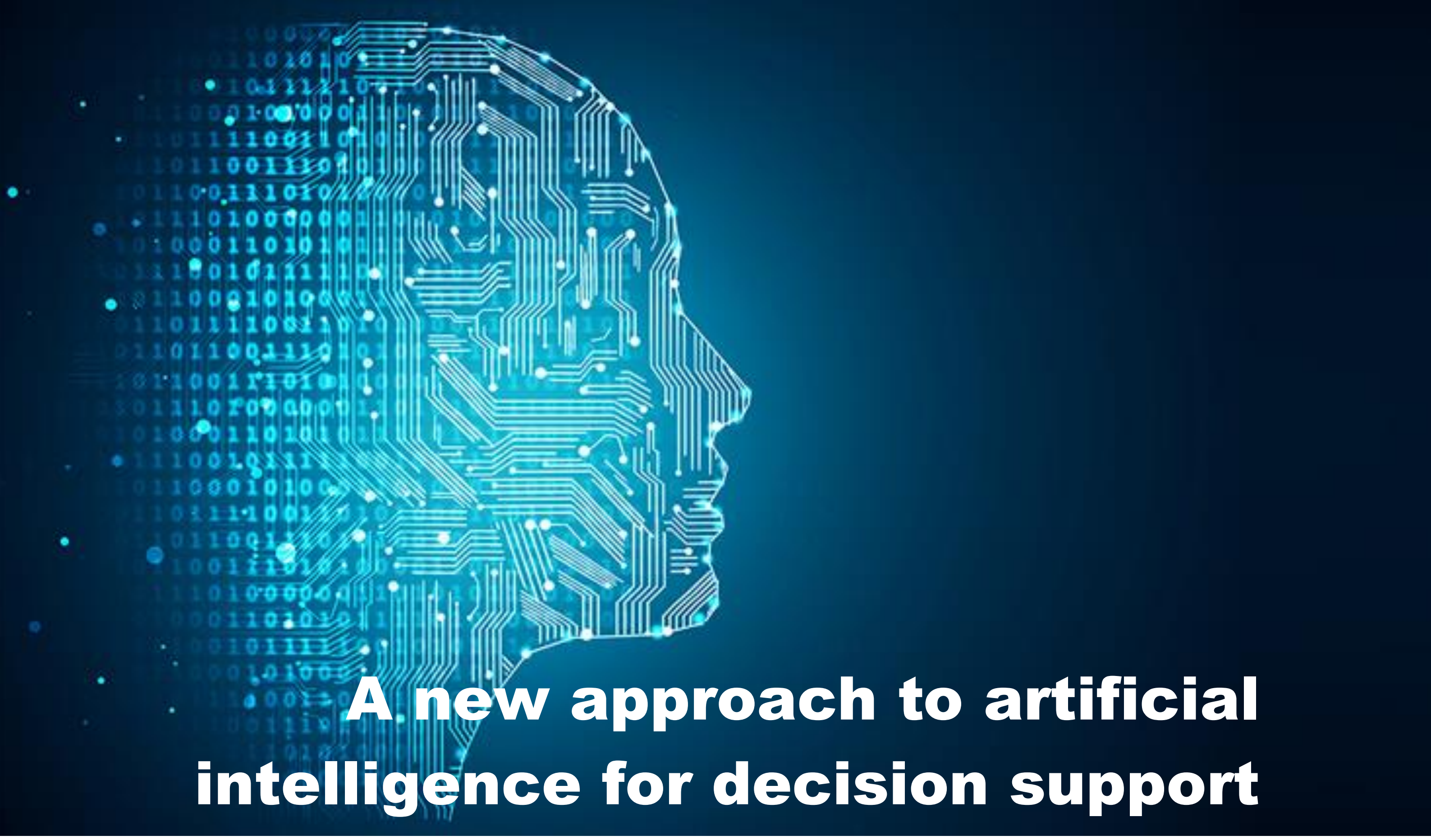


A new approach to artificial intelligence for decision support

Case study in the Neonatal Intensive Care Unit of the University Medical Centre of Groningen

Annebel ten Broeke
4973267

Master thesis submitted to Delft University of Technology
in partial fulfilment of requirements for the degree of
Master of Science
in **Complex System Engineering and Management**
Faculty of Technology, Policy, and Management

To be defended publicly on October 5th 2020.

Graduation committee

Chair: Dr. E.J.E. Molin	TU Delft, Section Engineering Systems and Services
First supervisor: Prof. Dr. Ir. C.G Chorus	TU Delft, Section Engineering Systems and Services
Second supervisor: Dr. H.G. Van der Voort	TU Delft, Section Multi-Actor Systems
External supervisor: N. Heyning (M.Sc.)	CEO Councyl

Acknowledgements

About nine months ago I started searching for a topic for my graduation project to complete my MSc Complex System Engineering and Management at the Delft University of Technology. I was incredibly fortunate that the company Councyl just launched who were willing to provide me with the opportunity to start as a graduate intern at their company and help conduct one of their first use-cases. It provided me with the chance to work on a topic that fascinated me. My enhanced interest for this topic turned out to be, especially, important as the COVID-19 pandemic took over the world, causing me, and the entire planet, to stay and work at home. Although this was sometimes challenging, my fascination and enthuse for this research still made me enjoy working on my graduation project.

Not only the exciting topic made my graduation project an enjoyable ride, but the people that supported and helped me throughout this study also contributed to that feeling.

Therefore, I would like to thank all members of my graduation committee. First of all, I would like to thank Caspar Chorus, who provided me with interesting perspectives on my research and from whom I greatly appreciated his commitment to this study. Second, I would like to thank Eric Molin, who provided me with detailed and critical feedback on my thesis. Your help and expertise on the design of choice experiments helped me a lot. Lastly, I would like to thank Haiko Van der Voort who time and time again provided me with interesting and new perspectives on my research.

Furthermore, I would like to thank all UMCG physicians that collaborated with me in this research. Especially, the physicians that aided in the design of the choice experiment and the set-up of this project.

Also, Nicolaas, thank you for your great help during this project. I could always turn to you for support, and your constructive feedback was valuable.

Lastly, I would like to thank my parents for just being my parents, and helping me throughout the way.

Annebel ten Broeke
Delft, September 2020

Summary

For decades researchers deliberated and continue to debate on how to support and assist humans in decision-making. This resulted in the development of Intelligent Decisions Support Systems (IDSSs). An IDSS is an application of artificial intelligence (AI) that desires to enhance and support decision making by enabling tasks to be performed by a computer while mimicking human capabilities. The two most generally classified types of IDSSs are knowledge-based and non-knowledge based systems. A knowledge-based system, also called an Expert system, directly translates domain knowledge into a set of rules or cases to support human decisions. In contrast, non-knowledge based IDSSs apply the rapidly growing branch of AI known as machine learning (ML), that grounds its decision-support on feature extraction of labelled training data.

Recently, a company called Councyl, in collaboration with the TU Delft, developed a new approach to AI that has the potential to constitute a novel type of IDSS for judgement purposes. The new approach to AI is called BAIT (Behavioural artificial intelligence technology). BAIT utilises discrete choice modelling (DCM) to codify the domain expertise of experts' in order to provide introspection on their decisions and support future judgments.

As BAIT is a new IDDS approach it requires testing in different settings to gain insight into the usefulness and effectiveness of this new method.

This research will explore the potential of BAIT by employing the system at the Neonatal Intensive Care Unit (NICU) of the University Medical Centre of Groningen (UMCG). It will utilise BAIT for the choice task of UMCG physicians on whether to provide parents with a recommendation against or in favour of surgery on a premature baby diagnosed with Necrotizing Enterocolitis (NEC), given the indication that surgery is required to sustain life. NEC is a severe intestinal disease that affects premature neonates.

This study desires to interpret the lessons learned in this case study to discuss the potential of BAIT as a novel IDSS in the medical sector. Therefore, this study aims to answer the following research question:

Does BAIT have potential to serve as a novel type of IDSS in the medical sector?

Methodology

This research applies a case study approach. The NICU of the UMCG is the setting in which BAIT is applied, on the choice task of the UMCG physicians discussed above, to gain insight into the usefulness and effectiveness of the approach. The lessons learned in this case study are used to examine the potential of BAIT as a novel IDSS in the medical sector

Additionally, as this research applies a case study approach by utilising BAIT on the choice task of the UMCG physicians, it also administers the research method that is inherent to BAIT's approach for decision support. Hence, DCM is applied in this research, as it is the method practised by BAIT, to codify domain expertise and support future judgments.

This research asked the group of UMCG physicians to conduct a choice experiment. The choice experiment consisted of 35 choice scenarios. Every choice scenario included a context with two choice options; either provide a recommendation against or in favour of surgery. The choice scenarios contain a set of attributes. The attributes that construct a choice scenario are variables

that UMCG physicians take into account when deciding what treatment is the best option for a child. The recommendations provided by the UMCG physicians on the choice scenarios entail what treatment an individual physician would prefer to recommend to the parents of the new-born based on the physicians professional and medical expertise. From the choices of the UMCG physicians on the hypothetical choice scenarios, a choice model can infer the weights, also called parameters, of decision variables on their recommendations, which provides introspection on their choice behaviour. In this study, a Binary Logit model was estimated from the choice data to estimate the parameters.

Furthermore, the choice model can be utilised for decision support. Also, a questionnaire was included to measure several professional and personal characteristics of the individual UMCG physicians. 15 UMCG physicians conducted the choice experiment.

Results

By estimating the choice model, this research can infer, the weights that the UMCG physicians attach to different attributes incorporated in the choice experiment. Comparing the weights of the variables to examine the importance of the decision variables on the recommendations of the UMCG physicians is tricky due to the different attribute ranges drafted for each attribute incorporated in the choice experiment. Therefore, the relative importance per attribute is calculated to provide introspection on the choice behaviour of the UMCG physicians. The relative importance illustrates the impact of the decision variables on the recommendation for surgery relative to each other. The results found that gestational age, the wish of parents, birth weight, the ultrasound of the brain, and the congenital co-morbidity nearly make up for 75% of the relative importance; hence, the recommendation on surgery of the UMCG physicians is primarily determined by these variables. The other nine attributes incorporated in the choice experiment have considerably less impact on the medical advice of the UMCG physicians. The variable gender demonstrates to have the least impact on the advice for a preferred treatment and portrays a relative importance of 0.01%.

Moreover, this study also researched the differences in choice behaviour of the UMCG physicians based on professional and personal characteristics. The most distinct observation found was the difference between the most impactful variables for child surgeons and neonatologists. While the variables gestational age and birth weight portray the largest relative importance for neonatologists, these variable are considerably less important for child surgeons. Whereas the congenital co-morbidity and the ultrasound of the brain are the most impactful on the recommendation for child surgeons, these variables are considerably less important for the medical advice of neonatologists.

Also, another noticeable observation was that the older and more experienced UMCG physicians more often provided recommendations against surgery compared to younger and less experienced physicians. This study found that this observation might be related to the level of confidence of the provided medical advice on the choice experiment. The UMCG physicians were also required to rate the level of confidence of their recommendations in the choice experiment. This study estimated that on average, the confidence level for the less experienced UMCG physicians was 69% while for more experienced physicians, it was calculated at 78%, which is a considerable difference. Since this research studies a very impactful dilemma, as it concerns an end-of-life decision, the lower confidence levels could explain that less experienced physicians are more hesitant to make such impactful decisions compared to more experienced physicians.

Conclusion & Discussion

This research aimed to investigate whether BAIT has potential to serve as a novel IDSS in the medical sector.

By examining the usefulness and potential of BAIT in this case study, comparing the currently deployed IDSSs with BAIT, and examining BAIT's trustworthiness, this research found that BAIT has a legitimate potential to serve as a novel IDSS in the medical sector. Nonetheless, before a new type of CDSS is implemented in an institutional environment, such as a hospital, it must also comply with many regulations and be approved by an ethical committee. These strict regulations help to prevent harm from arising to the patients impacted by a new CDSS as well as the physicians utilising the system and, hence, ensures that the principles for trustworthy and ethical AI are protected. This research also identified the bureaucracy of the institutional environment as a great challenge for the implementation of new CDSSs. Accordingly, for the successful implementation of BAIT, further research must be conducted on the legal requirements of CDSSs in health care.

Moreover, this research also identified the possible reduction of professional autonomy as a potential hurdle for successful implementation of BAIT in the medical sector. Physicians may worry that a CDSS reduces their professional autonomy as they feel they are expected to act by the judgment provided by a CDSS. A CDSS can, however, also enhance the collective professional autonomy of physicians since if experts have access to a system that enables them to support their judgments to patients and possibly third parties, when questioned about their decision, it can protect their professional autonomy. For his matter is it important that a CDSS provides explainable and transparent decision support, otherwise, the supported judgments can still not be transparently explained to patients or third parties. As BAIT provides explainable decision support, it is able to support the collective professional autonomy of medical experts. Therefore, it illustrates the trade-off between defending collective professional autonomy by limiting individual professional autonomy. The acceptance of a reduction of individual autonomy significantly differs per individual physician and the institutional environment an expert operates in. Hence, whether physicians are willing to trade off individual autonomy for an enhanced collective autonomy supported by BAIT is, yet, to be determined.

Further research

Moreover, this research provided several recommendations for further research. For the case study, it recommended conducting the same research in other hospitals on the same choice task to explore the differences and similarities between institutions. Moreover, this study advised to, also, utilise BAIT to investigate the importance of the factors that determine whether parents favour surgery or comfort care. As an improved understanding of which factors parents find most important while deliberating their wish on the preferred treatment for their child may support shared decision making. And, research shows that approximately 80% of the parents highly value shared or active decision-making and experience less regret with the enforced treatment when shared decision making is applied.

In conclusion, as explained above, as BAIT is a new IDDS approach, it requires testing in different settings to gain insight into the usefulness and effectiveness of this method. To further investigate the potential of BAIT in the medical sector, this study advises conducting more case studies to further investigate the potential and effectiveness of BAIT in the medical sector. And, ultimately, also in other sectors.

Contents

List of Figures

List of Tables

List of Appendix Figures

1 Introduction

Decision making is a fundamental process in our daily lives. Within the first hours of our day, we choose what we desire to eat for breakfast, or whether we would like to eat anything at all. Humans are consumed with making countless of such basic decisions per day. It is less often that we face complex dilemmas in need of deliberate decision making. For example, consider a business decision on whether to expand to global markets or an employee of the immigration service needing to decide whether to accept or decline the application of a refugee for citizenship. These dilemmas induce heavy decision burdens on the individuals that need to make these decisions. Moreover, we also know that individuals can make poor choices, which for high stake decisions can have great negative consequences. Hence, for decades researchers debated and continue to deliberate on how to support and assist people in making accurate and efficient decisions. In the 1970s, the term decision support system (DSS) emerged. A DSS is a computerised system designed to assist human decision making (Burstein, W. Holsapple, & Power, 2008). Since the introduction of DSSs, extensive research has been conducted to make them more advanced. The DSSs developed into a intelligent decision support system (IDSS) most commonly applied in today's society. An IDSS is an application of artificial intelligence (AI) that desires to enhance and support human decision making by enabling tasks to be performed by a computer while mimicking human capabilities (Yilmaz & Tolk, 2008).

1.1 Intelligent decision support systems

Making a decision requires the selection of a course of action between several alternatives to come to a solution for a problem (Chikwe, 2018). IDSSs desire to support decision making by employing prediction or judgement for decision support (Yilmaz & Tolk, 2008). Prediction entails utilising available information to generate information you do not have. It can, thereby, extrapolate insights from data that can help facilitate decision making. The health sector was one of the first domains that applied an IDSS for prediction purposes by employing an IDSS for medical diagnosis (Gulavani & Kulkarni, 2014). While judgement, in an environment with any degree of uncertainty, is the process of choosing a course of action based on the highest degree of expected outcome in consideration of the prediction (Agrawal, Gans, & Goldfarb, 2018). Agrawal, Gans, & Goldfarb (2019) describes the difference between prediction and judgement as follows:" while prediction can obtain a signal of the underlying state, judgment is the process by which the payoffs from actions that arise based on that state can be determined." Additionally, IDSSs exist that use both prediction and judgment by generating predictions and relying on those predictions to decide what to do next. This research will solely focus on IDSSs that use judgement to assist in decision making.

IDSSs are designed based on different methodologies. The two most generally classified types of IDSSs are knowledge-based and non-knowledge based systems (Abbasi & Kashiyarndi, 2010).

1.1.1 Current Intelligent Decision Support Systems

A knowledge-based system, also called an Expert system, was developed early on in the field of AI and uses domain knowledge as a frame of reference. The knowledge is commonly represented in the form of a concept, its intent, and the context (Ezhilarasu, Skaf, & Jennions, 2019). Figure 1.1 illustrates the components of an Expert system. An essential element of an Expert System is the knowledge base. The knowledgebase contains the domain knowledge written in the form of knowledge representation language (Abbasi & Kashiyarndi, 2010). The inference engine manipulates the domain knowledge captured in the knowledge base into a set of rules. This set of rules generates decisions and usually concern if-then statements. They can, however, also be substituted by cases for case-based reasoning, but most Expert systems employ rule-based reasoning (Aamodt, 1993). Finally, the inference engine interacts with the user interface. The user interface provides the opportunity for users to query the Expert system. The users of an Expert system can employ two actions: Ask or Tell. Ask is applied when users desire to extract information from the Expert system and utilise it for decision support. While when experts want to add new knowledge to the system, they use the action Tell (Hopgood, 2005).

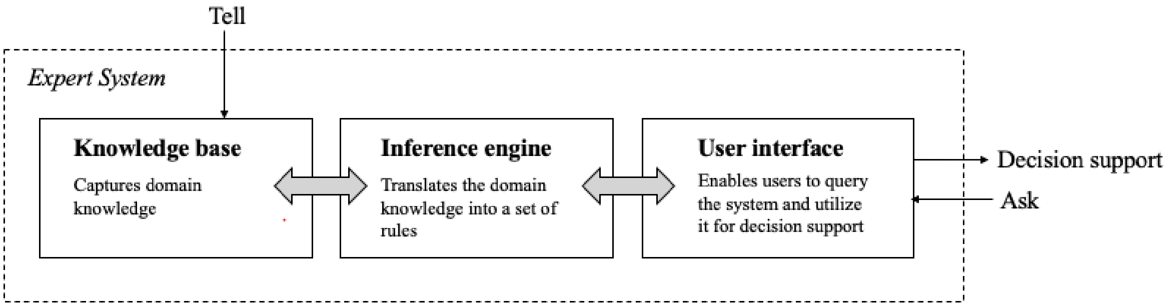

Tell
Expert System
Knowledge base
Captures domain knowledge
Inference engine
Translates the domain knowledge into a set of rules
User interface
Enables users to query the system and utilize it for decision support
Decision support
Ask

Figure 1.1: Workflow of a knowledge-based system

Contradictory to knowledge-based IDSSs, non-knowledge based IDSSs apply the rapidly growing branch of AI known as machine learning (ML). Figure 1.2 illustrates the workflow of a non-knowledge based system. Non-knowledge based IDSSs uses ML techniques, such as deep learning or super vector machines, to recognise and analyse patterns from unseen data to solve problems (Burrell, 2016). Deep machine learning, for example, applies an artificial neural network to solve problems by using nodes and weighted connections from the artificial neural network to recognise and analyse patterns (Ezhilarasu et al., 2019). To identify patterns from unseen data, the ML model is first trained by using training data. Therefore, unlike knowledge-based IDSSs, there is no need for input of experts and no necessity to write rules as input. The machine is trained by using labelled data that provides examples of desired input-output behaviour. This training data is, hence, labelled with the behaviour that the algorithm should conduct on its own to provide decision support (Sargent, 2001). Therefore, ML models base their recommendations on features that correspond to elements of training data that provided them with desired input-output behaviour. ML also allows the system to learn by including a feedback loop that re-uses the predicted outputs to train new versions of the model (Jordan & Mitchell, 2015).

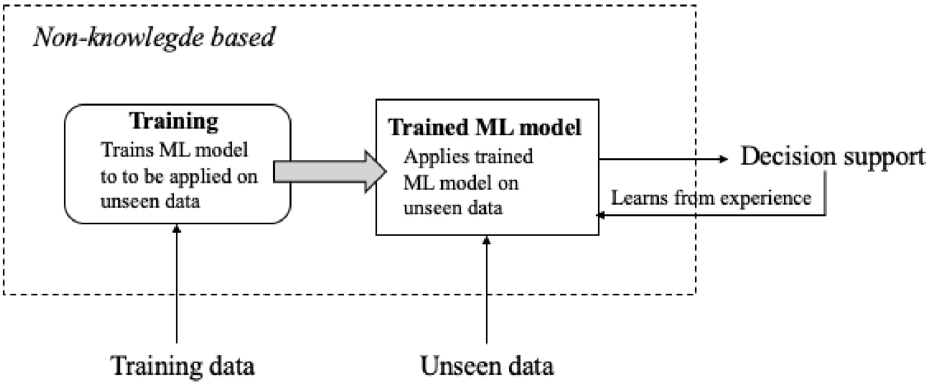


Figure 1.2: Workflow of a non-knowledge based system

1.1.2 A new approach to AI for decision support (BAIT)

The previous section described the two most generally classified types of IDSSs. Recently, a company called Councyl, in collaboration with the TU Delft, developed a new approach to AI that has the potential to constitute a novel type of IDSS for judgement purposes. The new approach to AI is called BAIT (Behavioural Artificial Intelligence Technology).

BAIT utilises discrete choice modelling (DCM) for decision support. DCM is used to analyse choice behaviour and predict future choices of individuals (Louviere, Flynn, & Carson, 2010). Through an appropriately designed choice experiment, the method can elicit individual's preferences by asking to state their choice over different choice sets. The choice alternatives captured in a choice set each contain a set of attributes. By estimating a choice model, from the observed choices, the weights that decision makers attach to different attributes can be determined (Louviere et al., 2010). Moreover, a choice model can estimate future choice probabilities for the alternative incorporated in the choice sets (Chorus, 2018). BAIT desires to apply these choice probabilities for decisions support of experts' decisions.

BAIT utilises choice modelling by asking a group of experts to conduct a choice experiment. The choice experiment reflects a decision that domain experts face in their line of work—for example, the choice of a surgeon to perform surgery. In the choice experiment, the domain experts face multiple hypothetical choice scenarios for a specific decision. Based on the decisions of the experts on these hypothetical choice scenarios, a choice model can capture the effects of decision variables on their choices, which provides introspection on their choice behaviour. After that, the choice model can be utilised for decision support and to possibly automate decisions (Van Wijnen, 2019). Therefore, the added value of BAIT does not only lie in providing decision support by utilising the choice model itself to predict future choices. Research shows that it is hard for experts to explain their logic behind decisions (Wagholikar, Sundararajan, & Deshpande, 2012). BAIT can aid in an improved understanding of experts' implicit decision-rules and behaviour, which can induce valuable discussions among experts or an enhanced understanding of how to improve decisions. Hence, the introspection by itself may already supports future decisions of experts. Figure 1.3 depicts the workflow of BAIT.

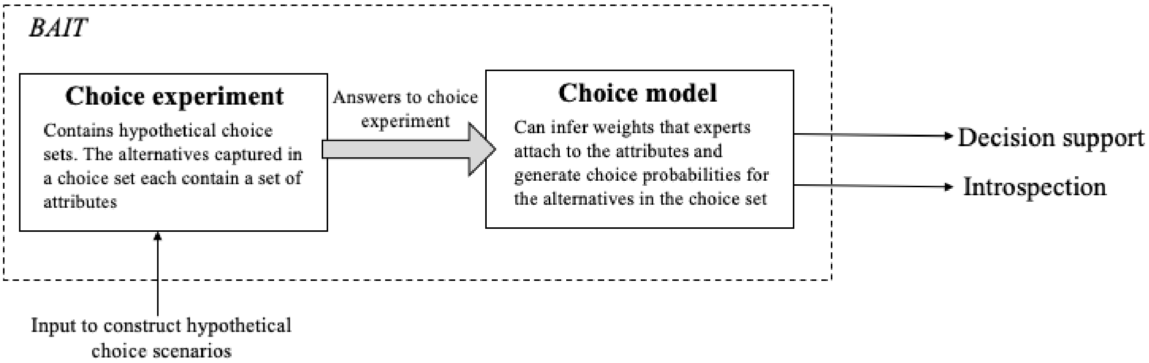

BAIT
Choice experiment
Contains hypothetical choice sets. The alternatives captured in a choice set each contain a set of attributes
Answers to choice experiment
Choice model
Can infer weights that experts attach to the attributes and generate choice probabilities for the alternatives in the choice set
Decision support
Introspection
Input to construct hypothetical choice scenarios

Figure 1.3: Workflow of BAIT

1.2 Objective

As BAIT is a new IDDS approach it requires testing in different settings to gain insight into the usefulness and effectiveness of this new method. This study will examine the potential of BAIT in the medical sector by employing BAIT at the Neonatal Intensive Care Unit (NICU) of the University Medical Centre of Groningen (UMCG). It will utilise BAIT for the choice task of UMCG physicians on whether to provide parents with a recommendation against or in favour of surgery on a premature baby diagnosed with Necrotizing Enterocolitis (NEC), given the indication that surgery is required to sustain life. NEC is a severe intestinal disease that affects premature neonates. It initiates an inflammatory process that can lead to intestinal tissue damage (Carr and Gadepalli, 2019). Section 1.3 and 4.1 provides a further explanation of the choice task investigated in this research.

As explained in Section 1.1.2, both the introspection on experts' decisions and the generated choice probabilities by BAIT have the potential to support future decisions of experts. Therefore, this research will first explore how: both the introspection of the UMCG physicians choice behaviour and the choice probabilities generated by BAIT can support future recommendations of UMCG physicians. After that, it will discuss whether BAIT can be effective in supporting the future recommendations of the UMCG physicians. This study will interpret the lessons learned in this case study to discuss the potential of BAIT as a novel IDSS in the medical sector.

1.3 The case study

This research will apply BAIT on the choice task of UMCG neonatologists, and neonatal surgeons indicated that surgery is required to sustain the life of a premature new-born with NEC, on whether to provide the parents of the new-born with advice in favour of or against surgery. A recommendation opposing operation means that the physician believes that comfort care should be initiated, resulting in the death of the child. Therefore, the choice task of the UMCG neonatologists and neonatal surgeons concerns an end-of-life (EoL) decision. In this research the UMCG neonatologists and neonatal surgeons are both referred to as UMCG physicians. Section 4.1 further elaborates on the choice task. The UMCG physicians use their judgement to decide whether surgery or comfort care is, in their opinion, the best treatment for the child. The UMCG physicians establish their recommendations on the values of decision variables that they take into account when deciding what treatment to recommend. These decision variables include personal and medical characteristics of the neonate, but, also include the physicians' opinion on the carrying capabilities of the parents and comprises the wishes of the parents on the preferred treatment. Hence, an UMCG physician establishes a final recommendation on what he or she believes is the best treatment for a neonate, however, when contemplating the advice, the perspective and capabilities of parents are taken into consideration. Although this research includes the wish of parents on the preferred treatment as a factor impacting the UMCG physicians' recommendations, it does not include a discussion on why parents favour surgery or comfort care due to time restrictions.

Moreover, the choice task focuses on what recommendation an UMCG physician will provide the parents of a premature on the preferred treatment rather than querying whether they would perform surgery: yes or no. Because in the institutional environment, surgery is not immediately executed after an individual physician voiced his or her opinion on the appropriate treatment. An explanation on the decision-making process towards the final decision for treatment is included in the discussion of this research. Conclusively, this research provides introspection on the medical recommendations of the UMCG physicians on surgery established on their own medical and professional expertise.

1.4 Research questions

This study aims to achieve its objectives. Therefore, the research questions are drafted such that answering them helps in accomplishing the goals of this study. The principal aim of this study is to explore the potential of BAIT in the medical sector. Therefore the main research question of this study is:

1. Does BAIT have potential to constitute a novel type of IDSS in the medical sector?

In order to investigate the potential of BAIT in the medical sector, this study utilises BAIT on the choice task of UMCG physicians on whether to provide parents with a recommendation against or in favour of surgery on a premature new-born diagnosed with Necrotizing Enterocolitis (NEC), given the indication that surgery is required to sustain life. This study aims to answer the following research questions specified for the case study:

2. How can BAIT support medical recommendations of the UMCG physicians on surgery?

3. Does BAIT have potential to support medical recommendations of the UMCG physicians on surgery in the future?

The answers to these research questions are interpreted and used to answer research question 1. Moreover, the potential of a new AI technology is predominately determined by user and societal acceptance for a new AI technology. Research shows that the level of trust significantly impacts AI acceptance (Gefen, Karahanna, & Straub, 2003). Therefore to determine the potential of BAIT to support future recommendations of UMCG physicians and to constitute as a novel IDSS in the medical sector, this research will include a discussion on the trustworthiness of BAIT.

1.5 Research approach

This research applies a case study approach. A case study approach allows a phenomenon to be studied in detail in a real-life context. A case study can if deliberately conceptualised and carefully carried out, yield into useful insights into the phenomenon (Stake, 2005). In this research, the specified choice task of the UMCG physicians, discussed in Section 1.3, is the case study on which BAIT is applied to retrieve insights into whether BAIT has potential to form a novel type of IDSS in the medical sector.

Moreover, a literature study is conducted to aid in answering the research questions. The literature review aids in answering the research questions by investigating the following aspects:

- What are the differences between BAIT and knowlegde and non-knowledge based IDSSs?

A thorough examination on the differences between BAIT and the currently deployed IDSSs helps to understand the system characteristics of BAIT compared to other IDSSs and further examines the characteristics of BAIT. Section 3.1 presents documentation on this matter.

- What factors impact the trustworthiness of a novel AI technology?

For an AI technology to be accepted by society and it's users the technology must be perceived as trustworthy as research shows that that the level of trust is a fundamental factor for AI acceptance (Gefen et al., 2003). An understanding of the principles and requirements for trustworthy AI, therefore, helps to determine the potential of BAIT. Section 3.2 discusses the literature on trustworthy AI.

- What supports and hinders the implementation of IDSSs in health care?

To examine the potential of BAIT in the health care sector it is essential to understand what supports and hinders the implementation of IDSSs systems in health care. Hence, researching this topic will aid in answering research question 1. Section 4.3 presents literature on this matter.

Moreover the literature research will also examine and explore literature on the case study. It will investigate the following topics:

- What are the characteristics of the choice task investigated in this research?

For BAIT to provide introspection, and possibly future decision support, on the choice task of UMCG physicians, it is essential to understand the characteristics of their choice task.

- What professional and personal characteristics of UMCG physicians may impact the medical recommendations on neonatal surgery?

Moreover, apart from providing introspection on the choice behaviour of the complete group of UMCG physicians, BAIT can also examine the differences between choice behaviour among the UMCG physicians. In order to examine the differences, it is essential to understand what personal or professional characteristics might explain differences in choice behaviour. Section 4.2 elaborates on this topic.

Additionally, as this research applies a case study approach by utilising BAIT on the choice task of the UMCG physicians in order to investigate the potential of BAIT, it also administers the research method that is inherent to BAIT's method for decision support. Hence, Discrete Choice Modelling (DCM) is applied in this research, as it is the method practised by BAIT, to codify domain expertise and support future judgments. Section 1.1.2 already provided a brief discussion on DCM. Chapter 2 will give a further explanation on DCM.

1.6 Project set-up together with the UMCG

This study is conducted in cooperation with the UMCG. To codify the domain expertise of UMCG child surgeons and neonatologists, the following research steps are executed together with the UMCG:

1. The design of the choice experiment. This project step includes interviews with the UCMG physicians to determine the decision variables for the choice task. Chapter 5 discusses the design of the survey.

2. The execution of the choice experiment by the UMCG physicians.

3. Discussion of the results, in a plenary meeting, with the UMCG physicians that executed the choice experiment.

1.7 Report outline

Figure 1.4 illustrates the report outline of this research. Chapter 2 explains DCM, as it is the method practised by BAIT, to codify domain expertise and to support future judgments. It describes how DCM is applied in this research for introspection of the UMCG physicians' choice task. Moreover, Chapter 3 and 4 present the literature review carried out in this study. Chapter 3 discusses literature on the differences between the currently deployed IDSSs and

BAIT and deliberates on the principles of ethical and trustworthy AI. Chapter 4 provides literature on the case study by discussing the choice task of UMCG physicians.
Additionally, Chapter 4 explains the possible influential personal and professional characteristics of the UMCG physicians on the choice task and provides literature on Clinical Decision Support Systems (CDSS)s. After that, Chapter 5 discusses how this research carried out the first procedure step of BAIT which is the design of the choice experiment. Furthermore, Chapter 6 presents the second procedure step of BAIT, which is the model estimation. Thenceforth, Chapter 7 presents the descriptive results of BAIT that provides insight into the choice behaviour of the UMCG physicians. After that, Chapter 8 will provide introspection on the UMCG physicians' choice task by analysing the model estimates. Subsequently, Chapter 9 provides a conclusion and discussion by answering the research questions. Finally, Chapter 10 reflects on the possible limitations of this study and includes recommendations for further research.

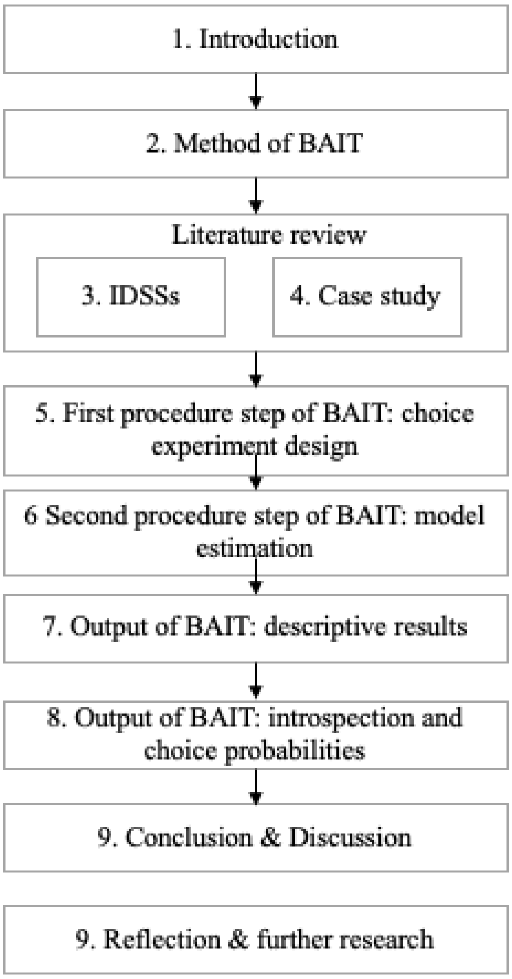

1. Introduction
2. Method of BAIT
Literature review
3. IDSSs
4. Case study
5. First procedure step of BAIT: choice experiment design
6 Second procedure step of BAIT: model estimation
7. Output of BAIT: descriptive results
8. Output of BAIT: introspection and choice probabilities
9. Conclusion & Discussion
9. Reflection & further research

Figure 1.4: Report outline

2 Method of BAIT

This chapter describes the methodology, discrete choice modelling (DCM), that BAIT applies to codify domain expertise in order to provide introspection on experts decisions and to support future judgments. This section specifies how DCM is applied in this research. Section 2.1 elaborates on DCM. Next, Section 2.2 explains how the codified domain expertise can help to provide introspection on the recommendation of UMCG physicians and how the generated choice probabilities can aid future recommendations.

2.1 Discrete choice modelling

Section 1.1.2 already provides a discussion on discrete choice modelling (DCM). DCM is used to analyse choice behaviour and predict future choices of individuals (Louviere et al., 2010). For this research, DCM is applied for introspection on the choice task of the UMCG physicians on whether to recommend against or in favour of surgery on a premature neonate with NEC, given the indication that surgery is required to sustain life. DCM can provide introspection on the choice task as it helps discover the weights of decision variables on the recommendations of UMCG physicians on surgery.

A choice experiment must be designed to elicit individuals' preferences as input for a discrete choice model. Section 2.1.1 further discusses the choice experiment and Chapter 5 describes the design of this choice experiment. Moreover, Section 2.1.2 explains how a choice model estimates the parameters, to provide introspection on the choice task, from the choice data.

2.1.1 Choice experiment

Choice modelling can utilise revealed preference (RP) or stated preference (SP) data. RP data focuses on what people did, while SP data exposes choices in a hypothetical context by using stated adaptation experiments (Abdullah, Markandya, & Nunes, 2011). RP data is retrieved by using experimental designs to compose a survey that consists of choice sets that construe two or more hypothetical choice alternatives, each incorporating several attributes (Abdullah et al., 2011). This research will apply SP data as this study desires to capture physiological reflective information about the behaviour of UMCG physicians on the choice task (Molin, 2018). It is, however, essential to keep in mind that SP data can induce hypothetical bias (Molin, 2018). Hypothetical bias concerns the question: "would an UMCG physician have made the same recommendation in real life?". This form of bias is commonly caused by the fact that in a choice experiment, perfect information is provided, which in real life is generally not the case. Another potential cause of hypothetical bias is that the consequences of the decisions are not felt (Molin, 2018).

This study designs a stated adaptation experiment to evaluate the choice preferences by employing an experimental design. An experimental design consists of hypothetical choice scenarios. For each choice scenario, the respondents are requested to make a choice. It is universally acknowledged that the more the hypothetical scenarios simulate real-world decisions, the higher the validity of the observed choices (Molin, 2010). For this research, valid choices are especially crucial since the encoded decision-rules are possibly used for future decision support. Therefore, the experimental design must seek to reflect choice scenarios that UMCG physicians face in real-life situations. The experimental design will consist of binary choice tasks, which include a context and two choice options; either provide a recommendation against or in favour of surgery. The recommendation entails what treatment an individual

physician would prefer to recommend to the parents of the new-born based the physicians own professional expertise.
The attributes that construct a choice scenario reflect variables that UMCG physicians take into account when deciding what treatment is the best option for a child. Additionally, for each choice task, the UMCG physicians are required to rate the level of certainty about their recommendation. Section 5.3.4 describes the reasons for including this question.

The attributes levels vary over the different hypothetical choice scenarios. An experimental design makes sure that enough variation between the choice scenarios exists to estimate the choice model, discussed in the next section. Moreover, the choice tasks must not exhaust the UMCG physicians as this can lead to unreliable parameters (Molin, 2018). In most stated adaptation experiments, typically, about ten choice scenarios are included. For this research, the choice tasks are, however, executed by experts. The choices they face in the choice-experiment are dilemmas that they make more often and, thus, they are assumed to be able to handle more choice tasks. This research will apply an efficient design. Section 5.3.3 provides a further explanation of the construction of the experimental design.

2.1.2 Discrete choice model

Two acknowledged choice models are Random Utility Maximization (RUM) and Random Regret Minimization (RRM). RUM is the most widely endorsed model and assumes that each decision-maker chooses the alternative that generates the outcome of which they experience the highest utility. The utility is an indicator value that determines the degree of relative content. While the RRM model assumes that a decision-maker chooses the alternative from which he or she expects to experience the least regret (Chorus, 2018).

To decide which model to apply, researchers can either estimate the parameters with both models and see which model fits the data better. Conversely, a model can also be determined beforehand based on other criteria. RUM models may be preferred as they are elegant models that are most commonly utilised, and empirical evidence shows that they perform well. Additionally, compared to RRM models, they are easier and quicker to estimate. While RRM models may be preferred when choices are perceived as difficult and when the decision-maker believes he or she will be held accountable for the decisions (Chorus, 2018).

This research consists of a binary choice task: either recommend against or in favour of surgery. For binary choice tasks, RRM is not suitable, as it requires at least three alternatives (Chorus, 2018). Therefore, this research will utilise a RUM model to estimate the parameters.

Utility function

The assumption underlying RUM models is that a decision-maker chooses the alternative from which it experiences the highest overall utility. In this research that entails that an UMCG physician chooses the treatment from which it experiences the highest utility.
The overall utility consists of a systematic utility and a random utility. For every attribute that is part of the systematic utility function, a parameter (β) will be estimated by the model. The parameter represents the weight of an attribute. By accumulating the parameter with the attribute value, it results in a contribution to the utility function. The systematic utility concerns the sum of all utilities of the attributes in an alternative.
The random utility also called the error term is considered as “noise” and cannot be predicted by the model. Therefore, even when the systematic utility is highest, the alternative might still not be chosen due to the random utility. Consequently, it is only possible to predict choices up to a probability. Equation 1 provides the linear additive utility function utilised in this research.

Equation 1: Linear additive utility function

$$U_i = V_i + \epsilon_i = \sum_m \beta_m \, x_{im} + \epsilon_i$$

Where,
i represents an alternative
m represents an attribute
U_i is the utility for alternative i
V_i is the systematic utility for alternative i
ϵ_i is the random utility for alternative i
x_{im} indicates the attribute value of attribute m for alternative i
β_m denotes the attribute weight for attribute m and is to be estimated by the model

Choice probabilities

As explained above, since the overall utility includes a random error, it is only possible to predict choices up to a probability. This research applies the Binary Logit model to calculate the choice probabilities. Equation 2 shows the binary logit formula.

Equation 2: Binary logit formula

$$p_i = \frac{e^{Vi}}{1 + e^{Vi}}$$

Where,
p_i is the probability that alternative i is chosen
Vi indicates the systematic utility of alternative i

The parameter are estimated based on the principle of maximum likelihood which is the set of parameters that make the data most likely (Chorus, 2018).

This study incorporates a binary choice task that includes two choice options; either provide a recommendation against or in favour of surgery. The systematic utility function for the recommendation in favour of surgery is computed with the linear additive utility function. The binary logit formula is used to calculate the probability for a recommendation in favour of surgery. As it is a binary choice task the probability for a recommendation against surgery is equal to one minus the probability of a recommendation in favour of surgery.

For this research, two types of models are estimated. The binary logit model will estimate parameters from the choice data of the question on the preferred treatment. Additionally, a second type of model is estimated by using the choice data on the preferred treatment and the confidence level of the recommendation of the UMCG physicians. The second model applied is a linear regression model. An explanation on why the linear regression model is estimated is explained in Section 5.3.4

2.2 How can BAIT provide introspection on the choice task of the UMCG physicians and support recommendations in the future?

This section briefly introduces how the estimated parameters by BAIT can provide introspection on the UMCG physicians choice task and considers how the choice model can be utilised to support future recommendations of the UMCG physicians. An in-depth discussion is, however, provided in the conclusion and discussion of this research.

2.2.1 Introspection

The estimated parameters determine the effect of an attribute on the utility function of a recommendation in favour of surgery. Therefore, for example, if the model estimates a parameter of 10 utils/kg for the attribute birth weight, the utility function for a recommendation in favour of surgery increases with 10 when the birth weight increases with 1 kg. The estimated parameters can be exploited to, for example, calculate the relative importance of the attributes incorporated in the choice experiment. A further discussion on how the estimated parameters can be exploited to provide introspection on the choice task of the UMCG physicians is provided in Section 8.1.

2.2.2 Future decision support

The estimated model can calculate the likelihood of a recommendation in favour of surgery. The likelihood of a recommendation in favour of operation can be interpreted as follows: for a neonate with specific characteristics, the estimated probability calculated by the model reflects the percentage of the group of UMCG physicians that would recommend in favour of surgery. In contrast, the rest of the group would advise against an operation. How this information can support future recommendations will be discussed in Chapter 9.

3 Intelligent Decision Support Systems

This chapter will present literature on Intelligent Decision Support Systems (IDSSs). As explained in Chapter 1 IDSSs are an application of AI that desire to enhance decision making by enabling tasks to be performed by computers while mimicking human capabilities. Section 3.1 will examine the differences between non-knowledge based IDSSs, knowledge-based IDSSs, and BAIT. Chapter 1 already illustrated the workflow of the IDSSs. Further analysis of the differences between the systems will help to comprehend the strengths and weaknesses of BAIT and aid in answering research question 1. Moreover, Section 3.2 will present literature on trustworthy AI to comprehend how to achieve users' acceptance for the new AI technology.

3.1 Comparison of the current IDSSs and BAIT

This section will examine the differences between the currently applied IDSSs and BAIT. Table 3.1 at the end of this section, provides an overview of the characteristics of the two most generally classified types of IDSSs and BAIT.

3.1.1 Comparison between the currently deployed IDSSs

As elucidated in Chapter 1, knowledge-based IDSSS, also called Expert systems, provide decision-support by directly translating domain knowledge into a set of rules or cases. While non-knowledge based systems apply ML learning that grounds decision-support on feature extraction of labelled training data. Both Expert and non-knowledge based IDSSs have strengths and limitations. The process of directly translating domain knowledge into decision rules leads to high transparency and interpretability of decisions generated by Expert systems (Waltl, Bonczek, & Matthes, 2018). In contrast to non-knowledge based systems that apply a so-called "black-box" technology. The black box characteristic entails the difficulty or sometimes even impossibility to explain the reasons for specific system outcomes (Burrell, 2016). This black box technology forms an obstacle for the interpretability and transparency of the systems' generated recommendations.

A drawback of a knowledge-based system is the inability to self-update. The missing learning capability implies that when the system encounters a problem for which no rules are designed, the system is not able to provide decision support. Therefore, the system is limited to its underlying rules. Consequently, if the system desires to offer solutions to new problems, additional rules must be added or removed, and this can be a stringent process. New rules might be conflicting with already existing rules, and eliminating rules might impact the operation of the entire system. Accordingly, the maintenance of an Expert system is a time-consuming and challenging process (Ezhilarasu et al., 2019). For these reasons, it is hard for Expert systems to deal with complex problems in dynamic environments because translating and adjusting rules is a challenging task (Prentzas & Hatzilygeroudis, 2007). The speed of execution of getting knowledge-based systems up and running is usually faster compared to non-knowledge based systems because of the need for training an ML model (Sargent, 2001).

Non-knowledge based IDSSs are most commonly utilised for complex problems in dynamic environments due to their competence to deal with a large number of features (Berner & La Lande, 2007). Moreover, the feedback loop in ML ensures that a non-knowledge based system is not restricted by underlying rules and does not require manual maintenance. In contradiction, the feedback loop provides the ML model with learning capabilities and keeps improving the model over time (Jordan & Mitchell, 2015).

Additionally, ML is greatly valued by its method to organise and search for patterns in big data sets to generate fast and precise decisions. In addition, ML systems are based on a probabilistic approach, which means that the system generates probabilities as output for decision support (Burrell, 2016).
ML, however, does need an extensive amount of training data to perform on unseen data adequately. The more complex a problem gets, the more training data is required to make the results viable. Thus, if data is unavailable, applying ML can be problematic. Also, if the data is available but lacks quality, the results on unseen data will be inadequate because feeding a model poorly will provide meaningless results (Jordan & Mitchell, 2015).
Moreover, another limitation of ML learning is called concept drift. As explained in Chapter 1, ML models base their recommendations on features that represent desired input-output behaviour, which could induce problems because unseen data might not always be generated in the same way as training data does. That means that ML models might not be able to manage changes in input data type, and therefore cannot be generalised (Ravi, 2020).

