Delft University of Technology

Master Thesis

The Value of System Dynamics for Healthcare Resource Modelling

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

Author:

Mark E. Lukacs - 5378559

Graduation committee:

Chairperson	-	Prof.dr. Cees van Beers, *Values, Technology and Innovation*
First Supervisor	-	Dr. Saba Hinrichs-Krapels, *Multi-Actor Systems*
Second Supervisor	-	Dr. Willem L. Auping, *Multi-Actor Systems*
External Advisors	-	Dr. Mart Stein and Berend Beishuizen, *RIVM - CIb - LCI*

10th December, 2022

To be defended in public on 19.12.2022

TU Delft Delft University of Technology

Acknowledgments

It's not the destination, it's the journey is one of the most cliché motivational phrases for a reason: The acquired knowledge, skills, and connections are often more important than the result itself. At least, this was definitely the case for this thesis. I gained a detailed understanding of how scientific research works. I gained more understanding of how qualitative and quantitative methods can be mixed in soft science, something I consider a very powerful mindset. On the other hand, I also gained practical skills, such as learning to type without looking at the keyboard or biking with hands in my pockets to guard against the cold (while the only thought keeping my spirit high is the promise of the coffee I will drink once I arrive).

Nevertheless, there are many people whom I want to thank for their support during this thesis. I would like to thank my first supervisor, **Saba Hinrichs-Krapels**, for her continuous support and availability, for being the first person to whom I could ask my questions, and for her infinite patience when it came to answering my many questions. Her advice and support is much appreciated. I would also like to thank my second supervisor **Willem L. Auping** for guiding me through the vast landscape of modelling and for pointing out relevant articles and information in the process. He tirelessly kept providing a forward-looking perspective and inspired me to do better. This research would not have been possible without the support from my external advisors from RIVM: **Mart Stein** and **Berend Beishuizen**, who offered me the opportunity to work within the PANDEM-2 innovation project, and for introducing me to RIVM and being my on-hands subject experts on the topic of epidemiology. I would also like to thank the people who participated in the workshop. Without them, this study (quite literally) could not have finished.

My parents **Mikolt Jelencsics** and **András Lukács** were constant promoters of my studies and provided me with a safe environment to learn. **Mónika Braun** was a constant pillar of support and motivation, and I cannot express enough gratitude for her. **Zoltán Boros**, **Miruna Beţianu**, and **Kelsey Franz** were always joining me whenever I wanted to relax, and take some time off from my thesis. The support of **Pranathi Srikrishna** and **Siraadj Salarbux** throughout my time as an MOT student. Finally, I would like to thank all my other friends and family who supported me in finishing my master's thesis.

Executive Summary

It is no secret to hospital and public health managers that resource shortages worsen pandemics. The importance of preparedness has long been recognized within the European Union. One of the current H2020 innovation projects in this domain is PANDEM-2, aiming to improve pandemic preparedness from the side of resource management and sharing by creating cutting-edge digital tools. As part of these tools, a system dynamics (SD) healthcare resource model is being developed, with the ultimate goal of embedding it in a dashboard accessible to pandemic managers. This is done in order to support managers in rapidly making evidence-based assessments and decisions, or as in this thesis shortened, to provide situational awareness[1]. In short, the specific problem we were tackling was the exploration of how can pandemic preparedness can be achieved via current healthcare resource models and how a specific resource model (developed by a previous intern) can be used. First, to gain a general understanding of the state-of-the-art models, we looked into the scientific literature from two directions: We looked at how existing resource models work, are validated, and are used via literature review. For another perspective, we looked at scientific frameworks describing modelling and validation to inform our methodology. Therefore, this thesis seeks to answer the question: *How to support healthcare resource managers in acquiring situational awareness via an SD model?* To gain a better understanding, we did a literature review first to understand how others approach the topic of healthcare resource modelling.

We first analyzed the existing scientific literature by a preliminary search, which was also used to construct a more detailed and refined second search. In this second search, we used the PubMed database to search for articles containing the keywords *hospital and healthcare resource*, *pandemic*, *model*, *validation*, and synonyms. Then the returned articles were screened for relevancy, resulting in a total of 25 healthcare resource models analyzed. Within these analyzed models (and articles), we found that the most common approach is using SD models, and the second most common approach is using regression models. Roughly two-thirds of the models fall into these two categories. Furthermore, we found that there is a stronger focus on hospital resources than public health resources and that no common approach is used for model validation. We also found that the articles demonstrating that the model is used to support real-life decision-making were usually not about SD models; therefore, examining how to use SD healthcare resource models for decision support is not mainstream. We also found that the model used in our research is novel in the sense that it encompasses resources on a more detailed level than existing published models.

To further our understanding, we decided to answer our research question by holding a workshop, where we examine how to communicate model outputs. While examining the relevant modelling methodological frameworks, we defined the tasks that need to be done in this thesis through the lens of the modelling cycle. We need to perform the tasks of verification, validation, and holding a workshop, which partly encompasses evaluation. Then examining the literature about verification and validation, we encountered the implication of a well-known philosophical problem of scientific theories' for modelling: It cannot be demonstrated whether the model (or the theory) is a truthful description of the phenomenon. To overcome this problem, in modelling, validation refers to building confidence that the model is fit for its purpose. In this study, the purpose of the model changed from describing the different mechanisms found important to generate semi-realistic outputs to be used in the workshop; therefore, it had to be revalidated. This was addressed by performing a particular set of relevant validation tests. The model passed verification and then the validation for this purpose, so we continued with the workshop. We decided that in the workshop, we would use a presentation to communicate intervention opportunities for the pandemic based on the model outputs. Then after each intervention, the participants were asked to evaluate the easiness of understanding the output and to talk about what actions the presented information inspire.

By holding the workshops, we found several relevant facts: First, it was found that the goal participants were searching for was to get rid of the perceived gap. This also meant they were searching for insights that could be used for operational planning purposes. Furthermore, the analysis does not need to stop at visualizing outputs. One of the participants indicated that further analyzing the graphs is not as easy for them as for an analyst working with the model. We have also seen that participants tend to augment the presented data with their experiences, which (unless explicitly presented) leads to assumptions about how the model works. Some participants also pointed out that the contact tracing part of the model is already outdated (in less than a year). We have identified some practical ways to avoid ambiguity while communicating about healthcare resource models. First, we found that despite the insights we gained by analyzing model outputs were not novel, the discovered

[1] While WP3 is also named situational awareness, in this thesis, we are not referring to the data aggregation analytics and visualization of epidemiological descriptors, but to the fact of being aware of how resources are affected by the pandemic

scenarios were still good discussion starters in the workshop. This is likely the mechanism of the scenarios acting as a reminder for passive knowledge, which participants subsequently shared. Furthermore, extra care should be taken to explain the context of how the data got generated, especially concerning the model. As the presented data left some space for interpretation, participants sometimes had different assumptions than the ones coded into the model. While these could be resolved in the workshop to some extent, this will not be the case for the dashboard. Given some familiarity with the audience, it is possible to expect some questions and misunderstandings, which could be proactively addressed in a description or in a 'frequently asked questions'. We also identified two presentation types that were easier to process than presenting key model outputs: The first option is to analyze key model outputs further than graphing and present the key insights (such as peak resource demand) in a tabular format. Alternatively, the second option is to build all visualization on the same template and explain that template on the first occurrence in detail. In subsequent occurrences, it should be enough to point out only the interesting parts and give participants time to process the information.

From another perspective, participants expressed a need for data that can be used for planning purposes. However, given the uncertainty about the system, these, as we call consolidative models, cannot be constructed yet. While exploratory modelling is an alternative SD technique for addressing deep uncertainty, it does not attempt to produce numerically accurate predictions. However, from a novel perspective, the consolidative and exploratory approaches can be viewed as two ranges on the spectrum of uncertainty about the modelled system. Viewed from this perspective, validation means reducing uncertainty about the system. Nevertheless, to achieve the consolidative models, datasets about resource usage are needed, but as far as we know, no such dataset exists. As data to create such datasets is probably already being collected for operational purposes, it is likely that the collection and aggregation of these data are not happening. However, creating such datasets comes with some challenges. There is a value trade-off between privacy and preparedness through data collection, and the current data collection techniques are unlikely to be unified. Overcoming these challenges would need quite a significant upfront investment. To answer our original question of how to support healthcare resource managers in acquiring situational awareness, this thesis argues that, by far, the biggest utility could be achieved by strengthening data collection and aggregation, as it enables the possibility to develop surrogate models. However, as this requires a significant upfront investment, question-driven exploratory models remain an alternative way to address these uncertainties.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The COVID-19 pandemic is the latest example that significant healthcare shortages can affect the mortality rate of a disease (Abdolhamid, Pishvaee, Aalikhani, & Parsanejad, n.d.; Olivieri, Palu, & Sebastiani, 2021). While the world kept a close eye on hospital and intensive care unit (ICU) occupancy, there have also been several reports of other resources having insufficient capacity, such as testing capacity or personal protective equipment (PPE) (Rijksoverheid, 2020; V&VN, 2020).

While a rushed production of these resources is possible in an emergency, it still takes time to re-organize (Vecchi, Cusumano, & Boyer, 2020). Moreover, while adapting the supply to the demand is relatively fast for simple resources like face masks, more complex production processes take more time to adapt. The slowest is the 'production' of hospital staffing, as it takes three and five years to train nurses and doctors. Here rushed production equals makeshift emergency solutions, like asking sufficiently trained people to volunteer, such as recently retired employees or ones being on unpaid leave (The Local, 2020), or such as asking for the help of health students (Operatív törzs, 2021). However, it is undisputed that the better solution is to have sufficient trained medical personnel. Therefore it is important to understand how a pandemic causes a surge in resource demand. This not only helps decision-makers by making a very-rough estimation of the size of the required resources, but it also enables them to recognize the early signals of it. The latter is especially important, as it gives more time to prepare, possibly avoiding the situation of insufficient supplies and the introduction of highly disruptive non-pharmaceutical interventions, such as lockdowns.

Within the European Union, the importance of improving pandemic preparedness has long been realized and formalized in the decision 1082/2013/EU (Council of the European Union, 2013). In agreement with this decision, several innovation projects are funded, one being the PANDEM-2 project. The project is funded by an H2020 grant and is organized as a consortium consisting of 21 partners (European commission, n.d.). This project "implements and demonstrates the most important novel concepts and IT systems to improve the capacity of European pandemic planning and response" (European Research Executive Agency, 2020, Annex 1, p. 3). An important detail is that despite the attention generated by the current COVID-19 pandemic, the project aims to develop cutting-edge solutions for the management and planning of all types of pandemics (e.g. influenza, Ebola) in accordance with the aforementioned 1082/2013/EU decision. Furthermore, as part of the whole EU approach, cross-border communication and resource sharing receive special attention (PANDEM-2, n.d.-a).

Within the PANDEM-2 consortium, as a specific part of Work Package Four (WP4), the Dutch National Institute for Public Health and the Environment (RIVM) took on the challenge to "assess the needs, feasibility and practicality of the software tools for national, regional and local stakeholders" In addition, RIVM will also aid in the development of a resource modelling system (PANDEM-2, n.d.-b). The consortium partner leading the development of WP4 and the resource modelling is the National University of Ireland, Galway (NUIG). Together with RIVM and other consortium partners, they are developing a system dynamics (SD) model to understand healthcare resource shortages better. The reason behind the SD paradigm was the previous experience gained from multiple similar projects, such as Stein et al. (2012); Yanez, Duggan, Hayes, Jilani, and Connolly (2017).

As the PANDEM-2 project started before this thesis, a version of the healthcare resource model has already been developed during a joint RIVM-TU Delft internship, with the goal of describing the health system from the perspective of COVID-19 treatment (de Schipper, 2022a). This model is referred to as *original model* in this thesis. Alongside this thesis, a modelling team from NUIG also utilized the original model, among other things, to further advance their model (*NUIG model*) for fulfilling parts of the WP4 deliverables: namely "Predictive

Pandemic Modelling"[1] and "Design and Implement Resource Planning System" (European Research Executive Agency, 2020, Annex 1, p. 24).

These two goals do not exist in a vacuum, as NUIG's model will be connected to the "common visual analytics service", which "will be instantiated in a suite of interactive data visualization and user interface components, within the WP3 Dashboard environment, for visual querying and interaction with pandemic data of high dimension and complexity to support rapid evidence-based assessment and decision making" (European Research Executive Agency, 2020, Annex 1, p. 24). In this thesis, the term *dahsboard* refers to the instantiation of this common visual analytics service. Similarly, this ability for rapid evidence-based assessment and decision making is shortened *situational awareness*[2]. Therefore this thesis was carried out, as part of an internship within RIVM, under WP4 of the PANDEM-2 project, with the aim of exploring what utility an SD model can provide.

From a scientific perspective, this is an interesting question because we perceive a lack of general knowledge about how to approach healthcare resource modelling via SD modelling, especially on the topics of model validity and evaluation. While there are well-established frameworks, these are generic to SD; therefore, it is interesting to see how these can be applied more specifically in the uncertain, multi-disciplinary environment that characterizes healthcare resource modelling during a pandemic. The uncertainty originates from the fact that validation via controlled experiments is not even worth considering and from the lack of formalized datasets (for many auxiliary but relevant resources, such as demand for public testing). Multi-disciplinarity originates from the different practices conducted by people working in public health and a hospital setting: Public health professionals tend to use more scientific literature in their day-to-day work than those working in a hospital setting.

There is a third interesting perspective: examining the role of modelling with respect to knowledge management. A modelling process always condenses the knowledge of the model builder (Bolt, Bayer, Kapsali, & Brailsford, 2021); therefore, modelling can be viewed as a way to formalize tacit knowledge about a real-world system. However, as the model will be enacted by people from multiple disciplines (i.e., the public health and hospital side), it can also be viewed as a means to achieve cross-disciplinary communication. This latter role classifies the model as a *boundary object* (Newell, Morton, Marabelli, & Galliers, 2019). In this sense, modelling can also be viewed as a technology, as it is an application of scientific knowledge for practical purposes. In this thesis, we investigated the directions a future technology (i.e., healthcare resource modelling with a dashboard) could take in an attempt to answer which modelling technologies we need and when.

Despite ultimately, the dashboard will utilize the model developed by NUIG, RIVM expressed their interest in exploring the whether the behaviour of the original model is realistic and how the results of such a model should be presented on the dashboard. Therefore this thesis addresses the research question:

How to support healthcare resource managers in acquiring situational awareness via an SD model?

The term *healthcare resource* should be understood as a reference to the two resource sides: hospital resources and public health resources. *An SD model* refers to the original model, where the healthcare resources are represented by two sub-models, plus a compartmental sub-model, to determine the speed of the pathogen spread in the population. We also defined two sub-research questions to support answering the main research question:

- *How can healthcare resource models be validated?* - Addressed by conducting a literature review on model validation, to explore how other researchers tackle the problem of validation. Given the method, this sub-research question also has the implicit aim of understanding which resources via what techniques they model.
- *How are, or can healthcare resource models be used?* - Addressed by continuous consultation with subject-matter experts and by conducting workshops to give recommendations on how to communicate the model outputs to the intended audience of the dashboard.

This thesis is structured in the following way: First, to gain a general understanding of healthcare resource modelling, we looked into the scientific literature from two directions: In Resource Modelling in Healthcare Literature (chapter 2), we look at how published resource models work, validated, and used, via two literature

[1] In retrospective, predictive pandemic modelling is not the best name, given the division between consolidative and exploratory modelling.

[2] WP3 of the PANDEM-2 project also named situational awareness, however, in this thesis, we deviate from this. We are not referring to the data aggregation, analytics, and visualization of epidemiological descriptors. Instead, we are referring to the awareness of a pandemic's effect on the different resources.

reviews. Then, for gaining a deeper level of understanding, another perspective is presented in Methods (chapter 3) via the scientific frameworks describing modelling and validation, which in turn informs our methodology for validating the model. The methodology for the workshop is also presented in this section. In Results of the Workshop (chapter 4), the findings of the workshop are described. Then, in Discussion and Recommendation (chapter 5), the implications of the results are described, and the research questions are addressed. Finally, in Conclusion (chapter 6), the work of this thesis is summarized, and the study is concluded.

Chapter 2

Resource Modelling in Healthcare Literature

To address the first sub-research question, namely *How can healthcare resource models be validated?*, we decided to conduct a literature review in model validation. First, a Preliminary Literature Review (section 2.1) was conducted to verify whether our insights about the scientific literature hold true and to gain a general idea of which dimensions the literature is worth analyzing along. After that, another, more specific literature review was done, specifically to explore how other healthcare resource models tackle the question of validation (section 2.2).

2.1 Preliminary Literature Review

The preliminary literature review was done in two steps. First, in Search Design, the appropriate keywords were selected and justified, then in Analysis of Results, the returned articles are analyzed in detail.

2.1.1 Search Design

For this search, the database selected was the Web of Science core collection (Clarivate, n.d.). The keywords were the following :

- *system dynamics* - The justification for this keyword is that the model which will be validated is an SD model.
- *pandemic* or *epidemic* or *avian influenza* - the keyword pandemic was included because this thesis researches resource usage related to pandemics, opposed to non-contagious diseases such as cardiovascular problems. The epidemic keyword was included, as this term is sometimes used interchangeably with pandemic. The keyword avian influenza was also included, as it was hypothesized that the related data quality might be better (B. Beishuizen, personal communication, 15th March, 2022).
- *hospital* - The decision to include this keyword separately was an unnecessary precaution in hindsight. The search algorithm of the Web of Science platform does not treat spaces as literals. If it would, then searching for *hospital resources* does not find the expressions like "resources of hospitals". However, this was only found out after performing the search.
- *resources* or *capacity* - Since the search should include healthcare resource models, including *hospital resources* was a trivial choice. As a synonym, *hopsital capacity* was included too.