Additionally, due to the expanding application of AI systems in sensitive areas, such as health care, a discussion emerged about AI bias. Unconscious social and individual inclinations commonly shape human decisions. AI has the potential to reduce human decision-making bias, but can also induce system bias. The definition of bias referred to is: "the systematic discrimination against groups or individuals based on inappropriate use of traits or characteristics" (Friedman & Nissenbaum, 1996). Knowledge and non-knowledge based IDSSs are prone to bias.

As mentioned earlier in this section, the potential to directly transfer knowledge into rules is an advantage for a knowledge-based system because it enhances the interpretability of the system's generated decisions. Directly translating knowledge into rules, however, also makes the system prone to bias as the output of the system is only as good as the expert judgments, and human decisions are known to contain implicit bias. Implicit bias entails that humans have pre-judgments towards groups or individuals without conscious knowledge. Research shows that personal characteristics influence the inherent bias of individuals (Silberg & Manyika, 2019). Additionally, an Expert system's straightforward approach in translating domain knowledge into decision rules may provoke intuitive results because it tends to be challenging for domain experts to explain their logic for obtaining specific decisions. Sometimes they may not even understand factors that influence their reasoning, which leaves room for unconscious bias (Wagholikar et al., 2012).

Non-knowledge based systems aim to reduce the bias of human decision making by training the ML algorithm to only consider variables that improve the decision-making capabilities of the system based on training data. Evidence shows that this method has the potential to reduce bias and increase the fairness of the decision-making process (Miller, 2018). ML can, however, also deploy unnoticed bias at a massive scale. The training data can unknowingly contain implicit racial or ideological biases. For example, recently, a technology company designed a hiring ML model that prejudiced women. It explored subtle correlations associated with women and dismissed them (Ming, 2019). Manyika et al,. (2018) states that minimising AI bias is critical for AI systems to reach their full potential and is essential for trustworthy AI.

Today, non-knowledge based IDSSs are commonly preferred over knowledge-based systems because of their learning capability and its method to organise and search for patterns in big data sets. The decision on an IDSS still, however, highly depends on the problem you want to solve. For example, when a problem does not require a firm understanding of the model, and enough qualitative training data is available, ML is preferred. However, when limited qualitative training data is available, and the problem context requires decisions to be thoroughly explained, ML is not suitable, and a knowledge-based system will be preferred.

3.1.2 Comparison between the currently deployed IDSSs and BAIT

The previous section discussed the characteristics of the current IDSSs systems and elucidated their limitations and strengths. The preference for one method over another depends on the kind of problem and the required decision. This section will elaborate on the differences between BAIT and non-knowledge and knowledge-based IDSSs.

For BAIT, similar to knowledge-based systems, knowledge of experts is the frame of reference. However, instead of directly translating domain knowledge into rules, the new approach studies the choices of experts. From these choices a choice models captures the expertise of experts for specific decisions and the model can eventually be enforced for decision support. Although the knowledge of experts is the frame of reference for decision support of both BAIT and knowledge-based systems, the way the systems' support decisions is different. Whereas BAIT grounds it's decision support on the trade-offs that experts' make between decision variables captured by the decision-rules, a knowledge-based system does not compromise the values of different decision-variables. Knowledge-based systems apply if-then-else statements for which clear threshold values determine whether a decision follows one path or another path. Also, as explained in Section 2.2.2, BAIT provides decision support by estimating the likelihood for a specific decision or outcome. The estimated probability represents the percentage of experts that would generate that particular decision or outcome. Whereas, a rule-based system provides decision support by offering the solution that the system believes is appropriate based on the if-then statements.

Moreover, similarly to non-knowledge based systems, BAIT establishes recommendations based on past choices of experts. ML, however, just grounds its recommendations on the characteristics of those choices rather than understanding the reasons for decisions. For example, when doctors use an ML model to decide on the best treatment for their patient, the model cross-references similar patients' data and compares the procedures and outcomes. This way, it can discover which characteristics indicate that an individual will have an appropriate response to treatment and, thereby, predicts the best response (treatment) for a patient. While BAIT also incorporates an understanding of why experts make choices. The weights for decision variables elicited from the choices establish this understanding. Thus, in the example explained above, the BAIT can explain why doctors chose a specific treatment. Therefore, this new approach to AI does have an understanding of cause effects compared to classic ML models that do not as it's a black box technology.

The extracted decision-rules by BAIT from experts choices, similarly to Expert systems, enhances the transparency and interpretability of decisions generated by the system. Similar to knowledge-based systems, the decision-support is, however, limited to its encoded decision rules captured by the choice model. Hence, in the future, BAIT desires to incorporate a feedback loop, that feeds back the lessons learned from decisions that it has supported. An analysis of the possibility of including a feedback loop is, however, outside of the scope of this research.

Furthermore, as mentioned in Section 3.1.1, for IDSSs to reach their full potential, it is crucial to minimise the systematic bias of the systems. Minimising system bias will maximise the fairness of the decision-making process and is vital for enabling users to trust the system (Silberg & Manyika, 2019). BAIT provides the opportunity to tackle system bias. Silberg et al.

(2019) describe that one way to address system bias it to "engage in fact-based conversating about potential bias in human decisions." BAIT can engage in faced-based conversating about potential bias as the choice model analysis the choice behaviour of experts. This provides an understanding of cause-effects and can, thereby, discover human decision-making bias. For example, in the case mentioned in Section 3.1.1, in which a hiring ML model prejudiced women, BAIT can identify such bias at an early stage because the estimated parameters for the decision variable from the choices of experts will recognise that there is an inclination for not hiring women. Hence, BAIT has the potential to make implicit human decision-making bias; explicit and filter out unwanted prejudgements.

DCM can, however, induce other potential forms of bias, as mentioned in Chapter 2, choice experiments can cause hypothetical bias. Hypothetical bias concerns the question of whether decision-makers would make similar choices in real life as compared to the choices they make in the stated adaptation experiment (Molin, 2018).

Furthermore, to construct the choice experiment, the input of experts is required to compose the choice scenarios. Therefore, the constructed choice experiment already contains choices of experts about which decision variables to include in the construction of the choice scenarios. Sometimes experts are unaware of factors that influence their decisions. Therefore, the elements included in a choice experiment, and results obtained from the choice experiment, are limited to factors that the specific group of experts find essential and can apprehend. Additionally, the number of attributes included in the choice scenarios is not infinite. When individuals face too many attributes, they are unable to make trade-offs between all attributes. In choice-experiments, the number of attributes included is usually limited to 7; however, since experts have extensive knowledge on the choice scenarios they face in a choice-experiment, expected is that they can handle a few more.
Nonetheless, it is expected that when they encounter more than 15 to 20 attributes, it is unlikely that they can make trade-offs valuations between all attributes. Consequently, the system is limited to the number of attributes, of which experts can still successfully conduct trade-off valuations between all attributes. Because adding more attributes will not provide valuable information as the experts will not include an attribute for their decision and will, thus, not provide information for introspection or to aid decisions in the future.

In conclusion, each decision-support system relies on distinct types of input data, manipulates the data differently and generates different output to support human decisions. Non-knowledge based systems are dependent on the quality of big unstructured data, and its probabilistic approach entails that it generates probabilities as output for decision support. In contrast, knowledge-based systems are dependent on structured data captured in a knowledge base. The knowlegde is translated into rules, which determine whether a decision follows one path or another path to reach a true value used for decision support.
Finally, BAIT self-generates data by designing a choice experiment that reflects a decision that domain experts face in their line of work. The answers to the choice experiment are used as input data for a choice model. The choice model manipulates the data to infer the weights that experts attach to the decision variables, and the generated choice probabilities that enclose the trade-offs that experts' make between decision variables can be used for decision support.

Table 3.1 provides an overview of the characteristics of the two most generally classified types of IDSSs and BAIT.

Table 3.1: Systems' characteristics

Characteristic	*Knowledge-based*	*Non-knowledge based*	*BAIT*
Operation	Directly translating domain knowledge into rules or cases for decision-support. Hence, clear threshold values determine the decision path.	Decision-support grounded on feature-extraction of labelled training data.	Utilises choice modelling to extract decision-rules for decision-support. The decision-rules enclose the trade-offs that experts' make between decision variables.
Type of input data	Dependent on structured data captured in a knowledge base	Dependent on unstructured big data	Dependent on self-generated choice data for a specific domain decision made by experts
Output	A true value	A probability	A probability
Transparency and decision-traceability	High, due to its comprehensive decision-rules	Opaque due to its black box technology	High, due to its encoded decision rules and cause-effect understanding
System bias	Implicit and unconscious human-decision making bias	Data bias	Hypothetical bias, but has the potential to make human decision-making bias explicit
Flexibility	Limited as the system cannot recognise beyond rules due to its inability to self-update	High as it can adapt to subtle cases but still defined as the model might not be able to manage changes in input data type and its ability to learn due to its feedback loop	Moderate as the model cannot recognise beyond the encoded decision-rules captured from experts' choices but the system has the opportunity to incorporate a feedback loop in the future
Ability to deal with complex problems	Low as designing rules for complex problems is a very challenging task	High due to its competence to deal with a large number of features	Moderate as it does not require the task of manually designing rules, but the system is limited by the number of attributes a choice experiment can capture

3.1.3 A need for IDSSs that can support ethical dilemmas

The Dutch Organization for Scientific Research (NWO) describes that for the development of future AI technologies, ethical and legal issues must be considered. By virtue of the fact that the AI technology must eventually function in our society that consists of social norms and values. Today multiply researchers are investigating ways to establish ethical and explainable AI. Currently, most efforts are, however, being carried out by Computer Science experts. NWO believes that more efforts from social and behavioural sciences are required to establish the evaluation of ethical AI technology ("Kunstmatig," n.d.).

BAIT is an example of a new AI technology that originates from behavioural science. BAIT applies choice modelling that for years has been practised to investigate moral and ethical considerations. Hence, BAIT can be an appropriate technique for decision support in sensitive areas, such as health care or criminal justice, in which explainable and ethical decision-making processes are crucial.

Furthermore, Van Harmelen, an AI professor at the University of Amsterdam, believes that AI technology will evolve to be considered as a colleague of humans rather than a replacement of humans. He believes that collaboration between human and the machine will produce improved results compared to individually conceived results by either the AI technology or a human. Van Harmelen explains that for AI to serve as a possible colleague of the humankind, the technology should be able to interpret and utilise the knowledge of humans. Today, NWO is conducting research to facilitate a competent collaboration between humans and AI technology (*Waar blijft de mens?* n.d.).

BAIT is an AI technology that has the potential to serve as a colleague for experts. By using the decision-rules captured by a choice model as the frame of reference for decision support, it interprets and utilises the knowledge of colleague experts. Utilising BAIT to aid decisions, thus, can be considered as asking a colleague for advice as the recommendations provided by BAIT are grounded on decision-rules from the experts themselves.

Recently, BAIT has been granted the take of subsidy of the NWO for building Human-inspired decision systems for Artificial Intelligence.

3.2 Trustworthy AI

The level of trust determines an individuals' behaviour towards AI, and research shows that the level of trust is a fundamental reason for AI acceptance (Gefen et al., 2003). Trust can be defined as a combination of trusting beliefs and trusting intention. Trusting beliefs concerns trusting the system's competence and integrity, and presuming that the system will behave as it promises. Trusting intention concerns the consent to trust the system in possibly risky situations (Siau, 2018).

Moreover, the impressive developments in the field of AI that constitutes significant opportunities but also induces risks have started a debate on the trustworthiness of AI. Numerous studies, therefore, investigate the conditions and requirements under which trust in AI is generated and reduced (Danks, 2019). The increased reflection on how to develop trustworthy AI resulted in several high-profile initiatives that drafted a set of guidelines for reliable AI.

For example, a high-profile initiative that drafted guidelines for trustworthy AI is the Montreal declaration. In November 2017, the University of Montreal put into motion a declaration for the trustworthy development of AI that resulted in multiple events and conferences on the responsible application of AI. The first principles included in the declaration were: well-being, autonomy, justice, privacy, knowledge, democracy and responsibility ("Context - Responsible AI Declaration," n.d.).

Over the past years, multiple similar high-profile initiatives evolved. All these initiatives aim to support and expedite the development of trustworthy AI. Some of the guidelines and principles designed by those initiatives overlap while others differ. The expanding number of proposed principles contribute to the evolution of trustworthy AI, but can also confuse because of the lack of regularity in those documents. Therefore, Harvard University conducted a comparative analysis of several high-profile initiatives (Floridi & Cowls, 2019). The comparative analysis resulted in a framework of five principles that considered the ethical implications of AI to establish trustworthy AI. The five principles included in the framework are:

1. Beneficence: The AI technology should prioritise the well-being of both planet and people.
2. Non-maleficence: AI should prevent harm from arising whether it's by intent or due to unpredicted behaviour of the AI technology.
3. Autonomy: Humans should always retain the power to decide which decision to take to protect the intrinsic value of personal choice.
4. Justice: The AI technology should correct unfair discrimination, ensure that the AI benefits are sharable and prevent new harms from arising.
5. Explicability: The AI technology must answer to the questions "How does it work" and "Who is responsible for the way it works".

AI4people, Europe's first global forum on the social impact of AI, adopted this framework to conceive twenty recommendations for a "Good AI Society" presented to the European Commission (Floridi et al., 2018). The European Commission embraced the twenty recommendations of AI4people that on April 8 2019 published a document drafted, by a group of high-level experts, on the ethics guidelines for trustworthy AI ("Ethics guidelines for trustworthy AI | Shaping Europe's digital future," n.d.).

Trust over time

Moreover building and maintaining trust for AI technology is a dynamic process. It involves a graduate alteration from an initial trust to continuous trust. Initial trust helps to tackle initial conceptions of uncertainty and risk (Li, Hess, & Valacich, 2008). Siau (2018) demonstrates multiple factors that impact initial trust formation. The article explains that the explainability and trialability of a new AI technology are crucial for initial trust-building. Trialability entails that the users of the technology must have the opportunity to access and try out the new AI technology to help establish a high initial trust level.

Moreover, explainable AI that can justify its procedures and conclusions helps enhance the initial level of trust for a new AI technology. Finally, the article states that representation, also, plays an essential role in initial trust-building. A technology that mimics the behaviour of humans helps to build a faster 'emotional' connection with the AI technology, and that way enhances the trust level (Siau, 2018).

Furthermore, continuous trust for an AI technology primarily depends on the performance of the AI technology. The technology must be reliable and accessible to enhance the level of trust. Also, data security plays a vital role in trusting a new AI technology (Kusumasondjaja, Shanka, & Marchegiani, 2012). Lastly, users tend to trust AI technology more when the system works in partnership with the user rather than independently taking over tasks. This reduces the fear of job replacement that hinders continuous trust for new AI technology (Siau & Shen, 2003). Hence, it also supports the argument of professor Van Harmelen explained in Section 3.1.2. He states that for AI technology to evolve, the technology must be considered as a colleague of humans rather than a replacement of humans.

As mentioned earlier in this section, the level of trust is a fundamental reason for AI acceptance, and without trust, it may hinder the uptake of an AI technology (Gefen et al., 2003). Additionally, trustworthy AI is a precondition for the responsible and ethical application of an AI technology which is especially essential for the implementation of AI in sensitive areas such as the health sector. This research will provide recommendations on how best to utilise BAIT such that initial and continuous trust is established.

4 The case study

This chapter discusses the components of the clinical setting of this case study. Section 4.1 presents a further explanation on the choice task, and Section 4.2 describes the professional and personal characteristics of physicians that may impact the medical recommendations of the UMCG physicians. Thenceforth, Section 4.3 examines Clinical Decision Support Systems (CDSSs), which is the term for IDSSs in health care.

4.1 Choice task

This research will examine the dilemma of UMCG physicians, given the indication that surgery is required to sustain life, on whether to recommend against or in favour of operation on a new-born diagnosed with Necrotizing Enterocolitis (NEC). NEC is a severe intestinal disease that affects premature neonates. It initiates an inflammatory process that can lead to intestinal tissue damage. The prevalence of NEC is 14% of the new-borns who weigh less than 1 kg (Carr & Gadepalli, 2019). Over the past years the incidence of NEC in the Netherland also significantly increased. This increase seems to be related to the new Dutch guidelines for active treatment of extremely premature new-borns with 24 and 25 of gestation (Heida et al., 2017). Additionally, the estimated death rates of neonates with NEC ranges from 20% to 30% with the highest death rates for new-borns that require surgery (Neu & Walker, 2011). Surgery, also, induces risks of severe neurological and physical deterioration (Rees, Pierro, & Eaton, 2007). The Dutch nation-wide approach is when a treatment is medically futile; the procedure should be stopped to prevent unnecessary suffering (Verhagen, Van Der Hoeven, Van Meerveld, & Sauer, 2007). The UMCG physicians, therefore, base their decision for surgery on the expected quality of life, which entails the value of a neonate's life related to present and future capacities. A surgeon, thus, faces the ethical question: I can perform neonatal surgery, but should I?

This is a very challenging and burdensome dilemma, primarily, because it concerns an end-of-life (EoL) decision. In the Netherlands, EoL decisions on neonates are carefully deliberated on, and must always be in the best interest of the child. Consequently, EoL decisions are only contemplated when the quality of life is expected to be so bad that continuing procedure will cause unnecessary suffering for the child. If unnecessary suffering is expected palliative or, also called comfort care, is initiated. Palliation wishes to make the neonate feel more comfortable and relief the suffering of the child, but does not desire to cure the new-born.

The reasons that make the dilemma of physicians to withdraw or proceed with surgery on a child with NEC very challenging are related to the components of a “wicked” problem. Wicked problems are commonly described as open-ended, which means that there is no endpoint or single “solution” for the dilemma. Additionally, for these problems, the essence of the ‘problem’ and the preferred ‘solution’ are strongly questioned (Hensher, Rose, & Greene, 2005). Hisschemöller & Hoppe (2018) explains that a wicked problem consists of two components. The first component concerns a scarcity of knowledge regarding the nature of the problem and consequences of the solution, which induces uncertainties. The second component involves divergence of perceptions and values for the issue and preferred solution. The dilemma of physicians to withdraw or proceed with surgery on a child with NEC possesses a certain degree of both elements.

Firstly, a scarcity of knowledge exists regarding the development of the health of a premature neonates, especially after surgery. Acknowledged is that surgery induces an increased risk of neurological and physical deterioration. Primarily neurodevelopmental outcomes are known to be weak after surgical treatment (Robinson et al., 2017). The exact progress of the health of the child after surgery is, however, very challenging to determine, this concerns short term outcomes, but especially long term outcomes, years after surgery. Currently, there is a lack of information on clinical parameters that can predict the progression of the health of the child after surgery. Therefore, further progress in this area is required to predict poorer outcomes after surgical procedures on neonates with NEC (Henry & Moss, 2005). To determine the expected quality of life, with a lack of knowledge regarding the development of the child's well-being , is very challenging. Therefore, the dilemma of physicians on whether to operate a new-born diagnosed with NEC becomes even more difficult.

Besides the difficulty of medically defining the quality of life after surgery, another factor complicates the task of determining a decent quality of life because the norm "quality of life" is implicit. Neonates cannot express the extent of their suffering, and possible adverse developments of the disease only become apparent at an older age. So how can an individual determine what a decent quality of life is for another human being? This is a burdensome quarrel for professionals and is driven by past experiences and personal norms and values which leads to controversies between institutions and individual surgeons on the decision for surgery on neonates with NEC (Carr & Gadepalli, 2019b). Section 4.2 discusses the personal characteristics variables of professionals that may impact EoL decisions.

Undoubtedly, physicians are not the only ones devoted to making the appropriate decision for surgery concerning new-borns with NEC; parents are too. Parents are invested in their child's well-being and may be doubtful about the recommendations of doctors for the treatment of their child, especially in a situation in which their child's life is in danger. While doctors might advise parents against surgery because of severe medical implications, parents often want to keep their child alive at all costs (Boland et al., 2019). The preference of parents for surgical procedures is also significantly impacted by personal norms and values. Hence, conflicting preferences may exist between parents and professionals, but as explained above, even between professionals themselves.

Conclusively, due to the reasons explained in the previous paragraphs, the decision of neonatologists and neonatal surgeons on surgery for a new-born with NEC is a complex and ethical dilemma that induces a heavy decision burden on the professionals.

4.2 Influential personal characteristics on choice task

Previous research shows that the judgements of physicians for EoL in intensive care units vary between specialists. Factors that influence the EoL decisions of physicians range from religious affiliation, culture, and geographical region to the personal characteristics of doctors (Sprung et al., 2007).

Religious affiliation is progressively identified as an essential factor that impacts physicians' care for seriously ill patients. Research shows that religion affects the importance that physicians attach to patients' wishes. Lawrence and Curlin (2009) found that highly religious specialists tend to attach less significance to patients' preferences compared to non-religious physicians. Additionally, the impact of religion on EoL decisions shows significant differences for different types of EoL decisions and patients (Sprung et al., 2007).
Furthermore, various studies on the impact of religious affiliation on EoL decisions provide conflicting results (Chakraborty et al., 2017). Studies on EoL decisions in NICU's also portray these contradictory results. A study that researched the effect of the religious affiliation of American neonatologists on the care for high-risk neonates observed no significant impact. For the majority of the neonatologists that participated in the study, the choice for treatment could

not be associated with religious beliefs (Donohue, Boss, Aucott, Keene, & Teague, 2010). Contradictory, another study that examined the impact of religious beliefs of a group of European neonatologists' on EoL decisions provides different results. This study found that religious affiliation significantly impacted neonatologists behaviour. The results found that non-religious and Protestant neonatologists primarily based their EoL decisions on the quality of life expected for the neonate. While neonatologists with other religious backgrounds preferred treatment that preserved life at all costs, also called a pro-life attitude (Rebagliato et al., 2000).

Moreover, a critical factor that influences EoL decisions is the cultural norm within or across countries. An increased number of studies exhibited that the way physicians make EoL decisions and communicate with the patient and patient's family is impacted by social and moral values (Blank, 2011). The study of Rebagliato et al. (2000) also identified that cultural norms and values were significant factors that affected neonatologists' attitudes on EoL decisions. The research examined that Dutch and British neonatologists favoured decisions based on the quality of life. While other European countries such as Hungary and Italy exhibited a pro-life attitude due to societal norms and values. The study also relates the pro-life perspective of these countries to the more significant influence of religion in these nations.

Additionally, multiple studies investigated the influence of physicians' personal and professional characteristics on EoL decisions. The most common questions asked in research that explored the impact of personal characteristics on EoL decisions are; the physicians' age, gender, religious background, already discussed above, and whether the doctor has children of his or her own. The most frequently asked question on professional practices concerned the number of years of professional experience (Cuttini et al., 2000; Dombrecht et al., 2020; Donohue et al., 2010; Rebagliato et al., 2000; Sprung et al., 2007). The results of the impact of personal and professional traits on EoL decisions, similar to religious affiliation, portray different results for various studies (Chakraborty et al., 2017).
For example, a study that researched EoL decisions in NICU's in seven European countries that whether the neonatologists had a child showed no significant impact on EoL decisions. Contrary, both the length of professional experience and age did portray a considerable correlation with EoL decisions. Older and more experienced neonatologists were more likely to make EoL decisions compared to younger and less experienced neonatologists (Cuttini et al., 2000). On the other hand, a similar study, published in the same year, that examined EoL decisions of neonatologists in NICU's in ten European countries obtained different results. This study demonstrated that the gender of physicians' and having children significantly influenced EoL decisions as the study found that these factors, as well as the length of professional experience, impacted neonatologists' attitudes towards EoL decisions. Female neonatologists and doctors without children more commonly based their EoL decisions on the quality of life rather than illustrating a pro-life perspective compared to men and physicians with children.

To summarise, understanding how personal and professional characteristics impact EoL decisions of doctors is a complicated matter as studies provide deviating results. Establishing universal agreements on how personal and professional characteristics impact EoL decisions around the world, therefore, remains elusive. This research will investigate, whether, in this specific sample of physicians, the personal and professional characteristics play a role regarding the recommendation for surgery on neonates with NEC. Section 5.4.1 discusses the personal and professional characteristic variables included in this research.

4.3 Clinical decision support systems

In health care, IDSS systems are called Clinical Decision Support Systems (CDSS) (Osheroff et al., 2007). The selection for a type of CDSS design depends upon parameters such as data availability, the cost of the system, the efficiency required and the complexity of the problem (Abbasi & Kashiyarndi, 2010). Over the past years, the non-knowledge based CDSSs have become more popular compared to knowledge-based CDSSs (Kwiatkowska, Atkins, Ayas, & Ryan, 2007). Predominantly, because knowledge-based CDSSs are practically unfeasible to deal with the high complexity problems in healthcare (Wagholikar et al., 2012). Non-knowledge based CDSSs can provide patient-specific evidence-based advice by recognising and analysing patterns in the Electronic Medical Records System (EMRs). Evidence shows that CDSS can help deal with an overload of complex clinical information, generate an avoidance of treatment errors and bring about practice improvement (Jaspers, Smeulers, Vermeulen, & Peute, 2011).

Regardless of the promising evidence on CDSSs, it still does not ensure their uptake by institutions and even if they are employed the physicians themselves often neglect their support (Heselmans et al., 2012). Understanding what hinders and supports the application of CDSSs is crucial for more effective implementation of CDSSs.

It is vital to understand the physicians perspective on CDSSs. Physicians worry that a CDSSs may reduce their professional autonomy and believe that a CDSS could be used against them when medical differences arise. Research shows that an increased sense of control over an CDSS helps to reduce these worries (Liberati et al., 2017). Moreover, technical and usability problems may form a barrier to the implementation of CDDSs. A wrong understanding of the technical obstacles of CDSSs might also prevent the use of the system. However, since the professionals' preferences and perspectives do not impact these barriers, they are likely easier to solve with sufficient technical support and a thorough explanation about how to use the system (Heselmans et al., 2012).

Moreover, most CDSSs do not include and reveal the decision-making processes of the professionals themselves which is an extra reason for physicians to be hesitant about CDSSs leading to suboptimal implementation (Khairat, Marc, Crosby, & Al Sanousi, 2018). Testing it with the physicians and observing it's use is essential for prosperous implementation and increases the acceptance of physicians towards CDSSs (Berner & La Lande, 2007).

If the UMCG desires to implement BAIT for future decision support, these barriers must be considered for an optimal implementation. As the recommendations of BAIT are based on the decision-rules of the group of UMCG physicians, it may already increase the acceptance of the system by the group of professionals because the recommendations are based on their own expertise. The discussion of this research that determines whether BAIT works in this context will consider what support and hinders the uptake of CDSSs.

5 First procedure step of BAIT: choice experiment design

To elicit UMCG neonatologists and neonatal surgeons' preferences on the recommendation for surgery and to investigate possible heterogeneity between UMCG physicians, a survey is designed. This chapter illustrates the design of the survey. Section 5.1 provides the structure of the survey, and Section 5.2 discusses the design approach. After that, Section 5.3 and 5.4 elaborate on the design of the components of the survey. Moreover, Section 5.5 discusses the sample group for this survey. The final survey can be found Appendix I.

5.1 Structure

The survey will consist of the following elements:

- An introduction to the survey. Before the UMCG physicians conduct the stated adaptation experiment and questionnaire, an introduction for the survey is provided. (Section 5.3.1)
- The stated adaptation experiment. (Section 5.3)
- A questionnaire on personal characteristics to investigate taste heterogeneity between UMCG physicians and questions about how the UMCG physicians experienced executing the stated adaptation experiment. (Section 5.4)

5.2 Design approach

Four UMCG physicians were involved in the process of designing the survey. This group included two UMCG neonatologists and two UMCG neonatal surgeons.

For this research, the experimental design must reflect the real-life choices of UMCG physicians. On the other hand, the construction of the experimental design must limit the correlations between attributes for the choice model to accurately estimate reliable parameters given the limited number of experts that will conduct the survey. Therefore, a trade-off must be made between including more attributes, levels, and constraints to make sure that the choice scenarios reflect real-life choices of UMCG physicians and secure that the design enables the choice model to estimate reliable parameters without effects that cannot be explained. This trade-off was carefully taken into consideration during the design of the survey. As stated in Section 2.1.1 to construct the choice sets, efficient design is applied. An efficient design requires a pilot study to obtain priors. Priors are the best guesses for the parameters (Molin, 2018). Therefore, first, a pilot survey will be designed of which the priors obtained from the pilot study are incorporated as input for the final survey. Section 5.3.3 provides a further explanation of the efficient design.

The construction of the pilot survey was conducted in four phases:

- Phase 1: Individual semi-structured interviews.

In phase 1, four individual interviews were conducted with the UMCG physicians that assisted in the design of the survey. During these interviews, the goal of this research and the required elements necessary to construct the experimental design were discussed. Appendix A illustrates the content of these interviews. At the end of each meeting, the UMCG physicians were asked to provide a list of attributes, ranges, and attribute levels. Appendix B incorporates these lists.

- Phase 2: Plenary meeting to compose a final list of attributes

The four lists of the UMCG physicians were merged into a single list. Phase 2 consisted of a plenary meeting with all the involved UMCG physicians to discuss the combined list of attributes and reduce the lists to several attributes and attribute levels that could be used to compose a prototype survey. Appendix C provides a discussion on this meeting.

- Phase 3: Discussion of the prototype survey

Based on the final list of attributes and attribute levels, a prototype survey was designed and sent to the four UMCG physicians. Phase 3 consisted of a plenary meeting that discussed the prototype survey. Based on this discussion, modifications to the survey were made and presented to the UMCG physicians. After that, through iterative email contact, the prototype survey was finalised. Finally, the survey was presented to Dr. Eric Molin and Prof. Dr. Caspar Chorus for methodological feedback, and the final experimental design was constructed. Appendix D incorporates the feedback given on the prototype survey.

- Phase 4: Discussion of the final (pilot) survey

In phase 4, a few choice scenarios of the final survey were simultaneously executed with one of the UMCG physicians to confirm that the design was qualitatively constructed. This meeting focused on whether the ranges of the attributes forced the experts to make trade-offs between attributes. Thus, to avoid incorporating attribute levels that would constitute a definite “yes” or “no” for surgery among the entire group of experts. Additionally, the meeting intended to discover whether the physician believed the experiment was ready to be executed by his or her UMCG colleagues. Appendix E illustrates a discussion of this meeting.

The prototype survey, pilot, and final survey are constructed with software engineered by Wem.io.

5.3 Stated adaptation experiment

This section will discuss the construction of the stated adaptation experiment. Section 5.3.1 deliberates on the introduction of the survey. Moreover, Section 5.3.2 elaborates on the selection of attributes and attributes levels for the choice experiment, and Section 5.3.3 discusses the construction of the choice scenarios. Finally, Section 5.3.4 describes the questions posed in the stated adaptation experiment.

5.3.1 Introduction

The survey consists of an introduction before the stated adaptation experiment and the questionnaire. The introduction was drafted in collaboration with the group of UMCG physicians. The UMCG physicians were asked to compose a paragraph that explained the dilemma questioned in the choice scenarios in medical vocabulary to assure that all colleagues thoroughly understood the question asked. The experts included an essential remark on the type of problem. The paragraph explained that the choice scenarios concerned the dilemma of whether to perform surgery after an operation indication. This remark was included to guarantee that the UMCG physicians would not mistake the question in the choice scenarios to be a diagnosis for an operation indication; thus, whether surgery is necessary to sustain life rather than an end-of-life decision after an operation indication is given. Furthermore, the introduction explains the objective of this research and discusses the structure of the survey.

5.3.2 Attributes and levels

The final selection of attributes and levels were established through iterative modifications over the four phases of the survey design. The selection and iterative improvements were all based on the feedback of the UMCG physicians. Table 5.1 provides the final list of attributes and levels for the stated adaptation experiment. The order of the list corresponds to the order of the attributes in the choice experiment. The group of physicians provided the preference of order.

The first and second phases of the survey design constitute the foundation of the list with attributes and levels included in the experiment. In the last two stages, minor adjustments to the list were made based on the guidelines discussed later in this section.

In the first stage of the survey design, the UMCG physicians were asked to construct a list of attributes and levels. While composing this list, the physicians had to include and focus on the following factors:

- Limit the list of attributes to, preferably, a maximum of twenty attributes. Because when individuals face too many attributes, they are unable to make trade-offs between all attributes and, therefore, tend to neglect some attributes while making a choice or stating a preference.
- The attributes should not overlap as this research desires to estimate the impact of individual attributes on the decision for surgery.
- Attempt to draft attribute ranges that match reality, but that also forces an individual to make trade-offs between attributes. For example, if the range of an attribute is determined too small, the attribute might still be essential but varies too little to have an impact on the choice.
- Preferably, choose three levels per attribute. This remark was included because, at the start of this research, it aimed to apply an orthogonal design of twelve attributes and three levels each. However, this study quickly discovered that an orthogonal design was not applicable. Section 5.3.3 explains the reasons why.
- Label the attributes with “ crucial,” “important,” or “nice to have” regarding the importance for of decision on surgery. Firstly, because this would help reduce the number of attributes if the list was too big and, secondly, to provide information on priors for the efficient design. Section 5.3.3 elaborates on how the labels are used to determine priors.

Appendix A further elaborates on the content of the individual interviews. After receiving the four personal lists of attributes and levels, these were merged into one file. The list was ordered according to the labels that the UMCG physicians attached to the attributes. The crucial attributes ordered on top of the list and the nice to have attributes on the bottom. Additionally, the attributes were grouped based on medical resembling. During the plenary meeting in the second phase, the combined list of 56 attributes was reduced to a list of 16 attributes based on a discussion among the four UMCG physicians. The group of physicians relatively fast agreed upon which attributes to include or neglect in the experiment. Appendix C provides a debate on this meeting. The guidelines for selecting the attributes and levels are the following:

- Each attribute must impact the decision for surgery and could be the decisive factor for the physicians to recommend surgery or comfort care. This does, however, not mean that each attribute should be equally important.
- The levels must be drafted such that the minimum and maximum range still forces the UMCG physicians to make trade-offs between other attributes. Therefore, this research desires to avoid incorporating attribute levels that will constitute a definite “yes” or “no” for surgery. Otherwise, the other attributes become insignificant. Moreover, BAIT desires to utilize the choice model for decision support. If among all experts, specific levels would generate a definite yes or not for surgery, the model is not of added value.

The model only becomes purposeful in situations where attribute levels constitute a quarrel for physicians on whether to perform surgery or initiate comfort care.

- Lastly, the attributes, attribute levels and combination of attribute levels included in the choice scenarios must match reality and provide enough information for the experts to make a considered decision. This research preferably desires the attributes range to capture at least 85 % of the bulk of observations in reality. However, a trade-off must be made between including a wide attribute range and removing attribute levels that constitute a definite yes or no for surgery to provide more information on trade-offs between attributes. This trade-off is deliberated in consultation with the UMCG physicians.

Finally, as explained at the start of this chapter, making sure that the choice scenarios reflect real-life choices of UMCG physicians and securing that the design enables the choice model to estimate trustworthy parameters is a complex trade-off.

As explained in Section 2.1.1, it is universally acknowledged that the more the hypothetical scenarios simulate real-world decisions, the higher the validity of the observed choices. Hence, including attributes and specific attribute levels that the UMCG physicians believe best reflect reality is a precondition. Therefore, a relatively extensive list of attributes was drafted compared to the number of attributes commonly incorporated in choice experiments. Caussade, et al (2005) investigated how individuals cope with an increased amount of information incorporated in a choice experiment. The results show that an increase in the number of attributes had an apparent impact on the choice behaviour of respondents as it increased the variance of the error term. This could be the consequence of respondents either making mistakes or adapting a simplifying technique based on partial information, hence, resulting in attribute non-attendance (ANA), due to the increased amount of information. Both consequences impact the consistency of the decision-making process of the respondents. Therefore, usually, a minimum of 7 attributes are included since research shows that incorporating more attributes may invoke ANA, which possibly generates parameter bias (Hensher et al., 2005). Hence, this research investigated literature that examined methods to deal with non-attendance (ANA). Literature, however, showed that the reliability of ANA approaches is still under consideration and further research must be conducted to analyse and improve the reliability of those approaches (Weller, Oehlmann, Mariel, & Meyerhoff, 2014). Therefore, this research neglects including a method to deal with possible ANA.

Additionally, constraints were required to filter out combinations between attribute levels that do not exist in reality. This research tried to minimize the number of restrictions to limit the correlations between attributes. For example, at the start of the survey design, the attribute "actual weight" was considered. This attribute was, however, changed into "growth." Because otherwise, many constraints were required between the attributes "birth weight", "days since birth" and "actual weight." To conclude, unfortunately, before conducting the stated adaptation experiment and estimating the parameters, it is not possible to, with high confidence, announce that the choice model will estimate reliable parameters without any effects the researcher cannot explain. Conducting a pilot study will hopefully increase the reliability of the parameters estimated for the final survey and reduce the chance of effects that cannot be explained.

Table 5.1 depicts the final list of attributes and levels included in the survey. The proposed survey for the UMCG physicians is, however, presented in Dutch. Together with the involved UCMG physicians, the appropriate terminology for the attributes and levels was established, which can be found in Appendix F. Furthermore, the levels are drafted such that level 1 constitutes the lowest likelihood for surgery and level 4 the highest probability for a recommendation in favour of operation per attribute. For the attribute gender, the UMCG physicians pronounced that girls might have a slightly bigger chance for advice in support of surgical procedure compared to boys, hence, the attribute girl is set as the second level.

Table 5.1: Attributes and levels

Attribute	Level 1	Level 2	Level 3	Level 4
Gender	Boy	Girl		
Gestational age	24 weeks	26 weeks	28 weeks	30 weeks
Birth weight	500 grams	650 grams	800 grams	1500 grams
Perinatal asphyxia	Yes	Dubious	No	
Congenital comorbidity	Present with high impact	Present with minor impact	Absent	
Progress since birth before a diagnosis of NEC	Serious complications	Minor complications	No complications	
Age since birth	0 – 7 days	7 – 14 days	14 - 21 days	
Growth since birth	Weak	Intermediate	Good	
Ultrasound of the brain	Bad prognosis	Intermediate prognosis	Good prognosis	
Lung function	Weak	Intermediate	Good	
Hemodynamic	Instable despite maximal support	Stable with support	Stable without support	
Cerebral oxygenation	40	60	80	
Wish of parents	In favour of comfort care	Doubtful about surgery	In favour of surgery	
The carrying capacity of parents	Weak	Intermediate	Good	

5.3.3 Experimental design

An experimental design is applied to construct the choice scenarios for the stated adaptation experiment. It combines attribute levels to form an alternative in a choice scenario. It also determines how to combine the alternatives into choice sets. In this study, the choice sets consist of just one choice alternative. Based on the single choice alternative, the UMCG physicians must decide whether to advise in support of surgery or comfort care. The final stated adaptation experiment consisted out of 35 choice scenarios.

Types of experimental designs

Different types of experimental designs exist. Experimental designs can either be full-factorial or fractional factorial. A full factorial design composes an experiment with all possible combinations of attribute levels. These designs usually result in too many choice tasks for respondents to conduct. Therefore, most commonly fractional factorial designs are practised. This type of design reduces the number of choice sets required for an stated adaptation experiment. Fractional factorial designs are either random, orthogonal, or efficient. Random fractional factorial designs select a random fraction of the full factorial design. The random selection, however, induces correlations between attributes resulting in higher standard errors and, therefore, in less reliable parameters. In orthogonal designs, a fraction is selected for which the correlations between attributes are zero (ChoiceMetrics, 2018). It, thereby, reduces standards errors and enhances the reliability of parameters. Lastly, efficient designs minimize standard errors and maximize information per choice tasks. Compared to orthogonal designs, it gets rid of dominant choice alternatives in choice tasks. Removing dominant choice alternatives is favourable as those choice tasks do not provide information about trade-offs between attributes. Efficient designs require a pilot study to obtain priors that are used as input for the experimental design. Priors are best guesses for parameters (Molin, 2018).

The decision for efficient fractional factorial design

This research will apply an efficient fractional factorial design. An efficient design is chosen because this study needs to include constraints. Constraints are required to filter out unrealistic combinations of attribute levels, which is essential for this research because the choice scenarios must reflect real-life choice tasks of UMCG physicians. Constraints, by definition, induce correlations between attributes and the design, thereby, loses its orthogonality. Additionally, because the design consists of a large number of attributes and levels that may generate many choice tasks, an efficient design helps to reduce the number of choice tasks compared to an orthogonal design.

Usually, a pilot study of approximately 30 respondents is conducted to determine priors. However, as the group of experts only consists of 15 experts, a pilot study with 30 respondents is not feasible. Therefore, 3 out of the 15 UMCG physicians participating in this study will execute the pilot study. Although it is understood that this is a small group of respondents for an stated adaptation experiment, it will provide more information on priors compared to no pilot study.

The pilot study will only estimate simple linear effects, as it desires to limit the number of parameters to be estimated with a very small group of respondents. The parameters estimated with the pilot study will provide information on the relative importance of the attributes. Similarly to the final survey, the pilot study also requires an efficient design because constraints are, likewise, needed for the pilot experimental design. Although at this point, there is no information on priors based on a pilot study yet, it will still include priors. If priors are chosen with the right sign and a modest magnitude, the design will increase in efficiency and will not be of worse quality compared to an orthogonal design. Therefore, the pilot study includes guesses for priors. The attribute levels are drafted such that an increase in attribute value increases the chance for surgery; therefore, the priors receive a positive sign. The magnitude is based on a fictional utility range. The priors for the crucial and important attributes included in the experiment are established on a utility range of 2. The nice to have attributes included are based on a utility range of 1. For the categorical attributes of which it is unsure what levels have a more significant positive impact on the decision for surgery, the priors are set to zero.

With the information obtained from the pilot study, the final experimental design is designed for the entire group of UMCG physicians. The final experimental design consists of 35 choice scenarios. Each choice scenario consisted of a single choice alternative for which the UMCG physicians had to decide whether to advise in support of surgery or comfort care.

The software package Ngene composes the experimental design. Appendix G provides the Ngene syntax for the pilot study and Appendix H for the final experimental design.

Moreover, both the pilot study and final experiment consists of two extreme choice scenarios manually added at the start of the survey. The first choice scenario includes all the highest levels of the attributes and the second choice scenario comprises all the lowest levels. In the introduction of the survey, a remark is added about the two included an extreme scenario with the purpose that the UMCG physicians are not put off by them.

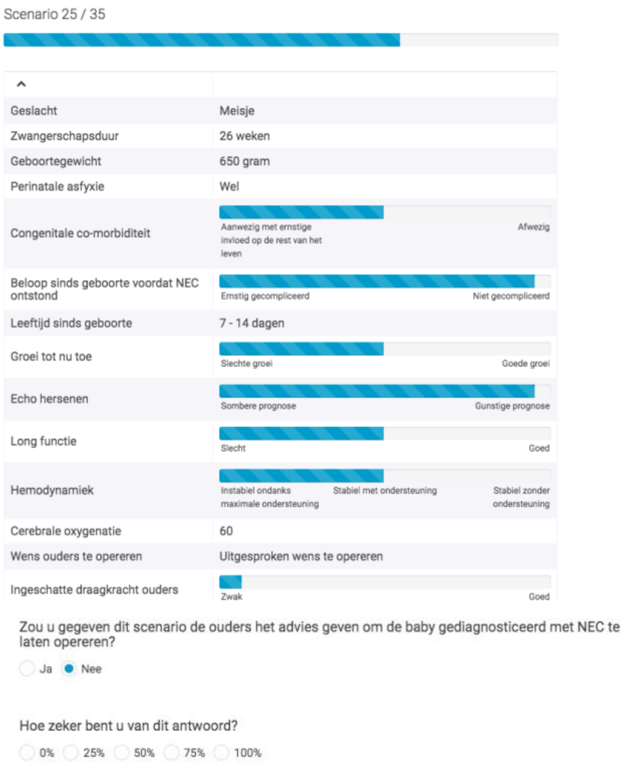
5.3.4 Questions

As mentioned in Section 2.1.1, the choice scenarios in the stated adaptation experiment will include the following questions:

1. Will you provide a recommendation in favour of or against surgery to parents?
2. How certain are you about your recommendation?

The group of respondents for both the pilot study as for the final survey is limited. Therefore, instead of solely asking the UMCG physicians to provide their answer to the dilemma of neonatal surgery, it will also include a question on the certainty of their decision. A Likert scale is provided to rate the certainty of their decisions. The scale consists of : 0%, 25%, 50%, 75% and 100% certainty. This question is incorporated because the single question on the recommended treatment, might provide too little information with the limited group of respondents for the model to estimate the parameters. The second question will, accordingly, provide more information and, therefore, if the first model is unable to estimate the parameters, the second model is used as fall back. The model that is applied to determine the parameters for the choice data of both questions is a linear regression model.

Figure 5.1 provides an example of one of the choice tasks included in the survey. Appendix I illustrates the final survey proposed to the UMCG physicians.



Scenario 25 / 35

Geslacht	Meisje
Zwangerschapsduur	26 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven — Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd — Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei — Goede groei
Echo hersenen	Sombere prognose — Gunstige prognose
Long functie	Slecht — Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning — Stabiel met ondersteuning — Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak — Goed

Zou u gegeven dit scenario de ouders het advies geven om de baby gediagnosticeerd met NEC te laten opereren?

○ Ja ● Nee

Hoe zeker bent u van dit antwoord?

○ 0% ○ 25% ○ 50% ○ 75% ○ 100%

Figure 5.1: Example of a choice task

5.4 Questionnaire

At the end of the survey, a questionnaire is included. The questionnaire comprises questions on personal characteristics variables to investigate heterogeneity between UMCG physicians. Additionally, the questionnaire also includes statements on how the UMCG physicians experienced executing the stated adaptation experiment. These statements are described in Section 5.4.2

5.4.1 Personal and professional characteristics variables

In the final part of the survey, questions on personal characteristics variables are included. The questions are based on the literature review conducted in Section 4.2. The literature review shows that most studies, that investigate the influence of personal characteristics variables on EoL decisions on neonates regularly consider the following variables: age, gender, religion, having children and the number of years of professional experience. The studies, however, provide deviating results on the impact of these personal characteristics on EoL decisions. This research will investigate whether, in this case, study, these personal characteristics variables impact the studied EoL decision. Table 5.2 depicts the personal characteristics variables enclosed in this research.

Table 5.2:Personal characteristics variables

Age?	Are you religious?	Do you have children?	Years of professional experience?	Gender?
25-35 years 35-45 years 45-55 years 55-65 years 65 > years	Yes No	Yes No	0-5 years 5-10 years 10-15 years 15-20 years 20 > years	Female Male

5.4.2 Statements to elicit opinion on the stated adaptation experiment

Additionally, in the questionnaire, different statements are proposed to determine how the UMCG physicians experienced conducting the stated adaptation experiment. A Likert scale is provided to rate the statements. The scale varies in levels: 1=Strongly disagree, 2=Disagree 3= Neutral, 4=Agree, 5=Strongly agree. Table 5.3 exhibits the proposed statements.

Moreover, at the end of the survey, a text block is included in which the respondents can provide their feedback or remarks on the survey. The answers to the statements and potential remarks are considered in the discussion of the results.

Table 5.3: Statements to elicit opinion on stated adaptation experiment

Proposed statements
1. Executing the choice experiment was challenging. 2. I enjoyed executing the choice experiment. 3. Executing the choice experiment was educational. 4. The choice scenarios in the choice experiment were realistic. 5. The choice scenarios in the choice experiment forced me to contemplate my decision thoroughly.

5.5 Sample and population

The sample for this study is the group of 15 UMCG physicians. The physicians are either neonatologists or child surgeons. The UMCG is known to be the only Dutch hospital that is recognized by the Ministry of Public Health as an NEC specialist ("Kinderchirurgie," n.d.). The UMCG physicians are, therefore, known experts in this field. This research desires to provide the UMCG neonatologists and neonatal surgeons with introspection on their recommendations. Therefore the estimated parameters in this study do not have to be tested for statistical significance. To provide an example of this argument, imagine asking a high school teacher to determine whether the average height of 13-year-old Dutch girls in her class is above 1.60 meters. The school teacher, consequently, measures the average height of 13-year-old Dutch girls in her class at 1.62 m. Hence, the teacher can state that the average height of 13-year-old Dutch girls in her class is above 1.60. If the teacher is asked to determine whether this effect, thus, whether 13-years-old Dutch girls are on average taller than 1.60 is true in the population, the measurement must be tested for statistical significance. Since this research is only interested in understanding how UMCG physicians make recommendations on surgery, it is not necessary to test the parameters for statistical significance.

6 Second procedure step of BAIT: model estimation

This chapter presents and evaluates the estimated choice models. This chapter will however not interpret the estimated parameters as Chapter 8 will analyse the parameters to provide introspection on the choice behaviour of the group of UMCG physicians. Firstly, to estimate the choice models, the data is coded. Section 6.1 discusses how the data is coded to estimate the models. Thenceforth, Section 6.2 presents the estimated choice models and, interprets and compares them.

6.1 Preparation of data

The group of 15 UMCG physicians executed the choice experiment between Friday the 26th of June and Friday the 17th of July 2020.

As mentioned in Section 5.3.4, this research applies two types of models. An binary logit model is estimated from the choice data of the question on preferred treatment, and a linear regression model is used to determine the parameters from the choice data of the preferred treatment and the certainty level of their recommendation.

Additionally, multiple binary logit models are estimated on the choice data of the preferred treatment. Firstly a binary logit model is estimated that dummy codes all variables, to investigate if the parameters are linear. Dummy coding is explained later in this section. Secondly, another binary logit model is estimated that incorporates multiple linear variables and several dummy coded variables. This model is used for introspection and is further discussed in Section 6.2.4.

To estimate the models, firstly the attributes included in the choice experiment are coded. Table 6.1 provides an overview of the coded attributes. For the linear parameters, the levels of all attributes received a numerical code. The numerical code 0 represents the lowest level, and each following level receives a code of 1, 2 and 3. A linear parameters assumes that the utility difference between attributes is similar.

During the design of the choice experiment, the physicians were asked whether they expected the attributes to be linear or non-linear. For most attributes the physicians were not certain about the utility course. For the attributes gestational age and birth weight, the physicians, however, with high confidence expected a non-linear utility course. For example, the UMCG physicians expected that the attribute value differences of gestational age between 24 weeks to 26 weeks would have a more significant impact on their recommendation for surgery compared to the attribute value differences of 27 to 30 weeks. Hence, expected is that the utility differences between the attributes levels are not similar.

As the physicians could not with certainty state which attributes included in the choice experiment were linear, this research applied dummy coding on all attributes. This study is able to investigate whether the parameters are non-linear by studying whether the attribute levels have diverse utility differences. In order to assess the utility contribution of the different levels, dummy coding is applied. As explained in Chapter 5, the levels of the attributes are drafted such that the lowest level constitutes the lowest likelihood for a recommendation in favour of surgery and the highest level the highest.

Hence, for this research, dummy coding is used since the lowest level establishes a reference level for all attributes. The utility contribution of the reference level is set to zero. The utility contribution for each other level of a variable discloses the difference in utility between the reference level and that attribute level. The only attribute that cannot be tested for linearity is gender as it consists of two levels.

Table 6.1: Coding of attribute variables

Variable	**Level**	**Coding**			
		Linear parameters	Dummy coded		
Gender	Boy	0	0		
	Girl	1	1		
Gestational age	24 weeks	0	0	0	0
	26 weeks	1	1	0	0
	28 weeks	2	0	1	0
	30 weeks	3	0	0	1
Birth weight	500 gram	0	0	0	0
	650 gram	1	1	0	0
	800 gram	2	0	1	0
	1500 gram	3	0	0	1
Perinatal asphyxia	Yes	0	0	0	
	Dubious	1	1	0	
	No	2	0	1	
Congenital comorbidity	Present with high impact	0	0	0	
	Present with minor impact	1	1	0	
	Absent	2	0	1	
Progress since birth before a diagnosis of NEC	Serious complications	0	0	0	
	Minor complications	1	1	0	
	No complications	2	0	1	
Age since birth	0 – 7 days	0	0	0	
	7- 14 days	1	1	0	
	14 -21 days	2	0	1	
Growth since birth	Weak	0	0	0	
	Intermediate	1	1	0	
	Good	2	0	1	
Ultrasound of the brain	Bad prognosis	0	0	0	
	Intermediate prognosis	1	1	0	
	Good prognosis	2	0	1	
Lung function	Weak	0	0	0	
	Intermediate	1	1	0	
	Good	2	0	1	
Hemodynamic	Instable despite maximal support	0	0	0	
	Stable with support	1	1	0	
	Stable without support	2	0	1	
Cerebral oxygenation	40	0	0	0	
	60	1	1	0	
	80	2	0	1	
Wish of parents	In favour of comfort care	0	0	0	
	Doubtful about surgery	1	1	0	
	In favour of surgery	2	0	1	

The carrying capacity of parents	Weak	0	0	0	
	Intermediate	1	1	0	
	Good	2	0	1	

Moreover, each personal and professional characteristic category per variable is coded with distinct numerical values. For each group of UMCG physicians in a specific personal characteristic category, a separate binary logit model, solely incorporating linear parameters, is estimated. Comparing the estimates of the individual models provides insight on the heterogeneity of the group of UMCG physicians for the recommendation on surgery. Table 6.2 presents the coded personal characteristics variables. Lastly, the UMCG physicians indicated to be interested in the difference in choice behaviour for the recommendation on surgery between child surgeons and neonatologists. Therefore, this research also estimated two binary logit models for those two groups

Table 6.2: Coding of personal characteristics variables

Variable	**Level**	**Coding**
Gender	Boy	0
	Girl	1
Religious	No	0
	Yes	1
Age	25-35 years	0
	35- 45 years	1
	45 - 55 years	2
	55-65 years	3
	65 > year	4
Kids	No	0
	Yes	1
Professional experience	0 – 5 years	0
	5 – 10 years	1
	10 – 15 years	2
	15 – 20 years	3
	20 > years	4

6.2 Estimated choice models

This section presents the estimated models. Section 6.2.1 describes and interprets the results of the linear regression model. It concludes and explains that the estimates of the linear regression model will not be further analysed in this research. After that, Section 6.2.2 describes the model evaluation metrics. Thereafter, Section 6.2.3 provides the estimates of the binary logit model with dummy coded variables. Lastly, Section 6.2.4 presents the binary logit model that incorporates multiple linear variables and several dummy coded variables. This model is used for introspection on the recommendation for surgery and is further discussed in Section 6.2.4. Chapter 8 analyses the parameters to provide introspection on the UMCG physicians choice behaviour.

The choice models are estimated using IBM SPSS Statistics 25.

6.2.1 Linear regression model

As explained in Section 5.3.4, a linear regression model is applied to estimate the parameters from the choice data of the questions on the preferred treatment and the degree of certainty. Also, it explained that the question on the degree of certainty was included as fall back in case the binary logit model could not converge with the choice data of the question on the recommended treatment alone. This, however, turned out not to be the case. Next section will show that the binary logit models did converge and generate parameters with the correct sign.
In contradiction, the estimated parameters by the linear regression model did display incorrect signs. Appendix J provides the estimated parameters of the linear regression model. As explained, the levels of the attributes are drafted such that an increase in attribute value increases the likelihood for surgery. Hence, the signs are expected to be positive. The estimated parameters of the linear regression model, however, show that, for example, the signs of the parameters gender, age since birth and growth since birth are all negative. In addition, the results also portray large discrepancies in magnitude of the parameters. For, example, the parameter wish of parents is estimated at 11.3 while age since birth has a value of -0.5. Also considering the high standard errors and the fact that the linear regression model was determined as fall back in case the binary logit model could not converge, which is not the case, this research will solely analyse the estimated parameters of the binary logit models.

6.2.2 Model evaluation metrics

There are multiple evaluation metrics and methods to assess the performance of the models. In this research, for the estimated binary logit models, the model fit parameters are calculated to compare the models and determine which model fits the observed choices best. The Log-Likelihood (LL) is a measure for the model fit. A LL that is closest to zero indicates the best model fit. Moreover, McFadden's rho-squared (ρ^2) is also a widely used measure to determine the goodness of fit.
Equation 3 provides the formula for the McFadden's rho-squared. The LL_0 represents the null-log-likelihood of the model for which all parameters are set to zero, . While the LL_β is the log-likelihood for the estimated model. The value of ρ^2 lies between 0 and 1 and can be interpreted as the percentage of uncertainty that is explained away by the model. A ρ^2 value closer to 1 represents a perfect model fit. There are, however, no universal agreements or guidelines on the value of ρ^2 that represents a good model fit. When comparing two models on the same data, the ρ^2 value will always be higher for a model with more variables. Hence, another metric exists that is called adjusted McFadden's rho-squared. The adjusted ρ^2 penalizes the model for including too many parameters. If the parameters are effective, the penalty is relatively smaller compared to the increase in LL. If the model contains parameters that do not add information to the model, the penalty becomes apparent. Equation 4 provides the formula for the adjusted ρ^2.

Equation 3: McFadden's rho-squared

$$\rho^2 = 1 - \frac{LL_0}{LL_\beta}$$

Where,
ρ^2 *is the McFadden's rho-squared*
LL_0 *denotes the null-log-likelihood*
LL_β *denotes the log-likelihood of the estimated model*

Equation 4: Adjusted McFadden's rho-squared

$$\rho^2 = 1 - \frac{LL_0 - K}{LL_\beta}$$

Where,
ρ^2 *is the McFadden's rho-squared*
LL_0 *denotes the null-log-likelihood*
LL_β *denotes the log-likelihood of the estimated model*
K represents the number of estimated parameters by the model

As explained in the previous paragraph, the LL is a measure of the model fit. A statistical test that is commonly applied to compare the model fit of different models is the likelihood ratio test. This test is practised, as often models are estimated based on sample data; hence it might be possible that a model retrieves a higher LL due to coincidence. The likelihood ratio test evaluates whether the differences in LL are significant. As explained in Section 5.5, this research, however, treats the sample as the population. Therefore, it is not necessary to test whether the differences in LL for the estimated models are significant.

Moreover, another way to evaluate the models is classification evaluation (CE). For CE, commonly, a contingency table is used that measures the classification performance by comparing the actual outcome of the model with the predicted outcome. For this research that entails that it compares the number of times, the UMCG physicians recommended against or in favour of surgery with the number of times the model predicted a preferred treatment. The software SPPS that is used to estimate the models also generates a contingency table. Multiple CE metrics exist that assess the model's discriminatory power. A CE metric that is most commonly utilized is the accuracy of the model (Beguería, 2006). The accuracy represents the percentage of correct predictions. Another widely used metric is sensitivity. Sensitivity presents the ratio between the predicted events and the actual events and, hence, demonstrates the models' ability to identify a recommendation against or in favour of surgery correctly.

This research will evaluate the binary logit models on the above-discussed evaluation metrics.

6.2.3 Binary logit model with dummy coded variables

This section will present the systematic utility function for the binary logit model that dummy codes all variables. Additionally, it will provide the model fit parameters and present the classification evaluation metrics.

Interaction effects

Firstly, this study will not include interaction effects between attributes. As explained in Appendix D during the design of the stated adaptation experiment, the UMCG physicians were asked whether interaction effects between attributes were plausible to exist. The UMCG physicians stated that they did not expect any interaction effects to be present between attributes. Moreover, during the meetings and interviews for the design of the stated adaptation experiment, it was analysed whether interaction effects could be discovered during the conversations with the UMCG physicians. The presence of interaction effects was, however, not found.

Moreover, it would be inaccurate to simply add all interaction effects between attributes to ensure that if interaction effects existed, they are captured by the model. This could lead to overfitting, and since this study includes a large number of attributes with 35 choice scenarios, it is simple, not possible to include all interaction effects between attributes. For these reasons, no interaction effects between attributes are included.

In Equation 5 the systematic utility function is displayed.

Equation 5: Systematic utility for dummy parameters

$$
\begin{aligned}
V_{Recommendationinfavorofsurgery}
&= \beta_{Gender1} * Gender1 + \beta_{Gestationalage1} * Gestationalage1 \\
&+ \beta_{Gestationalage2} * Gestationalage2 + \beta_{Gestationalage3} * Gestationalage3 \\
&+ \beta_{Birthweight1} * Birthweight1 + \beta_{Birthweight2} * Birthweight2 \\
&+ \beta_{Birthweight3} * Birthweight3 + \beta_{Perinatalasphyxia1} * Perinatalasphyxia1 \\
&+ \beta_{Perinatalasphyxia2} * Perinatalasphyxia2 + \beta_{Congenitalcomorbidity1} \\
&* Congenitalcomorbidity1 + \beta_{Congenitalcomorbidity2} \\
&* Congenitalcomorbidity2 + \beta_{Progresssincebirthbefore\ NEC1} \\
&* ProgresssincebirthbeforeNEC1 + \beta_{Progresssincebirthbefore\ NEC2} \\
&* ProgresssincebirthbeforeNEC2 + \beta_{Agesincebirth1} * Agesincebirth1 \\
&+ \beta_{Agesincebirth2} * Agesincebirth2 + \beta_{Growthsincebirth1} \\
&* Growthsincebirth1 \\
&+ \beta_{Growthsincebirth2} * Growthsincebirth2 + \beta_{Ultrasoundbrain1} \\
&* Ultrasoundbrain1 + \beta_{Ultrasoundbrain2} * Ultrasoundbrain2 \\
&+ \beta_{Lungfunction1} * Lungfunction1 + \beta_{Lungfunction2} * Lungfunction2 \\
&+ \beta_{Hemodynamic1} * Hemodynamic1 + \beta_{Hemodynamic2} * Hemodynamic2 \\
&+ \beta_{Cerebraloxygenation1} * Cerebraloxygenation1 + \beta_{Cerebraloxygenation2} \\
&* Cerebraloxygenation2 + \beta_{Wishofparents1} * Wishofparents1 \\
&+ \beta_{Wishofparents1} * Wishofparents1 + \beta_{Caringcapacityparents1} \\
&* Caringcapacity1 + \beta_{Caringcapacityparents2} * Caringcapacity2
\end{aligned}
$$

Table 6.3 presents the parameters estimated by SPSS. The model fit parameters for this binary logit model are:

1. LL= -243
2. McFadden’s rho-squared = 0.33
3. Adjusted McFadden’s rho-squared = 0.25

Table 6.3: Binary logit estimates for model with all variables dummy coded

Variable	Level	Parameter	Standard error	P-value
Gender	Boy	0		
	Girl	0.141	0.446	0.751
Gestational age	24 weeks	0		
	26 weeks	1.750	0.473	0.000
	28 weeks	2.082	0.431	0.000
	30 weeks	2.590	0.771	0.001
Birth weight	500 gram	0		
	650 gram	1.248	0.423	0.003
	800 gram	1.845	0.417	0.000
	1500 gram	2.225	0.925	0.016
Perinatal asphyxia	Yes	0		
	Dubious	0.400	0.371	0.281
	No	0.800	0.562	0.154

Congenital comorbidity	Present with high impact	0		
	Present with minor impact	0.900	0.336	0.007
	Absent	1.794	0.651	0.006
Progress since birth before a diagnosis of NEC	Serious complications	0		
	Minor complications	0.573	0.367	0.119
	No complications	0.320	0.497	0.519
Age since birth	0 – 7 days	0		
	7- 14 days	0.106	0.407	0.795
	14 -21 days	0.050	0.697	0.943
Growth since birth	Weak	0		
	Intermediate	0.094	0.423	0.824
	Good	0.184	0.472	0.696
Ultrasound of the brain	Bad prognosis	0		
	Intermediate prognosis	2.059	0.397	0.000
	Good prognosis	2.579	0.888	0.004
Lung function	Weak	0		
	Intermediate	0.264	0.345	0.444
	Good	0.333	0.466	0.474
Hemodynamic	Instable despite maximal support	0		
	Stable with support	0.567	0.416	0.173
	Stable without support	0.561	0.442	0.204
Cerebral oxygenation	40	0		
	60	0.146	0.381	0.703
	80	0.516	0.669	0.440
Wish of parents	In favour of comfort care	0		
	Doubtful about surgery	1.695	0.314	0.000
	In favour of surgery	2.209	0.489	0.000
The carrying capacity of parents	Weak	0		
	Intermediate	0.594	0.415	0.152
	Good	0.302	0.452	0.504
Constant		-8.473	1.852	0.000
Number of observations (n) = 525				

Table 6.3 illustrates that the parameters portray higher standard errors. This is not unexpected as limited choice data is used to estimate a large number of parameters. The results also presents some variables with counterintuitive results. For example, for the parameter age since birth, the parameter estimated for the middle level has a higher value compared to the maximum level. This is counterintuitive as the maximum level is expected to obtain the highest weight. The other variables that portray counterintuitive results are carrying capacity of parents and progress since birth before a diagnosis with NEC.

Table 6.4 presents the classification table for this model. The overall percentage correct represents the accuracy of the model and, thus, indicates the percentage of accurate predictions. A 75.8% prediction accuracy is considered high.

Table 6.4: Classification table for binary logit model with dummy coded variables

		Predicted		Percentage correct
		No	Yes	
Observed	No	197	70	73.9
	Yes	57	200	77.8
Overall percentage correct				75.8

Table 6.3 illustrated that multiple variables portray counterintuitive results and portray high standard errors for the parameters indicating that they are estimated with a considerate degree of uncertainty. Therefore, an additional binary logit model is estimated, including multiple linear and some dummy coded variables. As explained in Section 6.1, the physicians were very confident about the fact that birth weight and gestational age would portray a non-linear utility curve. Therefore, these variables are dummy coded in this binary logit model. As explained in Section 6.1, for the other attributes, the physicians were not certain about the utility courses for the other attributes.
Moreover, Table 6.3 portrays that the variables with the largest weights are: wish of parents, gestational age, birth weight, congenital co-morbidity, and ultrasound of the brain. A non-linear utility course compared to a linear utility course might significantly change the utility contribution of the attribute levels. For the variables with larger weights, this can have a considerable impact on the utility contributions, while for parameters with smaller values it will relatively have less impact. Therefore, this research estimated a new binary logit model that next to the birth weight and gestational age also dummy coded the variables: wish of parents, congenital co-morbidity, and ultrasound of the brain. Next section discusses the model evaluation metrics for this binary logit model.

6.2.4 Binary logit model applied for introspection of the UMCG physician's choice behaviour

This section will present the systematic utility function for the binary logit model used for the introspection of the UMCG physician's choice behaviour. Moreover, it provides the model fit parameters and displays the classification evaluation metrics.

Equation 6 displays the systematic utility function.

Equation 6: Systematic utility for the binary logit model applied for introspection

$$
\begin{aligned}
V_{Recommendationinfavorofsurgery}
&= \beta_{Gender} * Gender + \beta_{Gestationalage1} * Gestationalage1 + \beta_{Gestationalage2} \\
&* Gestationalage2 + \beta_{Gestationalage3} * Gestationalage3 + \beta_{Birthweight1} \\
&* Birthweight1 + \beta_{Birthweight2} * Birthweight2 + \beta_{Birthweight3} \\
&* Birthweight3 + \beta_{Perinatalasphyxia} * Perinatalasphyxia \\
&+ \beta_{Congenitalcomorbidity1} * Congenitalcomorbidity1 \\
&+ \beta_{Congenitalcomorbidity2} * Congenitalcomorbidity2 \\
&+ \beta_{Progresssincebirthbefore\ NEC} * ProgresssincebirthbeforeNEC \\
&+ \beta_{Agesincebirth} * Agesincebirth + \beta_{Growthsincebirth} * Growthsincebirth \\
&+ \beta_{Ultrasoundbrain1} * Ultrasoundbrain1 + \beta_{Ultrasoundbrain2} \\
&* Ultrasoundbrain2 + \beta_{Lungfunction} * Lungfunction + \beta_{Hemodynamic} \\
&* Hemodynamic + Hemodynamic + \beta_{Cerebraloxygenation} \\
&* Cerebraloxygenation + \beta_{Wishofparents1} * Wishofparents1 \\
&+ \beta_{Wishofparents1} * Wishofparents1 + \beta_{Caringcapacityparents} \\
&* Caringcapacity
\end{aligned}
$$

Table 6.5**Error! Reference source not found.** presents the parameters estimated by SPSS. The model fit parameters for this binary logit model are:

1. LL= -245
2. McFadden's rho-squared = 0.32
3. Adjusted McFadden's rho-squared = 0.27

Table 6.5:Binary logit estimates of model applied for introspection

Variable	Level	Parameter	Standard error	P-value
Gender	Boy	0		
	Girl	0.020	0.392	0.960
Gestational age	24 weeks	0		
	26 weeks	1.656	0.431	0.000
	28 weeks	1.851	0.368	0.000
	30 weeks	2.859	0.549	0.000
Birth weight	500 grams	0		
	650 grams	1.238	0.411	0.003
	800 grams	1.835	0.394	0.000
	1500 gram	2.507	0.731	0.001
Perinatal asphyxia		0.452	0.233	0.053
Congenital co-morbidity	Present with high impact	0		
	Present with minor impact	0.944	0.336	0.007
	Absent	1.752	0.651	0.006
Progress since birth before a diagnosis of NEC		0.230	0.201	0.252
Age since birth		0.250	0.231	0.279
Growth since birth		0.183	0.200	0.359
Ultrasound of the brain	Bad prognosis	0		
	Intermediate prognosis	1.798	0.332	0.000
	Good prognosis	2.782	0.571	0.000
Lung function		0.204	0.194	0.293
Hemodynamic		0.279	0.191	0.144
Cerebral oxygenation		0.430	0.215	0.046
Wish of parents	In favour of comfort care	0		
	Doubtful about surgery	1.729	0.308	0.000
	In favour of surgery	2.154	0.440	0.000
The carrying capacity of parents		0.216	0.202	0.284
Constant		-8.830	1.512	0.000
Number of observations (n) = 525				

The evaluation metrics of both models portray that the McFadden's rho-squared for the binary logit model with all dummy coded variables is higher compared to the binary logit model incorporating linear and dummy coded variables. The higher ρ^2 for the binary logit model that dummy codes all variables is coherent, as the ρ^2 will always be higher for a model with more parameters when comparing two models on the same data. As explained earlier in this chapter, the adjusted ρ^2 penalizes a model for including too many parameters. The penalty is relatively small when the parameters are effective, but the penalty becomes apparent when the added parameters add little information to the model. The adjusted ρ^2 for both the models illustrates a higher adjusted ρ^2 for the binary logit model incorporating linear and dummy coded variables (0.27) compared to the binary logit model that dummy codes all parameters (0.25). That means that 27% of the uncertainty is explained away by the model that incorporates linear and dummy coded variables and 25% by the other binary logit model. Hence, the binary logit model including linear and dummy coded variables explains away 2% of the uncertainty more. Accordingly, the additional parameters included in the binary logit model with all dummy coded variables adds little information to the model such that the penalty of the adjusted ρ^2 is more significant than the information it adds. Additionally, the model accuracy of the model incorporating multiple linear variables and several dummy coded variables improved with 0.2% to a 76% accuracy.

Therefore, the binary logit model that incorporates linear and dummy coded variables is used for introspection as it fits the observed recommendations of the UMCG physicians best compared to the model dummy coding all variables. Moreover, to determine if dummy coding more variables would improve the model fit and, hence, fit the observed recommendations better, multiple binary logit models are additionally estimated. These binary logit models dummy coded the five variables; wish of parents, gestational age, birth weight, congenital co-morbidity, and ultrasound of the brain, and dummy coded one of the other variables to determine if the model fit would increase. Appendix K provides the model fit calculations for all these individual binary logit models. The results show that the model fit for these models did not improve compared to the binary logit model dummy coding the five variables. Therefore, the binary logit model presented in this section is used for introspection of the UMCG physicians' choice behaviour.

Lastly, another metric that is estimated for the binary logit model used for introspection is the mean average deviation (MAD). The MAD determines the average deviation in percentage points between the likelihood for a recommendation in favour of or against surgery and the actual distribution of recommendations per choice scenario. Hence, the MAD determines to what extent the model accurately predicts the distribution of recommendations in favour of or against surgery. In comparison to the prediction accuracy of the model, the MAD, therefore, considers the fact that the choice scenarios were drafted such that they did not constitute a definite yes or no for a recommendation on surgery. The calculated MAD for the model is 5 percentage points and the calculation of the MAD is provided in Appendix S. This can be interpreted that on average, per choice scenario, the model predicts a "wrong" recommendation for less than one physician as one physician out of 15 physicians represents 7% (1/15). While presenting the results to the UMCG physicians for introspection, this measure is used to explain the model's performance.

7 Output of BAIT: descriptive results

This chapter presents the descriptive results of BAIT established on the choice data retrieved from the survey. Firstly, Section 7.1 presents the sample characteristics in terms of the personal and professional traits measured by the survey. Section 7.2 provides the feedback on the choice experiment given by the UMCG physicians. Lastly, Section 7.3 discusses the choice behaviour of the group of UMCG physicians on the choice experiment.

7.1 Sample characteristics

This section demonstrates the sample characteristics. Table 7.1 provides the distribution of the UMCG physicians in the sample group in terms of the personal characteristics measured by the survey. As explained in Section 5.5 in this study, the sample of UMCG physicians is treated as the population. Hence, the distribution of sample characteristics is not compared with the distribution of personal characteristic variables in the 'true' population.

As explained in the previous section, for each personal characteristic category, a binary logit model is estimated. Table 7.1 illustrates that there is an unequal distribution between categories for some personal characteristic variables. For example, out of the 15 UMCG physicians, only three physicians indicated that they were religious. That means that the estimated parameters for religious physicians in the sample group only includes the choice data of three physicians, representing a smaller sample of three respondents. Hence, for the estimated binary logit models per personal characteristic category, only linear parameters were included to minimize the number of parameters to be estimated with limited choice data. Moreover, as Table 7.1 shows, the sample distribution of age and length of professional experience is fairly distributed over all categories. For example, the sample group of 25 to 35 years old physicians only consist of two doctors. Therefore, to generate more reliable results, for both the variables age and years of professional experience, the groups are reduced to two categories. For the variable age, two binary logit models are estimated for an age category of 25-45 years and 45 years old and above. And, for the variable professional experience, two binary logit models are estimated for physicians with 0-10 years of experience and physicians with 10 years of experience or more.

Table 7.1: Sample characteristics

Personal characteristic	**Category**	**Distribution in the sample group**
Gender	Male Female	5/15 10/15
Religious	Yes No	3/15 12/ 5
Having children	Yes No	11/15 4/15
Age	25-35 years 35-45 years 45-55 years 55-65 years 65 > years	2/15 6/15 5/15 2/15 0/15
Years of professional experience	0-5 years 5-10 years 10-15 years 15-20 years 20 > years	5/15 2/15 3/15 3/15 2/15

Specialisation	Child surgeon Neonatologist	4/15 11/15

7.2 Feedback of the sample group on the survey

This section presents the answers of the UMCG physicians on the statements that were proposed at the end of the survey to determine how the UMCG physicians experienced conducting the stated adaptation experiment. Additionally, it discusses the general feedback provided by the sample group.

Table 7.2 demonstrates the average rating per statement. As presented in Section 5.4.2 the proposed statements are rated using a Likert scale. The scale varies in levels: 1=Strongly disagree, 2=Disagree 3= Neutral, 4=Agree, 5=Strongly agree. The average rating per statement is calculated by summing up the scores of each UMCG physician and dividing the value by the sample size.

The results presented in Table 7.2 show that the average for the proposed statements fluctuate around level four, which entails that the physicians agree with the statements. As explained earlier in this study the choice scenarios must simulate real-world choice tasks to enhance the validity of the observed choices. Hence, the fact that the UMCG physicians experienced the choice scenarios as realistic is considered positive, as it displays an increased validity of the provided recommendations on surgery.

Additionally, the high ratings on the difficulty of the stated adaptation experiment and the requirement of deep thought can be considered as favourable and unfavourable. Since as explained in Chapter 5 the experiment had to be drafted such that the minimum and maximum attribute range still forced the UMCG physicians to make trade-offs between other attributes. Therefore, this research desired to avoid incorporating attribute levels that would constitute a definite “yes” or “no” for surgery. Incorporating levels that establish a distinct “yes” or “no” for operation would have most likely displayed ratings disagreeing with statements 1 and 5. The next section will support this assertion as it will show a devoid of many choice scenarios that triggered a full agreement for or against surgery.

The considerably high ratings for statement 1 and 5 may, however, also indicate that the choice tasks were experienced as too challenging due to diverse reasons. It could, for example, be that the choice experiment incorporated too many attributes, such that the UMCG physicians’ were unable to make trade-offs between all attributes and, therefore, neglected some attributes while giving a recommendation. Moreover, it could also illustrate a lack of attributes or information on attribute levels. The general feedback provided by the UMCG physicians, discussed in the next paragraph, includes a comment on the fact that the levels of several attributes left room for too much own interpretation. Table 7.2 also presents that the UMCG physicians also experienced the stated adaptation experiment as fun and educational.

Table 7.2: Rating of choice experiment

Rating variable	Rating
1. Experiment was difficult	3.9
2. Experiment was fun	4.1
3. Experiment was educational	3.7
4. Experiment was realistic	3.9
5. Experiment encouraged deep thought	3.9

Moreover, the UMCG physicians had the opportunity to provide general feedback on the stated adaptation experiment. A gross fraction of the sample group mentioned that they believed some attribute levels should have been explained in more detail. The stated adaptation experiment indicated the levels of multiple attributes with weak, intermediate, and good. A fragment of the group believed this left too much room for own interpretation. The specific attributes mentioned were congenital comorbidity, the ultrasound of the brain, and growth since birth. One of the physicians' stated that a particular type of congenital comorbidity can be the decisive factor in providing a recommendation in favour of or against surgery, hence, not knowing the specific types of congenital comorbidity made the choice tasks more difficult. Another physician recommended adding an explanation at the beginning of the survey, explaining what is meant with, for example, a good ultrasound of the brain or an intermediate growth since birth. The physician stated that this would help establish a generalized interpretation of the attribute levels among all physicians executing the choice experiment, and it would make the choice tasks less complicated. Conclusively, these comments also illustrate why the group of physicians labelled the experiment as difficult.

7.3 Choice behaviour

For each choice scenario, the UMCG physicians were asked for a recommendation in favour of or against surgery. Additionally, for each answer, they were required to indicate the degree of certainty about their judgment.

In total, the group of physicians voted 265 times against surgery and 255 times in favour of operation. Hence, 51% of the votes were in favour of surgery, and 49% against surgery. Figure 7.1 provides the distribution of answers concerning their recommendation for neonatal surgery and the corresponding certainty level on their advice.

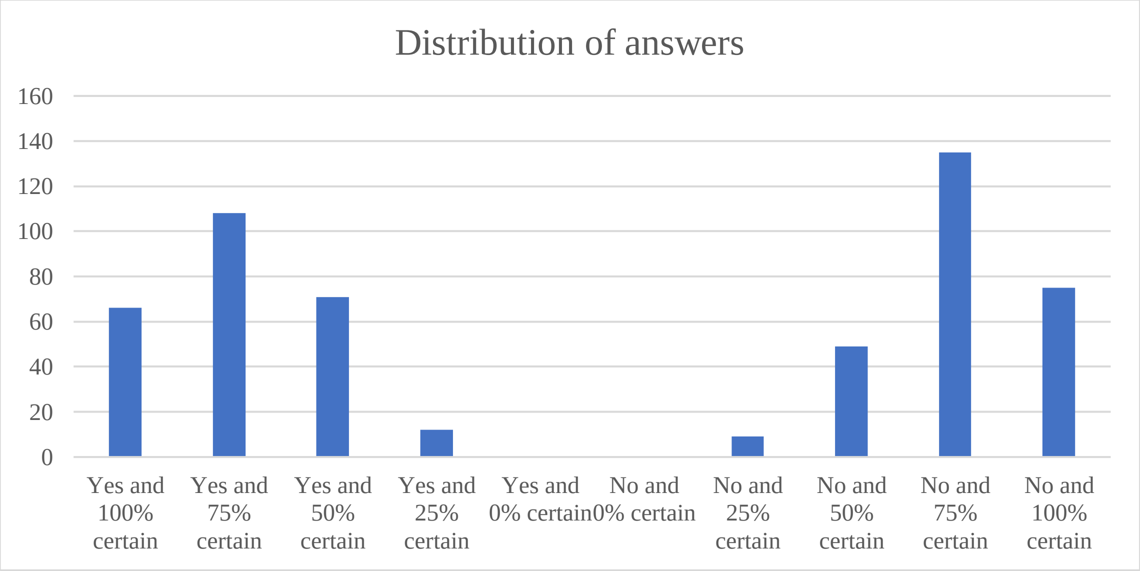


Figure 7.1: Distribution of answers

Figure 7.1 demonstrates the distribution of the degree of certainty for both a recommendation against or in favour of operation. It shows that zero UMCG physicians were 0% certain about their advice. This observation was probable as it would be troublesome if experts who are making such impactful EoL decisions state to be 0% certain about their judgement. A recommendation with 0% certainty would, also, make explaining and defending a physicians' recommendation to parents of the ill new-born very challenging. This could lead to conflicts between the physicians and the parents, resulting in a possibly difficult decision process. Hence, it is comprehensible that the UMCG physicians did not indicate to be 0% confident about their recommendation.

Moreover, the gross majority, with 73%, indicated to be 75% or 100% confident about their advice that supported or opposed surgery. The UMCG physicians most often indicated to be 75% confident about their advice and rarely showed a certainty level below 50%. The high confidence levels are somewhat remarkable. On the one hand, as a society, it is consoling to observe that physicians who provide such impactful recommendations are very confident about their judgement in most cases. On the other hand, as explained in Section 4.1, this choice task is very complex as the physicians deal with a high degree of uncertainty on the development of the health of the new-born after neonatal surgery and diverse opinions exist on what is an adequate expected quality of life. Hence, it is striking that recommendations made under such a high degree of uncertainty are given with such high confidence levels. The high confidence levels might be explained by the phenomena of overconfidence in clinical decision making. Overconfidence of physicians often occurs in the context of judgement and decision making in health care. It is recognized as a common cognitive bias (Pat Croskerry & Norman, 2008). A further deliberation on this topic is provided in the discussion of this research that relates this phenomenon to the potential of BAIT to support future recommendations of the UMCG physicians.

Furthermore, Figure 7.1 demonstrates that the distribution of the degree of certainty for both a recommendation against or in favour of operation is, to some extent, similar. The graph does show that, although it is a small difference, on average higher confidence levels are provided for recommendations against surgery. This seems plausible as a recommendation against surgery concerns ending the life of a new-born. Hence, a possible expectation is that the decision to end the life of another human being is made with a higher level of certainty compared to saving the life of a new-born. It would, however, also be reasonable to have made the assumption that a recommendation against surgery could also constitute in a lower level of confidence as ending the life of a new-born in an environment with a high level of uncertainty might be provided with a lower level of confidence compared to recommending in favour of surgery.

Furthermore, Figure 7.2 presents the distribution of choices per choice scenario for the question regarding the preferred treatment. As expected the first two choice scenarios demonstrate a unanimous yes and no for surgery as the first choice scenario consisted of all the highest levels of the attributes and the second choice scenario of all the lowest levels. Figure 7.2 illustrates that apart from choice scenario 1 and 2, for several more choice scenarios all UMCG physicians unanimously recommended in support of or against surgery. These choice scenarios are choice tasks; 13, 16, 19, 22 for which all UMCG physicians provided a recommendation in favour of surgery and choice scenario 33 for which all physicians advised against surgery. Figure 7.2 also portrays that for most choice tasks diverse recommendations against and in favour of surgery are provided. This illustrates that in the process of making medical decisions, as explained in Section 4.1, physicians base recommendations on their own personal and professional experience. This is also called “accumulated clinical knowlegde” defined as a physician’s own personal and professional knowledge acquired through years of education, experience and training (Uy, Sarmiento, Gavino, & Fontelo, 2014). The choice scenarios are, however, drafted such that they did not include attribute values that would definite constitute a “yes” or “no” for surgery and the choice tasks can, hence, be experienced as difficult. Therefore a certain degree of heterogeneity was to be expected.

Figure 7.3 displays the choices per choice scenario, including the answers on the degree of certainty per recommendation. The 0% certainty levels are removed from the graph since as explained earlier in this section nobody indicated a 0% certainty level. The graph depicts that for the choice tasks that show significant heterogeneity between the UMCG physicians such as choice scenario 14 for which eight physicians chose a recommendation in favour of operation, and seven opposing surgery, the certainty levels are more diverse. Moreover, for choice scenario 14 the certainty level of 50% was selected the most, which is low considering that in 73% of the scenarios a certainty level of 75% or 100% was selected. Figure 7.3 seems to portray a trend that for the choice tasks with a high level of heterogeneity between the UMCG physicians, such as choice scenario 14, lower level of certainties are provided compared to, for example, the choice scenarios for which unanimous recommendations that supported or opposed surgery were given.

For the interpretation of the generated choice probability by BAIT, that as discussed in Section 2.2.2 portray the number of UMCG physicians that would provide a recommendation in favour of surgery. It is essential to understand whether a choice probability of around, for example, 55% illustrates that the UMCG physicians are less certain about which recommendation to provide to parents. Or that 55% demonstrates that this percentage of the group of UMCG physicians are very confident that advice in favour of surgery is the most appropriate recommendation. In comparison to the other 45% who are confident that a recommendation opposing surgery is the best medical advice.

To possibly implement BAIT as decision support in the future, the correct interpretation of the choice probability is an important matter. Therefore an additional analysis is performed that determines the correlation between the generated choice probability by BAIT for the choice scenarios included in the choice experiment and the confidence levels.

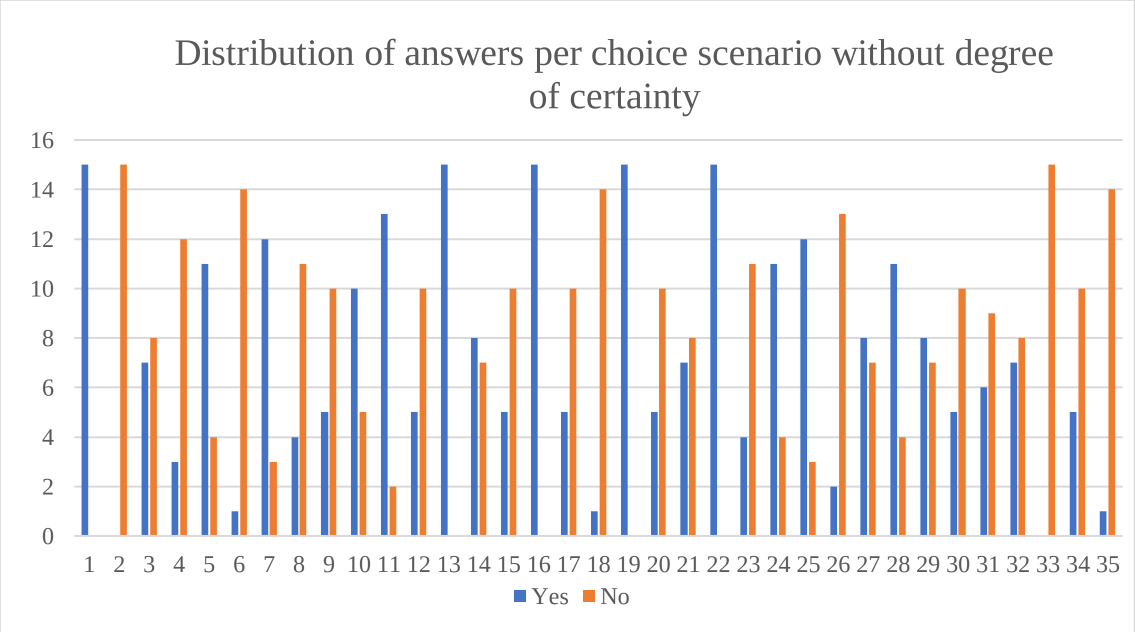


Figure 7.2: Distribution of answers per choice scenario excluding the degree of certainty

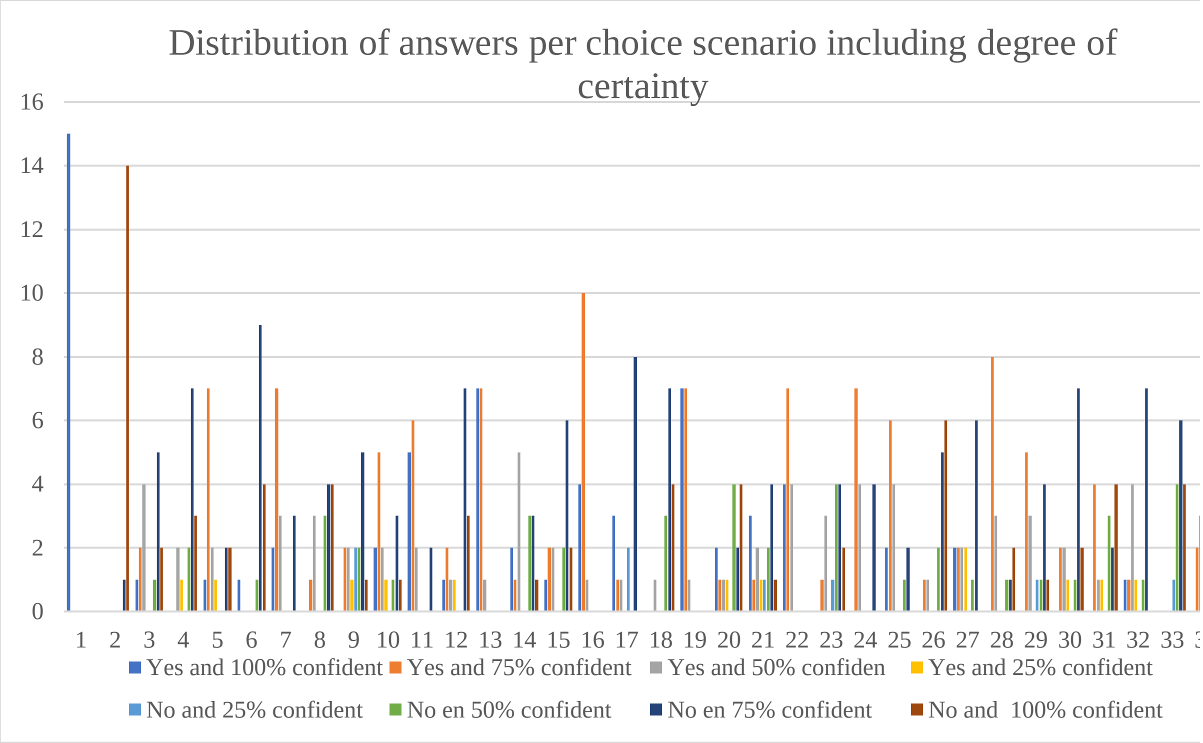


Figure 7.3: Distribution of answers per choice scenario including the degree of certainty

As mentioned in the previous paragraph, an additional analysis is conducted to determine the correlation between the generated choice probabilities of BAIT and the confidence levels provided by the UMCG physicians on the choice experiment to investigate how to interpret the generated choice probabilities. In order to conduct this analysis, the choice probabilities of BAIT are firstly estimated for all choice scenarios. After that, the 'spread' of the recommendations for each choice scenario is calculated based on the generated choice probabilities. In which a 0 % spread indicates that BAIT expects that all UMCG physicians would unanimously recommend for or against surgery. In contrast, a 100% spread suggests that 50% of the group of UMCG physicians are expected to recommend in favour of surgery and the other 50% against surgery. Also, the average confidence level per choice scenario is estimated. The calculations of the spread and the average confidence level per choice scenario are provided in Appendix T.

Firstly the choice scenarios were ordered from a 0% spread to a 100% spread. Figure 7.4 depicts a scatterplot for the ordered choice scenarios. The horizontal axis plots the spread and the vertical axis the confidence level. Figure 7.4 illustrates that there is indeed a relationship between higher spread values and lower confidence levels as it shows that on average, the level of confidence reduces when the percentage of spread increases. To further identify this relationship, the Pearson correlation is estimated. The correlation determines the extent to which two variables have a linear relationship with each other. The correlation can be useful as it can indicate a predictive relationship. It must, however, be noted that the correlation shows the strength of the relationship but does not completely characterise the relationship between two variables. The Pearson correlation is calculated by using SPSS and is estimated at -0.687. A value of -0.687 illustrates a strong linear relationship which indicates that for an increase of spread, the confidence level is likely to decrease.

In conclusion, although, as explained in the previous paragraph, the correlation does not completely characterise the relationship of spread and confidence level, it does illustrate a strong negative linear relationship between the two variables. Therefore, the best way to interpret the choice probabilities generated by BAIT is that choice probabilities equal to or close to 50% in most cases indicates that the group of UMCG physicians are less certain about which recommendation to provide. Hence, for the example provided in the last chapter, a choice probability of 55% generated by BAIT, most likely indicates that UMCG physicians are more doubtful about which recommendation to provide to parents rather than 55% of the group of UMCG physicians being very confident that advice in favour of surgery is the most appropriate recommendation. A further discussion on the implication of this observation on the ability of BAIT to serve as decision support is presented in the conclusion and discussion of this research.