The search fields were set to Title, Abstract, and Keywords, including non-author generated keywords. The non-English articles were excluded, and the remaining 88 results were analyzed. After an abstract screening, the following 23 articles seemed relevant to the research question of this thesis: (Abdolhamid et al., n.d.; Cakan, 2020; Cui, Qiu, Liu, & Hu, 2017; Ejigu et al., 2021; Garcia-Vicuna, Esparza, & Mallor, 2022; Ibarra-Vega, 2020; Joulaei, Honarvar, Zamiri, Moghadami, & Lankarani, 2010; Keeling et al., 2021; Liu, Cao, Liang, & Chen, 2020; Liu & Zhang, 2016; Mu, Wei, & Yang, 2019; Mu & Yang, 2018; Mugisha, Ssebuliba, Nakakawa, Kikawa, & Ssematimba, 2021; Pei, Yuan, Yu, & Li, n.d.; Pierce et al., 2021; Rocha et al., 2021; Tembine, 2020; Verma, Saini, Gandhi, Dash, & Koya, 2020; Vierlboeck, Nilchiani, & Edwards, 2020; A. Wang, Xiao, & Zhu, 2018; Weissman et al., 2020; Wood, McWilliams, Thomas, Bourdeaux, & Vasilakis, 2020; Zhao, Li, Wang, & Jiang, 2021).

2.1.2 Analysis of Results

Within this set of articles, 16 had a healthcare resource model, which included hospital resources. These models are summarized on Table 2.1. First, the types of modelled hospital resources were identified. Then it was examined whether the lack of resources is fed back to the model (e.g. by modifying the disease mortality). Finally, the type of the underlying epidemiologic model was identified. Here, compartmental models are models where people may progress between compartments, such as susceptible or infected. When these were based on differential equations, they received special attention due to (generally) being SD models. Lastly, the 'Disease type' column of Table 2.1 indicates which disease was modelled (if left empty, the authors did not specify).

Article	Generic hospital resource	Hospital beds	ICUs	Venti-lators	Testing capacity	Resource depletion feedback (yes/no)	Transmission model	Disease type
(Weissman et al., 2020)		x	x	x		no	SIR[1]	COVID-19
(Ibarra-Vega, 2020)	x					yes	SIRD	COVID-19
(Wood et al., 2020)			x			yes	queue model	
(Keeling et al., 2021)		x	x			no	SEIR	COVID-19
(Verma et al., 2020)		x	x	x	x	no	SEIR	COVID-19
(Mu & Yang, 2018)		x				no	SI + SEIR[2]	H7N9
(Abdolhamid et al., n.d.)		x	x			yes	SEIR	COVID-19
(Pei et al., n.d.)		x			x	no	queue model[3]	COVID-19
(Liu & Zhang, 2016)	x					yes	SEIR	COVID-19
(Cui et al., 2017)		x				yes	SIR	
(Mu et al., 2019)		x				yes	SI + SIR	H7N9
(Mugisha et al., 2021)		x				no	SEIHR(S)	COVID-19
(Garcia-Vicuna et al., 2022)		x	x			no	flow process[4]	COVID-19
(Ejigu et al., 2021)			x			no	SEIHR	COVID-19
(Pierce et al., 2021)		x				no	SEIHR(D)	COVID-19
(A. Wang et al., 2018)	x					yes	SIS	

Table 2.1: Overview of the preliminary literature review.

Generally speaking, there are two types of models: Ones that include a resource feedback loop and those that do not. The latter type is usually used to research questions like *When is resource depletion reached?*, or *Under what (policy) parameters are the hospital resources 'just enough'?*. The models which included resource feedback usually included hospital beds and ICU admissions. Some models use a 'generic hospital resource'. This is a mathematical construct which assumes that all hospital resources can be represented with a single variable, which is indifferently supplied to all hospitalized individuals. While the models without feedback tend to model more resources, those are also very focused on bed and ICU occupancy.

Most of the models are compartmental models, which are based on differential equations. This is not surprising, considering that one of the search keywords is *system dynamics*, which is also based on differential equations.

It is worth noting that nearly all of these models assume that infected individuals cannot get infected again. This is visible in the 'transmission model' column, as the last compartment (usually R or D) is not followed by an S compartment again. This is a clever simplification when the modelled time frame is relatively short compared to the time it takes immunity to be lost. However, in case of a longer time frame (e.g. years), several factors make the population susceptible again. For example, pathogen mutations occur, immunity naturally decreases in individuals, or over a decade, a significant percentage of the population consists of newborns. The model inherited by this thesis work has a SEIRS-like structure: (susceptible - exposed - infected - recovered/dead - then again: susceptible) (de Schipper, 2022b), clearly distinguishing it from these models.

[1] Where a compartmental model was used, the following letter code was used based on the existence of compartments: S - susceptible, E - exposed, I - infected, H - hospitalized, R - recovered/removed, D - dead.

[2] H7N9 is a zoonotic disease (avian influenza), the separate SI compartments model the disease spread in the animal population, and the SEIR compartments model the disease spread in the human population.

[3] The model utilizes the mathematical queueing theory, where patients arrive probabilistically to the hospital, based on a specified (time-dependent) distribution

[4] This model assumes that patients move between different hospital care types under a constant (average) time, and the pathway selection is defined probabilistically.

It should not come as a surprise that although the search was not explicitly aimed at COVID-19, most articles are about it. This is probably the result of the enormous research effort used to tackle the latest pandemic. However, the underlying transmission models with suitable parameterization could simulate other infectious diseases too, like influenza.

The validation of the model was tackled differently in each article, but the general approaches were the following:

- Explicitly writing down assumptions.
- Tackle uncertainty by sampling some parameters from a distribution, then display confidence intervals in the model predictions.
- Compare model predictions with real-life data, usually in a quantified manner, using mean-squared error.

Many of the articles included scenarios and scenario analysis. Most articles focused on the effect of different government policies, such as social distancing or the use of face masks. These scenarios modified the parameters but not the structure of the model. However, the approaches to analyzing these scenarios were not uniform. For example Ejigu et al. (2021) sampled the dimensions of policy space and conducted a full factorial experiment. In contrast, Keeling et al. (2021) were probably limited by computational power due to stochastic elements in their model and only explored a few scenarios.

No cases were found where the scenarios were used to validate the model. This was probably not documented in the articles if it was done this way. However, the lack of hard data for scenario-based validation suggests that these models' validation relied on qualitative approaches.

To conclude the findings of the preliminary research, quite some models utilize a resource feedback loop to simulate the depletion of resources, but these tend to oversimplify the resources in favour of accuracy. Furthermore, in past models, the population cannot get susceptible again. Lastly, there is no unified way to validate healthcare resource models. This last point is the research gap this thesis aims to address. Hopefully, a direction towards a more comprehensive validation approach can be found as a small step towards a universally applicable exploratory model validation theory.

2.2 Literature Review on Validation

While the results of the preliminary literature review were promising, we decided to conduct a more detailed literature review with the scope explicitly set on healthcare resource model validation. The literature review described in this section builds on the lessons learned from the preliminary literature review. The description of this literature review also follows the structure of the preliminary literature review: first, the Search Design is described, followed by the Analysis of Results.

2.2.1 Search Design

This time, PubMed was chosen as the search base. This platform is a bibliographic database of life sciences and biomedical information (Canese & Weis, 2013). These fields are very relevant for pandemics; therefore, with healthcare resource modelling, PubMed is a more suitable database to search on.

The first step of a literature review is to identify the keywords to be searched. While a set of relevant keywords were identified during the Preliminary Literature Review, we decided to improve by incorporating the newly acquired knowledge about the field in the following way:

- The keyword *system dynamics* was changed to *model*. The rationale behind this is that some of the relevant articles during the preliminary literature search were only found because their abstract incorporated the terms *system* and *dynamics* separately. A more general term: *model* can also find these articles while correcting this problem.
- Despite *hospital resources*, the model consists of *healthcare resources* too, such as PPE stock, contact tracing capacity, and vaccine stockpile. The keyword *healthcare* was added in a way that it creates a Cartesian product with the keyword *resource*. This Cartesian product approach was followed to include a greater set of synonyms (via Automatic Term Mapping).
- From a resource modelling perspective, all highly infectious respiratory diseases are similar enough to justify the addition of *influenza*, *SARS*, and *covid* beside the existing *pandemic* and *epidemic* keywords.
- To perform the actual scope change, the term *validation* was added.

For the first try, all terms were searched in the [Title/Abstract] fields. It means that the citation's title, abstract, and author-defined keywords were searched (National Library of Medicine, 2022b). Unfortunately, this search resulted in an unexpectedly high number of articles. Upon investigation, it was found that PubMed's Automatic Term Mapping included some very broad categories. In brief, the automatic term mapping is a core algorithm to the search, which ensures that users do not need to be 100% accurate and inclusive with their keywords (National Library of Medicine, 2022b). For example, searching for *covid* automatically includes *coronavirus*, *sars-cov-2*, and many other commonly used terms of the phenomenon. While automatic term mapping was generally helpful, it was counter-productive in some cases. For example, it included the keyword *Health care resources* (as a MeSH word), which in turn includes distributional activities, which is outside the scope of this study (National Library of Medicine, 2022a). Therefore a very manual approach was chosen, which enabled the exclusion or replacement of automatic term mapping, thereby granting finer control over the search. The final search string, including the exclusion and replacement of automatic term mapping, was:

> ("epidemiological models"[MeSH Terms] OR "model*"[Title/Abstract]) AND "validat*"[Title/Abstract] AND (pandemic*[tiab] OR epidemic*[tiab] OR covid19[tiab] OR covid-19[tiab] OR influenza OR SARS) AND ("health-care"[Title/Abstract] OR "health-care"[Title/Abstract] OR "healthcare"[Title/Abstract] OR "hospital"[tiab] OR "hospitals"[tiab]) AND (resource[tiab] OR capacity)

This resulted in a processable amount of 131 articles at the time of the search. The resulting articles were screened based on their title and abstract, with the criteria that the article needs to describe a model containing the expected number of sick people in a pandemic. In the next step, the selected articles were read, and the decision was made to exclude an additional six articles. An article was identified during project planning which was not returned by the literature search despite its relevance: (Stein et al., 2012). This article was added manually. Then, the underlying model(s) were analyzed for each article. Several articles included more than one model; however, in the case of (Araz, Bentley, & Muelleman, 2014; Smith et al., 2021), it was decided that these models differ to the extent that these should be analyzed separately. This selection process is described in detail on Figure 2.1.

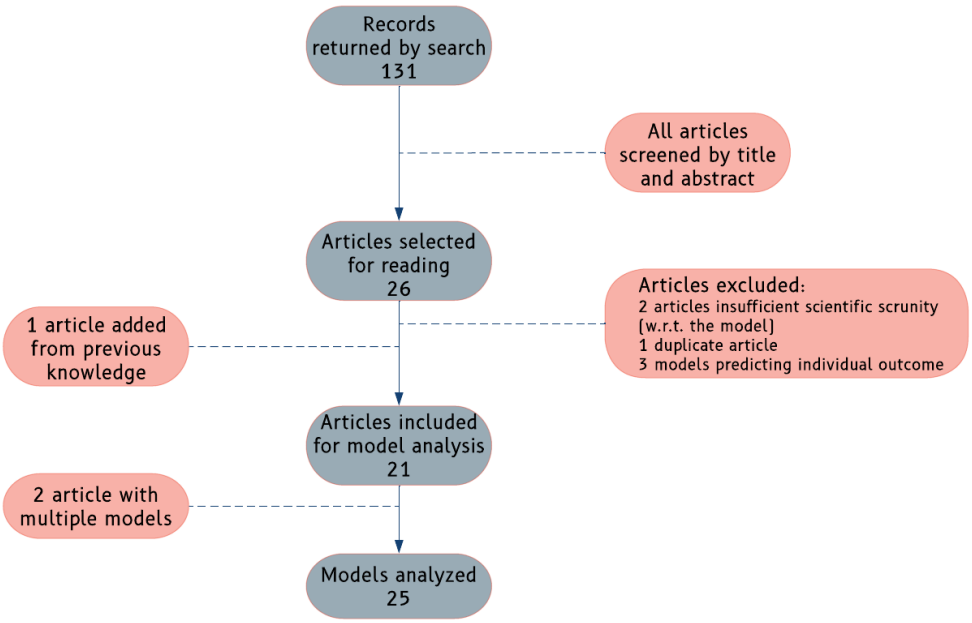


Figure 2.1: Overview of the model selection process

2.2.2 Analysis of Results

The 25 models returned by the search were analyzed, and the results can be seen Table 2.3. The analysis roughly followed three questions: What is the underlying mathematical model? (Model Columns); Which resources were modelled? (Resources Columns); How the validation process is approached? (Validation Columns). A more detailed explanation of these will follow:

Model Columns

As we did not only search for SD models, it is essential to identify the underlying mathematical model. There is not only a huge difference between a white-box, and black-box modelling approach (explained below), a connection between model type and other variables might uncover interesting relationships. Therefore these model columns give an overview of the mathematical model used in the articles. The first one is *Type*, and the second one is *Properties*, which consists of 3 sub-columns.

Column *Type*:
Given that the topic is epidemiology, most models are compartmental models. These were given a code according to which compartments are defined, based on the rules described in Table 2.2. As the primary form of a compartmental model consists of differential equations (Brauer, 2008), and the SD paradigm also utilizes differential equations heavily, I did not distinguish between the two categories. As most models were in this category, it is not indicated separately in this column. However, some less widespread approaches use a radically different model structure. If the model is one of these types, it is indicated in the table based on the following acronyms:

- ABM - Agent-Based Model
- ARIMA - AutoRegressive Integrated Moving Average
- VECM - Vector Error Correction Model
- m-IDEA - modified Incidence Decay and Exponential Adjustment
- GRM - generalized Richard's model
- Spatial - By default, the SD models assume the perfect mixing of people. This well-known limitation was tried to overcome in some cases by introducing some degree of spatial differentiation.
- Multi-model - the article covered multiple models. These were separated in the subsequent rows in the table.

Assigned letter	Compartment it denotes
S	Suspectible
E	Exposed
I	Infected
R	Recovered (or Removed)
D	Dead (when explicitly mentioned beside recovered)
Q	Quarantined
H	Hospitalized
V	Vaccinated

Table 2.2: Compartment abbreviations of 'Model type' column in Table 2.3

Column *Properties*:
This column is split into three sub-columns, each describing a specific property of the model. The first column describes explicitly if the model is SD-like ('sd') or another type ('o'). The second column describes explicitly if the model is a compartmental model ('c'), an autoregression model ('r'), or has another structure ('o'). This is interesting because there are fundamental differences between white- and black-box models. The difference between the two is that while the white-box models' internal structure is understood, the black-box models' structure is not. In this categorization, SD falls into the white box model category, as the modeller should understand the equations he put into the SD model. Therefore, the internal variables all should have a concrete meaning, as opposed to most of the machine learning approaches, such as autoregression models (TU Delft, 2020). Finally, the third column indicates if the model separates its population into age groups ('a'), if there is spatial differentiation ('s'), or if the spatial differentiation is only made for the resources ('sr'). This column is left empty if neither applies.

Resources Columns

One of the aims of the first sub-research question is to understand which resources are modelled. These columns are aimed at uncovering this, and resources are categorized either as hospital or as public health resources.

Column *Hospital*:
Similarly to the compartments, a code was constructed for each model. There were some models which used many resources. These are not indicated separately. The following abbreviations were used to create the code:

- U - 'unified hospital resource'. This is an abstraction of real-world resources. It can be translated to a real-world resource by finding the resource that prohibits providing more care in the hospital (bottleneck resource). While this approach makes the model simple, it also causes quite significant limitations.
- H - hospital census. This means how many patients are being treated in the hospital. Ward bed occupancy is another synonym of this concept.
- Icu - ICU occupancy.
- Ve - Number of ventilators in use.
- Edv - Emergency Department Visits.

Column *Public health*:
Similarly to the previous column, the following abbreviations were used to create the code:

- T - Testing. This testing could happen either as public testing, in a test on hospital admission, or a hospital exit situation.
- Va - Vaccinations
- C - Contact tracing

Validation Columns

Similarly to the previous columns, we are interested in how other researchers approach validation. We partially achieve this by looking at how these models were validated, hence were analyzed along the next columns:
Column *Calibration*:
This column indicates if the authors calibrated the model parameters or not. It is usually indicated by the presence of keywords 'fit' or 'fitting'. Since parameter-fitting is an essential step for the autoregression models, it is indicated with 'yes*'.

Column *Validation method*:
This column indicates which direction the authors chose to validate the model. An analytical (mathematical) approach is denoted by 'm'. This includes both the approaches of computing the model's various accuracy scores (e.g. 'root mean square error') and visualizing the results plus examining the graphs. If any other direction was chosen, it is denoted by 'o'. When there is a lack of discussion of model validation, it is denoted by 'N/A'.

Column *Validation by accuracy*:
This column indicates whether the authors suggested/implied that model accuracy justifies the usage of the model.

Column *Used in real life*:
The last column indicates whether the article demonstrated evidence of actual decision support for hospitals or public health agencies. This support includes both pandemic preparedness and response. Since most articles did not explicitly mention this, it is indicated with an asterisk ('*') if the categorization is disputable.

Insights

Now that we have a good overview of all the models, it is time to draw some conclusions from them. First of all: most models are about COVID-19. This is unsurprising, given the global effort and attention given to the latest pandemic. However, it is interesting that while we did not filter the results by publication date, almost all models were published in the years 2020-2022.