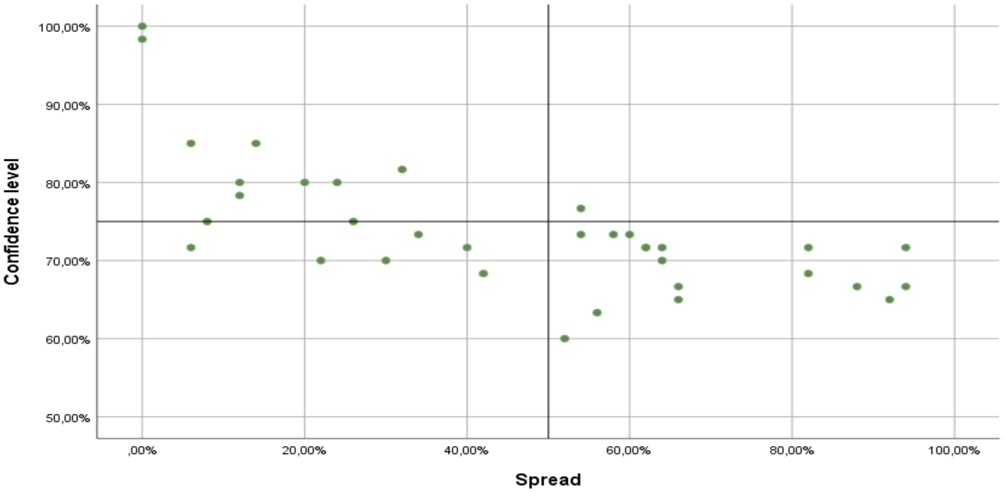


Figure 7.4: Scatterplot spread versus confidence level for the choice scenarios

8 Output of BAIT: Introspection on choice behaviour and generated choice probabilities

Chapter 6 codified the domain expertise of the group of UMCG physicians by estimating a binary logit model from the choice data retrieved from the stated adaptation experiment. The choice model estimates parameters, also called weights, for each decision variable. The parameters are analysed to provide the group of UMCG physicians with insight into their own choice behaviour.

This chapter provides the analyses of the estimated parameters to provide introspection on the UMCG physicians choice behaviour. Firstly, Section 8.1 presents how the parameters will be analysed to provide introspection on the choice behaviour. Section 8.2 offers a discussion on the estimated parameters and displays the utility courses. After that, Section 8.3 illustrates the maximum utility contribution and relative importance of the variables. Lastly, Section 8.4 demonstrates the results of the differences in choice behaviour between UMCG physicians established on diverse personal and professional characteristics.

8.1 Evaluation of model estimates to provide introspection

In order to provide introspection on the choice behaviour of the group of UMCG physicians, for each of the estimated parameters, the following factors are analysed.

- Parameter value:

The estimated parameter value presents the weight, or also called taste, of the attribute. After that, by accumulating the parameter with the attribute value, it results in a contribution to the utility function.

- Maximum utility contribution:

The maximum utility contribution determines the maximum impact of an attribute on the utility function of a recommendation in favour of surgery. It is calculated by estimating the difference between the lowest and highest utility contribution of the levels of an attribute for non-linear variables and for linear variables by multiplying the parameter with the number of levels.

- Relative importance:

The relative importance resembles the relative effect of a parameter on the maximum systematic utility of a choice alternative. It is calculated by dividing the maximum utility contribution per attribute by the maximum systematic utility of an alternative. The maximum systematic utility of the alternative supporting surgery is estimated by summing the maximum utility contributions of all attributes.

- Utility curve:

The utility curve visualizes the utility contribution per attribute level of an attribute. It helps to visualize whether a parameter has a linear effect or a distinct utility curve.

8.2 Parameters and utility curves

Table 6.5 in Section 6.2.4 presents the estimated parameters through SPSS that codify the domain expertise of the group of UMCG physicians. This section will first discuss the dummy coded variables of the model and demonstrate their utility course. Thenceforth it provides a summary of the rest of the variables.

Gestational age

Figure 8.1 displays the utility course for the variable gestational age. The utility course portrays a non-linear curve. As mentioned in Section 6.1, the UMCG physicians reflected on the utility course they expected for the attributes. The UMCG physicians explained that they expected the values of gestational age between 24 weeks to 26 weeks to have a more significant impact on their recommendation for surgery compared to the attribute values from 26 to 30 weeks. They mentioned that a difference in gestational age of 28 weeks compared to 26 weeks, would not significantly impact their recommendation, while a difference in gestational age of 26 weeks compared to 24 weeks would more significantly affect their advice. This is what the utility curve shows, as the utility difference between 24 weeks to 26 weeks is larger, namely 1.656 utils, compared to the utility difference of 26 weeks to 28 weeks, which is only 0.2 utils. Moreover, the utility contribution difference between 28 weeks and 30 weeks is also smaller compared to 24 to 26 weeks. Therefore, the utility curve matches the expectations of the UMCG physicians.

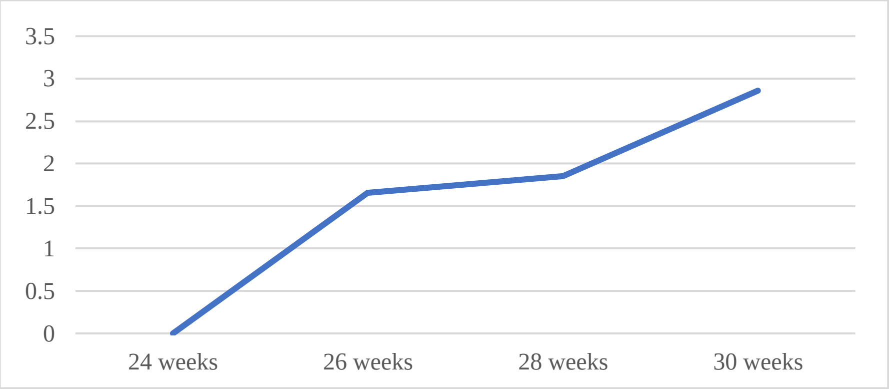


Figure 8.1: Utility course gestational age

Birth weight

Similarly to gestational age, the attribute birth weight shows a non-linear utility curve. Figure 8.2 the utility course for the variable birth weight. Likewise, the UMCG physicians expected the attribute value difference from 500 to 650 grams to have a more significant impact on their recommendation compared to the attribute value difference of for, example, 650 to 800 grams. This assumption is confirmed by the utility curve, as the utility curve is flatter for an attribute level increase from 650 to 800 grams compared to a rise from 500 to 650 grams.

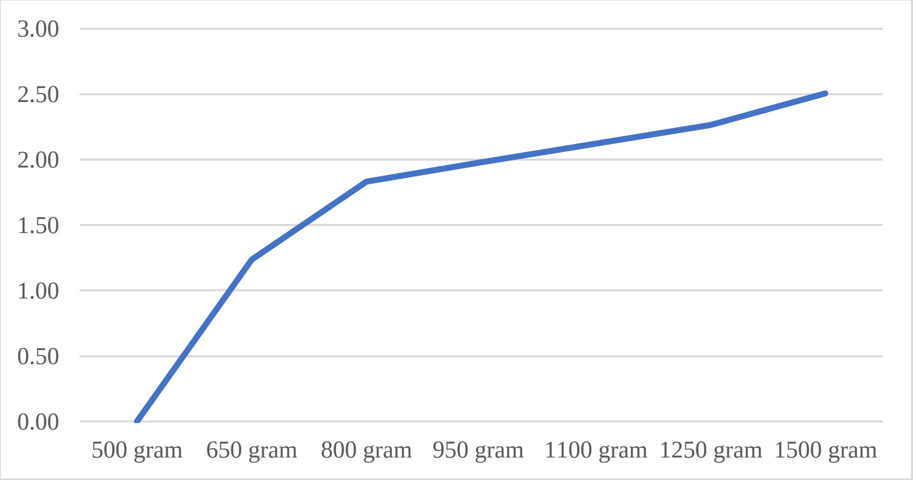


Figure 8.2: Utility course birth weight

Congenital co-morbidity

Figure 8.3 depicts the utility course of the variable congenital co-morbidity. The utility course indicates a linear utility curve as the difference between the utility of the attribute levels is almost similar. Because the utility difference between level 0 “present with high impact” and level 1 ‘present with minor impact’ is 0.94 and the utility difference between the level “present with minor impact” and “absent” is 0.81. Therefore, each level increase has an almost similar impact on the utility function which means that an attribute level increase from level 0 to 1 and level 1 to 2 has a nearly identical impact on the medical recommendations.

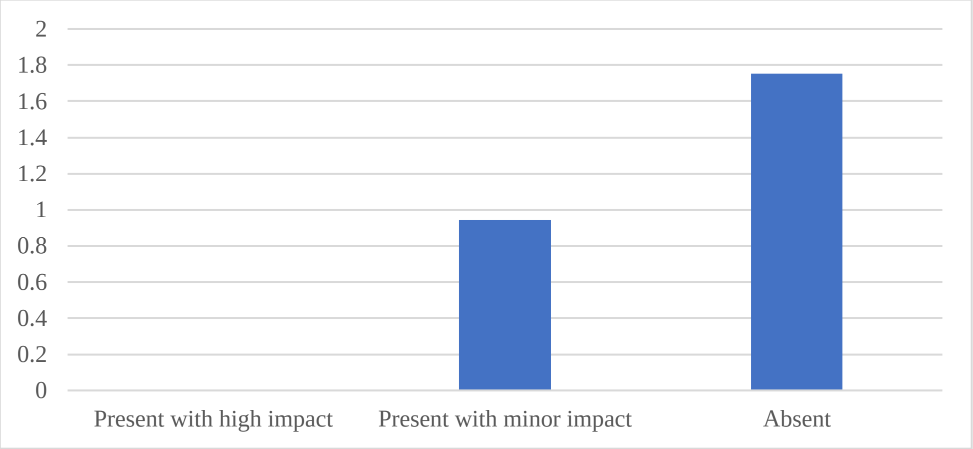


Figure 8.3: Utility course congenital co-morbidity

Ultrasound of brain

Figure 8.4 displays the utility course for the variable ultrasound of the brain. Similarly to the attribute congenital co-morbidity, the attribute ultrasound of the brain portrays a nearly linear utility curve. The utility course indicates that the utility contribution difference from a bad to intermediate prognosis is almost similar compared to the difference of an intermediate prognosis to a good prognosis. Hence, an increase from a bad to an intermediate prognosis has identical impact on the recommendation on surgery compared to an improvement from an intermediate prognosis to a good prognosis.

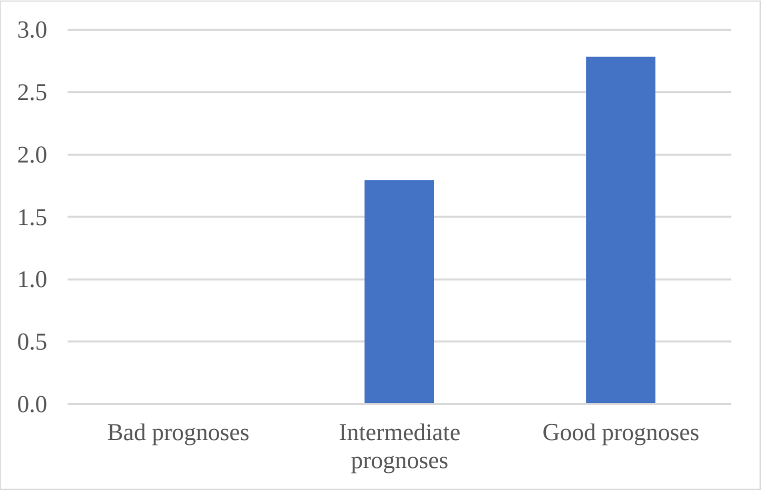


Figure 8.4: Utility course ultrasound of the brain

Wish of parents

Figure 8.5 illustrates the utility course for wish of parents. The utility course for the attribute wish of parents presents a non-linear utility course. The utility difference between in favour of comfort care and doubtful about surgery is larger compared to the difference between doubtful about surgery and in support of surgery. Hence, if for example, the wish of parents changes from doubtful about surgery to in favour of comfort care it would more significantly impact the recommendation of an UMCG physician compared to a wish that changes from doubtful about surgery to in favour of surgery. Furthermore, the parameters estimated for the levels of the variable wish of parents portray the lowest standard errors compared to the other dummy coded variables. Hence, the parameters for the wish of parents are most reliable compared to the parameters of the other dummy coded variables.

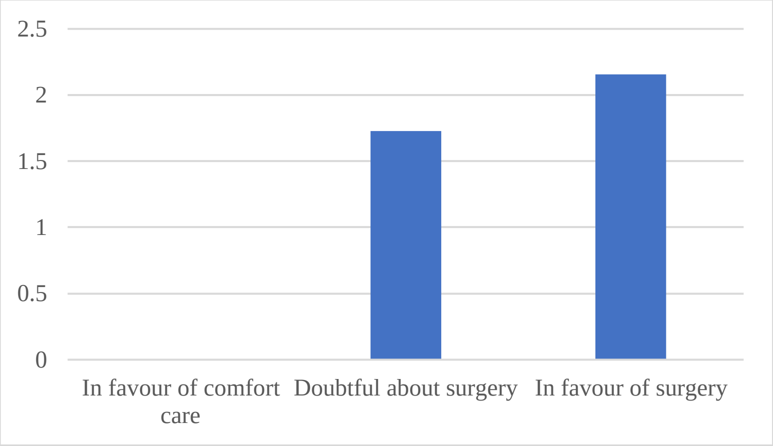


Figure 8.5: Utility course wish of parents

Other variables

The other parameters estimated by the binary logit model that are not yet discussed are perinatal asphyxia, progress since birth before a diagnosis of NEC, age since birth, growth since birth, lung function, hemodynamic, cerebral oxygenation, gender and the carrying capacity of parents. Compared to the estimated parameters of the dummy coded variables, the weights for these variables are considerably lower. Furthermore, the standard errors of these variables fluctuate around 0.2. Where gender is estimated with most uncertainty as it portrays a standard error of approximately 0.4 and hemodynamic with the least uncertainty as Table 6.5 illustrates the lowest standard error for hemodynamic. Comparing the parameters to examine the impact of the attributes on the recommendation for surgery is, however, tricky due to the different attribute ranges. Therefore, the maximum utility contribution and relative importance per variable is provided in the next section as the maximum utility contribution considers the range of an attribute.

8.3 Utility contribution and relative importance

This section illustrates the maximum utility contribution and relative importance of the attributes. These measurements are presented to the group of the UMCG physicians to provide introspection on their choice behaviour.

As explained in the previous section, the maximum utility contribution and relative importance per attribute are calculated because it is difficult to compare the utility contribution of the variables based on their parameter weights alone, due to the differences in attribute ranges. Figure 8.6 presents the maximum utility contribution per attribute, and Figure 8.7 shows the relative importance of each variable. It must, however, be taken into account that the maximum utility contribution and relative importance per attribute are still established on the range chosen for a variable. This research determined the variable ranges in consultation with the UMCG physicians. The ranges were drafted such that they capture at least 80% of the bulk of observations faced in reality. Therefore, although it must be taken into consideration that the maximum utility contribution and relative importance are established on the attribute ranges, the values comprised in the ranges are the values on which the UMCG physicians base their recommendations on in reality. Accordingly, the relative importance per variable is a relatively good representation of the importance per decision variable included in the stated adaptation experiment on the advice for neonatal surgery.

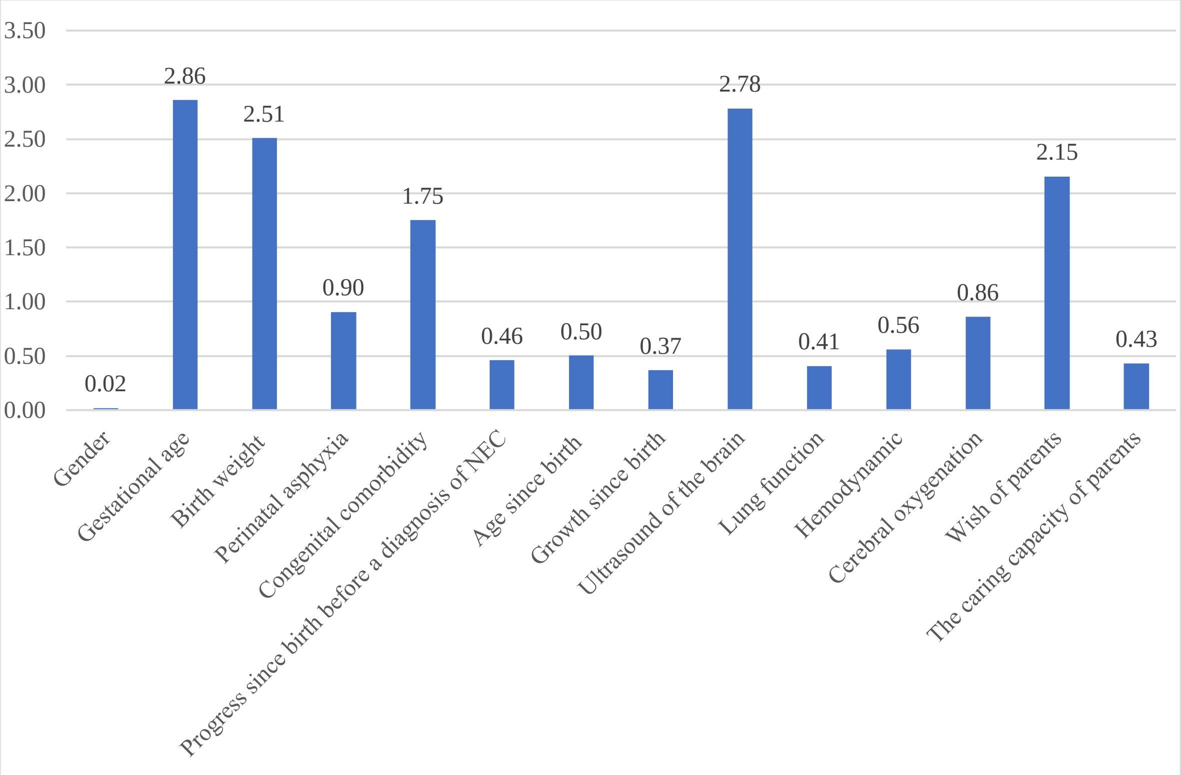


Figure 8.6: Maximum utility contribution per attribute

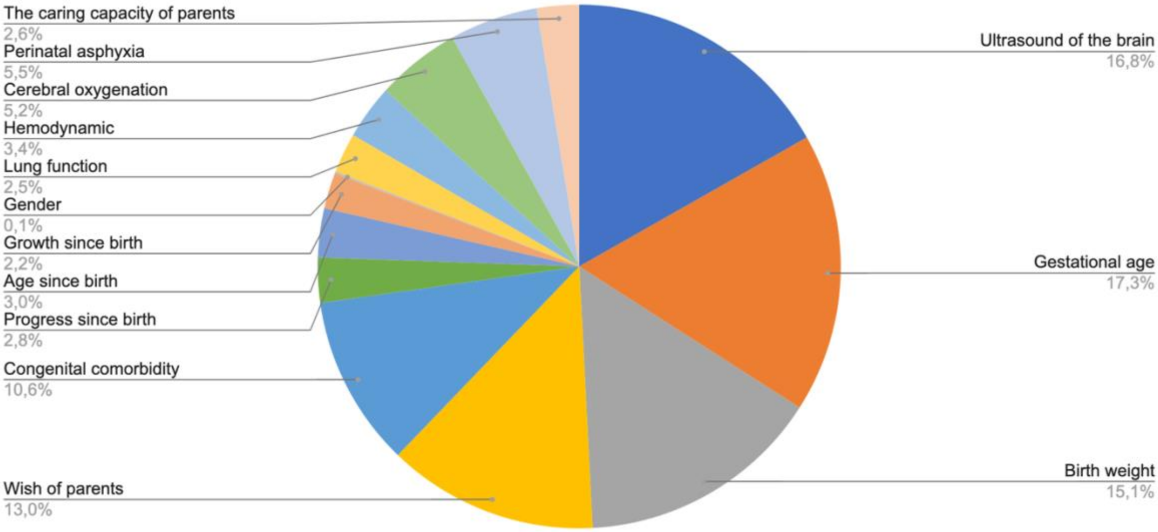


Figure 8.7: Relative importance per variable

Comparing the maximum utility contribution and relative importance of the variables provides introspection on the UMCG physicians choice behaviour. For, example, comparing the maximum utility contribution of the variables cerebral oxygenation and lung function portrays that the step from the lowest level (40) of cerebral oxygenation to the highest level (80) has more than two times the impact on the utility function then the step from a bad to a good lung function.
Moreover, Figure 8.7 depicts that gestational age, the wish of parents, birth weight, the ultrasound of the brain, and the congenital co-morbidity nearly make up for 75% of the relative importance; hence, the recommendation on surgery is largely determined by these variables. The other nine attributes have considerably less impact on the recommendation on surgery. The variable gender demonstrates to have the least impact on the advice for a preferred treatment and portrays a relative importance of 0.01%.

The most remarkable observation depicted in Figure 8.7 is that next to the four medical variables; gestational age, birth weight, the ultrasound of the brain, and the congenital co-morbidity, the wish of parents is also highly impactful on the recommendations of the UMCG physicians. During the discussion of the results, the UMCG physicians were not surprised by this result. One physician explained that the parents are the caretakers and must be willing and able to take care of the child. Therefore, if parents strongly prefer comfort care or surgery that plays an essential role for the medical recommendations of the physicians. The UMCG physicians were, therefore, rather surprised that the carrying capacity of the parents portrayed a considerable low relative importance.

8.4 Segmentation

In addition to providing the UMCG physicians with introspection on their choice task, this study also desires to determine whether professional and personal characteristics impact the experts' recommendations. Because an understanding of whether these characteristics, especially professional traits such as the length of professional experience or their specialism, impact their medical advice is, valuable information for the UMCG physicians.

As explained in Section 6.1 for each group of UMCG physicians in a specific personal characteristic category, a separate binary logit model, solely incorporating linear parameters, is estimated. Another possible way to investigate the impact of personal and professional characteristics is to include them as interaction effects and examine whether the model fit improves. This way, it is possible to investigate the impact of personal and professional characteristics together in one model compared to estimating an individual logit model per category. However, as the model already has to estimate 21 parameters for the attributes alone with limited choice data, this study has chosen to estimate a binary logit model per category.

Additionally, for the variable age, two binary logit models are estimated for an age category of 25-45 years old and 45 years old and above rather than for five groups which was asked for in the survey. The same applies to the variable length of professional experience for which two binary logit models are estimated for physicians with 0-10 years of experience and physicians with 10 years of experience or more. The categories are reduced to two instead of five categories because of the limited number of respondents. Since, estimating an binary logit model with two or three respondents per model will generate less reliable results, due to the small amount of choice data (information), compared to dividing the group of 15 physicians into two groups that are equally distributed.

Furthermore, for the personal characteristic variables gender, religion and adulthood, the distribution of physicians per category is also unequal. For example, the distribution for the variable religion shows that three physicians are religious, while twelve indicated not to be religious. Therefore, the number of physicians per category and the impact on the reliability of the estimated parameters is discussed later in this chapter.

This section will compare the distribution of recommendations in favour of and against surgery per category and it will compare the differences between the relative importance of the variables per category. The relative importance of the variables is the best measure to compare as it is calculated by dividing the maximum utility per attribute by the maximum systematic utility. Hence, it considers that each binary logit model obtains another maximum systematic utility. Therefore, for example, comparing the maximum utility contribution per variable will not be adequate as the maximum systematic utility differs per model. Thus, the maximum utility per attribute is incomparable between models. While, the relative importance is “relative” to the models’ maximum systematic utility and is, therefore, a better measure to compare.

8.4.1 Child surgeons versus neonatologists

This section describes the differences between the choice behaviour of UMCG child surgeons and UMCG neonatologists for the recommendation on surgery.

Firstly, it is essential to indicate that four UMCG physicians executed the stated adaptation experiment, while eleven neonatologists conducted the stated adaptation experiment. Appendix M presents the estimated parameters for both specialisms. Figure 8.8 presents the differences in relative importance per variable between child surgeons and neonatologists. This figure shows that the most considerable discrepancies exist between the variables gestational age, birth weight, congenital co-morbidity and the ultrasound of the brain. While the variables gestational age and birth weight portray the largest relative importance for neonatologists, these variable are considerably less important for child surgeons. Whereas the congenital co-morbidity and the ultrasound of the brain are the most impactful on the recommendation for child surgeons, these variables are considerably less important for the advice of neonatologists. Moreover, Figure 8.8 presents that for the less impactful variables, both groups, illustrate similar relative importance per variable.

Lastly, the percentage of recommendations that the UMCG child surgeons provided on the choice scenarios is 54% against surgery and 46% in favour of operation. In comparison, neonatologists recommended against surgery on 50% of the choice scenario’s and in support of surgery the other 50%.

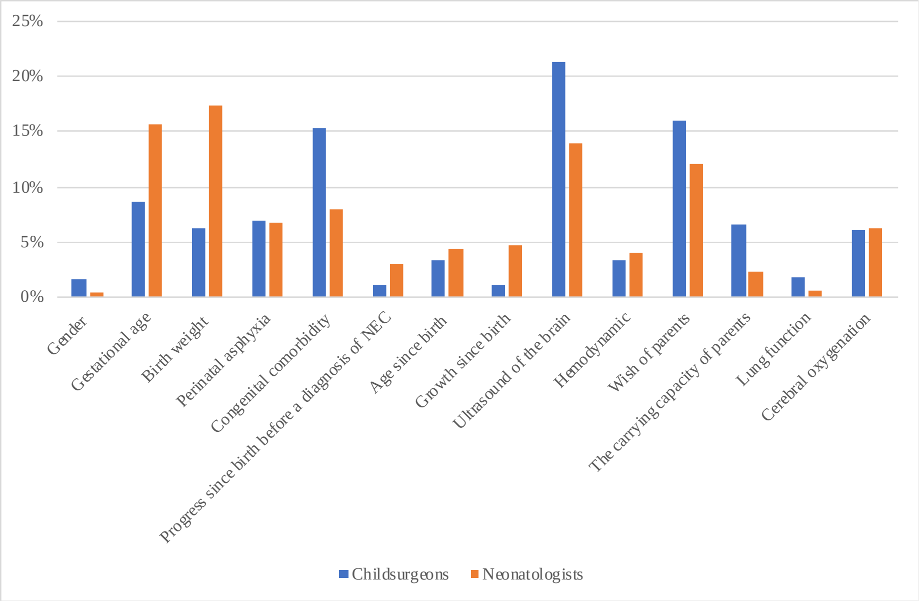


Figure 8.8: Difference in relative importance per variable between child surgeons and neonatologists

Impact of age and length of professional experience

Appendix N and Appendix O presents the estimated parameters for the age and length of professional experience categories. Figure 8.9 displays the relative importance per variable for the diverse age categories, and Figure 8.10 presents the relative importance per variable for the length of professional experience categories. The distribution of physicians over the categories for both variables are equally divided as there are 8 physicians between 25 and 45 years old and 7 physicians older than 45. While for the variable years of professional experience 7 physicians indicated to have 0-10 years' of experience and 8 indicated to have more than ten-year experience. It is likely that the choice data distribution over the categories for both variables is almost similar, with the exceptions of the choice data for one physician. The results support this assumption as the next paragraphs will show that the results are practically similar. Moreover, the standard errors for the parameters of the age and years of professional experience categories all fluctuate around 0.2 and 0.3, thus the parameters are approximately estimated with the same reliability.

Figure 8.9 and Figure 8.10 display, as was to be expected, a similar trend for the age categories and the professional experience categories on the relative importance per variable. The figures show that there are relatively small differences between the relative importance of variables between the age and years of professional experience categories. The most considerable distinction shown in the figures is that older and more experienced physicians portray a higher relative importance for the variable birth weight. In comparison, this variable is considered less important by younger and less experienced physicians.

Moreover, the percentage of recommendations that younger physicians provided on the choice scenario's is 46% against surgery and 54% in favour of operation. In contrast, older physicians recommended against surgery on 57% of the choice scenario's and in support of operation the other 43%. The same trend is detected for less experienced and more experienced physicians, as less experienced physicians recommended against surgery on 44% of the choice scenarios while more experienced physicians recommended against surgery on 57% of the choice scenarios. This displays that older and more experienced physicians are more inclined to recommend against surgery compared to younger and less experienced physicians.

The results that portray an inclination of older and more experienced UMCG physicians to recommend against surgery compared to younger and less experienced UMCG physicians match the results of the study that examined EoL decisions in seven European countries, including the Netherlands (Cuttini et al., 2000). This study found that more experienced physicians that were part of the study regardless of their origin more often decided in favour of ending the life of a new-born compared to less experienced physicians.

A possible explanation of this observation might be related to the confidence level that physicians attach to their medical recommendations. Research shows that less experienced physicians are often less confident about their decisions compared to more experienced physicians as findings suggest that more experience leads to an increased awareness of one's capabilities and, thereby, enhances a physician's confidence (Uy et al., 2014). Figure 7.1 in Chapter 7 shows that recommendations against surgery tends to be provided with slightly more certainty compared to a recommendation in favour of surgery, as explained in Section 7.3, plausible due to impact of recommending against surgery. Therefore, it is plausible to assume that the potential higher confidence of more experienced physicians might explain the higher inclination of these physicians to recommend against surgery compared to less experienced physicians. In order to examine this assumption, the average confidence level per category of the professional experience variable is calculated. The average confidence of less experienced UMCG physicians is estimated at 69% while for more experienced physicians, it is calculated at 78%, which is a considerable difference. This observation could, hence be part of the reason that more experienced physicians more often recommend against surgery.

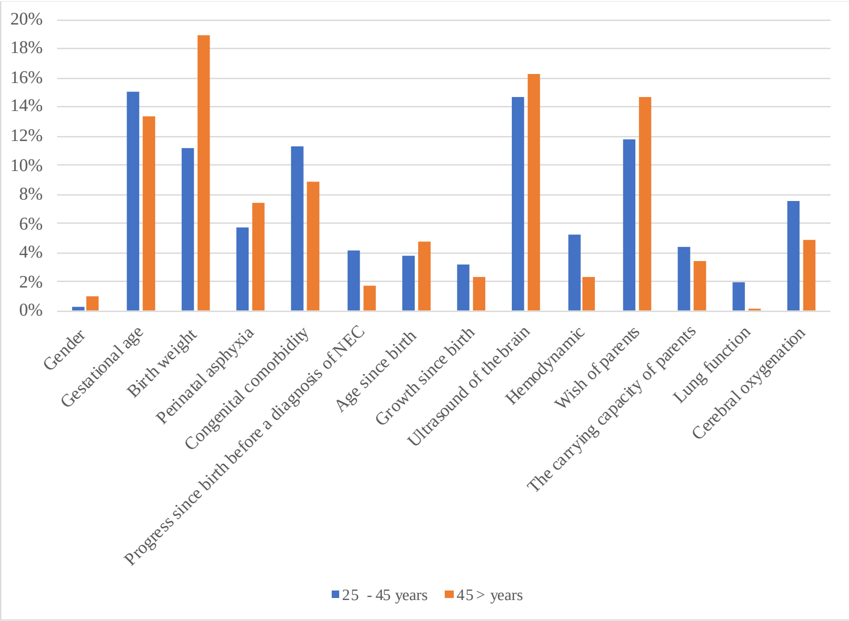


Figure 8.9: Difference in relative importance per variable between age groups

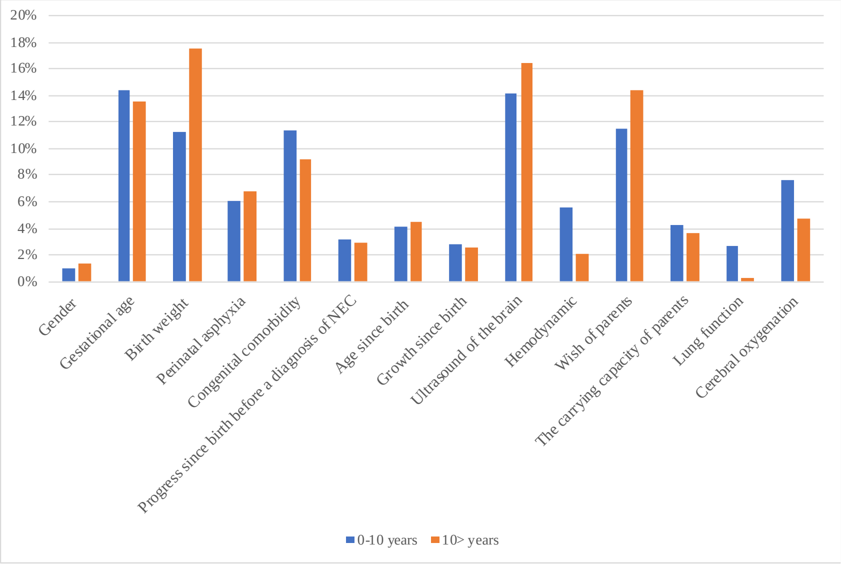


Figure 8.10: Difference in relative importance per variable between diverse years of professional experience categories

8.4.2 Gender

This section describes the differences between the choice behaviour of men and women in the group of UMCG physicians.

Five men and ten women executed the stated adaptation experiment. That entails that the parameters for the men are estimated with greater uncertainty compared to those for women. Appendix P illustrates the estimated parameters for the categories of men and women. Figure 8.11 demonstrates that the difference in relative importance per variable between men and women are relatively small. Additionally, the men recommended in favour of surgery at 45% of the choice scenarios while the women recommended in favour of surgery at 51% of the choice scenarios. Hence, the men in the group of UMCG physicians have a slightly larger inclination to advice against surgery compared to the women. The results must, however, be interpreted with care as the group of women that conducted the choice experiment is twice the size of the group of men.

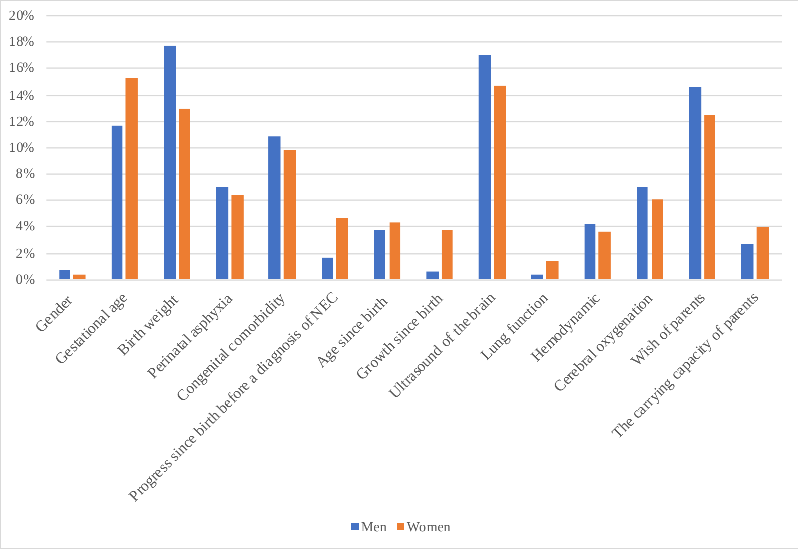


Figure 8.11:Difference in relative importance per variable between men and women

8.4.3 Religion

Appendix Q presents the estimated parameters for religious and non-religious UMCG physicians. The stated adaptation experiment was executed by three religious physicians and twelve non-religious physicians

Figure 8.12 depicts that the relative importance for the variables is in comparison for both groups small. The most considerable discrepancy is illustrated for the variable congenital co-morbidity. Congenital co-morbidity tends to be more important for the recommendations of non-religious physicians compared to religious physicians. Furthermore, the percentage of recommendations in favour of surgery that the religious UMCG physicians provided on the choice scenarios is 46% compared to 50% for non-religious physicians. The difference in percentage for the recommendations in favour of surgery is also a considerably small.

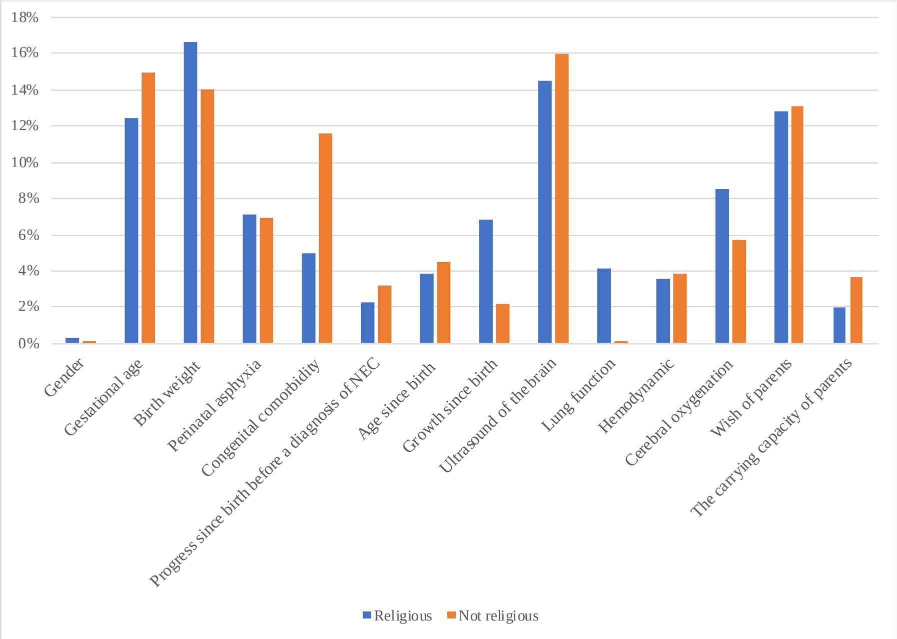


Figure 8.12: Difference in relative importance per variable between religion categories

8.4.4 Parenthood

Four physicians without children and eleven UMCG physicians with children executed the choice experiment. Appendix R illustrate the estimates for physicians with and without children.

Furthermore, Figure 8.13 portrays the relative importance of the variables for the two categories. Similarly to the variable religion and gender, there are relatively small differences between the two categories. The variable growth since birth is more important for the recommendations of physicians with children, while this variable is less important for physicians without children. Additionally, the variable growth since birth, also, portrayed one of the lowest standard errors for both binary logit models. Hence, this variable is estimated with greater reliability compared to multiple other parameters.
Moreover, the physicians with children recommended in favour of surgery at 51% of the choice scenarios while the physicians without children recommended in support of surgery at 44% of the choice scenarios. Therefore, the UMCG physicians with children show a slightly greater inclination to recommend in favour of surgery.

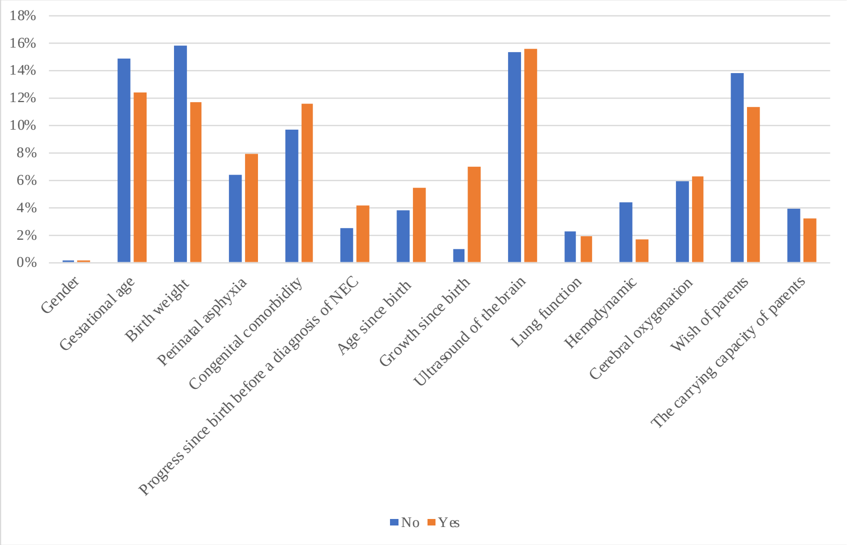


Figure 8.13: Difference in relative importance per variable between parenthood

As explained in Section 4.2, multiple studies illustrate deviating results on the influence of personal and professional characteristics on EoL decisions. Therefore, it is complicated to state that this particular group of child surgeons and neonatologists illustrated decision behaviour found in earlier studies.

One of the most distinct observations elicited in this section is the difference between the most impactful variables for child surgeons and neonatologists. Also, another noticeable observation is the fact that older and more experienced UMCG physicians more often provided recommendations against surgery compared to younger and less experienced physicians.

The results in this section, however, must be interpreted with care. Firstly, because the sample sizes for all categories are small, secondly, because, exempt from age and length of professional experience, the sample sizes between the categories significantly differ. Hence the parameters for one category are estimated with more uncertainty compared to the parameter for the other category.

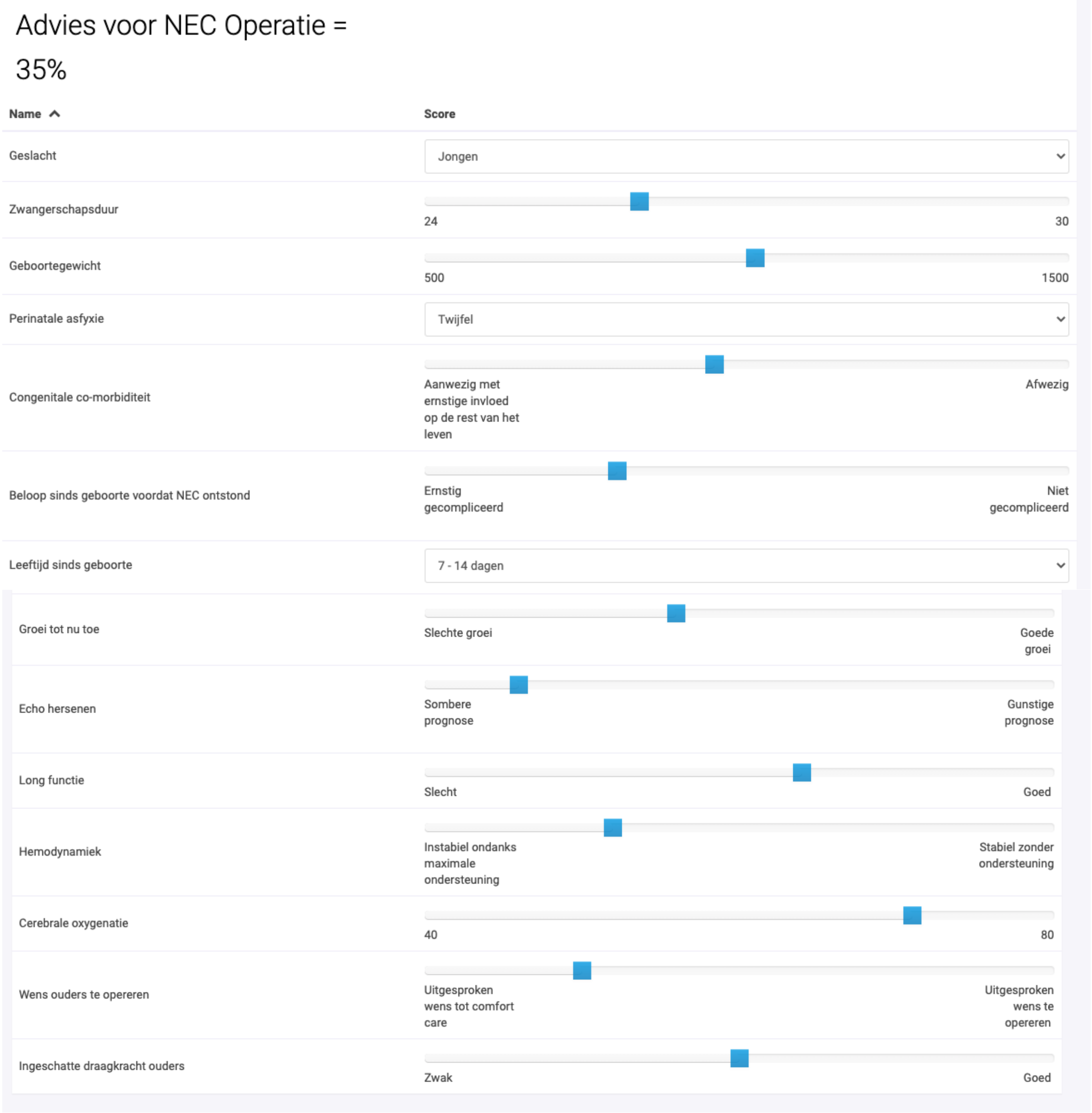
8.5 Interpretation of choice probabilities

As explained in Chapter 2, DCM is used to analyse choice behaviour and predict future choices of individuals. BAIT utilises the choice behaviour analysis to provide introspection on experts' decisions. This research conducted a choice behaviour analysis to provide introspection to the group of UMCG physicians on the recommendation for surgery on a premature neonate with NEC. The choice behaviour analysis was discussed in this chapter. This section explains how to interpret the choice probabilities estimated by the binary logit model.

For BAIT to be implemented for decision support, firstly the parameter weights are converted into utils per unit instead of utils per attribute level step. Appendix L illustrates the converted parameters into utils per unit. Appendix L shows that multiple parameters are converted into utils per percentage. These variables are converted into utils per percentage as, the UMCG physicians involved in the design of the choice experiment, explained that it's not as simple as, for example, a weak, moderate or good lung function. Sometimes the lung function is not definable as moderate or good but falls in-between a moderate or good lung function. Hence, the parameter values are converted into utils per percentage, such that 0% represents a weak lung function and 100% a good lung function. This way, for example, 80% indicates a lung function in-between moderate and good. The same argument holds for the other variables expressed in utils per percentage. However, Appendix L shows that for the variable perinatal asphyxia, the parameter is not converted into utils per percentage but is kept the same. This is because, unlike the other variables, for perinatal asphyxia, there are no in-between levels.

Figure 8.14 provides an example of a hypothetical future scenario and the corresponding probability for advice in support of operation calculated by the binary logit model. The probability presented in Figure 8.14 can be interpreted in multiple ways. It, firstly, indicates that 35% of the group of UMCG physicians would recommend in favour of surgery. Or, the chance that a randomly selected physician out of the group of UMCG physicians provides a recommendation in support of operation is 35 %.

In conclusion, as discussed in Section 3.1.2, implementing BAIT to aid decisions in the future, thus, can be considered as asking the entire group of colleague for advice since the likelihood in favour of surgery represents the percentage of colleagues that would provide a recommendation in favour of surgery.



Advies voor NEC Operatie =

35%

Name ^	Score
Geslacht	Jongen
Zwangerschapsduur	24 30
Geboortegewicht	500 1500
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	40 80
Wens ouders te opereren	Uitgesproken wens tot comfort care Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed

Figure 8.14: Example model calculation

A further discussion on how BAIT can aid decisions on the recommended treatment in the future is presented in the next chapter.

9 Conclusion and Discussion

This chapter aims to answer the research questions. Section 9.1 first discusses how BAIT can support future recommendations of the UMCG physicians to answer research question 3. Thenceforth, Section 9.2 deliberates on whether BAIT has potential to support the future recommendations of the UMCG physicians. Finally, Section 9.3 discusses whether BAIT has potential to serve as a novel CDSS in the medical sector.

9.1 Research question 3: How can BAIT support the recommendations of UMCG physicians on surgery?

This section discusses how BAIT can support the medical recommendations of the UMCG physicians. As explained in Section 1.3, it is essential to understand the decision-making process towards the final decision on surgery to comprehend how BAIT can best be implemented to support the recommendations of the UMCG physicians. Moreover, before arguing how BAIT could support future recommendations, it is essential to understand the purpose of decision-support. Accordingly, this chapter firstly elaborates on the decision-making process towards the final medical advice on surgery and then discusses the aim of implementing BAIT. Thenceforth, it will discuss how BAIT can support future recommendations.

When the UMCG physicians recognise that surgery is required to sustain the life of a new-born with NEC, the process of determining what final recommendation to provide to parents is initiated. Generally, after recognising that surgery is required to sustain life, the physicians have a few hours to a day to perform surgery. During the time that the ill new-born is hospitalised, the UMCG physicians are in close contact with the parents of the child. Therefore, the final judgement on surgery precedes a process of numerous conversations and consultations with the parents of the new-born. Hence, before giving final advice, the physicians are already aware of the wish of parents regarding surgery. In the plenary meeting that discussed the results, the UMCG physicians explained that in most cases, the final medical advice of the professionals is excepted by the parents and their judgement proceeds. Because of the ethical nature of the decision, an UMCG physician always consults a colleague before giving the final recommendation, even when convinced about his preferred medical advice. In addition, when an UMCG physician has doubts about performing surgery, multiple colleagues are consulted. The process of consulting colleagues results in a recommendation that is not solely dependent on the professional expertise of one doctor but incorporates the opinion of various experts to establish a thoroughly considered recommendation.

Before examining how BAIT can support future recommendations, it is essential to define the purpose or aim of the decision support that BAIT can offer. As explained in Chapter 4 the decision on surgery is a complex and ethical dilemma that induces a heavy decision burden on the professionals. Physicians solely provide a recommendation against surgery when the quality of life is expected to be so bad that continuing procedure will cause unnecessary suffering for the child. Determining the quality of life is, however, very challenging as the short term outcomes, but especially long term outcomes, years after surgery are hard to predict. Furthermore, the norm “quality of life” is implicit. What a doctor determines as an adequate quality of life is influenced by his norms and values and driven by past experiences.

Moreover, parents can also have different opinions on the definition of a good quality of life, and, finally, the neonate cannot express his or her opinion of an acceptable quality of life. Therefore, it is difficult to define what a ‘good’ recommendation on surgery is based on the norm “quality of life”. Hence, to state that BAIT should improve the quality of the medical recommendations is unattainable. Firstly, because a decision against surgery results in the death of the child, thus, it is impossible to determine whether performing surgery would have been the better option. And, secondly, when surgery is performed, and a neonate has grown up, the definition of a good quality of life can still not be defined due to different perceptions on what an adequate quality of life entails. Therefore, rather than stating that BAIT should improve the quality of a recommendation, it is more appropriate to state that BAIT should help to conceive highly contemplated and transparent recommendations.

Before deliberating how the generated choice probabilities by BAIT can aid future recommendations to help establish more considered and transparent recommendation, this paragraph will first discuss how the introspection on the choice task of the UMCG physicians by BAIT may already support future judgments. As explained in Section 1.6 the third phase of the project set-up with the UMCG was a plenary meeting to present the results of this study. Appendix U includes a brief discussion of this meeting. During this meeting, the results already triggered discussions among the experts. After the meeting, a few physicians declared that the discussions were valuable as the physicians started to reflect on their recommendations on surgery critically. Hence, the internal debates among professionals stimulated them to deliberate their recommendations and helped them to understand why differences in recommendations exist between colleagues. Besides triggering internal discussions among professionals, the introspection can also be utilised for educational purposes. For example, to show PhD students or starting physicians, on what grounds more experienced physicians provide recommendations on surgery.
Moreover, knowledge on the importance of the variables helps to determine for which variables is it highly valuable further to investigate the impact on the expected quality of life. For example, the results show that the ultrasound of the brain greatly influences the recommendations of the physicians. Hence, based on their medical opinion, the ultrasound of the brain is a critical indicator of the expected quality of life. Therefore, gaining a better understanding of the impact of, for example, a bad brain prognosis on the predicted quality of life helps to generate more considered recommendations on surgery.

Furthermore, the generated choice probability by BAIT can also support future recommendations. As discussed in Section 8.5, the choice probability indicates the percentage of UMCG physicians that would recommend in favour of surgery for a neonate with specific variable values entered into the model. Accordingly implementing BAIT to aid recommendations in the future, can be considered as asking the entire group of colleagues for medical advice. Thus, for every premature new-born in need of surgery to sustain the neonate's life, an UMCG physician can consult all colleagues without having to contact each physician physically. The opportunity to ask advice of all colleagues at once can aid in making thoroughly contemplated recommendations and supports the already practised decision-making process, as UMCG physicians always consult one or multiple colleagues before giving a final recommendation.

Section 7.3 further examined how to interpret the choice probabilities generated by BAIT based on the confidence levels provided by the UMCG physicians on the choice scenarios incorporated in the stated adaptation experiment. The analysis illustrated a strong linear relationship between the spread of BAIT's generated choice probabilities and the provided confidence levels by the UMCG physicians. This strong relationship entails that the most appropriate way to interpret the choice probability generated by BAIT is that a choice probability equal to 50%, which indicates a 100% spread, in most cases illustrates that the group of UMCG physicians are less certain about which recommendation to provide. In contrast to

50% of the group of UMCG physicians being very confident that advice in favour of surgery is the most appropriate recommendation and the other 50% that a recommendation against surgery is the best option.
Contradictory, a decrease in spread, hence, choice probabilities of, for example, 80% in favour of surgery and 20% against surgery are related to higher confidence levels. Later in this section, a discussion on the consequences of this finding for the decision support that BAIT can offer is provided.

Furthermore, both the ability to provide introspection on the UMCG physicians choice task by BAIT and the generated choice probabilities have the opportunity to mitigate overconfidence. Section 7.3 introduced the term overconfidence in clinical decision making. Overconfidence of physicians often occurs in the context of judgement and decision making in health care. It is recognized as a common cognitive bias (Pat Croskerry & Norman, 2008). Overconfidence increases the assumption of being correct, which decreases the seek for information or advise of colleagues that, otherwise, might have increased the chance of a better clinical judgement. Research illustrates that the more challenging a clinical task gets the higher chance of overconfidence, which is referred to as the "hard-easy effect" (Yang, Thompson, & Bland, 2012). As the choice task of the UMCG physicians studied in this research is a difficult choice task, due to the multiple reasons explained in Section 4.1, the chance of this effect occurring is probable.

Numerous approaches might be applied to mitigate overconfidence. One of the best ways to reduce overconfidence and improve decision making in health care is to provide the physicians with the clinical outcome of their decision to understand whether their medical judgement was, indeed, correct. Therefore, decision-making in health care would greatly benefit from reliable feedback on the clinical outcome of a physician's judgment. Yet feedback on clinical outcomes rarely exists in practise (P. Croskerry, 2000) During the final meeting that discussed the results, the UMCG physicians explained that they often do not receive feedback on the wellbeing of a child they have performed surgery on in the past.
Since feedback on the clinical outcome of physicians' judgement is usually not provided, it is reasonable to expect that the experts are apprehensive of how they make clinical judgements. Hence, it is somewhat disconcerting to observe that physicians are often unaware of their implicit decision rules. A study examined that physicians are unable to explain what their thought process was while making a clinical judgment in the past (Pat Croskerry & Norman, 2008). This study, however, explains that this is hardly shocking. It states that " experts are experts in part precisely because they have solved most problems before and need only recognize and recall a previous solution" (Pat Croskerry & Norman, 2008). In other word, experts make a habit of matching new problems with situations they solved in the past. This habit fits with what was explained earlier in this research, which illustrated that a physician greatly determines his or her recommendation on surgery based on past experiences. This pattern, however, entails that the decision behaviour of physicians remain invisible and habits of overconfidence are kept cloaked. Therefore Croskerry et al. (2008) states that physicians would benefit from insight in their decision behaviour by, for example, explicating their decision rules to understand if their judgment and decision-making process is correct. This information could help reduce cognitive bias and result in more contemplated decisions.

The ability of BAIT to provide introspection on experts' decisions by making implicit decision rules of experts explicit, hence, offers the opportunity to mitigate overconfidence and enhance more contemplated judgements. As explained earlier in this section, the provided introspection in this research already induced valuable discussions among the UMCG physicians and

provoked an enhanced reflection on their medical recommendations. Moreover, the generated choice probabilities also have the opportunity to mitigate overconfidence. Firstly, generated choice probabilities close to, for example, 80% in favour of surgery and 20% against surgery, can conceivably deal with this cognitive bias. It will, first off, make a physician who believes surgery is the appropriate treatment for this specific case feel strengthened in his judgement.
In contrast, it could result in a physician who believes that not performing surgery is the appropriate treatment but observes that 80% of his or her colleagues feel confident that performing surgery is the best treatment to further contemplate his or her own judgement. This realisation can mitigate overconfidence as it shows that his or her decision might not be the appropriate treatment. The UMCG could, for example, introduce a threshold which determines that if BAIT generates a choice probability of, for example, 70% in favour of surgery and a physician believes comfort care is the most appropriate treatment. The physician must at least contact three or more colleagues such that it results in thoroughly considered medical advice.

Choice probabilities closer to 50% in favour of surgery and 50% against surgery are a bit trickier to support a physician’s judgement directly. Since, if a UMCG physician doubts his or her recommendation, a choice probability of 50% will just illustrate that his or her colleagues are also doubtful about what medical advice to give. As explained earlier choice probabilities close to 50% are related to more uncertain recommendations. A choice probability of 50% does demonstrate that it is essential to discuss the case with multiple colleague physicians. Furthermore, since BAIT provided information about what variables are most impactful on the UMCG physicians recommendations a choice probability close to 50% can also indicate to examine the impactful variables further in order to achieve a recommendation in favour of or against surgery with more certainty. As explained earlier in this chapter, it would, therefore, also be beneficial to further investigate the impact of, for example, a bad brain prognosis on the predicted quality of life such that better judgements can be provided.

Furthermore, the fact that BAIT can provide transparent and explainable decision support due to its system characteristics also supports the ethical duty of physicians to transparently explain their medical recommendations to patients, with or without a CDSS (Lysaght, Lim, Xafis, & Ngiam, 2019). Explicating the expertise of experts, however, goes against Freidsons’ theory. Freidsons’ theory explains that experts instead desire their expertise to remain implicit as it provides them with professional autonomy and describes that professional autonomy is the defining characteristic of professional power and prestige (Freidson, 1970). Therefore, the next sections will briefly discuss BAIT’s possible impact on the professional autonomy of the medical experts and the potential effect on the acceptance of BAIT. The systems must, after all, be used by the physicians themselves and, hence, must be accepted.

Next section discusses whether BAIT has potential to support future recommendations of the UMCG physicians based on the deliberation of trustworthy AI and CDSSs provided in Chapter 3 and 4.

9.2 Research question 2: Does BAIT have potential to support medical recommendations of the UMCG physicians on surgery?

The previous section discussed how the introspection and the generated choice probabilities by BAIT could support the medical recommendations of the UMCG physicians. This section elaborates on whether BAIT has potential to support the medical recommendations in the future. It examines it’s potential based on its system characteristics and through a discussion on the systems trustworthiness and possible user acceptance of the UMCG physicians.

Firstly, the main focus of this research was codifying the domain expertise of the UMCG physicians through choice modelling to provide introspection on their choice task.
The last section explained that the introspection on the UMCG physicians choice task can aid in a possible reduction of overconfidence and result in more thoroughly contemplated recommendations. As explained earlier in this study, the result of this study were presented in a plenary meeting to a large number of the physicians that executed the stated adaptation experiment. In the meeting, the results triggered discussions among the professionals and provoked a reflection on their own expertise. Moreover, the illustrated differences between, for example, child surgeons and neonatologists provided valuable insights to the UMCG physicians and helped physicians to understand why colleagues might have contradicting believes on what the best recommendation on surgery is for specific cases. Overall, the physicians valued the introspection on their expertise. As the UMCG physicians valued the introspection, Frauds theory on experts rather keeping their expertise implicit, therefore, does not seem relevant for this particular group of medical experts. It must, however, be pronounced that not all UMCG physicians that executed the stated adaptation experiment were present at the meeting to discuss the results. Therefore, this discussion does not include their opinion on the added value of the introspection. Moreover, Section 9.1 discussed how the introspection could support future recommendations, whether it will actually help to generate more considered recommendations cannot be concluded at this moment in time, but can only be determined in the future.

Furthermore, as explained in Section 9.1, the generated choice probabilities by BAIT can also be utilised for decision-support. Several UMCG physicians, during the plenary meeting, pronounced that they would accept and appreciate the support of BAIT for future recommendations. This research, however, solely used BAIT to provide introspection on the choice task of the UMCG physicians. The following paragraphs will, hence, discuss the potential of BAIT for future decision support for this choice task. Still, it will not declare whether BAIT can be prosperous for decision support, as it is impossible to make that assertion without having implemented BAIT for decision support.

The following paragraph will first deliberate whether BAIT has the potential to be implemented as decision support on the choice task investigated in this research based on its system characteristics. Chapter 3 described the system characteristics of BAIT and of the currently deployed knowledge and non-knowledge based IDSSs. It presented the differences and similarities between the systems and explained how the systems establish decision support. It illustrated that knowledge-based systems translate knowledge of experts captured in a knowledge base into a set of rules. Hence, the decision support is established on rules for which threshold values determine whether a decision follows one path or another path. Chapter 3 explained that in the health care sector, non-knowledge based systems are more commonly applied due to the complex nature of the problems faced in health care for which rule-based systems are unsuitable. Take, for example, the choice task studied in this research. A rule-based system would provide a recommendation in favour of or against surgery based on threshold values, for example, a premature neonate that weighs less than 600 grams should not be operated. Such an if-then statement does not work in this context as the UMCG physicians establish recommendations by making trade-offs between decision variables.
In contradiction, BAIT does ground it’s decision support on the trade-offs that experts’ make between decision variables. Hence, it reflects how the physicians make recommendations and can, thereby, mirror the decision behaviour of the UMCG physicians.

Moreover, similarly to non-knowledge based systems, it provides decision support in the form of a likelihood instead of a true value like that of a knowledge-based system. It is reasonable to assume that decision support in the form of a probability is more likely to be accepted by the physicians compared to decision support that would either state: operate or comfort care. Because contracting advice compared to a physician's preferred treatment without an indication of the extent of the difference might results in scepticism for the system. Furthermore, unlike non-knowledge based systems for which the decisions generated are opaque due to its black box technology, BAIT generates explainable decision support. And, as explained earlier in this chapter, physicians have the ethical responsibility to transparently explain their medical recommendations to patients or, in this case, the parents of the patient. Therefore, the explainability of the system will help them achieve this duty.

Accordingly, the system characteristics of BAIT portray that for decision support in this context, BAIT seems promising, especially, compared to knowledge and non-knowledge based systems.

Moreover, as explained in Section 3.2, the level of trust determines an individuals' behaviour towards AI. And, research shows that the level of trust is a fundamental reason for AI acceptance, especially in sensitive areas such as health care. Hence, to increase the likelihood for a successful implementation of BAIT for future decision support of the UMCG physicians recommendations, the system must be perceived as trustworthy. Therefore, the next paragraphs will discuss the trustworthiness of BAIT.

Section 3.2 illustrated that trust is a dynamic process that involves a graduate alteration from an initial trust to continuous trust. Initial trust helps to tackle initial conceptions of uncertainty and risk, while continuous trust aids to establish the continued implementation of the AI technology. Section 3.2 explained that initial trust for an AI technology highly depends on the explainability, trialability and representation of the technology. The explainability of the system entails to what extent the users of the system understand the procedures of a technology. The procedure steps of BAIT are relatively simple to explain and understand. It involves the design of an stated adaptation experiment, thenceforth, the execution of the stated adaptation experiment by the group of experts to generate the choice data on which BAIT grounds its decision support. And, lastly, it manipulates the data by using a validated technique into decision-rules that can support future decisions. The simplicity of BAIT's method helps to enhance the explainability of the system but, primarily, the involvement of the experts themselves during the process steps of BAIT provides the future users (the experts) with an understanding on the procedures of BAIT.

Moreover, the fact that the decision support is based on the recommendations of the UMCG physicians themselves helps to enhance the level of representation of the experts. Section 3.2 explained that a technology that mimics the behaviour of humans helps to establish an initial trust level. Since BAIT is dependent on the choices of the UMCG physicians on the stated adaptation experiment, thus, dependent on the behaviour of the physicians themselves, it enhances the level of representation that aids in building an initial trust level.

Lastly, Section 3.2 illustrated that trialability of the technology, also, determines the initial trust for AI technology. Therefore, for prosperous implementation of BAIT it is essential that the experts have to opportunity to test the system.

In conclusion, BAIT's system characteristics have the potential to establish a high level of initial trust that will help to reduce the uncertainty and risk conceptions of the UMCG physicians on the system.

Furthermore, Section 3.2 explained that the system's performance primarily impacts the continuous trust for AI technology. The system must be reliable and accessible to enhance the level of trust. Reliable decision support is, especially, essential for the choice task of the UMCG physicians investigated in this research, as it concerns an EoL decision. Accordingly, BAIT must establish reliable decision support. Therefore, the reliability of the system must first be tested. A possible way to test the reliability is to utilise BAIT, after a recommendation on surgery is provided, and determine whether BAIT estimated the probability on surgery correctly. To be able to determine whether BAIT identified the probability correctly, the group of UMCG physicians must all determine what medical advice they would have provided given that specific case. This can be considered as time-intensive for the UMCG physicians. Section 3.2, however, also elicited that testing the system will enhance the acceptance of a CDSS by its users; hence, by the UMCG physicians. Thus, if testing the system the way described above determines that the system is reliable, it will aid in an enhanced acceptance of the UMCG physicians when the system might eventually be implemented for decision support. Contradictory, if BAIT, turns out to be unreliable as the choice probabilities do not mirror the recommendations of the UMCG physicians as good as they expected BAIT to reflect their choice behaviour that will, of course, cause the UMCG physicians not to trust BAIT. Users that do not trust an AI technology has, as explained earlier, severe consequences on the prosperous implementation of a novel AI technology as it will negatively impact the acceptance of users for the technology.

Moreover, as explained in Section 4.3 and which was briefly discussed in the last section, is that physicians worry that a CDSS may reduce their professional autonomy and could be used against them when medical differences arise. The physicians worry that a CDSS might impact their professional autonomy as they feel they are expected to act by the judgment provided by a CDSS. A CDSS can, however, also enhance the collective professional autonomy of physicians since if experts have access to a system that enables them to support their judgments to patients and possibly third parties, when questioned about their decision, it can protect their professional autonomy. For his matter is it important that a CDSS provides explainable and transparent decision support, otherwise, the supported judgments can still not be transparently explained to patients or third parties. As BAIT provides explainable decision support, it is able to support the collective professional autonomy of medical experts. Therefore, it illustrates the trade-off between defending collective professional autonomy by limiting individual professional autonomy. The acceptance of a reduction of individual autonomy significantly differs per individual physician and the institutional environment an expert operates in (Armstrong, 2002). Hence, whether physicians are willing to trade off individual autonomy for an enhanced collective autonomy supported by BAIT is, yet, to be determined.

As explained earlier, the UMCG physicians valued the introspection on their expertise. Hence this contradicts Frauds theory on experts rather keeping their expertise implicit. An introspection on their own choice behaviour is, however, significantly different from appreciating the aid of a new CDSS that will support their medical recommendations. Therefore, it is trickier to state that the UMCG physicians will accept the assistance of BAIT and not worry about a breach on their individual autonomy.

Hence for successful implementation of BAIT, it is essential to stress that the system solely desires to help establish more considered recommendations and to ease their decision burden compared to substituting the expert. This is essential because the UMCG physicians are the future "users" of BAIT. Hence, its aid must be accepted by these professionals.

The discussion in the paragraphs above illustrates that BAIT does have a legitimate potential to support the future recommendations of the UMCG physicians. Although the debate on BAIT in the paragraphs above illustrate that BAIT has potential to be implemented for decision support on the choice task investigated in this research, the implementation of BAIT for decision support can only determine whether the UMCG physicians will accept and appreciate the aid of BAIT. Lastly, while applying BAIT for introspection on the UMCG physicians' choice task, this study also discovered several hurdles that BAIT needs to tackle before being implemented for decision support. The next chapter discusses these hurdles and explains the limitations of this research. Firstly, the next section discusses whether BAIT has potential to constitute a novel type of CDSS in the medical sector.

9.3 Research question 3: Does BAIT have potential to constitute a novel type of IDSS in the medical sector?

The previous section described the potential of BAIT to support future recommendations on surgery of the UMCG physicians. This section will interpret the lessons learned in this case study and research to discuss the potential of BAIT as a novel CDSS in the medical sector.

Firstly, the system characteristics of BAIT illustrate BAIT's significant potential to serve as a novel CDSS in the medical sector, especially, compared to the currently deployed CDSSs. Most problems faced in the medical sector are too complex and impossible to be supported by knowledge-based systems. As the procedure of decision support of knowledge-based systems follows rules to derive at a true value for decisions support. This procedure is not applicable for most medical judgments, as usually, medical experts make decisions based on trading off decisions variables to arrive at the best possible solution. That's why non-knowledge based systems are, currently, more often enforced in the medical sector compared to knowledge-based systems as it can organise and search for patterns in big data sets to generate fast and precise decisions. The black box characteristic of a non-knowledge based system, however, is a large disadvantage, primarily, in the medical sector as physicians have the ethical duty to explain their clinical judgements to their patients.

In contrast, BAIT can support medical judgements and explain how the system derived at its conclusions; hence, it provides explainable decision support. Additionally, as the decision support encloses the trade-offs that experts' make between decision variables, it mimics the decision behaviour of medical experts. As explained in Section 4.3, currently, most CDSSs do not include and reveal the decision-making processes of the medical professionals themselves, which is a reason for physicians to be hesitant about the implementation of CDSSs. BAIT grounds its decision support on the decisions and decision rules of experts themselves which enhances the level of representation of the professionals. The illustrated system characteristics of BAIT, hence, seems fitting for a novel CDSS in the medical sector.

Moreover, the discussion on the trustworthiness of BAIT in the previous section illustrates that BAIT has the potential to constitute a high level of initial trust due to the system's explainability and the high level of representation of the experts. As explained earlier, the level of trust determines an individuals' behaviour towards a novel AI technology, and research shows that the level of trust is a fundamental reason for AI acceptance. Additionally, the trustworthiness of an AI technology is, incredibly, important in sensitive areas, such as the medical sector, that deal with ethical and impactful decisions, such as the EoL decision investigated in this research. Hence, the fact that BAIT seems to portray the potential of achieving a high level of initial trust, also, illustrates that it has legitimate potential to serve as a novel CDSS in the medial sector.

Furthermore, as elucidated the continuous trust of an AI system is significantly impacted by the performance and reliability of the system. Hence, to achieve continuous trust that enables BAIT to be prosperous for decision support in the medical sector, BAIT must first prove that the system is reliable by testing its performance with its users. Moreover, although BAIT has the potential to support the collective professional autonomy, it does impact the professional autonomy of the individual physicians. The last section described that this could possibly form a hurdle for the prosperous implementation of BAIT as personal autonomy is the defining characteristic of professional power and prestige that physicians might not desire to lose.

As explained in the last section, for this case study, the UMCG physicians did appreciate the introspection on their expertise. Also, they seemed willing to accept the aid of BAIT for decision support in the future. The acceptance for a novel CDSS and worries about the possible reduction of professional autonomy, however, greatly differs per individual. Hence, it is likely that this might be a significant hurdle BAIT must tackle for successful implementation.

Additionally, a study that examined the digital transformation in health care across Europe found that the Netherlands is a precursor in the digitalisation of health care compared to other European countries ("Nederlandse gezondheidszorg is digitale voorloper binnen Europa - Emerce," n.d.). It, also described what Dutch physicians believe to be the greatest challenges for the digitalisation of health care. These challenges include bureaucracy, the costs of the system, finding the right technology and the harm of sharing patients records. As BAIT does not require patients records, it already solves the challenge of the possible liability of sharing patients records. Besides, the study explains that Dutch institutions and Dutch physicians are more willing to experiment with new technologies compared to other countries. These findings illustrate a valid potential of BAIT to at least be tested as a novel CDSS in the Dutch medical sector.

The deliberation on the trustworthiness of BAIT and the positive feedback received by the UMCG physicians in this case study indicate that BAIT does have a legitimate potential to serve as a novel CDSS in the medical sector.

Nonetheless, before a new type of CDSS is implemented in an institutional environment, such as a hospital, it must also comply with many regulations and be approved by an ethical committee. These strict regulations help to prevent harm from arising to the patients impacted by a new CDSS as well as the physicians utilising the system and, hence, ensures that the principles for trustworthy and ethical AI discussed in Chapter 4 are protected. It is also, as stated in the previous paragraph, one of the most significant challenges of new digital technologies in the medical sector as a novel technology must conform with the rules of the bureaucracy. Accordingly, for the successful implementation of BAIT, further research must be conducted on the legal requirements of CDSSs in hospitals.

Next chapter describes the limitations of this research and provides recommendation for future research.

10 Reflection & further research

In this section, the limitations of BAIT, and this research, are identified and discussed. Furthermore, this section also provides recommendations for further research.

Firstly, the most prominent feedback provided by the UMCG physicians in the plenary meeting as well as written in the feedback section of the survey was that some experts believed that a few attribute levels should have been defined in more detail. The specific attributes mentioned were congenital comorbidity, the ultrasound of the brain, and growth since birth. The levels of these attributes are defined as “bad”, “intermediate”, and “good”. A few physicians believed that this left room for too much own interpretation and made the choice tasks more difficult. Instead, these physicians suggested defining, for example, a bad or good brain prognosis in terms of medical conditions.

During the design of the choice experiment, the levels for these variables were, however, intentionally drafted as bad, intermediate, and good to establish a generalised interpretation of the levels among all physicians. The levels were drafted this way as the physicians involved in the design of the choice experiment stated that there are no joint agreements on, for example, a bad, intermediate or good brain prognosis. One physician might interpret an ultrasound of the brain as bad while another physician might interpret it as intermediate. Therefore, to establish shared opinions on the attribute levels, the physicians were free to generate their own interpretation of bad or good, but at least all physicians would interpret the levels the same way.

Although the above paragraph explains the intention of drafting the attribute levels the way they were incorporated in the experiment, it still impacted how the physicians experienced conducting the stated adaptation experiment as a few physicians believed it made the already difficult recommendations even more challenging. Therefore, for future research, it is insightful to redefine the attribute levels based on defined medical conditions and determine whether it impacts the results; hence, the importance of variables on the recommendation for surgery.

Furthermore, the definition of the attribute levels also impacts the decision support of BAIT for future recommendations of the UMCG physicians. As mentioned in the previous chapter, the generate choice probabilities of BAIT indicate the percentage of UMCG physicians that would recommend in favour of surgery for a neonate with specific variable values entered into the model. Therefore, the generated likelihood of BAIT is dependent on the variable values entered into the model. Currently, as explained earlier in this chapter and indicated by the UMCG physicians, the levels are drafted such that it provides room for own interpretation. For example, the physicians differ in opinion on the definition of a good or bad brain ultrasound. Therefore, if one physician believes the ultrasound of the brain is bad. In contrast, another physician might define it as an intermediate ultrasound. Also, as presented in the results, the ultrasound of the brain has a large impact on the medical recommendations of UMCG physicians. Therefore, these different perceptions of the variable values significantly impact the calculated probability of BAIT for advice in favour of surgery. Hence, this entails that BAIT will generate different choice probabilities when distinct physicians utilise BAIT for the same new-born.

The fact that BAIT is subject to different opinions of the UMCG physicians on the appropriate attribute value and, hence, generates diverse choice probabilities for the same case when utilised by different UMCG physicians can reduce the mitigating effect on overconfidence and even enhance this bias. As the currently drafted levels enhances confirmation bias that may boost overconfidence. Confirmation bias reflects the inclination of seeking information that confirms one’s own opinion (Uy et al., 2014). Thus, in this case study, seeking a choice probability that confirms the initial preferred treatment of a UMCG physician. As the currently drafted levels support confirmation bias, it likely has an optimistic implication for the use of BAIT as the UMCG physicians can use the system as confirmation for their own opinion.

There are different possibilities to reduce this confirmation bias. The first possibility is to redraft the attribute levels and redo the experiment, hence, following the feedback provided by some of the UMCG physicians. However, as explained in Chapter 3, BAIT is limited to the number of attributes for which decision-makers can successfully conduct trade-off valuations. Therefore, including all possible medical conditions is unattainable. An option is to include bridging experiments by applying hierarchical information integration (HII). HII is based on the idea that individuals group similar attributes of choice alternatives into higher-order decision construct when dealing with complex problems (Bos, Van der Heijden, Molin, & Timmermans, 2004). This approach can be applied to further define, for example, the ultrasound of the brain. Moreover, instead of modifying the choice experiment or applying HII, it is also possible to obligate the UMCG physicians to collaboratively utilise BAIT to prevent the ability to use the system optimistically.

In conclusion, it is questionable whether the UMCG physicians themselves mind whether BAIT is prone to optimistic use as it provides them with an increased feeling of autonomy. It does, however, reduce the mitigating effect on overconfidence and might therefore result in less effective decision support that helps to generate more considered recommendations.

Furthermore, another possible limitation that may have impacted the reliability of the results is the large number of choice scenarios incorporated in the choice experiment. Chapter 5 explained that a choice experiment should not include too many choice tasks as it can exhaust the respondents. This study assumed that the UMCG physicians could handle a considerable large amount of choice tasks as the choice scenarios represent dilemma’s on which the UMCG physicians are experts. This assumption might have been incorrect as during the plenary meeting, one of the physicians, mentioned that the 35 scenarios, were indeed exhausting. And, stated that his recommendations at the last choice scenarios of the experiment might be provided with less consideration compared to the medical advice on the first choice scenarios. Other physicians voiced that they agreed with this comment. There is a possibility that this might have impacted the reliability of the results as an increase of choice scenarios tend to increase the error term variance. Accordingly, for future application of BAIT, this study would recommend, firstly, to randomize the order of the choice scenarios for the different experts conducting the choice experiment. And, secondly, stress that the experts should not conduct the choice experiment at once but in parts. Another possibility is to reduce the number of choice scenarios, but, if only a small amount of experts are available to conduct the choice experiment, this might be a less favourable option.

Moreover, hypothetical bias might also have played a role. As explained in Chapter 2, hypothetical bias concerns the question of whether the physicians would provide the same recommendations in reality as they did on the choice scenarios. The choice scenarios served as a hypothetical new-born with specific personal and medical characteristics. The choice experiment, of course, does not provide the opportunity to build up an emotional bond with the new-born or with the parents of the new-born. This bond might considerably impact the final recommendation of a physician. It is, however, impossible to determine whether the UMCG physicians would give other medical advice in real-life compared to their recommendations the choice scenarios. Therefore, it is difficult to estimate the impact of hypothetical bias on the reliability of the results.

Additionally, as mentioned at the end of Chapter 8, the results must be interpreted with care because of the limited number of UMCG physicians that executed the choice experiment. Since this research does not desire to generalise the results and just aspires to provide the UMCG physicians with introspection on their choice behaviour, the estimates do not have to be statistically significant. The limited number of respondents still, however, impacts the reliability of the estimated parameters because the choice model is required to estimate a vast number of parameters with limited information. Moreover, Chapter 5 explained that the extensive number of attributes included in the experiment may have also impacted the consistency of the decision making of the UMCG physicians as an increased amount of information, leads to an attribute processing strategy that portrays elements of relevancy and coping by neglecting attributes. Hence, it is reasonable to assume that in some cases attribute non-attendance might have played a role and, hence, might have impacted the results.

Moreover, this study would also advise Councyl to apply K-fold validation for future projects. When non-knowledge based models are trained, the training data is usually split into training and validation sets. The training data is applied to train the data, while the validation set is used to validate the prediction accuracy of the trained model. This is, however, very inconvenient or impossible when dealing with limited data. This inconvenience can be solved with the K-fold validation technique. For K-folds validation, the data is split into K parts. After that, K different models are built, and each model is trained on K-1 of the data parts and tested on one part. Hence, we can make predictions on all our data results ("Why and how to Cross Validate a Model? | by Sanjay.M | Towards Data Science," n.d.). Since BAIT works with a limited amount of data, K-fold validation is an appropriate technique to enhance the initial trust level of future users for BAIT as it can provide prove of adequate performance.

Additionally, this research solely investigated the choice behaviour of the UMCG physicians, but it is insightful to study the choice behaviour of physicians in other hospitals on the same choice task to explore the differences and similarities. Therefore, this study recommends executing the same research in different hospitals in the Netherland or outside the Netherland.

Finally, this research solely included the wish of the parents as a factor influencing the UMCG physicians recommendations. Therefore, it did not include an analysis of what parents find important when voicing their preferred treatment. For future research applying BAIT to investigate the importance of factors that determine whether parents favour surgery or comfort care might be insightful. Primarily, because research shows that to improve the decision-making process of such ethical and difficult decisions, shared decision making between physicians and parents on the appropriate treatment procedure gained a lot of interest and popularity. Research shows that approximately 80% of the parents highly value shared or active decision-making and experience less regret with the enforced treatment when shared decision making is applied (Soltys, Philpott-Streiff, Fuzzell, & Politi, 2020). An improved understanding of which factors parents find most important while deliberating their wish on the preferred treatment for their child may support shared decision making and is, thus, interesting to investigate.

In conclusion, as BAIT is a new IDDS approach, it requires testing in different settings to gain insight into the usefulness and effectiveness of this method. To further investigate the potential of BAIT in the medical sector, this study advises conducting more case studies to further investigate the potential and effectiveness of BAIT in the medical sector. And, ultimately, also in other sectors.

11 References

Aamodt, A. (1993). A Case-Based Answer to Some Problems of Knowledge-Based Systems. *Scandinavian Conference on Artificial Intelligence*, (February 1970), 168–182.

Abbasi, M. M., & Kashiyarndi, S. (2010). Clinical Decision Support Systems: A discussion on different methodologies used in Health Care. *Report*, 1–15. Retrieved from http://www.idt.mdh.se/kurser/ct3340/ht10/FinalPapers/15-Abbasi_Kashiyarndi.pdf

Abdullah, S., Markandya, A., & Nunes, P. A. L. D. (2011). Introduction to economic valuation methods. In *Research Tools in Natural Resource and Environmental Economics*. https://doi.org/10.1142/9789814289238_0005

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *The Economics of Artifi cial Intelligence*.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2018). *Exploring the Impact of Artificial Intelligence: Prediction versus Judgment*.

Armstrong, D. (2002). Clinical autonomy, individual and collective: The problem of changing doctors' behaviour. *Social Science and Medicine*, *55*(10), 1771–1777. https://doi.org/10.1016/S0277-9536(01)00309-4

Beguería, S. (2006). Validation and evaluation of predictive models in hazard assessment and risk management. *Natural Hazards*, *37*(3), 315–329. https://doi.org/10.1007/s11069-005-5182-6

Berner, E. S., & La Lande, T. J. (2007). *Overview of Clinical Decision Support Systems*. https://doi.org/10.1007/978-0-387-38319-4_1

Blank, R. H. (2011). End-of-Life decision making across cultures. *Journal of Law, Medicine and Ethics*, *39*(2), 201–214. https://doi.org/10.1111/j.1748-720X.2011.00589.x

blijft de mens? (n.d.). 13.

Boland, L., Graham, I. D., Légaré, F., Lewis, K., Jull, J., Shephard, A., … Stacey, D. (2019, January 18). Barriers and facilitators of pediatric shared decision-making: A systematic review. *Implementation Science*, Vol. 14, pp. 1–25. https://doi.org/10.1186/s13012-018-0851-5

Bos, I. D. M., Van der Heijden, R. E. C. M., Molin, E. J. E., & Timmermans, H. J. P. (2004). The choice of park and ride facilities: An analysis using a context-dependent hierarchical choice experiment. *Environment and Planning A*, *36*(9), 1673–1686. https://doi.org/10.1068/a36138

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Burstein, F., W. Holsapple, C., & Power, D. J. (2008). Decision Support Systems: A Historical Overview. In *Handbook on Decision Support Systems 1* (pp. 121–140). https://doi.org/10.1007/978-3-540-48713-5_7

Carr, B. D., & Gadepalli, S. K. (2019a). Does Surgical Management Alter Outcome in Necrotizing Enterocolitis? *Clinics in Perinatology*, *46*, 89–100. https://doi.org/10.1016/j.clp.2018.09.008

Carr, B. D., & Gadepalli, S. K. (2019b, March 1). Does Surgical Management Alter Outcome in Necrotizing Enterocolitis? *Clinics in Perinatology*, Vol. 46, pp. 89–100. https://doi.org/10.1016/j.clp.2018.09.008

Caussade, S., Ortúzar, J. de D., Rizzi, L. I., & Hensher, D. A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B: Methodological*, *39*(7), 621–640. https://doi.org/10.1016/j.trb.2004.07.006

Chakraborty, R., El-Jawahri, A. R., Litzow, M. R., Syrjala, K. L., Parnes, A. D., & Hashmi, S. K. (2017, October 1). A systematic review of religious beliefs about major end-of-life issues in the five major world religions. *Palliative and Supportive Care*, Vol. 15, pp.

609–622. https://doi.org/10.1017/S1478951516001061

Chikwe, J. E. (2018). *DECISION-MAKING FEASIBILITY AND TECHNIQUES: A PSYCHOLOGICAL AND STRATEGIC EVALUATION IMPERATIVES*.

ChoiceMetrics. (2018). *Ngene 1.2 USER MANUAL & REFERENCE GUIDE The Cutting Edge in Experimental Design*. Retrieved from www.choice-metrics.com

Chorus, C. (2018). *Statistical choice behaviour*.

Context - Responsible AI Declaration. (n.d.). Retrieved June 22, 2020, from https://www.montrealdeclaration-responsibleai.com/context

Croskerry, P. (2000). The feedback sanction. *Academic Emergency Medicine*, Vol. 7, pp. 1232–1238. https://doi.org/10.1111/j.1553-2712.2000.tb00468.x

Croskerry, Pat, & Norman, G. (2008). *Overconfidence in Clinical Decision Making*. https://doi.org/10.1016/j.amjmed.2008.02.001

Cuttini, M., Nadai, M., Kaminski, M., Hansen, G., De Leeuw, R., Lenoir, S., … Saracci, R. (2000). End-of-life decisions in neonatal intensive care: Physicians' self-reported practices in seven European countries. *Lancet*, *355*(9221), 2112–2118. https://doi.org/10.1016/S0140-6736(00)02378-3

Danks, D. (2019). *The Value of Trustworthy AI*. https://doi.org/10.1145/3306618.3314228

Dombrecht, L., Deliens, L., Chambaere, K., Baes, S., Cools, F., Goossens, L., … Beernaert, K. (2020). Neonatologists and neonatal nurses have positive attitudes towards perinatal end-of-life decisions, a nationwide survey. *Acta Paediatrica*, *109*(3), 494–504. https://doi.org/10.1111/apa.14797

Donohue, P. K., Boss, R. D., Aucott, S. W., Keene, E. A., & Teague, P. (2010). The Impact of Neonatologists' Religiosity and Spirituality on Health Care Delivery for High-Risk Neonates. *Journal of Palliative Medicine*, *13*(10), 1219–1224. https://doi.org/10.1089/jpm.2010.0049

Ethics guidelines for trustworthy AI | Shaping Europe's digital future. (n.d.). Retrieved June 22, 2020, from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Ezhilarasu, C. M., Skaf, Z., & Jennions, I. K. (2019). The application of reasoning to aerospace Integrated Vehicle Health Management (IVHM): Challenges and opportunities. *Progress in Aerospace Sciences*, *105*(February), 60–73. https://doi.org/10.1016/j.paerosci.2019.01.001

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and tam in online shopping: AN integrated model. *MIS Quarterly: Management Information Systems*, *27*(1), 51–90. https://doi.org/10.2307/30036519

Heida, F. H., Stolwijk, L., Loos, M. L. H. J., van den Ende, S. J., Onland, W., van den Dungen, F. A. M., … Bakx, R. (2017). Increased incidence of necrotizing enterocolitis in the Netherlands after implementation of the new Dutch guideline for active treatment in extremely preterm infants: Results from three academic referral centers. *Journal of Pediatric Surgery*, *52*(2), 273–276. https://doi.org/10.1016/j.jpedsurg.2016.11.024

Henry, M. C. W., & Moss, R. L. (2005). Surgical therapy for necrotizing enterocolitis: Bringing evidence to the bedside. *Seminars in Pediatric Surgery*, *14*(3), 181–190.

https://doi.org/10.1053/j.sempedsurg.2005.05.007

Hensher, D. A., Rose, J., & Greene, W. H. (2005). The implications on willingness to pay of respondents ignoring specific attributes. *Transportation*, *32*(3), 203–222. https://doi.org/10.1007/s11116-004-7613-8

Heselmans, A., Aertgeerts, B., Donceel, P., Geens, S., Van De Velde, S., & Ramaekers, D. (2012). Family physicians' perceptions and use of electronic clinical decision support during the first year of implementation. *Journal of Medical Systems*, *36*(6), 3677–3684. https://doi.org/10.1007/s10916-012-9841-3

Hisschemöller, M., & Hoppe, R. (2018). Coping with intractable controversies: The case for problem structuring in policy design and analysis. In *Knowledge, Power, and Participation in Environmental Policy Analysis* (pp. 47–72). https://doi.org/10.4324/9781351325721-4

Hopgood, A. A. (2005). The State of Artificial Intelligence. *Advances in Computers*, *65*(December 2005), 1–75. https://doi.org/10.1016/S0065-2458(05)65001-2

Jaspers, M. W. M., Smeulers, M., Vermeulen, H., & Peute, L. W. (2011). Effects of clinical decision-support systems on practitioner performance and patient outcomes: A synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association*, *18*(3), 327–334. https://doi.org/10.1136/amiajnl-2011-000094

Jordan, M. I., & Mitchell, T. M. (2015, July 17). Machine learning: Trends, perspectives, and prospects. *Science*, Vol. 349, pp. 255–260. https://doi.org/10.1126/science.aaa8415

Khairat, S., Marc, D., Crosby, W., & Al Sanousi, A. (2018, April 1). Reasons for physicians not adopting clinical decision support systems: Critical analysis. *Journal of Medical Internet Research*, Vol. 20. https://doi.org/10.2196/medinform.8912

Kinderchirurgie. (n.d.). Retrieved March 12, 2020, from https://www.umcg.nl/NL/UMCG/Afdelingen/KC-UMCG/professionals/kinderheelkunde/Paginas/default.aspx

Kunstmatig. (n.d.). Retrieved June 23, 2020, from https://www.nwo.nl/algemeen/actueel/social-media/onderzoek-online/2020-1-kunstmatig

Kusumasondjaja, S., Shanka, T., & Marchegiani, C. (2012). Credibility of online reviews and initial trust. *Journal of Vacation Marketing*, *18*(3), 185–195. https://doi.org/10.1177/1356766712449365

Kwiatkowska, M., Atkins, M. S., Ayas, N. T., & Ryan, C. F. (2007). Knowledge-based data analysis: First step toward the creation of clinical prediction rules using a new typicality measure. *IEEE Transactions on Information Technology in Biomedicine*, *11*(6), 651–660. https://doi.org/10.1109/TITB.2006.889693

Lawrence, R. E., & Curlin, F. A. (2009). Autonomy, religion and clinical decisions: Findings from a national physician survey. *Journal of Medical Ethics*, *35*(4), 214–218. https://doi.org/10.1136/jme.2008.027565

Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *Journal of Strategic Information Systems*, *17*(1), 39–71. https://doi.org/10.1016/j.jsis.2008.01.001

Liberati, E. G., Ruggiero, F., Galuppo, L., Gorli, M., González-Lorenzo, M., Maraldi, M., … Moja, L. (2017). What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. *Implementation Science*, *12*(1), 113. https://doi.org/10.1186/s13012-017-0644-2

Louviere, J. J., Flynn, T. N., & Carson, R. T. (2010). Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling*, *3*(3), 57–72. https://doi.org/10.1016/S1755-5345(13)70014-9

Lysaght, T., Lim, H. Y., Xafis, V., & Ngiam, K. Y. (2019). AI-Assisted Decision-making in Healthcare: The Application of an Ethics Framework for Big Data in Health and

Research. *Asian Bioethics Review, 11*(3), 299–314. https://doi.org/10.1007/s41649-019-00096-0

Manyika, J., Bughin, J., Chui, M., Silberg, J., & Gumbel, P. (2018). *the Promise and Challenge of the Age of Artificial Intelligence Briefing Note Prepared for the Tallinn Digital Summit October 2018*. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Featured Insights/Artificial Intelligence/The promise and challenge of the age of artificial intelligence/MGI-The-promise-and-challenge-of-the-age-of-artificial-intelligence-in-brief-Oct-2018.ashx

Miller, A. P. (2018). Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review*. Retrieved from https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

Ming, V. (2019). Human insight remains essential to beat the bias of algorithms. *Financial Times*. Retrieved from https://www.ft.com/content/59520726-d0c5-11e9-b018-ca4456540ea6

Molin, EJE. (2010). Context-dependent stated choice experiments. *Journal of Choice Modelling, 3*(3), 39–56. Retrieved from http://repository.tudelft.nl/view/ir/uuid:d4ab739c-a1f1-40bd-ab27-2d272b9e70c2/

Molin, Eric. (2018). *Statistical choice behaviour*.

Mrs. S. S. Gulavani & R. V. Kulkarni. (2014). a Review of Knowledge Based Systems in. *International Journal of Information Technology and Knowledge Management, 2*(August 2002), 271.

Nederlandse gezondheidszorg is digitale voorloper binnen Europa - Emerce. (n.d.). Retrieved September 20, 2020, from https://www.emerce.nl/achtergrond/nederlandse-gezondheidszorg-digitale-voorloper-binnen-europa

Neu, J., & Walker, W. A. (2011, January 20). Necrotizing enterocolitis. *New England Journal of Medicine*, Vol. 364, pp. 255–264. https://doi.org/10.1056/NEJMra1005408

Osheroff, J. A., Teich, J. M., Middleton, B., Steen, E. B., Wright, A., & Detmer, D. E. (2007). A Roadmap for National Action on Clinical Decision Support. *Journal of the American Medical Informatics Association*, *14*(2), 141–145. https://doi.org/10.1197/jamia.M2334

Prentzas, J., & Hatzilygeroudis, I. (2007). Categorizing approaches combining rule-based and case-based reasoning. *Expert Systems, 24*(2), 97–122. https://doi.org/10.1111/j.1468-0394.2007.00423.x

Professional Dominance: The Social Structure of Medical Care - Eliot Freidson - Google Boeken. (n.d.). Retrieved September 20, 2020, from https://books.google.nl/books?hl=nl&lr=&id=yl6DMv0lTNwC&oi=fnd&pg=PR10&ots=Wa8VN628Rk&sig=uRBbXOiQKOfiuBcjENVpmtvpf-c&redir_esc=y#v=onepage&q&f=false

Ravi, P. (2020). Overview of causal inference in machine learning - Ericsson. Retrieved April 23, 2020, from https://www.ericsson.com/en/blog/2020/2/causal-inference-machine-learning

Rebagliato, M., Cuttini, M., Kaminski, M., Persson, J., Reid, M., & Saracci, R. (2000). *Neonatal End-of-Life Decision Making*. *284*(19), 2451–2459.