From the Model columns, two groups stand out. Roughly half of the models utilize an SD-like approach (44%), and a quarter of the models are autoregression models (24%). The rest utilizes other, simpler approaches except for the ABM models (32%).

From the Resources columns, it is visible that most models consider the number of hospitalized people an important resource, therefore, included in the model. The second most modelled resource is ICU occupancy, clearly stating where the priorities lie. The rest of the resources are modelled very sporadically. Furthermore, there is a stronger focus on hospital resources than healthcare resources. Here we can see that the model we are using is novel, as it encompasses resources on a much more detailed level, especially by including testing and contact tracing capacity and the related compartments (e.g., isolated) in the compartmental model.

While it is not inferrable from the Validation columns alone, after reading the articles, it was clear that each article tackled the validation question very differently. Furthermore, a significant portion of the articles does not explicitly talk about validation or state that it depends on things not in the control of the authors (e.g. data quality in subsequent use). The articles which tackle the question of validity mainly employ a mathematical approach. The outputs are usually graphed and compared against historical data, or an accuracy score is defined (analytical approach). A minority of article suggests that model accuracy is some form of validation. However, this 'validity by accuracy' is a questionable approach when the model was calibrated. While over-fitting is primarily a problem of black-box models, calibration blurs the line between white- and black-box approaches; hence it is possible to over-fit these models too. Contrary to this, none of the included articles addressed whether the perceived accuracy results from correct generalization or from over-fitting. Interestingly, only a minority of the models were used to support actual decision-making, and these are usually not SD-like models. Therefore it is also a somewhat novel direction in this study to examine how SD models can be used for decision support.

Article	Disease	Model				Resources		Cali-	Validation	Validation	Used in
		Type	Properties			Hospital	Public health	bration	method	by accuracy	real life
(Sarkar, Pramanik, Maiti, & Reniers, 2021)	COVID-19	SIR Q	sd	c	sr	U		no	o		no
(Nguyen, Turk, & McWilliams, 2021)	COVID-19	VECM	o	r		H		yes*	m		yes
(Kuzdeuov et al., 2020)	COVID-19	Spatial SEIR	sd	c	s	Icu		yes	N/A	yes	no
(Roy, Dutta, & Ghosh, 2021)	COVID-19	Spatial SEIRD	sd	c	s	H		no	o		no
(Pierce et al., 2020)	COVID-19	SEIR[5]	o	c	a	H Icu Ve		yes	N/A		yes
(Watson et al., 2021)	COVID-19	SEIR	sd	c	a	H Icu		yes	N/A		no
(Picchiotti, Salvioli, Zanardini, & Missale, 2020)	COVID-19	SEIR	sd	c		none	T	yes	m	yes	no
(Tran et al., 2021)	COVID-19	SEIRV	sd	c		H Icu Ve	Va	yes	o[6]		no
(Abramovich et al., 2017)	influenza	underlying model un-accessible				7 in total		no	o[7]	yes	yes*
(X. Wang et al., 2021)	COVID-19	SEIR QHC	sd	c		H	C	yes	N/A		no
(Galbraith, Li, Rio-Vilas, & Convertino, 2022) [8]	COVID-19	ARIMA	o	r	sr	H		yes*	o		no
(Abdin et al., 2021)	COVID-19	SIR QH	sd	c	s	H U	T	yes	m	yes	no
(Campillo-Funollet et al., 2021)	COVID-19	SEIRH	sd	c		H		yes	m	yes	no*[9]
(Yin et al., 2021)	COVID-19	ABM SEIR Q	o	c	s+a	H		yes	m	yes	no
(Berta, Paruolo, Verzillo, & Lovaglio, 2020)	COVID-19	VECM	o	r		H Icu		yes*	m	yes	no*
(Smith et al., 2021)		multi-model									
	COVID-19	logistic growth model	o	o		none[10]		yes	m		yes*
	COVID-19	m-IDEA	o	o		none		yes	m		yes*
	COVID-19	GRM	o	o		none		yes	m		yes*
(Kamerlin & Kasson, 2020)	COVID-19	ABM	o	o	s+a	Icu		no	N/A	yes	no
(Du, Fox, Ingle, Pignone, & Meyers, 2022)	covid + influenza	SEIR H V	sd	c	a	H Icu	Va[11]	yes	N/A		no
(Earnest, Chen, Ng, & Sin, 2005)	SARS	ARIMA	o	r		H		yes*	m	yes	yes*
(Araz et al., 2014)[12]		multi-model									
	ILI[13]	Seasonal ARIMA	o	r		Edv		yes*	N/A[14]		no
	ILI	Holt Winters methods	o	r		Edv		N/A[15]	N/A		no
	ILI	linear regression[16]	o	o		Edv		yes*	N/A		no
(Stein et al., 2012)	influenza	SEIRD H	sd	c		28 resrouces in total		no[17]	N/A		yes

Table 2.3: Overview of the literature review

[5] This compartment model uses queueing theory instead of SD.
[6] The original article references Wikle et al. (2022), who empirically validated the model.
[7] Delegates the problem to the data acquisition quality.
[8] This article focuses more on social media monitoring than modelling. It examines 4 separate model but these are similar from this literature review's perspective, and treated as one.
[9] By the looks of the electronic supplementary material (website) I think there was an attempt at starting a spin-off company, but I found no signs of actual customers.
[10] This article was a false hit in the sense that it didn't model any resources, only the number of cases. Nevertheless I found it relevant, so I included it.
[11] In this model vaccinations are only available for influenza.
[12] This article focuses less on the model, and more on the reliability of the data used in the model.
[13] ILI abbreviates influenza-like-illness
[14] The article claims to validate google flu trends as a data source, and delegates the validation by asking for statistical validation from each model individually.
[15] I could not access a the full-text paper which describes the method in details, but based on common practice it was probably calibrated
[16] There are 3 separate linear regression models, but these are structurally similar enough to be treated as the same model
[17] (M. Stein, personal communication, 24th August, 2022)

2.3 Summary

For each literature review, the search design was first explained, including the identification of the keywords. After this, the results are presented, where several key insights have been realized. We have seen that COVID-19 dominates the healthcare resource modelling landscape compared to other respiratory diseases. Furthermore, roughly half of these models are SD (or similar) models. We have seen that while these model resource usage during a pandemic outbreak, most only include a few resources, and only a handful consider public healthcare resources. It is also visible that there is no single approach to the validation of these models, and the validation processes are barely reported in these articles.

Chapter 3

Methods

The literature informing our methods consist of several topics. First, a high-level description of the practical tasks is presented and explained through the lens of the Modeling Cycle (section 3.1). Then, to address the first research question (How can healthcare resource models be validated?), we explore the topic of Verification and Validation (section 3.2) and validate the model for the purpose of producing scenarios for the workshop. After this, we examine the literature related to the Workshop (section 3.3) and define the interventions presented in the workshop. Given that this methodology was created by merging ideas from many topics, an overview can be found in section A.1 to aid understanding.

3.1 Modeling Cycle

To understand where this thesis started, first, a modelling cycle will be used as a set of lenses to describe the prior and planned work in scientific terms.

3.1.1 Description of the Modeling Cycle

Creating simulation models is generally considered to be part of a *modelling cycle*. As modelling is a powerful method, and it is widely used across many disciplines (Bankes, 1993), various versions of the modelling cycle exist, emphasizing different parts of the modelling process. There is no 'cycle to rule them all'; therefore, this thesis presents a rather generic white-box modelling cycle (van der Wal & Nikolic, 2022) instead of an SD-specific one (Auping, d'Hont, van Daalen, Pruyt, & Thissen, 2022), as the planned work can be understood more clearly via this one. This modelling cycle consists of 7 steps:

- **Conceptualization** - In this step, the modeller(s) identify the research questions or make a problem formulation, then identify the real-world phenomenon to be modelled and the mechanism associated with this phenomenon. They also determine the model boundaries (i.e., what to model and what not to model) and the primary model outcomes (i.e., which outputs of the model to investigate). More detailed explanations of this step exist, such as the XLRM framework (Lempert, Popper, & Bankes, 2003); however, these fall outside the focus of this thesis. In SD, this step typically includes the activities up until the creation of a causal loop diagram.

- **Formalization** - In this step, the modeller translates the conceptual model into a rule-based form. While in practice, almost all of the rules take on the form of a $\mathbb{R}^n \rightarrow \mathbb{R}^m$ mathematical function augmented with boolean logic, it is also possible to use other technologies or languages to describe the exact rules. The importance lies in the fact that the formalized model leaves zero ambiguity about how the model should work. Describing these rules via mathematics is just a choice of convenience, as it is widely used yet concise language.

- **Implementation** - In this step, the modellers implement the formal model in their chosen modelling program. This step typically ends with a model which is capable of producing outputs. In SD, it is common practice to put another layer of abstraction over the formal model and to use a stock-flow diagram instead of writing down the differential equations. It is also worth noting that the boundary between formalization and implementation is quite fuzzy in SD, to the extent that Auping et al. (2022) does not consider these separate steps. For example, gathering the exact parameter values of the stock-flow diagram could be done in either the formalization or the implementation step.