Rees, C. M., Pierro, A., & Eaton, S. (2007). Neurodevelopmental outcomes of neonates with medically and surgically treated necrotizing enterocolitis. *Archives of Disease in Childhood: Fetal and Neonatal Edition, 92*(3), F193–F198. https://doi.org/10.1136/adc.2006.099929

Robinson, J. R., Rellinger, E. J., Hatch, L. D., Weitkamp, J. H., Speck, K. E., Danko, M., & Blakely, M. L. (2017, February 1). Surgical necrotizing enterocolitis. *Seminars in Perinatology*, Vol. 41, pp. 70–79. https://doi.org/10.1053/j.semperi.2016.09.020

Sargent, D. J. (2001). Comparison of artificial neural networks with other statistical approaches. *Cancer*, *91*(S8), 1636–1642. https://doi.org/10.1002/1097-0142(20010415)91:8+<1636::AID-CNCR1176>3.0.CO;2-D

Siau, K. (2018). *Building Trust in Artificial Intelligence, Machine Learning, and Robotics Supply Chain Management View project*. Retrieved from www.cutter.com

Siau, K., & Shen, Z. (2003, April 1). Building customer trust in mobile commerce. *Communications of the ACM*, Vol. 46, pp. 91–94. https://doi.org/10.1145/641205.641211

Silberg, J., & Manyika, J. (2019). Tackling bias in artificial intelligence (and in humans) McKinsey. *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*, 1–8. Retrieved from https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

Sprung, C. L., Maia, P., Bulow, H. H., Ricou, B., Armaganidis, A., Baras, M., … Thijs, L. G. (2007). The importance of religious affiliation and culture on end-of-life decisions in European intensive care units. *Intensive Care Medicine*, *33*(10), 1732–1739. https://doi.org/10.1007/s00134-007-0693-0

Stake, R. E. (2005). Qualitative Case Studies. In *The Sage handbook of qualitative research, 3rd ed.* (pp. 443–466). Thousand Oaks, CA: Sage Publications Ltd.

Uy, R. C. harle., Sarmiento, R. F. ranci., Gavino, A., & Fontelo, P. (2014). Confidence and Information Access in Clinical Decision-Making: An Examination of the Cognitive Processes that affect the Information-seeking Behavior of Physicians. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2014*, 1134–1140. Retrieved from /pmc/articles/PMC4419936/?report=abstract

Van Wijnen, J. . (2019). De computer die braver is dan de mens. *Financieel Dagblad*, pp. 5–10.

Verhagen, A. A. E., Van Der Hoeven, M. A. H., Van Meerveld, R. C., & Sauer, P. J. J. (2007, July). Physician medical decision-making at the end of life in newborns: Insight into implementation at 2 Dutch centers. *Pediatrics*, Vol. 120. https://doi.org/10.1542/peds.2006-2555

Wagholikar, K. B., Sundararajan, V., & Deshpande, A. W. (2012). Modeling paradigms for medical diagnostic decision support: A survey and future directions. *Journal of Medical Systems*, *36*(5), 3029–3049. https://doi.org/10.1007/s10916-011-9780-4

Waltl, B., Bonczek, G., & Matthes, F. (2018). Rule-based information extraction: Advantages, limitations, and perspectives. *Jusletter IT*, (February).

Weller, P., Oehlmann, M., Mariel, P., & Meyerhoff, J. (2014). Stated and inferred attribute non-attendance in a design of designs approach. *Journal of Choice Modelling*, *11*(1), 43–56. https://doi.org/10.1016/j.jocm.2014.04.002

Why and how to Cross Validate a Model? | by Sanjay.M | Towards Data Science. (n.d.). Retrieved September 20, 2020, from https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

Yang, H., Thompson, C., & Bland, M. (2012). The effect of clinical experience, judgment task difficulty and time pressure on nurses' confidence calibration in a high fidelity clinical simulation. *BMC Medical Informatics and Decision Making*, *12*(1), 113. https://doi.org/10.1186/1472-6947-12-113

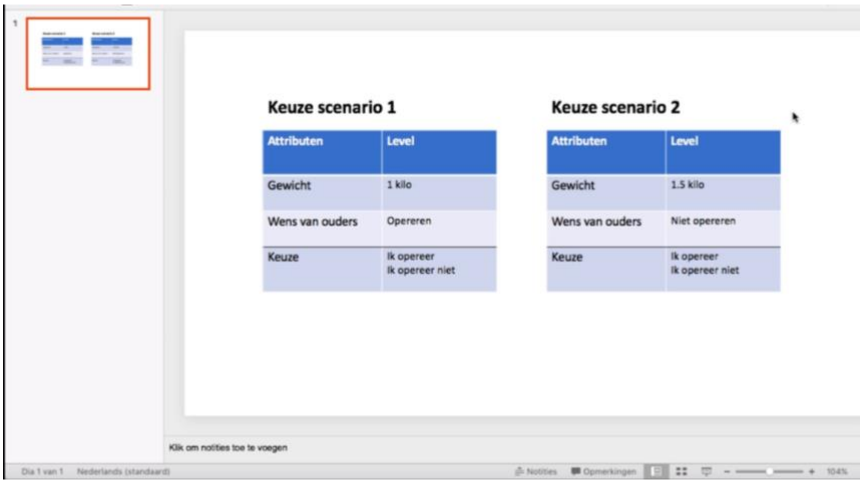
Yilmaz, L., & Tolk, A. (2008). Intelligent Decision Making: An AI-Based Approach. *Intelligent Decision Making: An AI-Based Approach*, *97*(January 2008), 193–226. https://doi.org/10.1007/978-3-540-76829-6

Appendix A Semi-Structured interviews

This Appendix summarizes the semi-structured interviews conducted with the four UMCG physicians to discover the decision-variables (attributes) for their choice task. It firstly illustrates the structure of the meeting. Thenceforth, it discusses the takeaways of this meeting.

Structure

1. Introduction about myself and this study.
2. Explanation on BAIT and how it will provide an introspection on their choice task.
3. An outline of the design process for the stated adaptation experiment, and an indication of the amount of effort and time each design step will consume.
4. An illustration of a simple stated adaptation experiment to provide a practical example of the theory that was just explained. The example shown to the UMCG physicians' is depicted below:



Keuze scenario 1

Attributen	Level
Gewicht	1 kilo
Wens van ouders	Opereren
Keuze	Ik opereer Ik opereer niet

Keuze scenario 2

Attributen	Level
Gewicht	1.5 kilo
Wens van ouders	Niet opereren
Keuze	Ik opereer Ik opereer niet

Klik om notities toe te voegen

5. A clarification that it is not feasible to include a limited number of attributes in the choice experiment due to :
   a. Attribute non attendance
   b. The number of choice tasks included in the experiment must not exhaust the experts and be practically feasible
6. An explanation about the requirements for the individual lists of attributes, ranges and levels. The conditions described are explained in Section 5.3.2.
7. A kind request to draft a maximum number of twenty attributes due to the reasons elucidated earlier in the interview.
8. A conclusion for the interview explaining that after the meeting, an Excel sheet will be sent in which the experts' can fill out the attributes, ranges and levels. The email includes a recap of the requirements for the list of attributes, ranges, and levels.

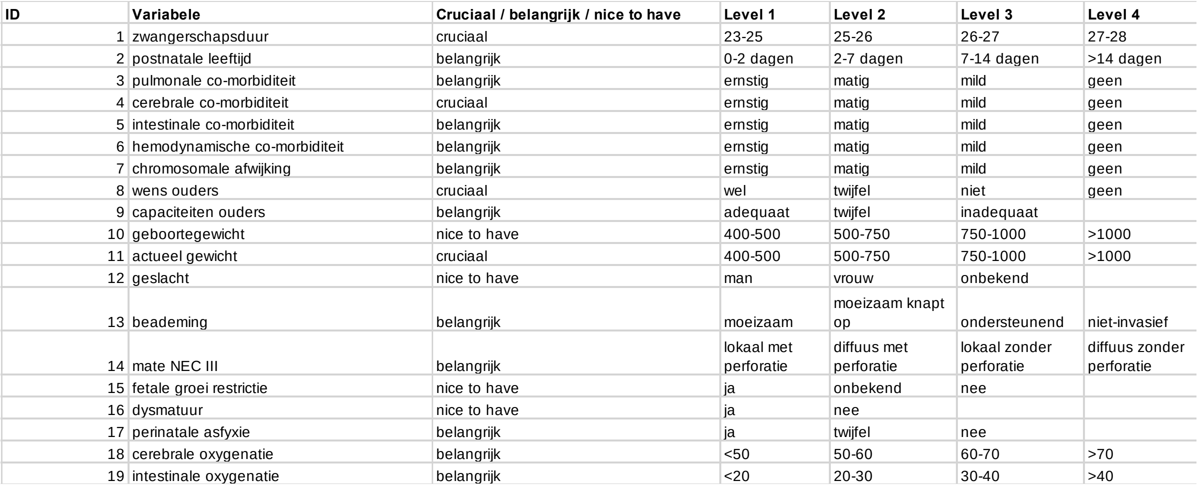
Takeaways

- In the first interview, the requirements for the attributes, ranges, and levels were thoroughly explained based on the theory of choice modelling. For the other interviews, the conditions were stated rather than telling why the requirements were drafted. This resulted in shorter discussions and less confusion. Accordingly, this research would advise, future employees of Councyl, to limit themselves to stating the requirements rather than providing a description of the requirements based on theory exempt when experts ask for an explanation.

- Make sure that the terms applied for the choice experiment such as ‘attributes’ for decision variables are described in terms of the experts’ vocabulary to prevent confusion.
- Some confusion existed among the UMCG physicians on whether this study researched a diagnosis for an operation indication or an end-of-life decision after an operation indication. Therefore, it is essential to clarify if there is a clear understanding of the choice task at the start of the interview since it is important that all the involved experts’ draw up a list of attributes for the same choice task.

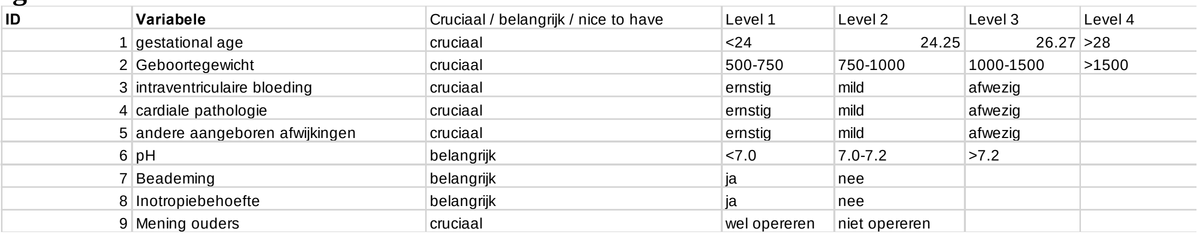
Appendix B Initial lists of attributes and levels

Figure B.1 till B.4 illustrate the lists of attributes that the four UMCG physicians, involved in the design of the stated adaptation experiment, individually drafted.



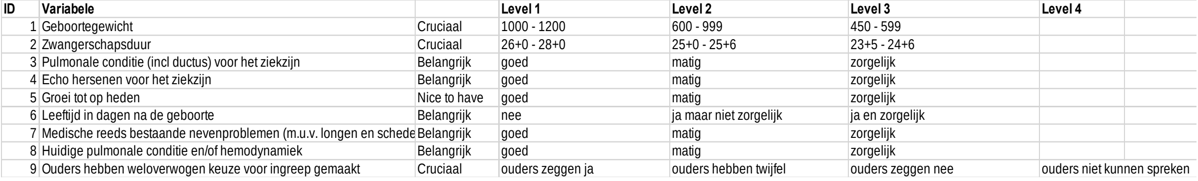
ID	Variabele	Cruciaal / belangrijk / nice to have	Level 1	Level 2	Level 3	Level 4
1	zwangerschapsduur	cruciaal	23-25	25-26	26-27	27-28
2	postnatale leeftijd	belangrijk	0-2 dagen	2-7 dagen	7-14 dagen	>14 dagen
3	pulmonale co-morbiditeit	belangrijk	ernstig	matig	mild	geen
4	cerebrale co-morbiditeit	cruciaal	ernstig	matig	mild	geen
5	intestinale co-morbiditeit	belangrijk	ernstig	matig	mild	geen
6	hemodynamische co-morbiditeit	belangrijk	ernstig	matig	mild	geen
7	chromosomale afwijking	belangrijk	ernstig	matig	mild	geen
8	wens ouders	cruciaal	wel	twijfel	niet	geen
9	capaciteiten ouders	belangrijk	adequaat	twijfel	inadequaat	
10	geboortegewicht	nice to have	400-500	500-750	750-1000	>1000
11	actueel gewicht	cruciaal	400-500	500-750	750-1000	>1000
12	geslacht	nice to have	man	vrouw	onbekend	
13	beademing	belangrijk	moeizaam	moeizaam knapt op	ondersteunend	niet-invasief
14	mate NEC III	belangrijk	lokaal met perforatie	diffuus met perforatie	lokaal zonder perforatie	diffuus zonder perforatie
15	fetale groei restrictie	nice to have	ja	onbekend	nee	
16	dysmatuur	nice to have	ja	nee		
17	perinatale asfyxie	belangrijk	ja	twijfel	nee	
18	cerebrale oxygenatie	belangrijk	<50	50-60	60-70	>70
19	intestinale oxygenatie	belangrijk	<20	20-30	30-40	>40

Figure B.1: First list of initial attributes



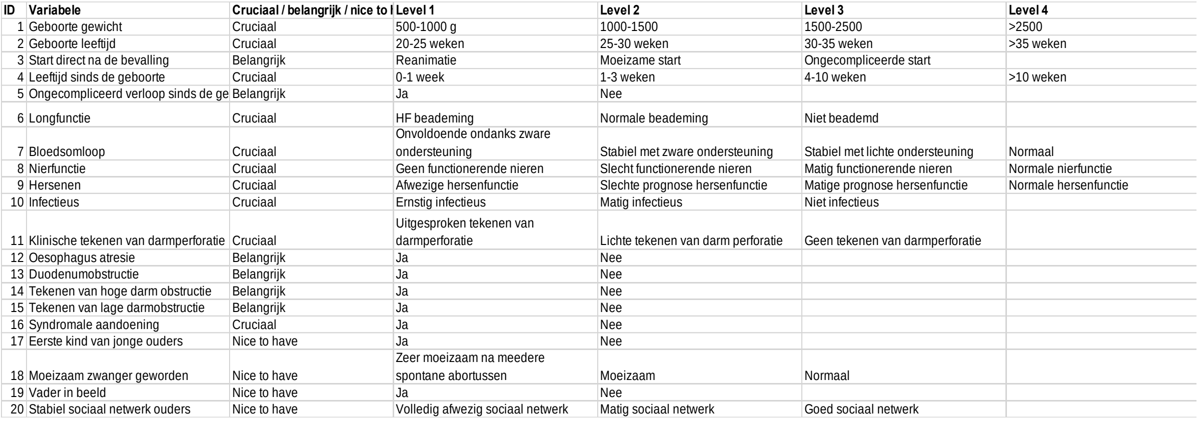
ID	Variabele	Cruciaal / belangrijk / nice to have	Level 1	Level 2	Level 3	Level 4
1	gestational age	cruciaal	<24	24.25	26.27	>28
2	Geboortegewicht	cruciaal	500-750	750-1000	1000-1500	>1500
3	intraventriculaire bloeding	cruciaal	ernstig	mild	afwezig	
4	cardiale pathologie	cruciaal	ernstig	mild	afwezig	
5	andere aangeboren afwijkingen	cruciaal	ernstig	mild	afwezig	
6	pH	belangrijk	<7.0	7.0-7.2	>7.2	
7	Beademing	belangrijk	ja	nee		
8	Inotropiebehoefte	belangrijk	ja	nee		
9	Mening ouders	cruciaal	wel opereren	niet opereren		

Figure B.2: Second list of initial attributes



ID	Variabele		Level 1	Level 2	Level 3	Level 4
1	Geboortegewicht	Cruciaal	1000 - 1200	600 - 999	450 - 599	
2	Zwangerschapsduur	Cruciaal	26+0 - 28+0	25+0 - 25+6	23+5 - 24+6	
3	Pulmonale conditie (incl ductus) voor het ziekzijn	Belangrijk	goed	matig	zorgelijk	
4	Echo hersenen voor het ziekzijn	Belangrijk	goed	matig	zorgelijk	
5	Groei tot op heden	Nice to have	goed	matig	zorgelijk	
6	Leeftijd in dagen na de geboorte	Belangrijk	nee	ja maar niet zorgelijk	ja en zorgelijk	
7	Medische reeds bestaande nevenproblemen (m.u.v. longen en schede	Belangrijk	goed	matig	zorgelijk	
8	Huidige pulmonale conditie en/of hemodynamiek	Belangrijk	goed	matig	zorgelijk	
9	Ouders hebben weloverwogen keuze voor ingreep gemaakt	Cruciaal	ouders zeggen ja	ouders hebben twijfel	ouders zeggen nee	ouders niet kunnen spreken

Figure B.3: Third list of initial attributes



ID	Variabele	Cruciaal / belangrijk / nice to l	Level 1	Level 2	Level 3	Level 4
1	Geboorte gewicht	Cruciaal	500-1000 g	1000-1500	1500-2500	>2500
2	Geboorte leeftijd	Cruciaal	20-25 weken	25-30 weken	30-35 weken	>35 weken
3	Start direct na de bevalling	Belangrijk	Reanimatie	Moeizame start	Ongecompliceerde start	
4	Leeftijd sinds de geboorte	Cruciaal	0-1 week	1-3 weken	4-10 weken	>10 weken
5	Ongecompliceerd verloop sinds de ge	Belangrijk	Ja	Nee		
6	Longfunctie	Cruciaal	HF beademing	Normale beademing	Niet beademd	
7	Bloedsomloop	Cruciaal	Onvoldoende ondanks zware ondersteuning	Stabiel met zware ondersteuning	Stabiel met lichte ondersteuning	Normaal
8	Nierfunctie	Cruciaal	Geen functionerende nieren	Slecht functionerende nieren	Matig functionerende nieren	Normale nierfunctie
9	Hersenen	Cruciaal	Afwezige hersenfunctie	Slechte prognose hersenfunctie	Matige prognose hersenfunctie	Normale hersenfunctie
10	Infectieus	Cruciaal	Ernstig infectieus	Matig infectieus	Niet infectieus	
11	Klinische tekenen van darmperforatie	Cruciaal	Uitgesproken tekenen van darmperforatie	Lichte tekenen van darm perforatie	Geen tekenen van darmperforatie	
12	Oesophagus atresie	Belangrijk	Ja	Nee		
13	Duodenumobstructie	Belangrijk	Ja	Nee		
14	Tekenen van hoge darm obstructie	Belangrijk	Ja	Nee		
15	Tekenen van lage darmobstructie	Belangrijk	Ja	Nee		
16	Syndromale aandoening	Cruciaal	Ja	Nee		
17	Eerste kind van jonge ouders	Nice to have	Ja	Nee		
18	Moeizaam zwanger geworden	Nice to have	Zeer moeizaam na meedere spontane abortussen	Moeizaam	Normaal	
19	Vader in beeld	Nice to have	Ja	Nee		
20	Stabiel sociaal netwerk ouders	Nice to have	Volledig afwezig sociaal netwerk	Matig sociaal netwerk	Goed sociaal netwerk	

Figure B.4: Fourth list of initial attributes

Appendix C Discussion on the merged list of attributes

This Appendix discusses the first plenary meeting with all UMCG physicians. This meeting aimed to reduce the number of attributes to be able to design a prototype experiment. The lists depicted in Appendix B were merged into one file. The crucial attributes were placed on top and the 'nice to have' attributes at the bottom; additionally, the attributes were grouped according to medical resembling's. An Excel document with the merged list of attributes was sent to the four UMCG physicians before this meeting. This Excel file included a sheet with the combined list of attributes and a sheet called "final attributes". During this meeting, the attributes that all the UMCG physicians would agree upon were placed in the sheet called: final list of attributes. Moreover, at the beginning of this meeting, several attributes were already placed in this sheet as these attributes were comparable in all the lists of the UMCG physicians and labelled as crucial or important.

Structure

1. An introduction with an explanation of the goal and structure of the meeting.
2. A discussion on the already drafted attributes and levels in the "final attributes" sheet.
3. A discussion on the other attributes in the merged list.

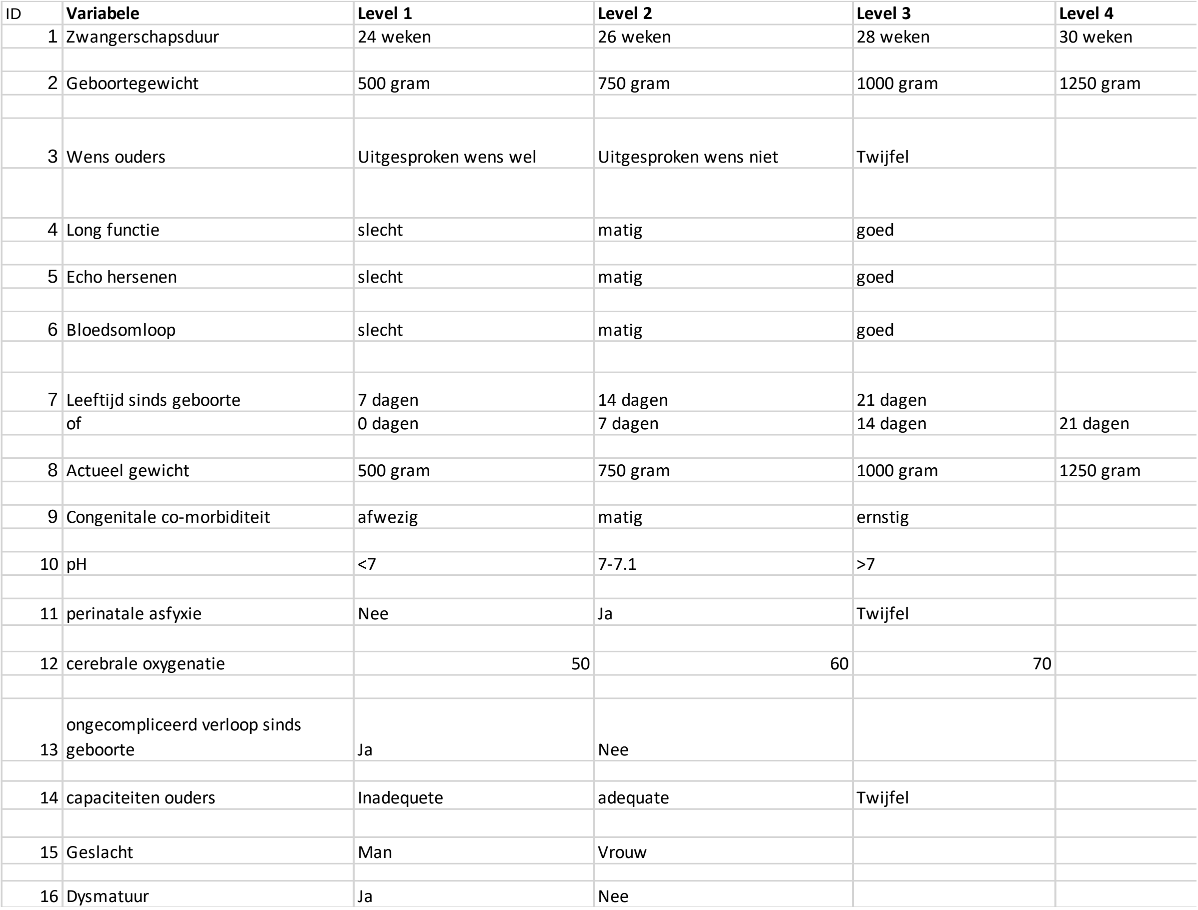
Per attribute, it was first determined whether to include the attribute in the final list. After that, the range of the attribute was defined. The attribute range should preferably capture at least 85 % of the bulk of observations in reality but also force the UMCG physicians' to make trade-offs between the attributes. Lastly, per attribute included in the final list, it was asked whether the UMCG physicians believed the attribute had a linear or non-linear effect on their decision and, accordingly, the levels of the attributes were drafted.

4. Lastly, the physicians were asked whether certain combinations of attribute levels did not occur in reality.

Figure C.1 depicts the list of attributes and levels derived after this meeting.

Takeaways:

- Although the individual lists of the UMCG physicians were distinct, during the plenary meeting, all experts relatively fast agreed upon the attributes they believed should be included in the choice experiment. Moreover, the final attributes included in the choice experiment were attributes that the group of UMCG physicians all labelled as crucial or important in their initial individual lists. Therefore, it may be more practical for both Councyl and the group of experts to, instead of individually, collaboratively draft an initial list of attributes and levels. This way, more effort can be put in improving the quality of the initial list rather than removing attributes.



ID	**Variabele**	**Level 1**	**Level 2**	**Level 3**	**Level 4**
1	Zwangerschapsduur	24 weken	26 weken	28 weken	30 weken
2	Geboortegewicht	500 gram	750 gram	1000 gram	1250 gram
3	Wens ouders	Uitgesproken wens wel	Uitgesproken wens niet	Twijfel	
4	Long functie	slecht	matig	goed	
5	Echo hersenen	slecht	matig	goed	
6	Bloedsomloop	slecht	matig	goed	
7	Leeftijd sinds geboorte	7 dagen	14 dagen	21 dagen	
	of	0 dagen	7 dagen	14 dagen	21 dagen
8	Actueel gewicht	500 gram	750 gram	1000 gram	1250 gram
9	Congenitale co-morbiditeit	afwezig	matig	ernstig	
10	pH	<7	7-7.1	>7	
11	perinatale asfyxie	Nee	Ja	Twijfel	
12	cerebrale oxygenatie	50	60	70	
13	ongecompliceerd verloop sinds geboorte	Ja	Nee		
14	capaciteiten ouders	Inadequete	adequate	Twijfel	
15	Geslacht	Man	Vrouw		
16	Dysmatuur	Ja	Nee		

Figure C.1: List of attributes after first plenary meeting

Appendix D Discussion prototype experiment

This Appendix elaborates on the meeting that discussed the prototype experiment which was sent to the UMCG physicians before the meeting.

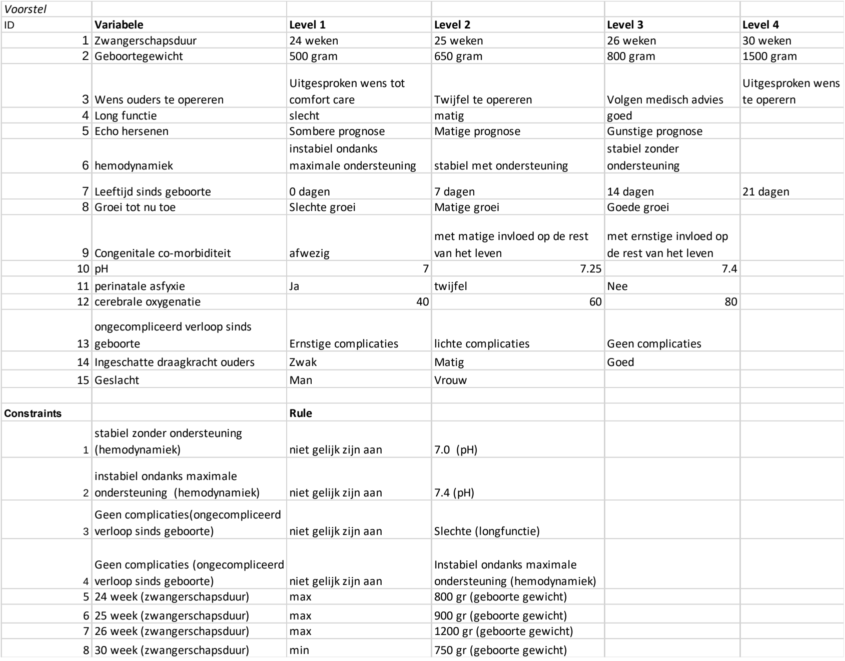
Structure

1. Introduction on the goal and structure of the meeting. The list of attributes, levels, and constraints generated at the end of this meeting is used to construct the pilot study experimental design.
2. In this meeting, we checked every attribute, and it's corresponding levels of the list depicted in Appendix C based on the guidelines for selecting attributes described in Chapter 5. Simultaneously, the proper vocabulary for the attributes and levels was checked.
3. Thenceforth, the constraints for the design were inspected.
4. After that, an explanation on interaction effects was provided, and the group of UMCG physicians discussed whether interaction effects should be included in the choice experiment. No interaction effects were included in the experiment.

After the meeting, a new prototype experiment was sent to the UMCG physicians. Based on the feedback of this meeting and some final comments provided per email, the list of attributes, levels, and constrains was established shown in Figure D.1.

Takeaways

- Based on the prototype experiment, the group of UMCG physicians provided additional feedback because they were forced to execute some choice tasks rather than discuss the attributes and levels with an Excel document only. For future projects, it therefore, might be of added value to already send a prototype experiment based on the initial list of attributes such that the group of experts understand the practical implementation of the list they are constructing.
- In future projects, it is of added value to provide an adequate example of an interaction effect based on the use-case since in the meeting some confusion existed about the interpretation of a interaction effect.



Voorstel					
ID	**Variabele**	**Level 1**	**Level 2**	**Level 3**	**Level 4**
1	Zwangerschapsduur	24 weken	25 weken	26 weken	30 weken
2	Geboortegewicht	500 gram	650 gram	800 gram	1500 gram
3	Wens ouders te opereren	Uitgesproken wens tot comfort care	Twijfel te opereren	Volgen medisch advies	Uitgesproken wens te operern
4	Long functie	slecht	matig	goed	
5	Echo hersenen	Sombere prognose	Matige prognose	Gunstige prognose	
6	hemodynamiek	instabiel ondanks maximale ondersteuning	stabiel met ondersteuning	stabiel zonder ondersteuning	
7	Leeftijd sinds geboorte	0 dagen	7 dagen	14 dagen	21 dagen
8	Groei tot nu toe	Slechte groei	Matige groei	Goede groei	
9	Congenitale co-morbiditeit	afwezig	met matige invloed op de rest van het leven	met ernstige invloed op de rest van het leven	
10	pH	7	7.25	7.4	
11	perinatale asfyxie	Ja	twijfel	Nee	
12	cerebrale oxygenatie	40	60	80	
13	ongecompliceerd verloop sinds geboorte	Ernstige complicaties	lichte complicaties	Geen complicaties	
14	Ingeschatte draagkracht ouders	Zwak	Matig	Goed	
15	Geslacht	Man	Vrouw		
Constraints		**Rule**			
1	stabiel zonder ondersteuning (hemodynamiek)	niet gelijk zijn aan	7.0 (pH)		
2	instabiel ondanks maximale ondersteuning (hemodynamiek)	niet gelijk zijn aan	7.4 (pH)		
3	Geen complicaties(ongecompliceerd verloop sinds geboorte)	niet gelijk zijn aan	Slechte (longfunctie)		
4	Geen complicaties (ongecompliceerd verloop sinds geboorte)	niet gelijk zijn aan	Instabiel ondanks maximale ondersteuning (hemodynamiek)		
5	24 week (zwangerschapsduur)	max	800 gr (geboorte gewicht)		
6	25 week (zwangerschapsduur)	max	900 gr (geboorte gewicht)		
7	26 week (zwangerschapsduur)	max	1200 gr (geboorte gewicht)		
8	30 week (zwangerschapsduur)	min	750 gr (geboorte gewicht)		

Figure D.1: List of attributes after second plenary meeting

Appendix E Discussion of pilot study design

This Appendix discusses the structure of the last meeting for the design of the stated adaptation experiment. In this meeting, some choice tasks of the designed pilot study were discussed together with one UMCG physician.

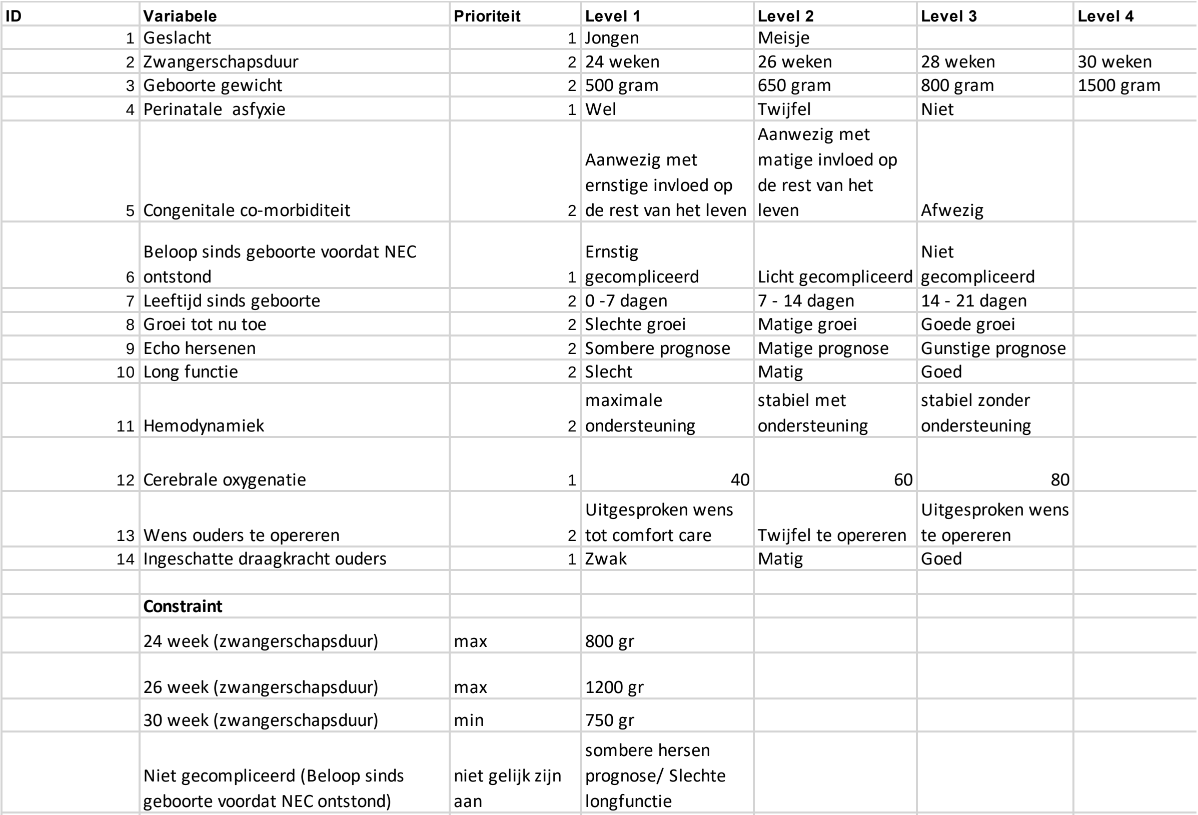
Structure

1. Introduction about the structure and goal of the meeting.
2. The last inspection on the Excel sheet of attributes, levels, and constraints, especially, to confirm that the lowest attribute level generates the largest probability for a recommendation against surgery and the highest level the greatest likelihood for a recommendation in favour of operation. Moreover, the UMCG physician was asked to execute five-choice tasks and reflect on each decision to confirm that the choice tasks were constructed such that they forced him or her to make trade-offs between the attributes. Thus, to assure that no specific attributes or attributes levels constituted a definite yes or no for surgery. Furthermore, this meeting also asked the UMCG physician whether colleagues possibly would make other decisions for the choice tasks. Lastly, the meeting intended to check whether the choice tasks were not too challenging.
3. Finally, after the UMCG physician executed several choice tasks and the above-mentioned requirements were confirmed, the meeting ended with some practical deliberations on the rest of the project.

After this meeting, some final modifications on the list of attributes, levels and constraints were made that resulted in the final list depicted in Appendix F.

Appendix F The final list of attributes and levels

Figure F.1 displays the final list of attributes, levels and constraints included in the pilot and final stated adaptation experiment.



ID	Variabele	Prioriteit	Level 1	Level 2	Level 3	Level 4
1	Geslacht	1	Jongen	Meisje		
2	Zwangerschapsduur	2	24 weken	26 weken	28 weken	30 weken
3	Geboorte gewicht	2	500 gram	650 gram	800 gram	1500 gram
4	Perinatale asfyxie	1	Wel	Twijfel	Niet	
5	Congenitale co-morbiditeit	2	Aanwezig met ernstige invloed op de rest van het leven	Aanwezig met matige invloed op de rest van het leven	Afwezig	
6	Beloop sinds geboorte voordat NEC ontstond	1	Ernstig gecompliceerd	Licht gecompliceerd	Niet gecompliceerd	
7	Leeftijd sinds geboorte	2	0 -7 dagen	7 - 14 dagen	14 - 21 dagen	
8	Groei tot nu toe	2	Slechte groei	Matige groei	Goede groei	
9	Echo hersenen	2	Sombere prognose	Matige prognose	Gunstige prognose	
10	Long functie	2	Slecht	Matig	Goed	
11	Hemodynamiek	2	maximale ondersteuning	stabiel met ondersteuning	stabiel zonder ondersteuning	
12	Cerebrale oxygenatie	1	40	60	80	
13	Wens ouders te opereren	2	Uitgesproken wens tot comfort care	Twijfel te opereren	Uitgesproken wens te opereren	
14	Ingeschatte draagkracht ouders	1	Zwak	Matig	Goed	
	Constraint					
	24 week (zwangerschapsduur)	max	800 gr			
	26 week (zwangerschapsduur)	max	1200 gr			
	30 week (zwangerschapsduur)	min	750 gr			
	Niet gecompliceerd (Beloop sinds geboorte voordat NEC ontstond)	niet gelijk zijn aan	sombere hersen prognose/ Slechte longfunctie			

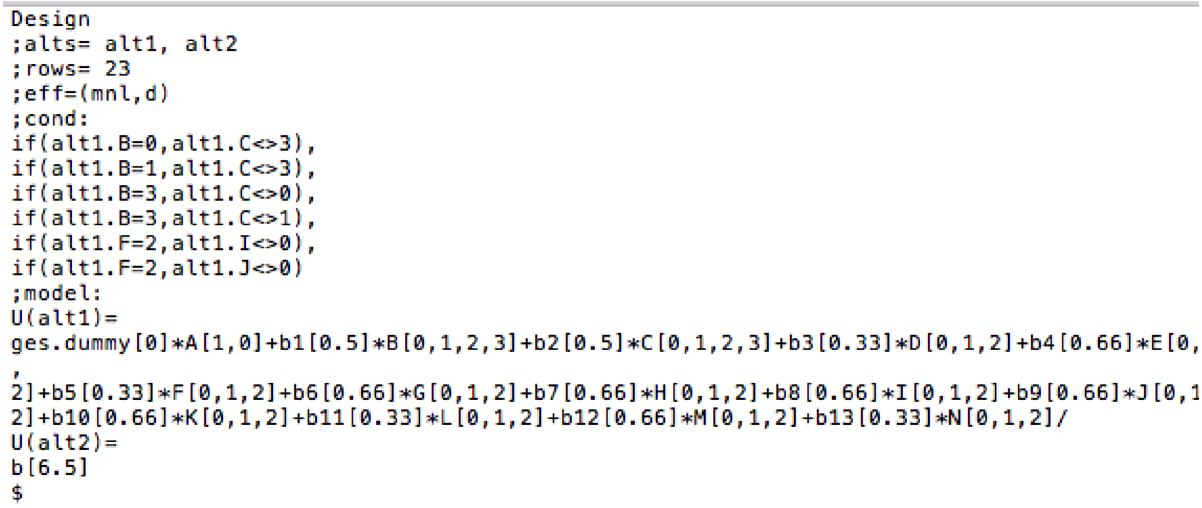
Figure F.1: Final list of attributes

Appendix G Pilot study design and results

This Appendix includes the Ngene syntax for the pilot study and illustrates the results of the pilot study. The priors for the final survey design are based on the binary logit model that solely incorporates the question on the preferred treatment.

Ngene syntax

Figure G.1 depicts the Ngene syntax for the pilot study that was constructed based on the list of attributes, levels, and constraints shown in Appendix F.



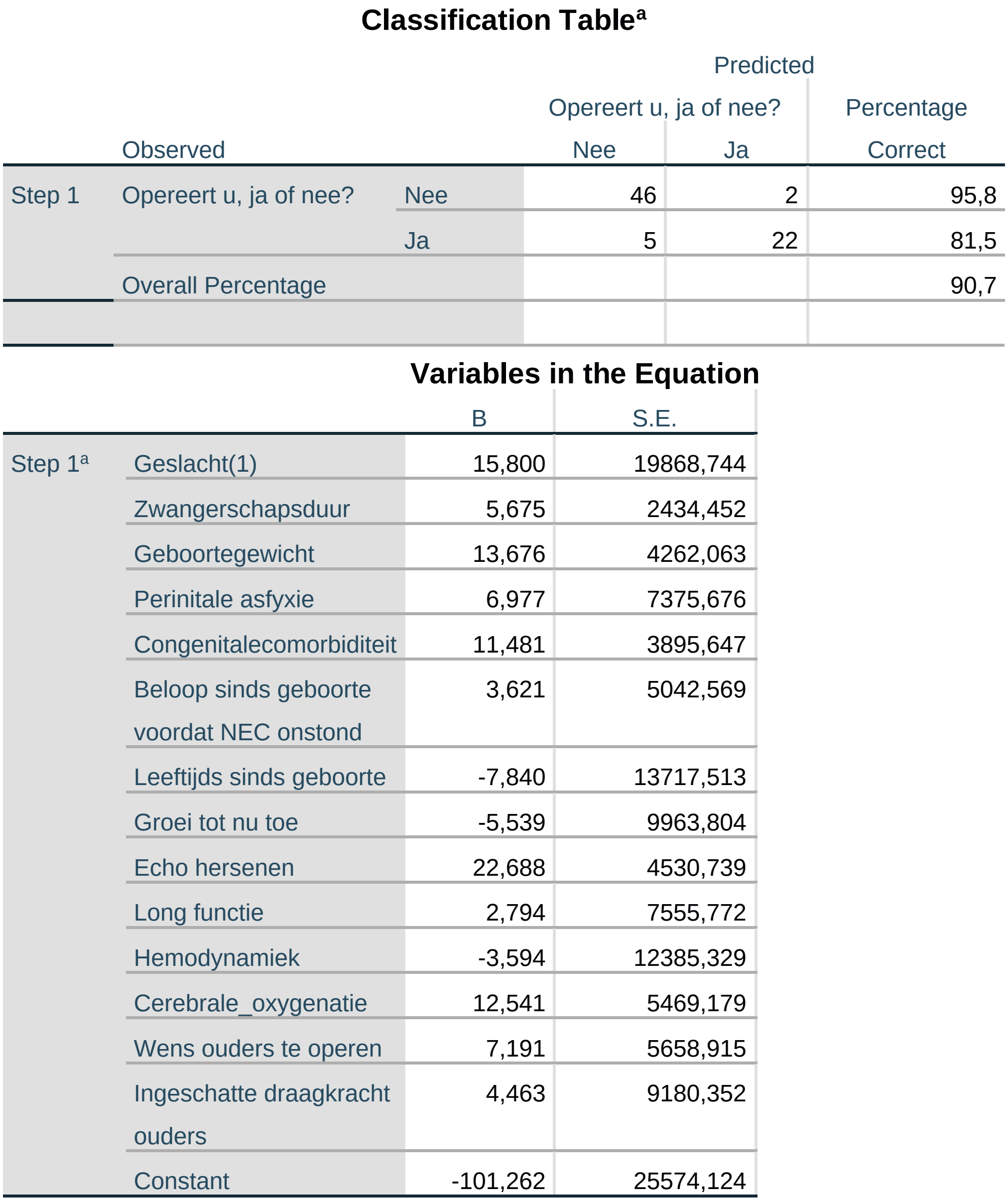
```
Design
;alts= alt1, alt2
;rows= 23
;eff=(mnl,d)
;cond:
if(alt1.B=0,alt1.C<>3),
if(alt1.B=1,alt1.C<>3),
if(alt1.B=3,alt1.C<>0),
if(alt1.B=3,alt1.C<>1),
if(alt1.F=2,alt1.I<>0),
if(alt1.F=2,alt1.J<>0)
;model:
U(alt1)=
ges.dummy[0]*A[1,0]+b1[0.5]*B[0,1,2,3]+b2[0.5]*C[0,1,2,3]+b3[0.33]*D[0,1,2]+b4[0.66]*E[0,
,
2]+b5[0.33]*F[0,1,2]+b6[0.66]*G[0,1,2]+b7[0.66]*H[0,1,2]+b8[0.66]*I[0,1,2]+b9[0.66]*J[0,1
2]+b10[0.66]*K[0,1,2]+b11[0.33]*L[0,1,2]+b12[0.66]*M[0,1,2]+b13[0.33]*N[0,1,2]/
U(alt2)=
b[6.5]
$
```

Figure G.1: Ngene syntax for the pilot study

Results pilot study

Figure G.2 provides the overall results of the pilot study, while Figure G.3 demonstrates the results per individual UMCG physician. These results show that the model was unable to estimate parameters for one of the individual experts' as the answers presumably provided too little information. Additionally, the overall results display some unexpected signs. As explained in Chapter 5, the priors for all attributes should have a positive sign because the attribute levels are drafted such that an increase in attribute value enhances the likelihood for a recommendation in favour of surgery. Therefore, the parameters for the attributes: Leeftijd sinds geboorte, Groei tot nu toe, and Hemodynamiek are not incorporated as priors for the final design. The priors for these attributes are seized from the parameters of the first individual UMCG physician as the estimated parameters for this individual did portray positive signs.

Moreover, another significant observation is that the estimated parameters are very large. The most logical explanation for this observation is that the model incorporates the preferences of a very small group of individuals. Results showed that the answers to the choice tasks were largely similar for this small group of physicians, which, thereby, generates a high prediction accuracy as demonstrated in the results. Expected is that the model for the final survey will estimate smaller parameters as the group of physicians is considerably larger.