- **Verification** - In this step, the modeller checks whether the implementation of the model is in line with the conceptual and formal description of the model. Usually, due to the sheer size of the SD models, a few mistakes happen during implementation, which alter model behaviour. Various methods exist to verify an SD model, ranging from peer-checking to unit testing. However, as this step is concerned with checking whether any error was made during formalization and implementation, conceptual errors remain unnoticed here.
- **Validation** - In this step, the modeller checks whether the implementation of the model is in line with the modelled real-world phenomenon. This step is aimed explicitly at noticing conceptual errors. While it is possible, validating an unverified model makes less sense, as the errors discovered during validation could also be implementation errors. Despite the easiness of describing the output of the validation step, it is surprisingly hard to perform this step. Therefore, this will be further addressed in section 3.2.
- **Simulation** - In this step, the modeller runs the model and saves the results. For SD, this is a tiny step, as performing a single model run is usually under a few seconds.
- **Evaluation** - In this step, the simulation results are evaluated to answer the research questions. In a simple case, this means examining the plots of key variables, though in more complex cases, any post-processing of the results also falls under this step. Nevertheless, the final output of this step should be interpretable to the problem owners, which usually means translating the mathematics back into English (or another) language. Also note that under evaluation in this modelling cycle, we mean a very different thing than what Auping et al. (2022) mean in their (SD specific) modelling cycle: In this cycle, evaluation means evaluating the results with respect to the research question, while in their cycle it means verification and validation.

At this point, this is only a sequence and not a cycle. The cycle part comes from the fact that based on the insights gained during evaluation, a better conceptual model can be built, and a new cycle can be started. However, it should be noted that this theory should be interpreted loosely, as, in practice, the process of building a model is not this sequential. For example, formalization errors may be discovered and corrected during implementation, or partial evaluation could be performed on an unfinished model as part of verification.

3.1.2 Describing Prior and Planned Work with the Modelling Cycle

Prior to the start of this thesis, there were two quasi-separate steps already done: First, somewhere during the PANDEM-2 project discussions, it was decided that there is a need for an SD model to simulate the resource usage during an epidemic (M. Stein, personal communication, 1st March, 2022). This essentially corresponds to problem formulation, therefore, part of the conceptualization step. As the SD paradigm was chosen due to prior experience with the AsiaFluCap model (Stein et al., 2012), and this is the model that de Schipper (2022b) used as a starting point of her work, this can be viewed as a start of a new modelling cycle.

Developing the original model corresponds to conceptualization, formulation and implementation steps. Unfortunately, the documentation of the model does not include a clear distinction between these steps, and the author did not answer our requests to elaborate on the model. However, in the reflections de Schipper (2022a) mentions that an extensive discovery of the model behaviour is needed, as extending the model was prioritized over experimenting with the already complete parts. This need for the discovery of model behaviour equals the need for validation. Furthermore, after inspecting the original model, we found that it does not follow the TU Delft modelling conventions. Therefore we decided to *check the model for errors*, to discover whether this is a source of errors or just a different convention, which is essentially the verification step.

However, to answer our second sub-research question, it is not enough to validate the model. We used the model to generate outputs, which were presented in the workshop as examples to generate discussion about how healthcare resource models can be used. While the modelling cycle cannot be used to describe all tasks related to holding a workshop, the part of the preparatory work related to post-processing model results falls under the evaluation step.

3.2 Verification and Validation

The first step to perform is verification. Verification in SD is relatively straightforward compared to validation; therefore, these will be discussed together. As many validation tests are described by the scientific literature, Barlas (1996) recommends choosing an appropriate, most crucial set of tests. However, to make this choice in an informed manner, an understanding of why and how validation works is required, which will be discussed in this section.

On a philosophic level, model validation stems from the concern of whether the model is the true description of the modelled phenomenon, which implies reliability for decision-making. However, the term validation does not necessarily mean that the model is true, but that the model is legitimate, that it does not contain obvious errors, or has not proven to be false (Oreskes, Shrader-Frechette, & Belitz, 1994). In line with this, J. D. Sterman (1984, Table 1) collected a set of tests for "Building Confidence in System Dynamics Models".

On a more practical level, examining the literature review written by Tsioptsias, Tako, and Robinson (2016) is helpful. The article analyses validation methods proposed by operational research, computer science, and modelling & simulation approaches (SD belongs to the last category). The authors define *verification* as the process concerned with "building the model right", or "to ensure the model runs as intended" (Tsioptsias et al., 2016, p 6:3-4). Furthermore, they also define *validation* as: "a process and evidence for building the right model" (Tsioptsias et al., 2016, p. 6:3). This idea is also present in the article of Barlas (1996, p. 188), "adequacy with respect to a purpose", or as we refer to it: "fit for purpose" (Auping et al., 2022, p. 60). What these articles imply is that, while keeping in mind the purpose of the original model in this research, the model has to be adequate for the purpose of producing example outputs for the workshop.

From another angle, model validation is a broad topic, going beyond SD in many aspects. As there is a big difference between white-box and black-box models, there is also a big difference between the validation of white-box and black-box models. A black-box model is very hard to validate; essentially, the only measurable indicator is the accuracy of the model (Barlas, 1996). While given sufficient free parameters and computational resources, these models can be extremely good at this. For example, in categorizing images (He, Zhang, Ren, & Sun, 2015). However, these models often fail to grasp the underlying context, which can result in painfully obvious failures, given a carefully crafted input, as shown by Szegedy et al. (2013). The problem of not grasping the underlying context correctly does not affect white-box models, as these contexts are built-in by the modeller. Although, for SD models, defining indicators for accuracy makes little sense for two reasons: First, the structure of the model affects the model behaviour more than the exact parameter values (W. Auping, personal communication, 24th May, 2022). Second, Barlas (1989b) demonstrated that behavioural similarity is hard to quantify even in the case of relatively simple outputs. Furthermore, as already pointed out in subsection 2.2.2, validation by accuracy is fundamentally flawed when learning or fine-tuning mechanisms are involved for the parameters.

In the article of Tsioptsias et al. (2016), a list of common validation tests are described for modelling and simulation. Keeping in mind the purpose of the model and that there is no experimental data available, the relevance of each validation test is indicated in Table 3.1.

Validation tests	Relevance of test
Comparison with (existing) data	No (data unavailable)
Statistical tests	No (data unavailable)
Face validity	Yes
Turing test[1]	No (domain experts are not capable of this[2])
Graphics or animation	Yes (though time-consuming due to the size of the model)
Qualitative analysis[3]	Yes (through thesis supervisors and workshop)

Table 3.1: List of common validation tests and their relevance for this thesis.

A systematic validation testing framework is also defined by Barlas (1996). The main idea in this framework is that different types of validities build on each other. The three stages identified are structural-, structure-oriented behaviour-, and behaviour validity. Structural validity includes low-level tests aimed at verifying if the conceptual model was implemented correctly. Structure-oriented behaviour validity aims at a middle level, with tests to discover errors in the equations between the different model variables. The highest level, behavioural validity, is aimed at examining model behaviour and comparing that with the expected behaviour of the real-world system.

3.2.1 Dealing with Uncertainty

There is a term in the modern Bayesian epistemology which is highly relevant for SD models: *epistemic uncertainty* refers to the lack of knowledge about the underlying system (Shariatmadar, Wang, Hubbard, Hallez, & Moens, 2022). Using this term, Bankes (1993) addressed the problem of epistemic uncertainty by describing two

[1] Do not confuse with the more famous version of the Turing test. In this setting, Turing tests refer to verification by the ability of knowledgeable people to distinguish between the real system and model outputs (Barlas, 1989a).

[2] (M. Stein, personal communication, 17th August, 2022)

[3] Qualitative analysis as peer-reviews, subject-matter expert evaluations, face validations, and similar methods (Pace, 2004).

approaches: consolidative modelling and exploratory modelling (if these terms sound unfamiliar, it is further elaborated in section A.2). In this categorization, the original model should be approached as an exploratory model due to its many uncertainties. This is relevant because, on a methodical level, some validation tests are conducted differently, despite the term 'validity' referring to the same concept for both consolidative and exploratory modelling. Auping (2018) made a fairly extensive comparison of these methodical differences. For example, structural validation changes because there is no 'single, best structure' due to the presence of deep uncertainty. Therefore in the exploratory approach, it is necessary to assess whether the relevant uncertainties have been taken into account.

3.2.2 Conceptual Description of the Refactored Model

While verifying the model, it quickly became apparent that the model equations would need to be changed. To differentiate the two versions of the model, the term *refactored model* will be used for the changed model (i.e., the model after the verification and validation), and the term original model will keep refer to the version finished before the start of this thesis. To better understand the steps done during verification and validation, it is worth understanding the refactored model on the conceptual level. While this section will present the conceptual description from the perspective of the refactored model, it is relevant for the original model, as it consists of the same concepts, albeit implemented differently. This model consists of 3 sub-models, as indicated on Figure 3.1. First, the 'Epidemiological compartment model' models the disease spread within the population. The possible compartments are presented in Figure 3.2. These compartments are within a vector model (i.e., the compartments are subscripted), meaning that each compartment is subdivided into three age groups. In addition, the 'Exposed', 'Infectious', and 'Recovering' compartments have a second type of division: isolation status. Therefore these are sub-divided into six groups.

The 'Public health resource model' is responsible for modelling three things: Testing, Contact tracing, and vaccination supply and demand. Testing (such as PCR) and contact tracing determine the flow of people between the 'Susceptible' to 'Quarantined' and between the isolation status sub-compartments. Vaccination determines the flow of 'Susceptible' to 'Vaccinated'.

The 'Hospital resource model' is responsible for modelling the hospital ward and ICU. For both, the occupancy is modelled (how many ward beds are available, incl ventilators for ICU), Staff availability, PPE, and medication. One of Lisette's simplifications was not to model the events beyond hospital resource scarcity. Therefore, the hospital resource model can only affect the compartmental model through medication availability. These medications, such as Remdesivir, speed up the recovery of patients; therefore, they spend less time in the respective hospital compartment. Furthermore, note the lack of direct interaction between the 'Public health resource model' and the 'Hospital resource model'.

The 'Aftercare' part of the original model consisted of rehabilitation, home care, and long COVID resource usage. However, these got deleted to limit the scope. The original and the corrected model can be downloaded from the GitHub repository accompanying this thesis[4].

[4] https://github.com/vioSpark/PANDEM-2-resource-management-under-different-scenarios
auxiliary published material/NL-Pandem-2_original.mdl for the original model, and
auxiliary published material/NL-Pandem-2_refactored.mdl for the refactored model

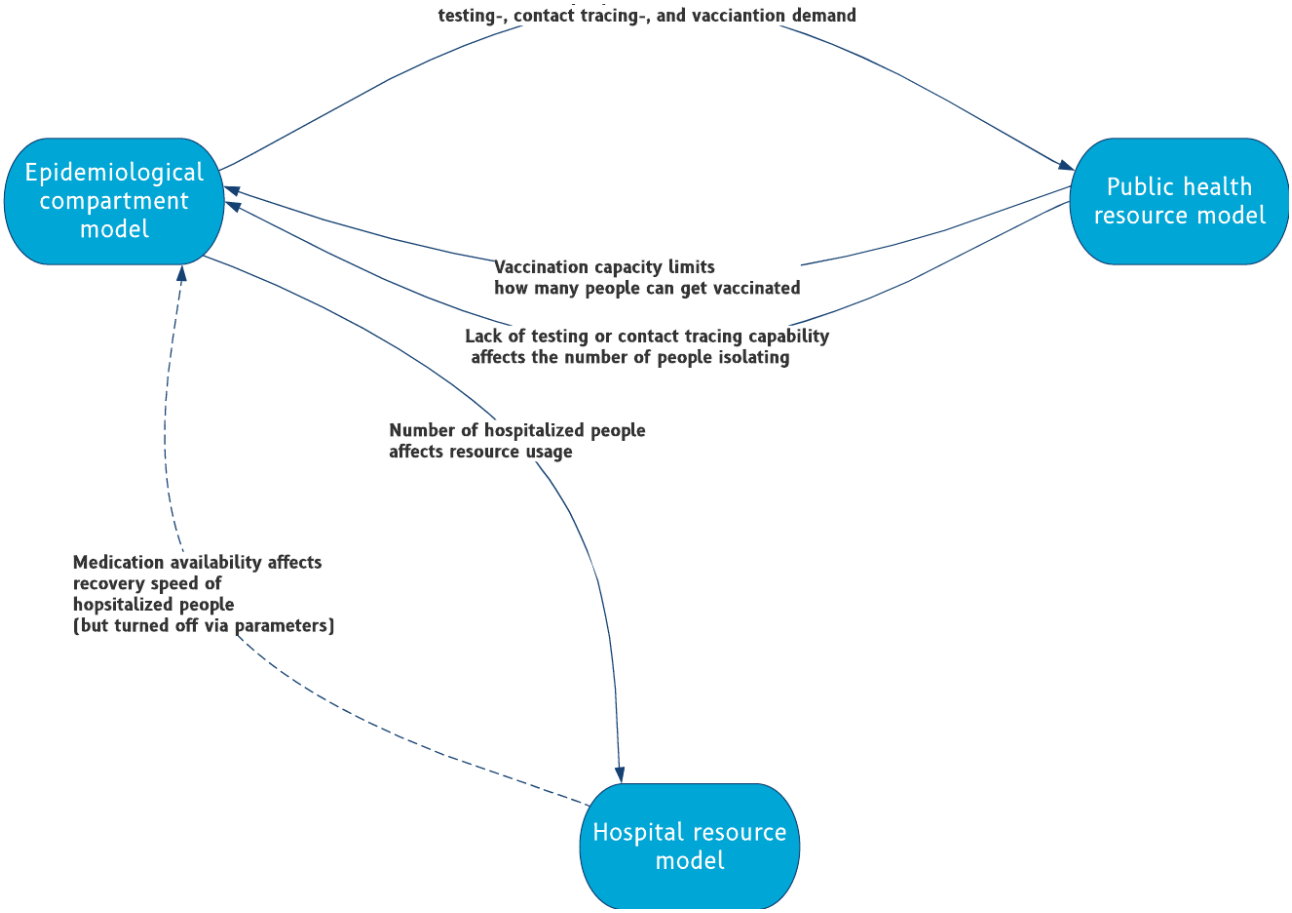


Figure 3.1: Subsystem diagram of the refactored model

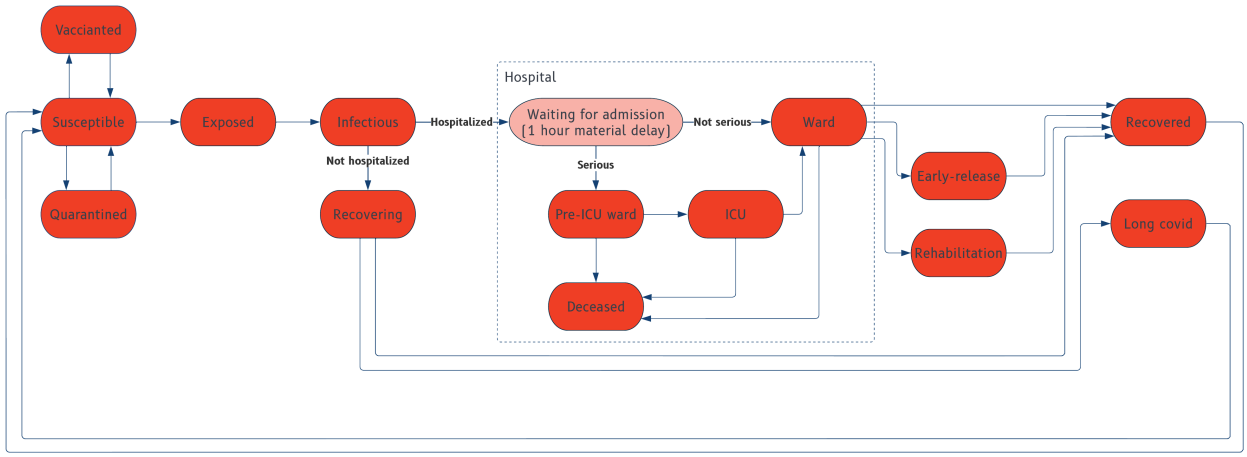


Figure 3.2: Visualization of the compartments in the refactored model.

3.2.3 Set of Verification and Validation Tests Used in this Study

As we can see, validation is not a straightforward process. However, we cannot avoid performing it, as validation is essentially determining whether the model is fit for purpose, and the purpose of the original model changed in this study: Our purpose with the model was to generate semi-realistic outputs that might appear in the dashboard, instead of the original purpose of codifying and communicating findings about the resource utilization during the pandemic. As discussed earlier, model validity can be classified into three types. For each SD model Barlas (1996) recommends defining a most crucial set of tests. From the many types of validity and validation

tests, we chose the following set to address the three levels of validation (structural validity, structure-oriented behaviour validity, and behaviour validity).

Structural Validity

Various direct-structure tests can address structural validity. These tests assess the model's validity by direct comparing the model relationships with knowledge about the actual system (Barlas, 1996). In this study, we have chosen to assess this via *face-checking* (Tsioptsias et al., 2016) and via *dimensional consistency check*, followed by *parameter confirmation* (Forrester & Senge, 1980).

For face-checking, each model equation was viewed individually and compared with the existing knowledge about the system. In this process, the meanings of the internal model variables were also examined both from the mathematical side and from the real-world system meaning side. While in the documentation of the original model, there are no clearly separated conceptual and formal descriptions, a somewhat clear conceptual image is embedded in those documents. Therefore, though to a limited extent, this step can also be called verification of the conceptual model. During this step, several model errors were discovered, which led to the decision to refactor the model. The refactored model was made sure to be error-free and pass the dimensional consistency check by using the built-in Vensim unit check (VENTANA systems inc., 2022a).