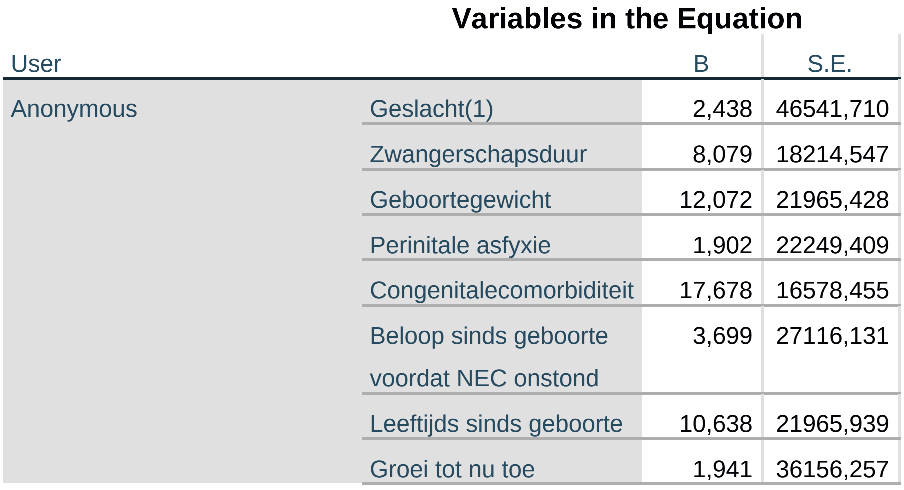
Classification Table[a]

Observed			Predicted Opereert u, ja of nee? Nee	 Ja	 Percentage Correct
Step 1	Opereert u, ja of nee?	Nee	46	2	95,8
		Ja	5	22	81,5
	Overall Percentage				90,7

Variables in the Equation

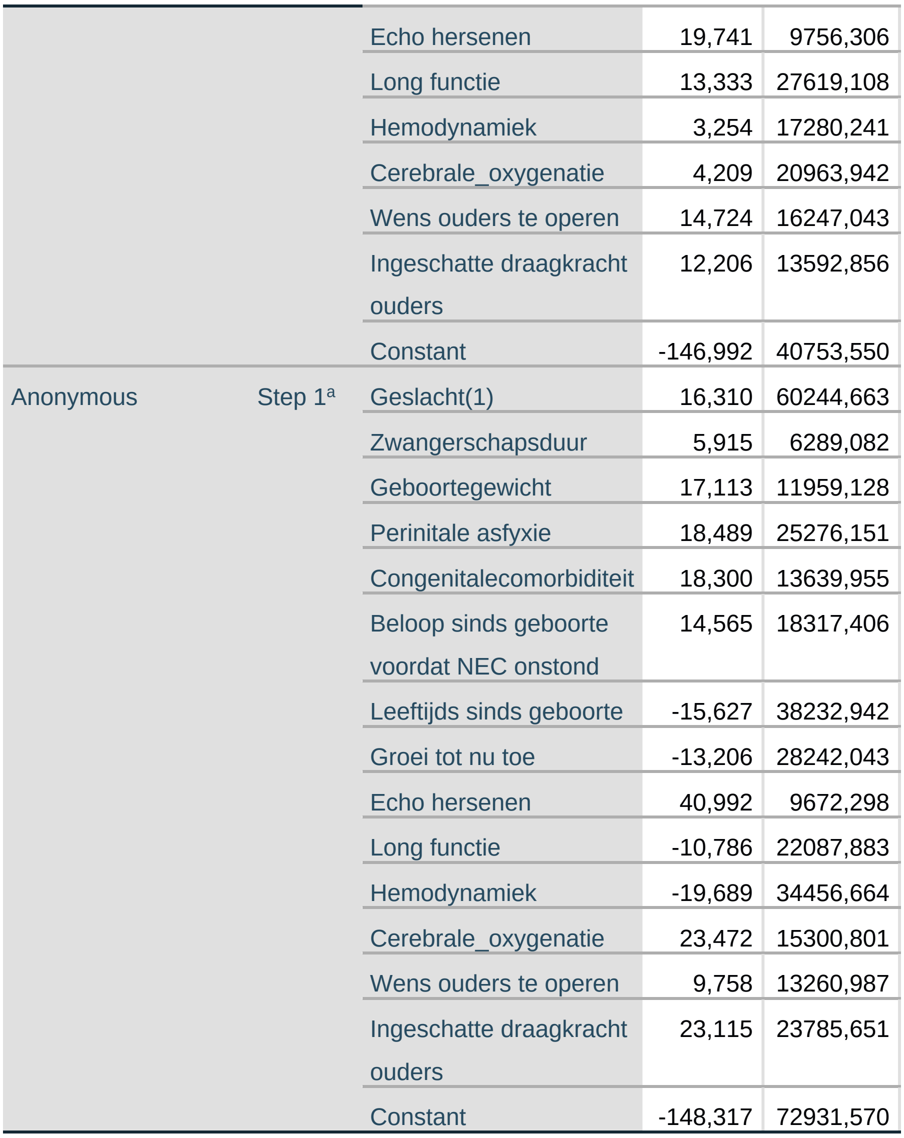
		B	S.E.
Step 1[a]	Geslacht(1)	15,800	19868,744
	Zwangerschapsduur	5,675	2434,452
	Geboortegewicht	13,676	4262,063
	Perinitale asfyxie	6,977	7375,676
	Congenitalecomorbiditeit	11,481	3895,647
	Beloop sinds geboorte voordat NEC onstond	3,621	5042,569
	Leeftijds sinds geboorte	-7,840	13717,513
	Groei tot nu toe	-5,539	9963,804
	Echo hersenen	22,688	4530,739
	Long functie	2,794	7555,772
	Hemodynamiek	-3,594	12385,329
	Cerebrale_oxygenatie	12,541	5469,179
	Wens ouders te operen	7,191	5658,915
	Ingeschatte draagkracht ouders	4,463	9180,352
	Constant	-101,262	25574,124

Figure G.2: Overall results of the pilot study



Variables in the Equation

User		B	S.E.
Anonymous	Geslacht(1)	2,438	46541,710
	Zwangerschapsduur	8,079	18214,547
	Geboortegewicht	12,072	21965,428
	Perinitale asfyxie	1,902	22249,409
	Congenitalecomorbiditeit	17,678	16578,455
	Beloop sinds geboorte voordat NEC onstond	3,699	27116,131
	Leeftijds sinds geboorte	10,638	21965,939
	Groei tot nu toe	1,941	36156,257

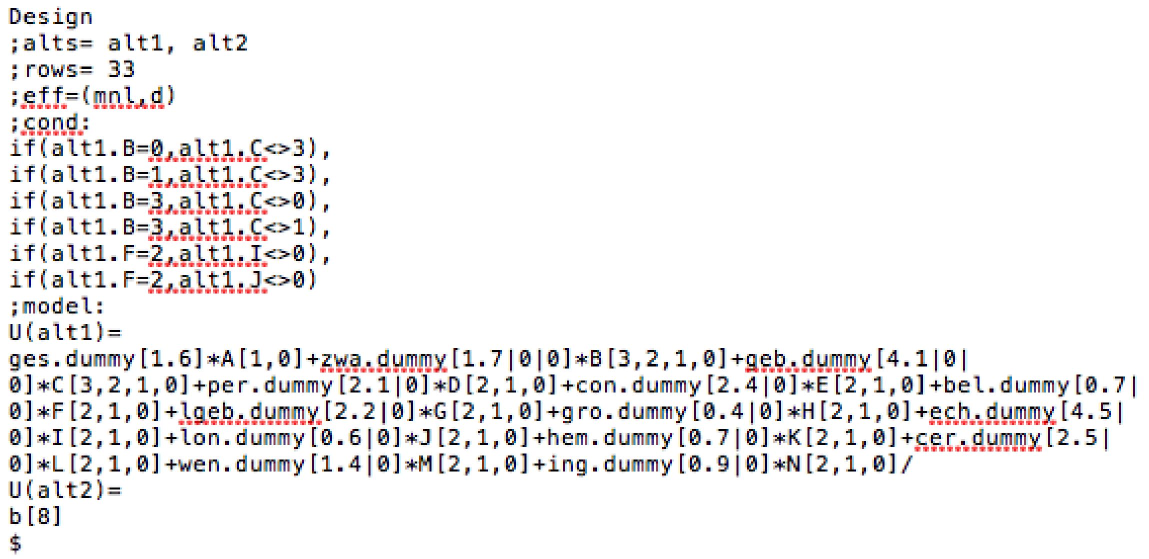


		Echo hersenen	19,741	9756,306
		Long functie	13,333	27619,108
		Hemodynamiek	3,254	17280,241
		Cerebrale_oxygenatie	4,209	20963,942
		Wens ouders te operen	14,724	16247,043
		Ingeschatte draagkracht ouders	12,206	13592,856
		Constant	-146,992	40753,550
Anonymous	Step 1[a]	Geslacht(1)	16,310	60244,663
		Zwangerschapsduur	5,915	6289,082
		Geboortegewicht	17,113	11959,128
		Perinitale asfyxie	18,489	25276,151
		Congenitalecomorbiditeit	18,300	13639,955
		Beloop sinds geboorte voordat NEC onstond	14,565	18317,406
		Leeftijds sinds geboorte	-15,627	38232,942
		Groei tot nu toe	-13,206	28242,043
		Echo hersenen	40,992	9672,298
		Long functie	-10,786	22087,883
		Hemodynamiek	-19,689	34456,664
		Cerebrale_oxygenatie	23,472	15300,801
		Wens ouders te operen	9,758	13260,987
		Ingeschatte draagkracht ouders	23,115	23785,651
		Constant	-148,317	72931,570

Figure G.3: Individual results of the pilot study

Appendix H Design of Final survey

This Appendix includes the Ngene syntax for the final survey. The priors of the syntax are established on the estimates of the pilot study. As explained in Appendix G, the priors for the pilot study are considerably large. Hence, the parameters are divided by a factor of 10 to serve as priors for the final survey. For each attribute, the priors were multiplied by the number of levels to determine the utility range. All attributes are dummy coded to test for non-linear effects, as the pilot study provided no information on non-linear effects these priors are set to zero.

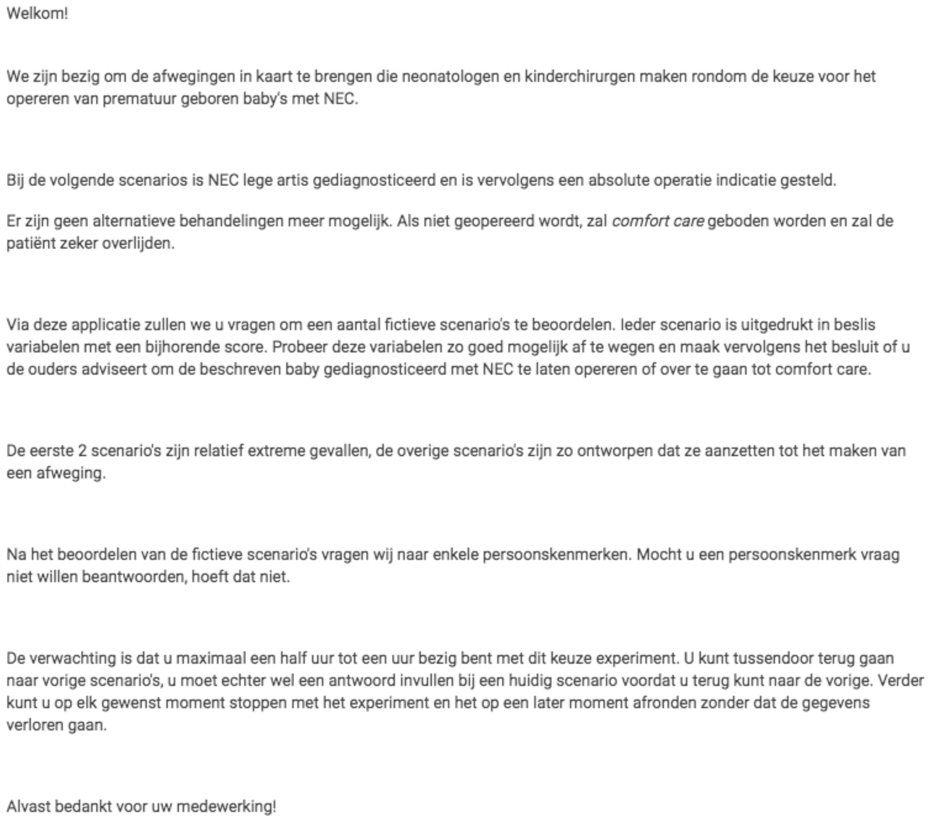


```
Design
;alts= alt1, alt2
;rows= 33
;eff=(mnl,d)
;cond:
if(alt1.B=0,alt1.C<>3),
if(alt1.B=1,alt1.C<>3),
if(alt1.B=3,alt1.C<>0),
if(alt1.B=3,alt1.C<>1),
if(alt1.F=2,alt1.I<>0),
if(alt1.F=2,alt1.J<>0)
;model:
U(alt1)=
ges.dummy[1.6]*A[1,0]+zwa.dummy[1.7|0|0]*B[3,2,1,0]+geb.dummy[4.1|0|
0]*C[3,2,1,0]+per.dummy[2.1|0]*D[2,1,0]+con.dummy[2.4|0]*E[2,1,0]+bel.dummy[0.7|
0]*F[2,1,0]+lgeb.dummy[2.2|0]*G[2,1,0]+gro.dummy[0.4|0]*H[2,1,0]+ech.dummy[4.5|
0]*I[2,1,0]+lon.dummy[0.6|0]*J[2,1,0]+hem.dummy[0.7|0]*K[2,1,0]+cer.dummy[2.5|
0]*L[2,1,0]+wen.dummy[1.4|0]*M[2,1,0]+ing.dummy[0.9|0]*N[2,1,0]/
U(alt2)=
b[8]
$
```

Figure H.1: Ngene syntax for the final experiment

Appendix I Final survey

This Appendix includes the final survey shown in Figure I.1. The first two choice tasks include the questions per choice scenario, while the other choice tasks just incorporate the choice scenarios. The choice experiment consists of 35 choice tasks.



Welkom!

We zijn bezig om de afwegingen in kaart te brengen die neonatologen en kinderchirurgen maken rondom de keuze voor het opereren van prematuur geboren baby's met NEC.

Bij de volgende scenarios is NEC lege artis gediagnosticeerd en is vervolgens een absolute operatie indicatie gesteld.

Er zijn geen alternatieve behandelingen meer mogelijk. Als niet geopereerd wordt, zal *comfort care* geboden worden en zal de patiënt zeker overlijden.

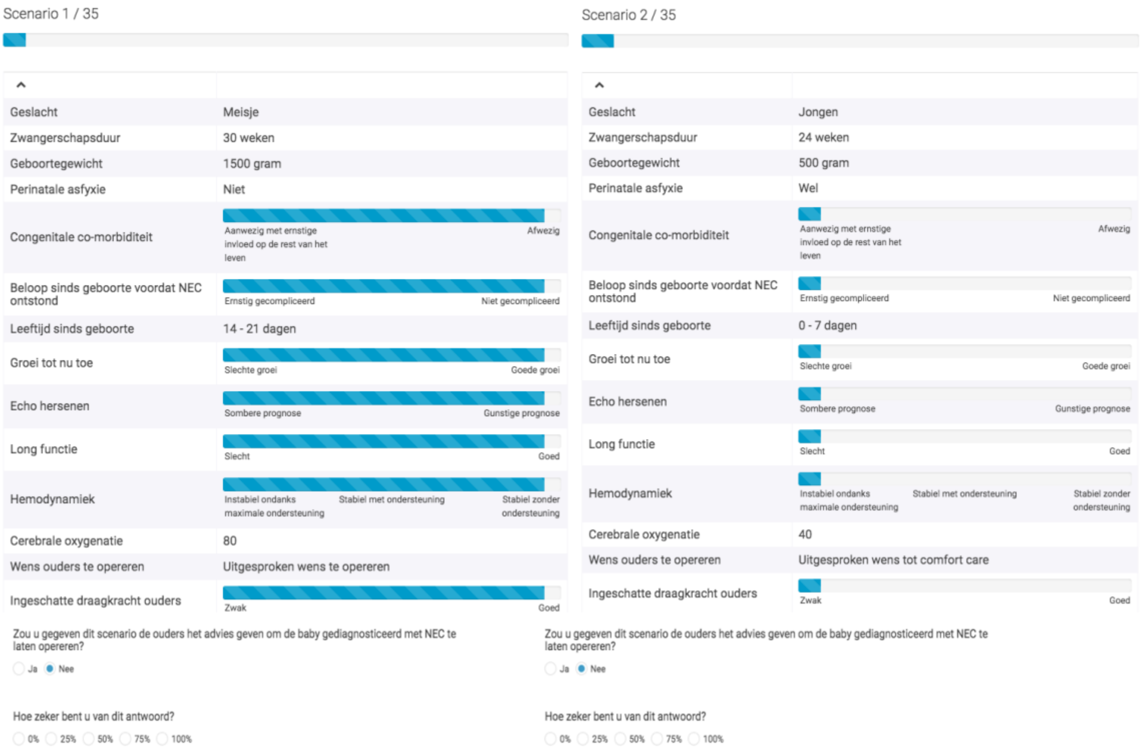
Via deze applicatie zullen we u vragen om een aantal fictieve scenario's te beoordelen. Ieder scenario is uitgedrukt in beslis variabelen met een bijhorende score. Probeer deze variabelen zo goed mogelijk af te wegen en maak vervolgens het besluit of u de ouders adviseert om de beschreven baby gediagnosticeerd met NEC te laten opereren of over te gaan tot comfort care.

De eerste 2 scenario's zijn relatief extreme gevallen, de overige scenario's zijn zo ontworpen dat ze aanzetten tot het maken van een afweging.

Na het beoordelen van de fictieve scenario's vragen wij naar enkele persoonskenmerken. Mocht u een persoonskenmerk vraag niet willen beantwoorden, hoeft dat niet.

De verwachting is dat u maximaal een half uur tot een uur bezig bent met dit keuze experiment. U kunt tussendoor terug gaan naar vorige scenario's, u moet echter wel een antwoord invullen bij een huidig scenario voordat u terug kunt naar de vorige. Verder kunt u op elk gewenst moment stoppen met het experiment en het op een later moment afronden zonder dat de gegevens verloren gaan.

Alvast bedankt voor uw medewerking!



Scenario 1 / 35

^	
Geslacht	Meisje
Zwangerschapsduur	30 weken
Geboortegewicht	1500 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed

Zou u gegeven dit scenario de ouders het advies geven om de baby gediagnosticeerd met NEC te laten opereren?

○ Ja ● Nee

Hoe zeker bent u van dit antwoord?

○ 0% ○ 25% ○ 50% ○ 75% ○ 100%

Scenario 2 / 35

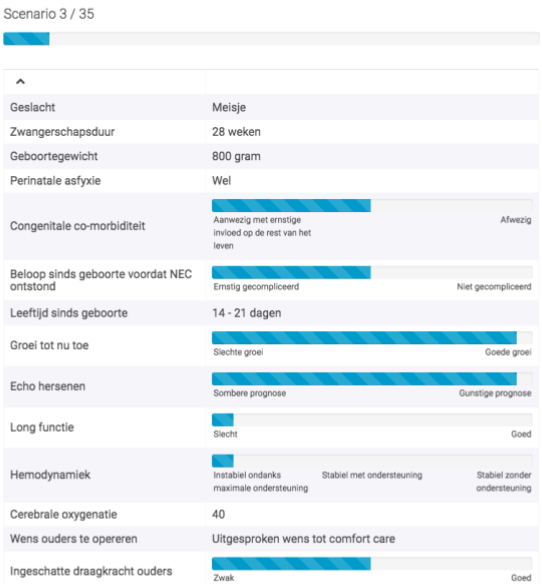
^	
Geslacht	Jongen
Zwangerschapsduur	24 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak Goed

Zou u gegeven dit scenario de ouders het advies geven om de baby gediagnosticeerd met NEC te laten opereren?

○ Ja ● Nee

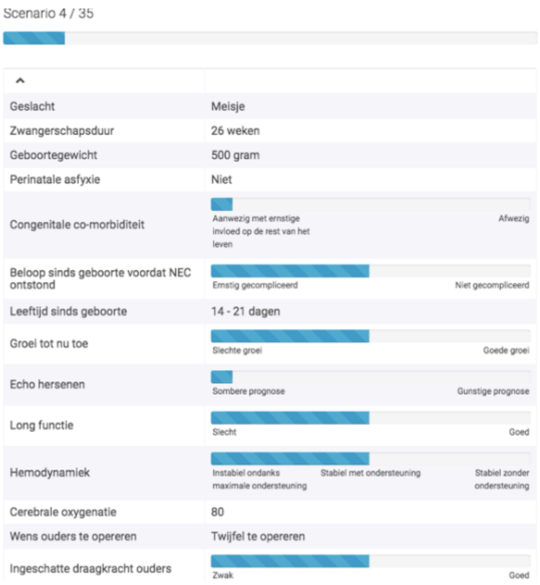
Hoe zeker bent u van dit antwoord?

○ 0% ○ 25% ○ 50% ○ 75% ○ 100%



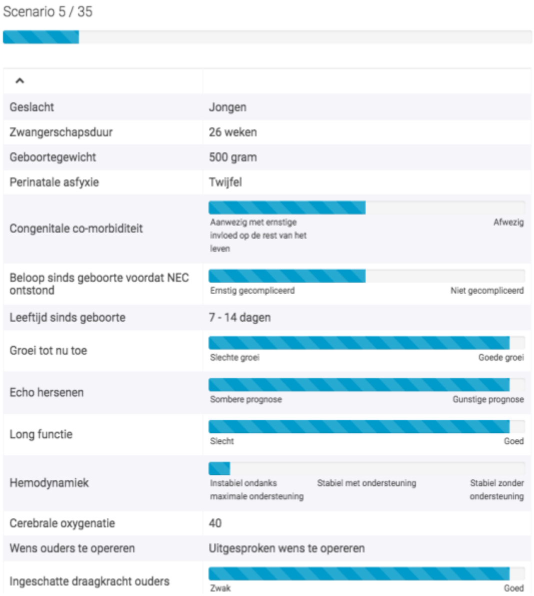
Scenario 3 / 35

Geslacht	Meisje
Zwangerschapsduur	28 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven / Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd / Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei / Goede groei
Echo hersenen	Sombere prognose / Gunstige prognose
Long functie	Slecht / Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning / Stabiel met ondersteuning / Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak / Goed



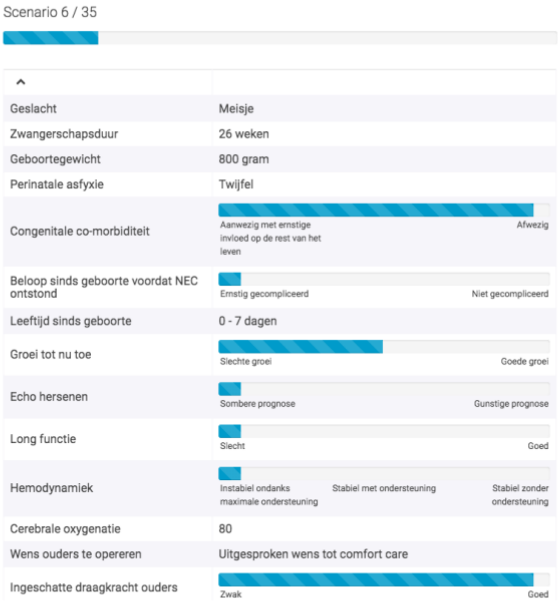
Scenario 4 / 35

Geslacht	Meisje
Zwangerschapsduur	26 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven / Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd / Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei / Goede groei
Echo hersenen	Sombere prognose / Gunstige prognose
Long functie	Slecht / Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning / Stabiel met ondersteuning / Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak / Goed



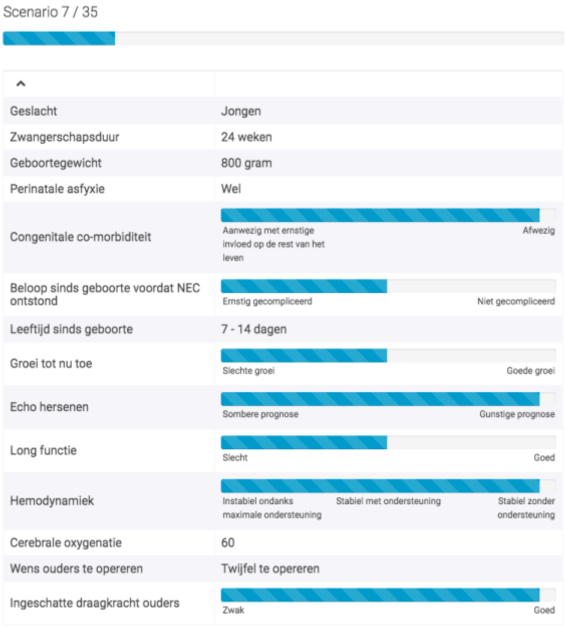
Scenario 5 / 35

Geslacht	Jongen
Zwangerschapsduur	26 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven / Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd / Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei / Goede groei
Echo hersenen	Sombere prognose / Gunstige prognose
Long functie	Slecht / Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning / Stabiel met ondersteuning / Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak / Goed



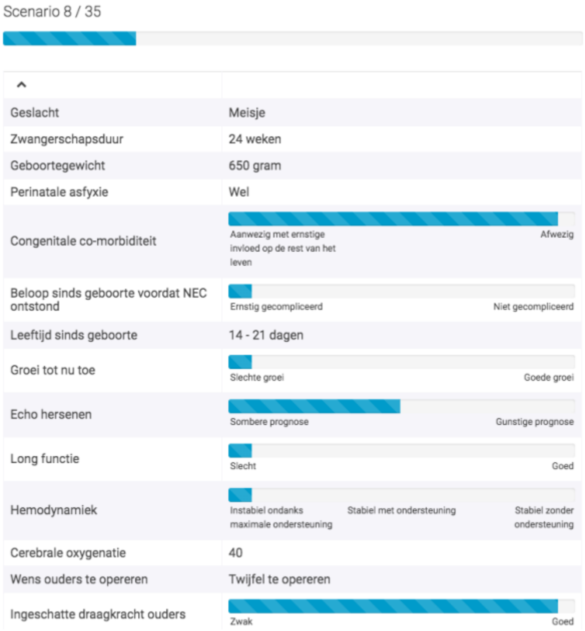
Scenario 6 / 35

Geslacht	Meisje
Zwangerschapsduur	26 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven / Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd / Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei / Goede groei
Echo hersenen	Sombere prognose / Gunstige prognose
Long functie	Slecht / Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning / Stabiel met ondersteuning / Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak / Goed



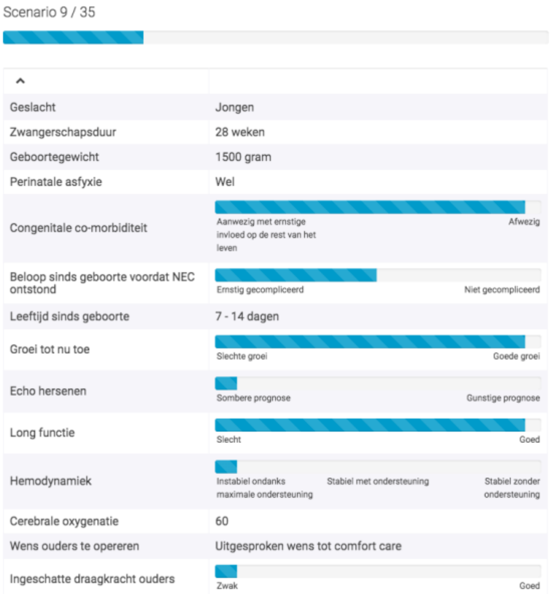
Scenario 7 / 35

Geslacht	Jongen
Zwangerschapsduur	24 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven / Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd / Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei / Goede groei
Echo hersenen	Sombere prognose / Gunstige prognose
Long functie	Slecht / Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning / Stabiel met ondersteuning / Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak / Goed



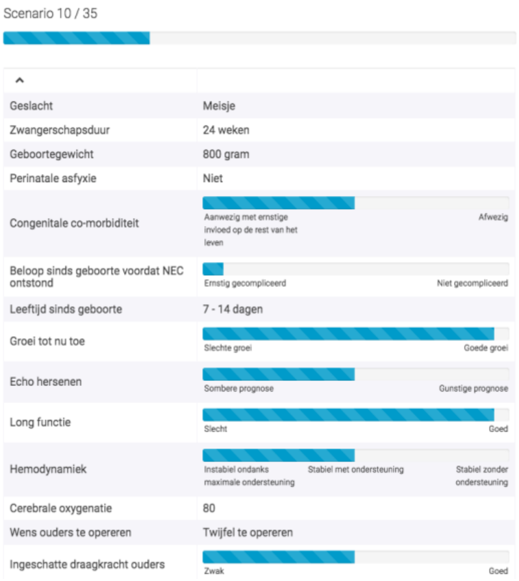
Scenario 8 / 35

Geslacht	Meisje
Zwangerschapsduur	24 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven / Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd / Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei / Goede groei
Echo hersenen	Sombere prognose / Gunstige prognose
Long functie	Slecht / Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning / Stabiel met ondersteuning / Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak / Goed



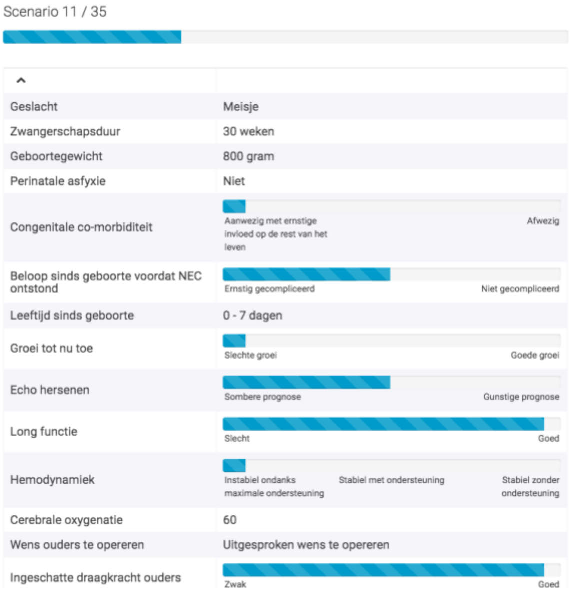
Scenario 9 / 35

Geslacht	Jongen
Zwangerschapsduur	28 weken
Geboortegewicht	1500 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak Goed



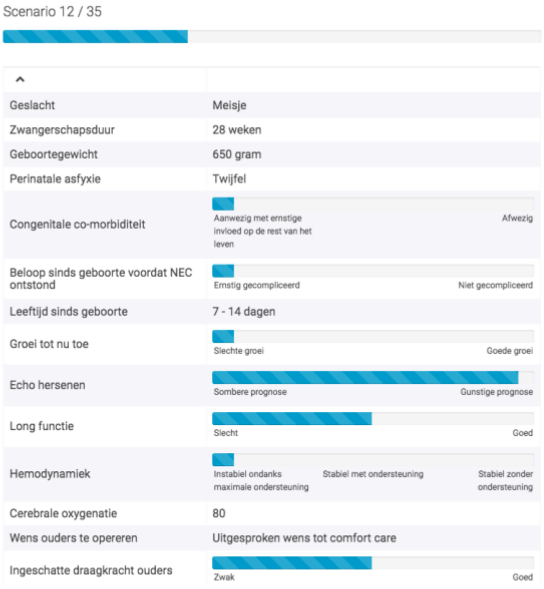
Scenario 10 / 35

Geslacht	Meisje
Zwangerschapsduur	24 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak Goed



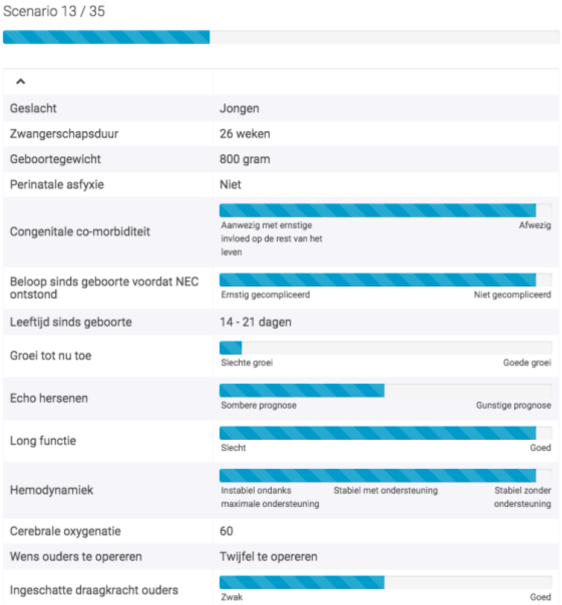
Scenario 11 / 35

Geslacht	Meisje
Zwangerschapsduur	30 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed



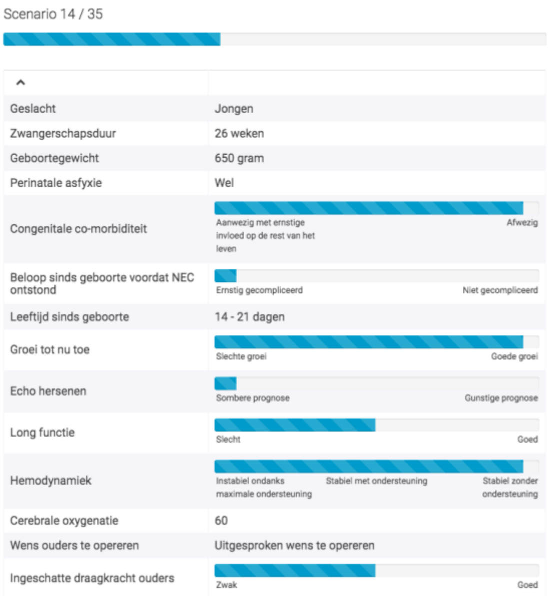
Scenario 12 / 35

Geslacht	Meisje
Zwangerschapsduur	28 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak Goed



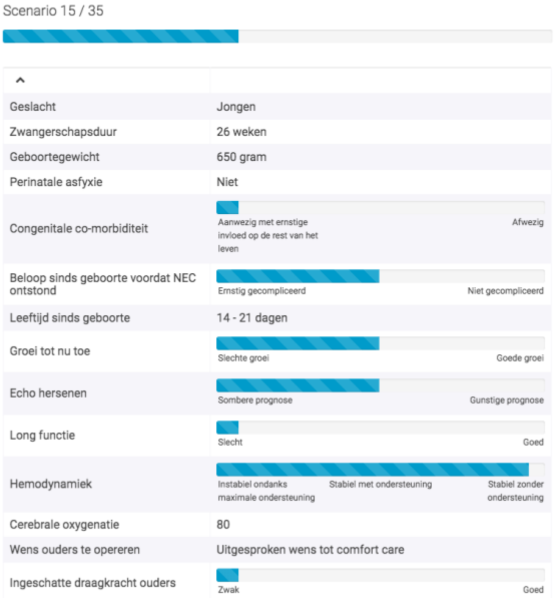
Scenario 13 / 35

Geslacht	Jongen
Zwangerschapsduur	26 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak Goed



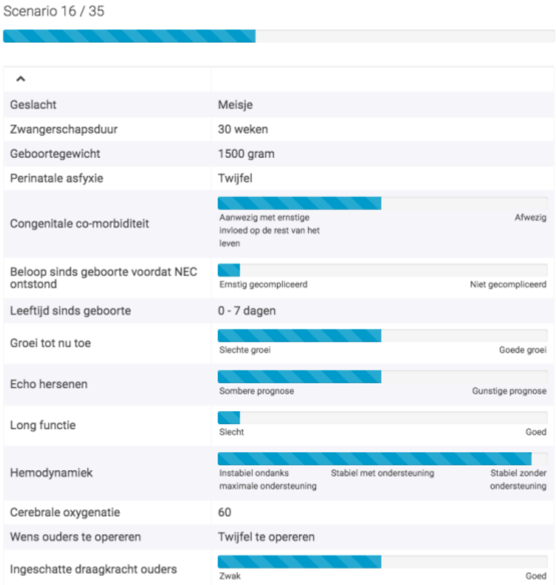
Scenario 14 / 35

Geslacht	Jongen
Zwangerschapsduur	26 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed



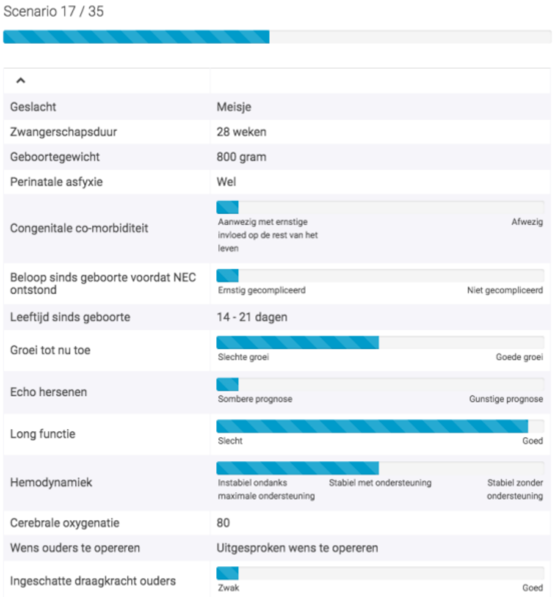
Scenario 15 / 35

Geslacht	Jongen
Zwangerschapsduur	26 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven – Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd – Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei – Goede groei
Echo hersenen	Sombere prognose – Gunstige prognose
Long functie	Slecht – Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning – Stabiel met ondersteuning – Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak – Goed



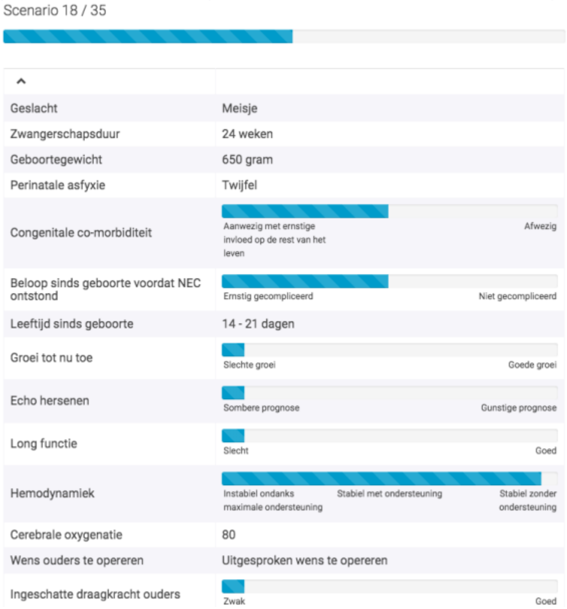
Scenario 16 / 35

Geslacht	Meisje
Zwangerschapsduur	30 weken
Geboortegewicht	1500 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven – Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd – Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei – Goede groei
Echo hersenen	Sombere prognose – Gunstige prognose
Long functie	Slecht – Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning – Stabiel met ondersteuning – Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak – Goed



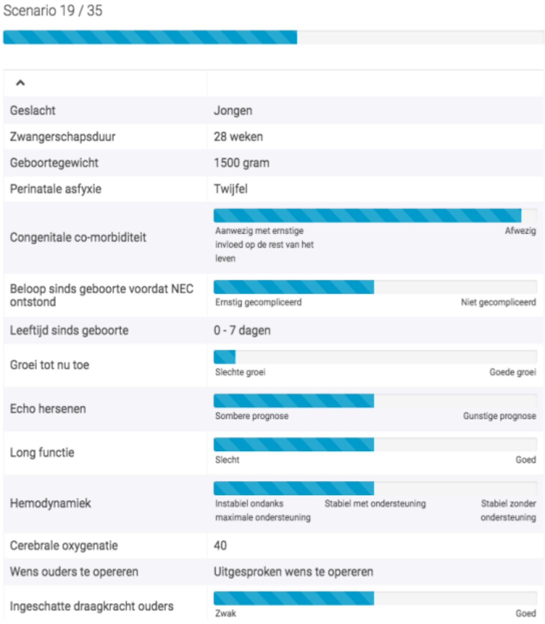
Scenario 17 / 35

Geslacht	Meisje
Zwangerschapsduur	28 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven – Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd – Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei – Goede groei
Echo hersenen	Sombere prognose – Gunstige prognose
Long functie	Slecht – Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning – Stabiel met ondersteuning – Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak – Goed



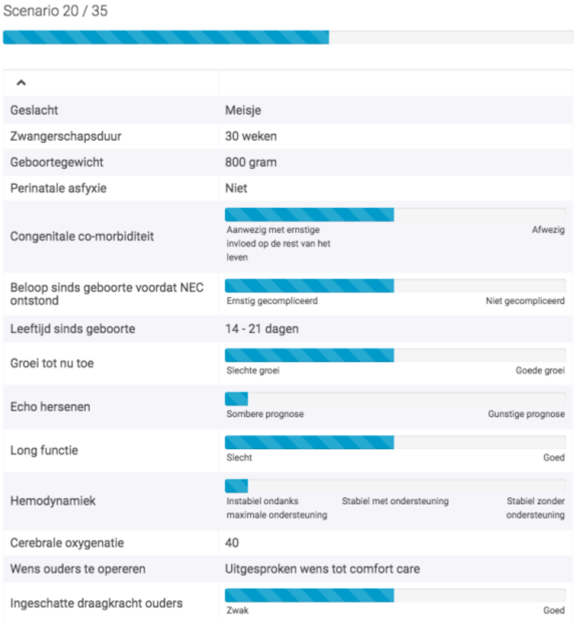
Scenario 18 / 35

Geslacht	Meisje
Zwangerschapsduur	24 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven – Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd – Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei – Goede groei
Echo hersenen	Sombere prognose – Gunstige prognose
Long functie	Slecht – Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning – Stabiel met ondersteuning – Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak – Goed



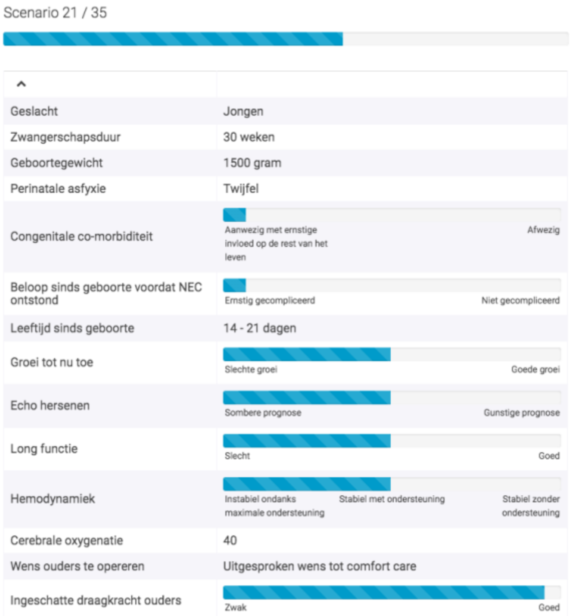
Scenario 19 / 35

Geslacht	Jongen
Zwangerschapsduur	28 weken
Geboortegewicht	1500 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven – Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd – Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei – Goede groei
Echo hersenen	Sombere prognose – Gunstige prognose
Long functie	Slecht – Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning – Stabiel met ondersteuning – Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak – Goed



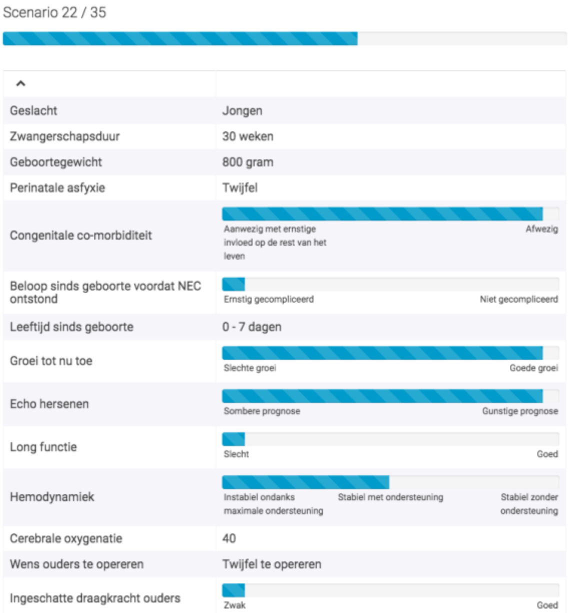
Scenario 20 / 35

Geslacht	Meisje
Zwangerschapsduur	30 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven – Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd – Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei – Goede groei
Echo hersenen	Sombere prognose – Gunstige prognose
Long functie	Slecht – Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning – Stabiel met ondersteuning – Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak – Goed



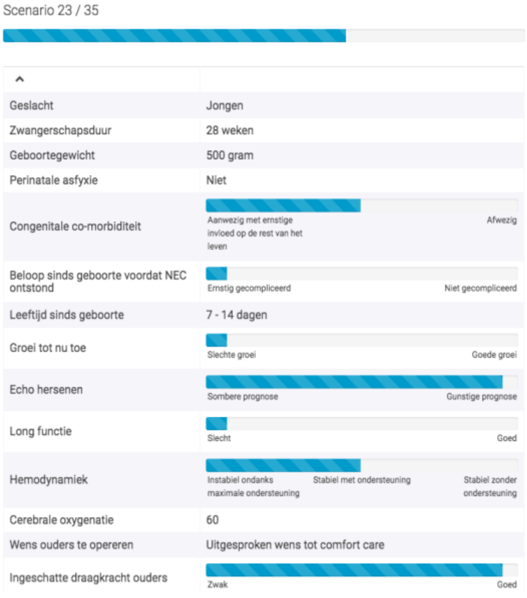
Scenario 21 / 35

Geslacht	Jongen
Zwangerschapsduur	30 weken
Geboortegewicht	1500 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak Goed



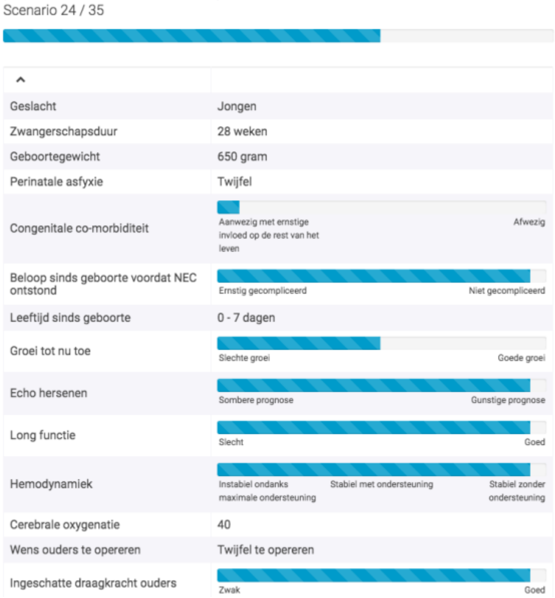
Scenario 22 / 35

Geslacht	Jongen
Zwangerschapsduur	30 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak Goed



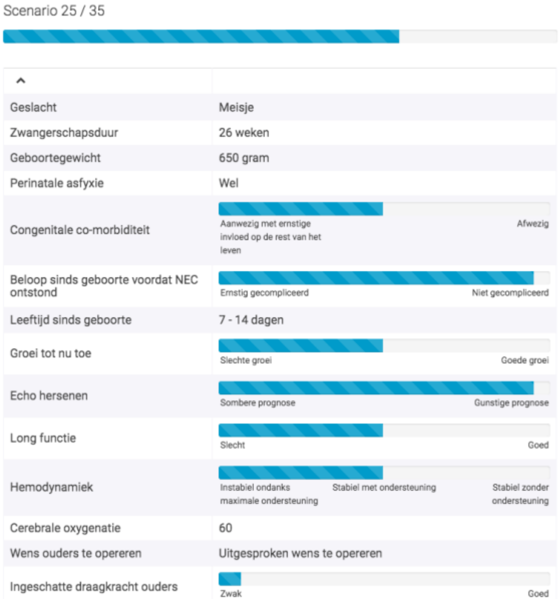
Scenario 23 / 35

Geslacht	Jongen
Zwangerschapsduur	28 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak Goed



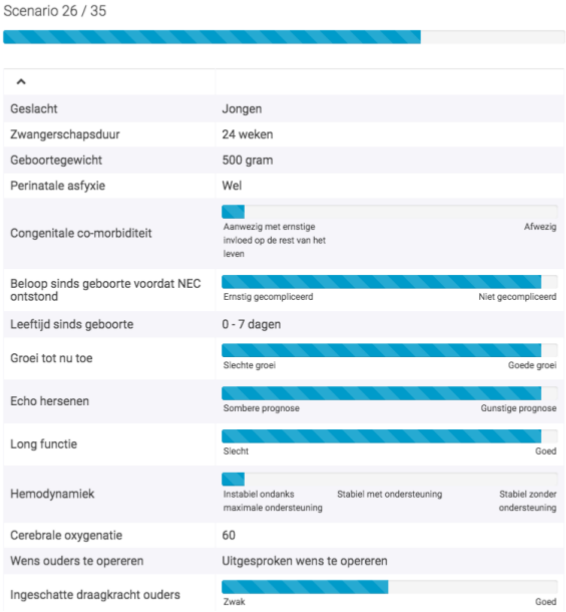
Scenario 24 / 35

Geslacht	Jongen
Zwangerschapsduur	28 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak Goed



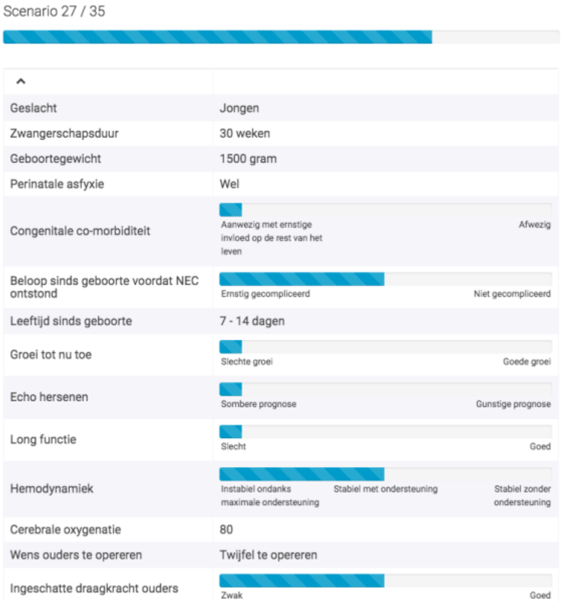
Scenario 25 / 35

Geslacht	Meisje
Zwangerschapsduur	26 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed



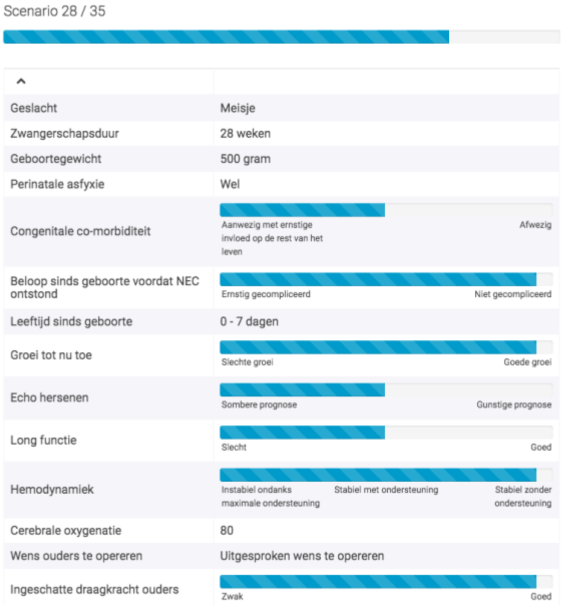
Scenario 26 / 35

Geslacht	Jongen
Zwangerschapsduur	24 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed



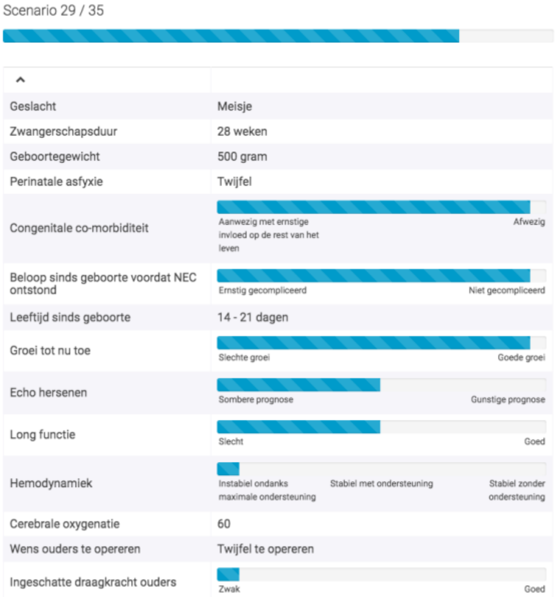
Scenario 27 / 35

Geslacht	Jongen
Zwangerschapsduur	30 weken
Geboortegewicht	1500 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak Goed



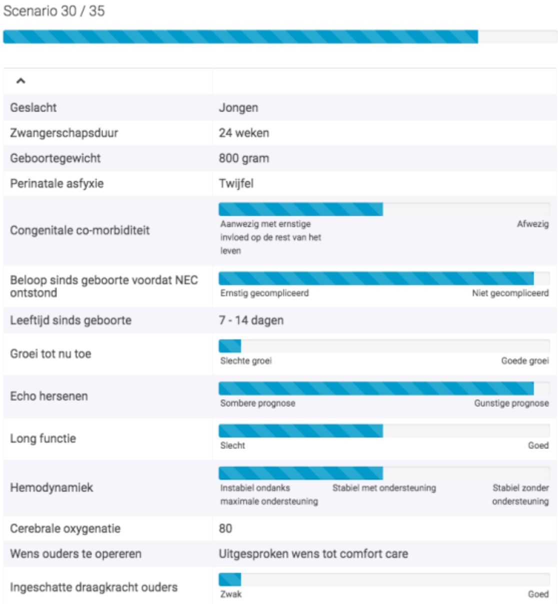
Scenario 28 / 35

Geslacht	Meisje
Zwangerschapsduur	28 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed



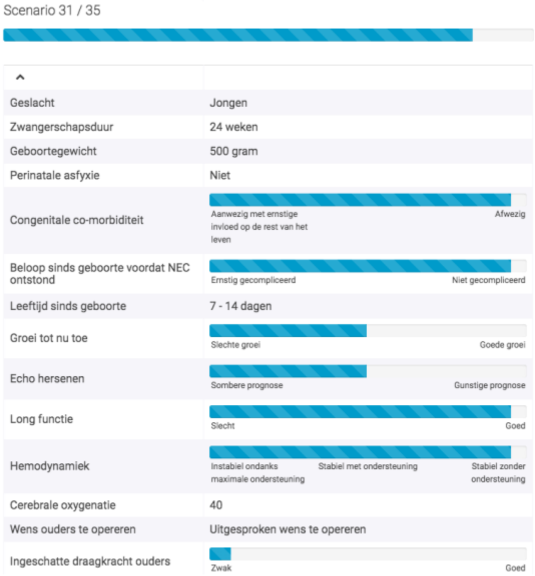
Scenario 29 / 35

Geslacht	Meisje
Zwangerschapsduur	28 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	14 - 21 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak Goed



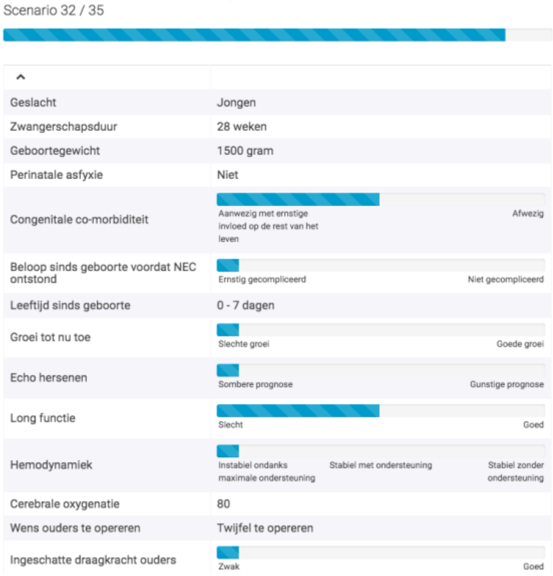
Scenario 30 / 35

Geslacht	Jongen
Zwangerschapsduur	24 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Twijfel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak Goed



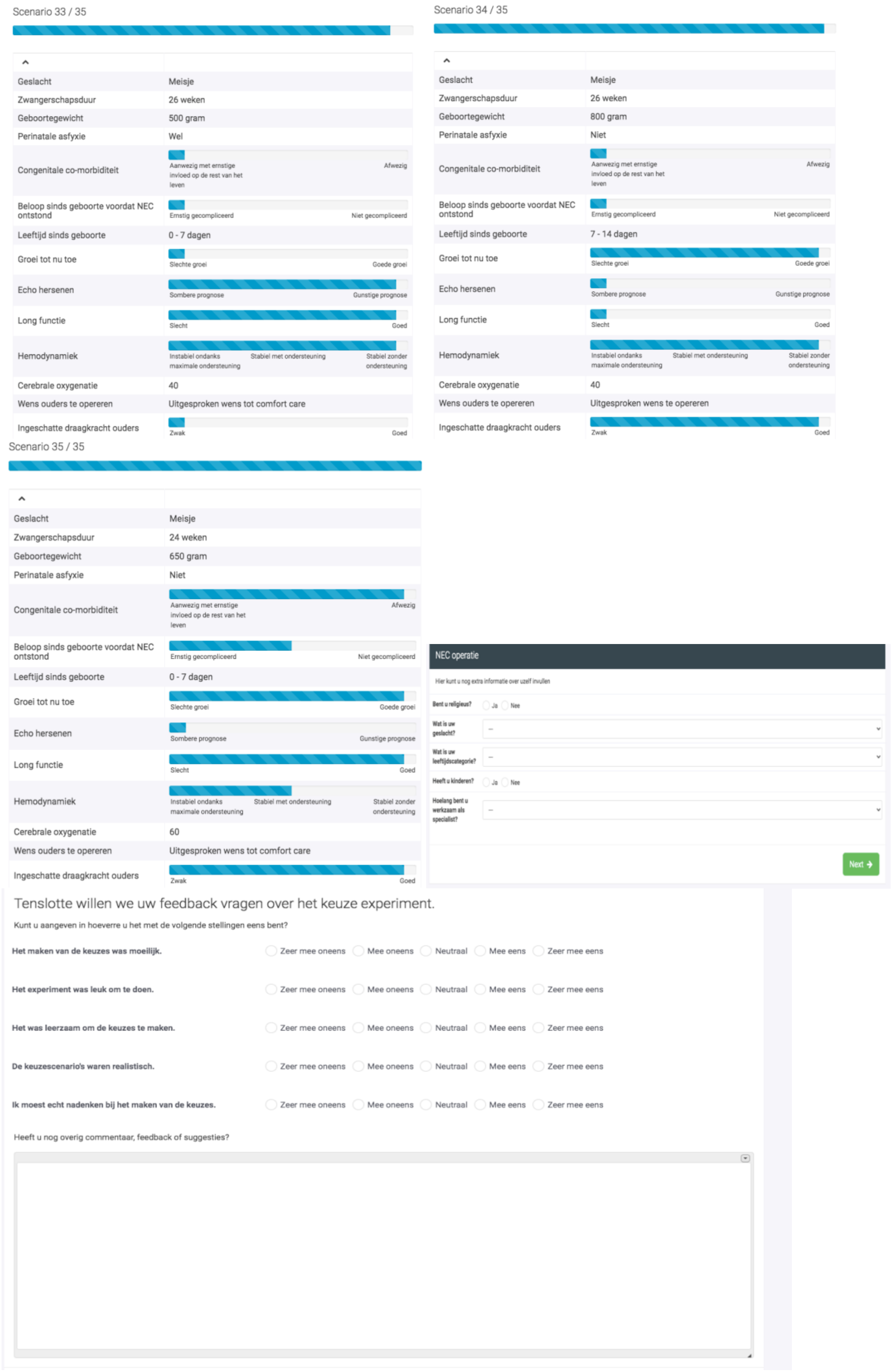
Scenario 31 / 35

Geslacht	Jongen
Zwangerschapsduur	24 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak Goed



Scenario 32 / 35

Geslacht	Jongen
Zwangerschapsduur	28 weken
Geboortegewicht	1500 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven Afwezig
Beloop sinds geboorte voordat NEC ontstond	Ernstig gecompliceerd Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei Goede groei
Echo hersenen	Sombere prognose Gunstige prognose
Long functie	Slecht Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning Stabiel met ondersteuning Stabiel zonder ondersteuning
Cerebrale oxygenatie	80
Wens ouders te opereren	Twijfel te opereren
Ingeschatte draagkracht ouders	Zwak Goed



Scenario 33 / 35

Geslacht	Meisje
Zwangerschapsduur	26 weken
Geboortegewicht	500 gram
Perinatale asfyxie	Wel
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven — Afwezig
Beloop sinds geboorte voordat NEC ontstond	Emstig gecompliceerd — Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei — Goede groei
Echo hersenen	Sombere prognose — Gunstige prognose
Long functie	Slecht — Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning — Stabiel met ondersteuning — Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak — Goed

Scenario 34 / 35

Geslacht	Meisje
Zwangerschapsduur	26 weken
Geboortegewicht	800 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven — Afwezig
Beloop sinds geboorte voordat NEC ontstond	Emstig gecompliceerd — Niet gecompliceerd
Leeftijd sinds geboorte	7 - 14 dagen
Groei tot nu toe	Slechte groei — Goede groei
Echo hersenen	Sombere prognose — Gunstige prognose
Long functie	Slecht — Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning — Stabiel met ondersteuning — Stabiel zonder ondersteuning
Cerebrale oxygenatie	40
Wens ouders te opereren	Uitgesproken wens te opereren
Ingeschatte draagkracht ouders	Zwak — Goed

Scenario 35 / 35

Geslacht	Meisje
Zwangerschapsduur	24 weken
Geboortegewicht	650 gram
Perinatale asfyxie	Niet
Congenitale co-morbiditeit	Aanwezig met ernstige invloed op de rest van het leven — Afwezig
Beloop sinds geboorte voordat NEC ontstond	Emstig gecompliceerd — Niet gecompliceerd
Leeftijd sinds geboorte	0 - 7 dagen
Groei tot nu toe	Slechte groei — Goede groei
Echo hersenen	Sombere prognose — Gunstige prognose
Long functie	Slecht — Goed
Hemodynamiek	Instabiel ondanks maximale ondersteuning — Stabiel met ondersteuning — Stabiel zonder ondersteuning
Cerebrale oxygenatie	60
Wens ouders te opereren	Uitgesproken wens tot comfort care
Ingeschatte draagkracht ouders	Zwak — Goed

NEC operatie

Hier kunt u nog extra informatie over uzelf invullen

Bent u religieus?	○ Ja ○ Nee
Wat is uw geslacht?	–
Wat is uw leeftijdscategorie?	–
Heeft u kinderen?	○ Ja ○ Nee
Hoelang bent u werkzaam als specialist?	–

Tenslotte willen we uw feedback vragen over het keuze experiment.

Kunt u aangeven in hoeverre u het met de volgende stellingen eens bent?

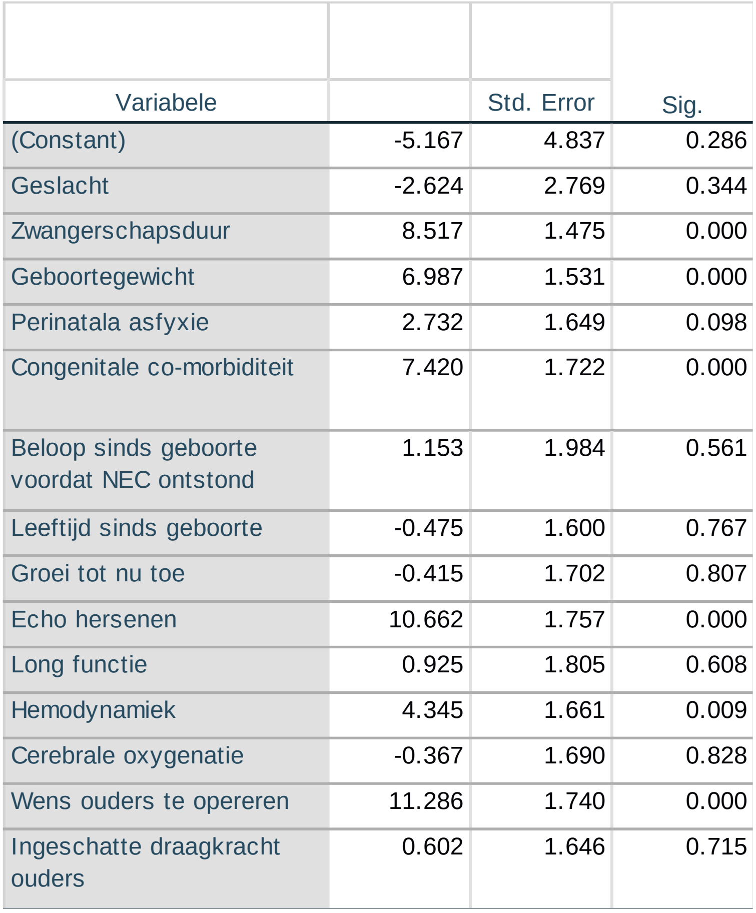
Het maken van de keuzes was moeilijk.	○ Zeer mee oneens	○ Mee oneens	○ Neutraal	○ Mee eens	○ Zeer mee eens
Het experiment was leuk om te doen.	○ Zeer mee oneens	○ Mee oneens	○ Neutraal	○ Mee eens	○ Zeer mee eens
Het was leerzaam om de keuzes te maken.	○ Zeer mee oneens	○ Mee oneens	○ Neutraal	○ Mee eens	○ Zeer mee eens
De keuzescenario's waren realistisch.	○ Zeer mee oneens	○ Mee oneens	○ Neutraal	○ Mee eens	○ Zeer mee eens
Ik moest echt nadenken bij het maken van de keuzes.	○ Zeer mee oneens	○ Mee oneens	○ Neutraal	○ Mee eens	○ Zeer mee eens

Heeft u nog overig commentaar, feedback of suggesties?

Figure I.1: Final survey

Appendix J Linear regression estimates

This appendix presents the parameters of the linear regression model estimated with the choice data of the recommendations on surgery and the certainty levels.

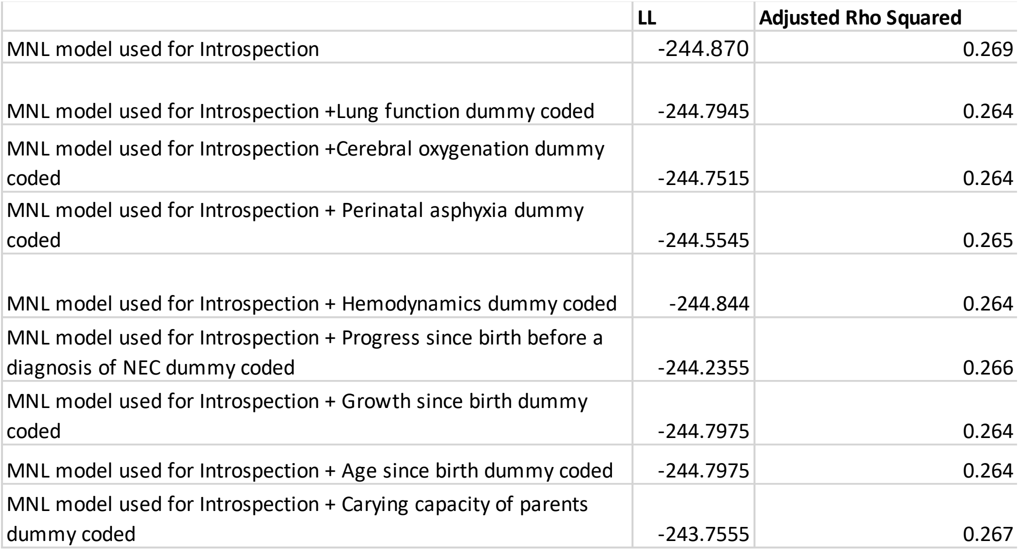


Variabele		Std. Error	Sig.
(Constant)	-5.167	4.837	0.286
Geslacht	-2.624	2.769	0.344
Zwangerschapsduur	8.517	1.475	0.000
Geboortegewicht	6.987	1.531	0.000
Perinatala asfyxie	2.732	1.649	0.098
Congenitale co-morbiditeit	7.420	1.722	0.000
Beloop sinds geboorte voordat NEC ontstond	1.153	1.984	0.561
Leeftijd sinds geboorte	-0.475	1.600	0.767
Groei tot nu toe	-0.415	1.702	0.807
Echo hersenen	10.662	1.757	0.000
Long functie	0.925	1.805	0.608
Hemodynamiek	4.345	1.661	0.009
Cerebrale oxygenatie	-0.367	1.690	0.828
Wens ouders te opereren	11.286	1.740	0.000
Ingeschatte draagkracht ouders	0.602	1.646	0.715

Figure J.1: Linear regression estimates

Appendix K Model fit parameters for individual binary logit models

Appendix K illustrates the model fit parameters estimated for multiple binary logit models to test whether dummy coding more variables would improve the model fit and, hence, fit the observed recommendations better. Figure K.1 depicts the model fit parameters of the binary logit models and shows that the model fit did not improve compared to the binary logit model dummy coding the five variables: wish of parents, gestational age, birth weight, congenital co-morbidity, and ultrasound of the brain.



	LL	Adjusted Rho Squared
MNL model used for Introspection	-244.870	0.269
MNL model used for Introspection +Lung function dummy coded	-244.7945	0.264
MNL model used for Introspection +Cerebral oxygenation dummy coded	-244.7515	0.264
MNL model used for Introspection + Perinatal asphyxia dummy coded	-244.5545	0.265
MNL model used for Introspection + Hemodynamics dummy coded	-244.844	0.264
MNL model used for Introspection + Progress since birth before a diagnosis of NEC dummy coded	-244.2355	0.266
MNL model used for Introspection + Growth since birth dummy coded	-244.7975	0.264
MNL model used for Introspection + Age since birth dummy coded	-244.7975	0.264
MNL model used for Introspection + Carying capacity of parents dummy coded	-243.7555	0.267

Figure K.1: Model fit parameters for the individual binary logit models

Appendix L Converted parameters to utils per unit

This Appendix illustrates the conversion of the parameters into utils per unit instead of utils per level. Figure L.1 present the converted parameters.

Attribute	**Parameter per level (utils /level step)**	**Parameter per unit (utils / unit)**
Gender (girl)	0.020	0.02 utils
Gestational age (24 weeks)		
Gestational age (26 weeks)	1.656	0.83 utils/week (24-25 weeks)
Gestational age (28 weeks)	1.851	0.1 utils/week (26-28 weeks)
Gestational age (30 weeks)	2.859	0.5 utils/week (28-30 weeks)
Birth weight (500 grams)		
Birth weight (650 grams)	1.238	0.01 utils/gram (500-650 grams)
Birth weight (800 grams)	1.835	0.004 utils/gram (650-800 grams)
Birth weight (1500 grams)	2.507	0.001 utils/gram (800-1500 grams)
Perinatal asphyxia	0.452	0.452 utils/level step
Congenital comorbidity (present with high impact)		
Congenital comorbidity (present with minor impact)	0.944	0.0189 utils/% (0-50%)
Congenital comorbidity (absent)	1.752	0.0162 utils/% (50-100%)
Progress since birth before a diagnosis of NEC	0.230	0.00230 utils/ %
Age since birth	0.250	0.00250 utils/%
Growth since birth	0.183	0.00183 utisl/%
Ultrasound of the brain (bad prognosis)		
Ultrasound of the brain (intermediate prognosis)	1.798	0.0360 utils/% (0-50%)
Ultrasound of the brain (good prognosis)	2.782	0.0164 utils/% (50-100%)
Lung function	0.204	0.00204 utils/%
Hemodynamic	0.279	0.00279 utils/%
Cerebral oxygenation	0.430	0.02 utils/SaO$_2$
Wish of parents (in favour of comfort care)		
Wish of parents (doubtful about surgery)	1.729	1.729 utils
Wish of parents (in favour of surgery)	2.154	2.154 utils
The carrying capacity of parents	0.216	0.0026 utils/%

Figure L.1: Converted parameters into utils per unit

Appendix M Specialisation estimates

This appendix presents the estimates of the binary logit models for child surgeons and neonatologists, as well as their model fit parameters. Figure M.1 illustrates the estimated parameters for the two specialisations. Figure M.1 illustrates that the standard errors for the parameters of the child surgeons fluctuate around 0.5. In contrast, the standard errors for the parameters of the neonatologists alter around 0.2 and, hence, are estimated with less uncertainty compared to parameters for child surgeons.

The binary logit model for the child surgeons does demonstrate a good model fit as the calculated adjusted ρ^2 is 0.83. The adjusted ρ^2 illustrates that the four UMCG physicians made consistent recommendations on the choice scenarios. The adjusted ρ^2 between the two binary logit models for are, however, incomparable due to the diverse choice data.

Moreover, the standard errors for the parameters illustrate that for both specialisms the variables gestational age and birth weight are most reliable as these variables portray the lowest standard errors while the variable gender is estimated with the most uncertainty.

Specialisation	Child surgeons	Standard error	Neonatologists	Standard error
Gender	-0.427	0.631	0.089	0.308
Gestational age	0.713	0.295	1.021	0.182
Birth weight	0.514	0.377	1.124	0.190
Perinatal asphyxia	0.868	0.428	0.664	0.210
Congenital comorbidity	1.895	0.444	0.772	0.209
Progress since birth before a diagnosis of NEC	0.143	0.432	0.295	0.221
Age since birth	0.422	0.465	0.437	0.234
Growth since birth	-0.143	0.358	0.460	0.216
Ultrasound of the brain	2.627	0.552	1.358	0.241
Lung function	0.416	0.383	0.405	0.206
Wish of parents	1.973	0.550	1.174	0.238
The carrying capacity of parent	0.826	0.493	0.232	0.223
Lung function	0.224	0.462	0.072	0.203
Cerebral oxygenation	0.765	0.471	0.620	0.238
Constant	-12.332	2.928	-9.551	1.511

Figure M.1: Estimated parameters for the group of child surgeons and neonatologists

Child surgeons	Neonatologists
LL: - 48	LL: -182
Adjusted rho squared: 0.83	Adjusted rho squared: 0.46

Appendix N Age group estimates

This appendix illustrates the estimates of the binary logit models for the two age categories. Additionally, it present the model fit parameters for both models. Figure N.1 presents the estimated parameters for the two age categories.

Age	25-45 years old	Standard error	45> years old	Standard error
Gender	-0.059	0.375	0.165	0.379
Gestational age	1.170	0.228	0.742	0.206
Birth weight	0.873	0.241	1.053	0.226
Perinatal asphyxia	0.676	0.261	0.622	0.247
Congenital comorbidity	1.326	0.274	0.738	0.252
Progress since birth before a diagnosis of NEC	0.479	0.286	0.148	0.260
Age since birth	0.436	0.304	0.403	0.271
Growth since birth	0.368	0.266	0.196	0.248
Ultrasound of the brain	1.718	0.295	1.358	0.288
Lung function	0.607	0.267	0.195	0.244
Wish of parents	1.375	0.297	1.231	0.296
The carrying capacity of parent	0.508	0.307	0.286	0.254
Lung function	0.236	0.248	-0.003	0.251
Cerebral oxygenation	0.878	0.318	0.413	0.277
Constant	-11.166	1.976	-8.658	1.709

Figure N.1: Estimated parameters for the age categories

25-45 years old:

LL = -121

Adjusted r squared = 0.63

45> years old:

LL = -120

Adjusted r squared = 0.63

Appendix O Length of professional experience estimates

This appendix illustrates the estimates of the binary logit models for the length of professional experience categories. Additionally, it present the model fit parameters for both models. Figure O.1 presents the estimated parameters for two models.

Length of professional experience	**0-10 years**	**Standard error**	**10> years**	**Standard error**
Gender	-0.224	0.392	0.254	0.368
Gestational age	1.092	0.234	0.838	0.200
Birth weight	0.853	0.258	1.081	0.220
Perinatal asphyxia	0.688	0.273	0.628	0.242
Congenital comorbidity	1.289	0.290	0.848	0.244
Progress since birth before a diagnosis of NEC	0.355	0.299	0.270	0.252
Age since birth	0.473	0.319	0.414	0.264
Growth since birth	0.314	0.275	0.236	0.242
Ultrasound of the brain	1.607	0.308	1.517	0.282
Lung function	0.641	0.280	0.190	0.234
Wish of parents	1.312	0.305	1.333	0.293
The carrying capacity of parent	0.486	0.322	0.340	0.249
Lung function	0.310	0.261	-0.023	0.244
Cerebral oxygenation	0.876	0.331	0.438	0.269
Constant	-10.562	2.071	-9.534	1.673

Figure O.1: Estimated parameters for length of professional experience categories

0-10 years
LL = -108
Adjusted R squared = 0.66

10> years
LL = -131
Adjusted R squared = 0.60

Appendix P Gender estimates

This appendix presents the estimates of the binary logit models for the group of females and male UMCG physicians, as well as their model fit parameters. Figure P.1 illustrates the estimated parameters for the two categories.

Gender	Man	Standard error	Women	Standard error
Gender	0.109	0.431	-0.085	0.333
Gestational age	0.610	0.237	1.119	0.195
Birth weight	0.924	0.255	0.952	0.207
Perinatal asphyxia	0.545	0.276	0.714	0.231
Congenital comorbidity	0.851	0.297	1.079	0.230
Progress since birth before a diagnosis of NEC	-0.128	0.296	0.514	0.244
Age since birth	0.292	0.307	0.473	0.260
Growth since birth	0.051	0.278	0.409	0.234
Ultrasound of the brain	1.326	0.334	1.626	0.258
Lung function	0.028	0.287	0.160	0.220
Hemodynamic	0.328	0.279	0.407	0.227
Cerebral oxygenation	0.550	0.312	0.671	0.267
Wish of parents	1.136	0.323	1.375	0.269
The carrying capacity of parents	0.216	0.288	0.443	0.256
Constant	-7.678	1.884	-10.770	1.683

Figure P.1: Estimated parameters for the gender categories

Male

LL = -88

Adjusted rho squared = 0.72

Female

LL = -156

Adjusted rho squared = 0.53

Appendix Q Religion estimates

This appendix illustrates the estimates of the binary logit models for religious and non-religious physicians. Additionally, it present the model fit parameters for both models. Figure Q.1 depicts the estimated parameters for the two categories. The parameters for the group of religious physicians are estimated with more uncertainty than those for the group of non-religious physicians, as illustrated by the higher standard errors.

Religious	No	Standard error	Yes	Standard error
Gender	0.016	0.283	0.084	0.715
Gestational age	0.885	0.157	1.243	0.437
Birth weight	0.829	0.174	1.654	0.492
Perinatal asphyxia	0.613	0.188	1.068	0.538
Congenital comorbidity	1.028	0.199	0.736	0.426
Progress since birth before a diagnosis of NEC	0.287	0.203	0.339	0.512
Age since birth	0.403	0.214	0.572	0.533
Growth since birth	0.197	0.187	1.022	0.595
Ultrasound of the brain	1.419	0.220	2.159	0.620
Lung function	0.010	0.191	0.622	0.446
Hemodynamic	0.345	0.187	0.538	0.439
Cerebral oxygenation	0.504	0.212	1.267	0.600
Wish of parents	1.166	0.219	1.920	0.593
The carrying capacity of parents	0.329	0.206	0.298	0.506
Constant	-8.740	1.329	-15.177	4.091

Figure Q.1: Estimated parameters for religious and non-religious physicians

Religious

LL= -41

Adjusted rho squared= 0.85

Not religious

LL = -205

Adjusted rho squared = 0.39

Appendix R Parenthood estimates

This appendix presents the estimates of the binary logit models for the UMCG physicians with and without children, as well as their model fit parameters. Figure R.1 illustrates the estimated parameters for the two categories. Due to the smaller sample size, the parameters for the physicians without children illustrate higher standard errors and, hence, are estimated with more uncertainty compared to the parameters for the group of physicians with children

Parenthood	**No**	**Standard error**	**Yes**	**Standard error**
Gender	-0.001	0.533	-0.025	0.304
Gestational age	0.883	0.290	0.948	0.171
Birth weight	0.834	0.312	1.005	0.189
Perinatal asphyxia	0.853	0.363	0.607	0.203
Congenital comorbidity	1.235	0.348	0.930	0.215
Progress since birth before a diagnosis of NEC	0.439	0.394	0.241	0.214
Age since birth	0.578	0.382	0.358	0.231
Growth since birth	0.742	0.367	0.097	0.205
Ultrasound of the brain	1.670	0.398	1.468	0.238
Lung function	-0.202	0.369	0.223	0.203
Hemodynamic	0.178	0.334	0.420	0.203
Cerebral oxygenation	0.677	0.400	0.564	0.228
Wish of parents	1.218	0.394	1.324	0.241
The carrying capacity of parents	0.339	0.371	0.378	0.223
Constant	-10.592	2.468	-9.342	1.456

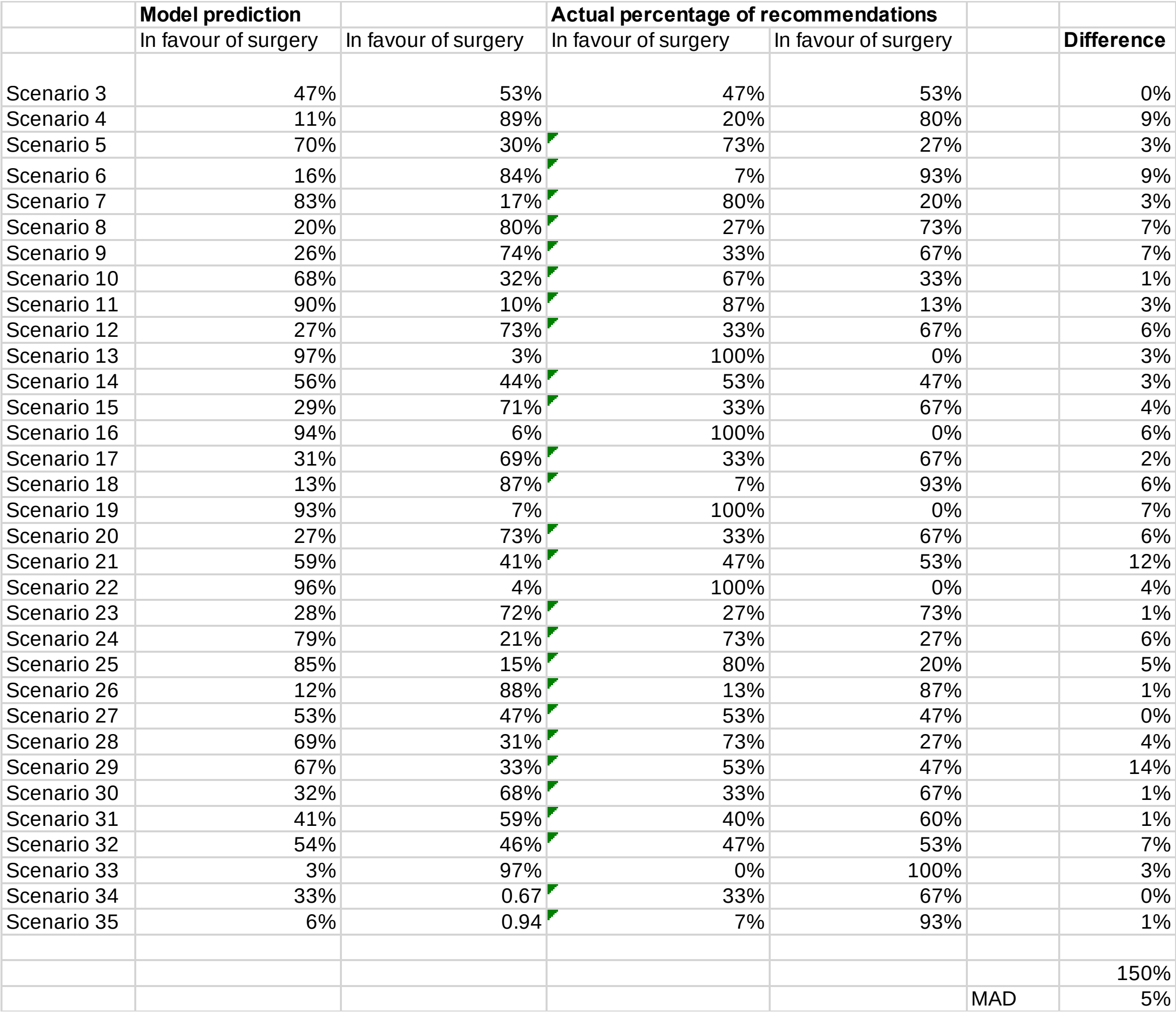
Figure R.1: Estimated parameters for the parenthood categories

No
LL = -65
Adjusted rho squared = 0.78

Yes
LL = -180
Adjusted rho squared = 0.46

Appendix S Mean absolute deviation calculation

This appendix presents the calculation of the mean absolute deviation. Figure S.1 provides the calculation. The MAD determines the average deviation in percentage points between the likelihood for a recommendation in favour of or against surgery and the actual distribution of recommendations per choice scenario. Hence, the MAD determines to what extent the model accurately predicts the distribution of recommendations in favour of or against surgery. The calculation neglects the first two choice scenarios as these were designed to constitute a definite yes and no for surgery among all physicians.

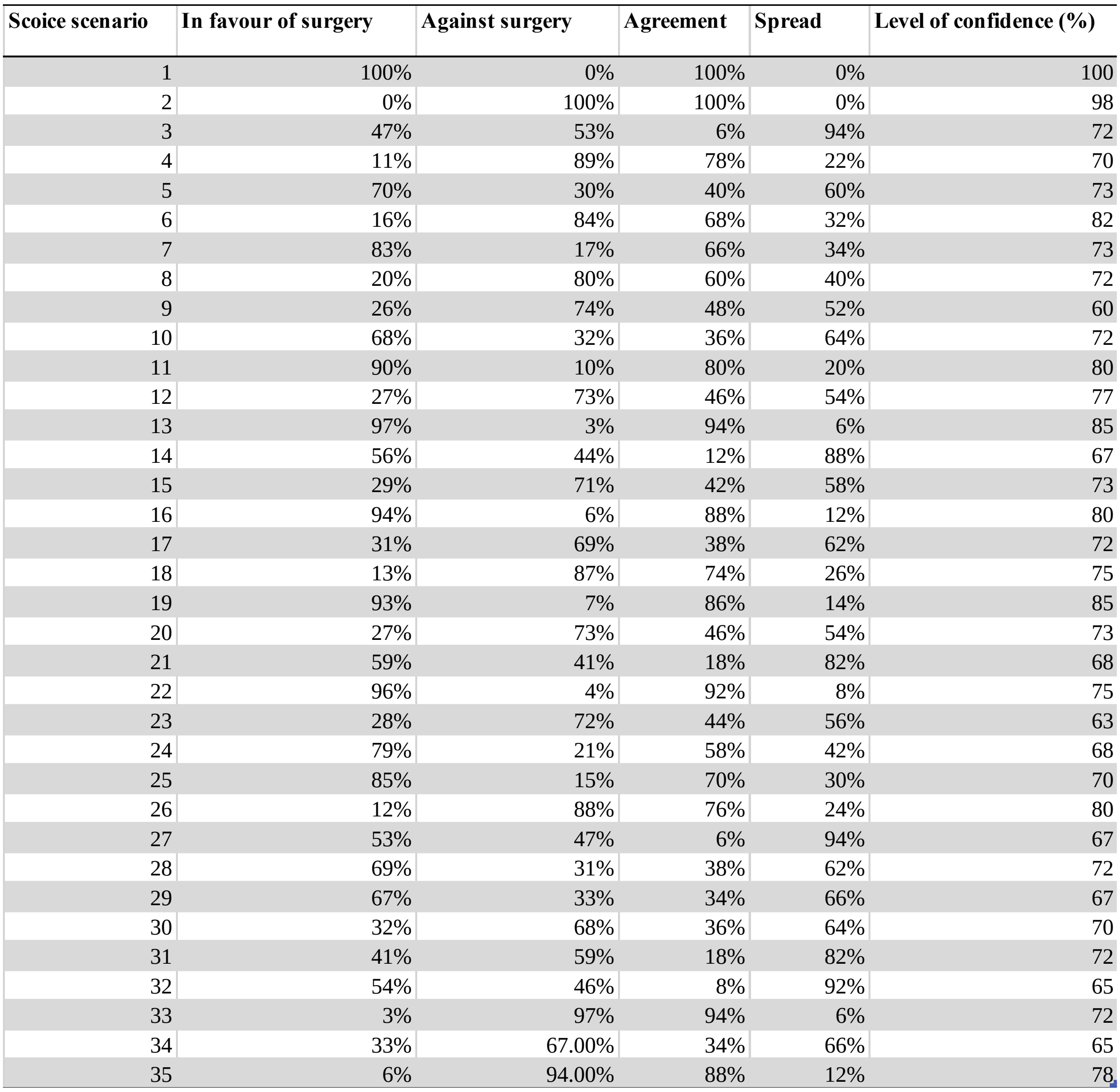


	Model prediction		Actual percentage of recommendations			
	In favour of surgery	In favour of surgery	In favour of surgery	In favour of surgery		Difference
Scenario 3	47%	53%	47%	53%		0%
Scenario 4	11%	89%	20%	80%		9%
Scenario 5	70%	30%	73%	27%		3%
Scenario 6	16%	84%	7%	93%		9%
Scenario 7	83%	17%	80%	20%		3%
Scenario 8	20%	80%	27%	73%		7%
Scenario 9	26%	74%	33%	67%		7%
Scenario 10	68%	32%	67%	33%		1%
Scenario 11	90%	10%	87%	13%		3%
Scenario 12	27%	73%	33%	67%		6%
Scenario 13	97%	3%	100%	0%		3%
Scenario 14	56%	44%	53%	47%		3%
Scenario 15	29%	71%	33%	67%		4%
Scenario 16	94%	6%	100%	0%		6%
Scenario 17	31%	69%	33%	67%		2%
Scenario 18	13%	87%	7%	93%		6%
Scenario 19	93%	7%	100%	0%		7%
Scenario 20	27%	73%	33%	67%		6%
Scenario 21	59%	41%	47%	53%		12%
Scenario 22	96%	4%	100%	0%		4%
Scenario 23	28%	72%	27%	73%		1%
Scenario 24	79%	21%	73%	27%		6%
Scenario 25	85%	15%	80%	20%		5%
Scenario 26	12%	88%	13%	87%		1%
Scenario 27	53%	47%	53%	47%		0%
Scenario 28	69%	31%	73%	27%		4%
Scenario 29	67%	33%	53%	47%		14%
Scenario 30	32%	68%	33%	67%		1%
Scenario 31	41%	59%	40%	60%		1%
Scenario 32	54%	46%	47%	53%		7%
Scenario 33	3%	97%	0%	100%		3%
Scenario 34	33%	0.67	33%	67%		0%
Scenario 35	6%	0.94	7%	93%		1%
						150%
					MAD	5%

Figure S.1: Calculation of Mean Absolute Deviation

Appendix T Spread and average confidence level calculations

This appendix presents the calculation of spread and average confidence level per choice scenario included in the choice experiment. Figure T.1 presents the calculations.



Scoice scenario	In favour of surgery	Against surgery	Agreement	Spread	Level of confidence (%)
1	100%	0%	100%	0%	100
2	0%	100%	100%	0%	98
3	47%	53%	6%	94%	72
4	11%	89%	78%	22%	70
5	70%	30%	40%	60%	73
6	16%	84%	68%	32%	82
7	83%	17%	66%	34%	73
8	20%	80%	60%	40%	72
9	26%	74%	48%	52%	60
10	68%	32%	36%	64%	72
11	90%	10%	80%	20%	80
12	27%	73%	46%	54%	77
13	97%	3%	94%	6%	85
14	56%	44%	12%	88%	67
15	29%	71%	42%	58%	73
16	94%	6%	88%	12%	80
17	31%	69%	38%	62%	72
18	13%	87%	74%	26%	75
19	93%	7%	86%	14%	85
20	27%	73%	46%	54%	73
21	59%	41%	18%	82%	68
22	96%	4%	92%	8%	75
23	28%	72%	44%	56%	63
24	79%	21%	58%	42%	68
25	85%	15%	70%	30%	70
26	12%	88%	76%	24%	80
27	53%	47%	6%	94%	67
28	69%	31%	38%	62%	72
29	67%	33%	34%	66%	67
30	32%	68%	36%	64%	70
31	41%	59%	18%	82%	72
32	54%	46%	8%	92%	65
33	3%	97%	94%	6%	72
34	33%	67.00%	34%	66%	65
35	6%	94.00%	88%	12%	78

Figure T.1: Spread and average confidence level calculations

Appendix U Discussion of the meeting that presented the model results of this study to the UMCG physicians

This Appendix provides a brief discussion on the final meeting that presented the results of this study to part of the group of UMCG physicians that executed the stated adaptation experiment.

The following bullet points illustrate comments and feedback provided by the UMCG physicians during the meeting:

- The presented choice behaviour matched the way the UMCG physicians expected to provide medical recommendations.
- The physicians were most surprised about the difference in choice behaviour between child surgeons and neonatologists.
- Some of the physicians stated that they would value and accept the aid of BAIT if BAIT would to be implemented to support future recommendations.
- The number of choice scenarios included in the choice experiment (35) was experienced as a lot. Some physicians announced that their recommendations at the last choice scenarios of the experiment might, therefore, be provided with less consideration compared to their medical advice on the first choice scenarios.

Overall the group of physicians present at the presentation valued the introspection on their own expertise. It must, however, be pronounced that not all UMCG physicians that executed the stated adaptation experiment were present at the meeting to discuss the results. Therefore, this discussion does not include their opinion on the added value of the introspection.

Moreover, during this meeting, the results triggered discussions among the experts. After the meeting, a few physicians declared that the discussions were valuable as the physicians started to reflect on their medical recommendations and choice behaviour critically.