On a high level, the following changes were made during structural validation: The compartment model got a complete overhaul. The compartments were connected via flows instead of the auxiliary variable connections. A new subscript was introduced to account for the isolating people instead of using separate compartments. Furthermore, many variables were renamed to be more expressive. Finally, some unconventional approaches were corrected to conform to the principles of the SD paradigm. The stock-flow diagram also received a visual overhaul. Most notably, the parameters were visually separated from the model variables via text colouring. The full list of identified errors can be found in section A.3 in the appendix. Unfortunately, not all of these errors were corrected due to time limits. The most notable error is the incorrect calculation of the disease's mortality (further elaborated in the appendix).

After passing these checks, *parameter confirmation* tests followed. This test is also called parameter verification (Barlas, 1996). Auping (2018) presented that this test is not done the same way for consolidative and exploratory modelling. In consolidative modelling, the parameter confirmation test is concerned with whether model parameters are conceptually and numerically match compared to the existing knowledge (Senge & Forrester, 1980). In this case, each parameter is thought to have a best value. In contrast, in exploratory modelling, the parameters have a bandwidth at best. Therefore, it can be assumed that the parameters' value can be anything in this range, and it is more beneficial to examine whether this range is correct than to examine the best value for the parameter (Auping, 2018).

While the parameters were conceptually verified, given the nearly 100 parameters of the model and that the measurement of many medical definitions differs from country to country, properly identifying the possible parameter ranges was deemed an effort to be beyond this thesis. Instead, as many parameters in the original model were left with placeholder values or marked as assumptions or as guesstimates; it was decided to spend effort on finding a better value for those instead of verifying the parameter's ranges. Guesstimated (guessed and estimated) parameters refer to the case when the parameter's value is calculated from other known or guessed values, using some, sometimes quite liberal, assumptions.

Despite the effort, not all assumed values could be replaced with higher certainty ones, as the admittedly superficial search did not find many parameter values. Since the construction of the original model, a very relevant literature review was completed by Beishuizen and et al. (2022b), aiming to identify the exact values of how much resources are used per patient during the treatment of COVID-19 and influenza. The findings of this search were included in the refactored model. The complete list of parameter values of the refactored model, including the sources of the parameters, can be found in the repository[5].

Structure-oriented Behaviour validity

The structure-oriented behaviour tests examine the next level of validity. The chosen approach was *unit testing*, which is unconventional, as it is not from the toolbox of SD modelling, but an approach adapted from test-driven development (confidential presentation, 2018). It was chosen as an alternative to *extreme-condition tests*. The difference between the two tests is that while extreme-condition testing is concerned whether a single model relationship between two variables behaves plausibly under extreme conditions (Forrester & Senge, 1980), unit testing is concerned with whether selected model variables stay in normal operating ranges.

Due to the extensive effort going into the lower-level structural validation, not many of these unit tests were performed. These were checking whether the total population (minus the deceased) remain the same in

[5] /auxiliary published material/NL-Pandem-2-Cap_new_parameters.xlsx

the model or that a compartment (in the compartmental sub-model) cannot have less than 0 population at any given timestep. Although the reality checks feature of Vensim is intended to implement extreme-condition testing, it was easily adapted to support unit testing (VENTANA systems inc., 2022b). Also, note that *unit testing* and *unit checking* (the dimensional consistency test) are entirely different but incredibly easy to confuse.

Behaviour Validity

While the original idea for this thesis was to follow a more consolidative approach and to validate the model based on historical data, we found out that the dataset needed for this is non-existent (RIVM internal team meeting, personal communication, 30th March, 2022). Unfortunately, the literature is not well-established about how to validate transient model behaviour (such as a one-time resource shortage). Barlas (1996) recommends the use of visualizations of typical behaviour features, although this is more of a recommendation than an established framework. To some extent, this is also something we did during the face-checking of the model, though this was done on an 'as-needed basis' rather than systematically. The lack of more extensive behaviour validity was a consequence of the fact that the previous two levels of validation had already uncovered many errors.

3.2.4 Conclusion of Validation

As we have seen by performing the validation tests, the refactored model still consists of errors. Worse, as the behaviour validation was not systematic, there is a reasonable suspicion that further tests would discover more errors. It is trivial that the refactored model should not be used for predictive purposes, as structural errors remain, and some parameter values are incredibly uncertain. However, we have to ask the question: Does our model need to be this accurate to produce outputs for our workshop? For that, the answer is a definite 'no': we can generate hypothetical scenarios, which do not have to be accurate, only realistic. This means that as long as basic domain-specific intuition works (such as resource shortages causing the disease to get noticeably worse or the correlation between the number of infected and deceased is positive), we should be able to define internally consistent scenarios. Internal consistency is a nice feature of using a mathematical model for scenario generation: examining model outputs and model equations together unambiguously explains what happened in the model, and relevant facts can be compiled into a scenario. Therefore we can say that the refactored model is fit for the purpose of scenario generation; therefore, it passed validation.

3.3 Workshop

To address our second research question (How are, or can healthcare resource models be used?), we planned to consult subject-matter experts and then hold a workshop to reduce individual and disciplinary bias. The consultations with subject-matter experts happened in weekly meetings (as they were the external advisors of this thesis), where the concept and the material of the workshop were refined over the approximate period of 2022 August - mid-October. Nevertheless, part of the scientific literature was consulted on how to hold a workshop, which will be presented in the following sections.

3.3.1 Workshop as a Research Tool

Workshops, as a research methodology, are not bound to any particular research approach. The word 'workshop' has many different meanings depending on the context. As a result, there is no common definition nor a common purpose for workshops. Regardless, there is a general notion that a workshop is a qualitative data-gathering process where participants interact with each other. A common theme emerges despite the lack of a common definition by analyzing the different definitions of workshops (Freytag & Young, 2018; Ørngreen & Levinsen, 2017; Thoring, Mueller, & Badke-Schaub, 2020). A workshop has multiple participants, a specific goal, and a pre-allocated space and time. The participants usually share domain expertise or an agenda/focus of a specific problem. The specific goal of the workshop could vary between a wide range of activities: information sharing and collective learning, problem investigation, problem solving, idea generation, innovation, or artefact evaluation. The pre-allocated space and time are solely required to ensure the interaction between the participants happens. However, to safeguard against the disruption of operational activities, a time limit is imposed on most workshops.

There are closely related qualitative approaches to workshops. Two are highlighted in this paragraph to shed more light on what a workshop in this thesis means. Firstly, a *focus group* is a technique where the researcher collects data about a specific topic via group interaction. These interactions between participants are likely

to bring forward opinions which otherwise would have been missed. However, focus groups are often used as a complementary research method to individual interviews (Freytag & Young, 2018). Secondly, *participatory design* originates from the approach where (industrial) workers are included in the design of the machinery they will end up using to prevent some problems from arising later. The second approach is also related to action research, which is an approach where despite the generation of (scientific) knowledge, the researchers also aim to provide the required insight for participants to change their own situation (Freytag & Young, 2018).

Thoring et al. (2020) identified two goals of workshops: Either creating new output, such as designs, ideas, or solutions, or evaluating specific aspects of interest, such as testing the usefulness of a process, product, or tool. Furthermore, evaluation workshops are often part of a broader action research initiative or an action design research project.

Given the planned dashboard of PANDEM-2 WP4, there is a clear direction to follow based on the literature mentioned above. We are designing a sort of virtual machinery: the model with the dashboard. While the intended target audience is very broad, healthcare- and hospital resource managers are among the primary audience. Applying the idea of a participatory design workshop to this situation, we should involve end-users from the primary target audience to find better communication techniques about the model and recommendations for the dashboard design.

3.3.2 Workshops in System Dynamics

Holding workshops is a technique also employed by SD. Among others, Rouwette and Vennix (2020) describes that the Group Model Building (GMB) approach utilizes workshops for knowledge elicitation, mainly in the conceptualization step of the modelling cycle, as a supporting technique besides interviews. A more elaborate description of group model building can be read in section A.4.

While some of the GMB workshops reportedly had around 40 participants (Leerapan et al., 2020), the general group size is 5-10 people (Bolt et al., 2021). The preference for this group size balances two effects. While including everyone's opinion suggests using a big group, the more participants attend, the less time they have to express their opinions in detail. For example, assuming that everyone speaks equally, a group size of 10 people over a 2 hours workshop gives 12 minutes per participant. On the other hand, a group size of 40 people over the same 2 hours gives every participant 3 minutes, which may be enough to communicate one thought but not enough to communicate an entire concept.

Another benefit of this small workshop style is that it allows participants to communicate with each other (Bolt et al., 2021). This is useful for two reasons: Firstly, if the participants have a different mental image of the real-world system, they can find this out and resolve these disagreements during the workshop. Secondly, the workshop provides a unique opportunity for participants to share their experiences with each other, which probably would not have happened otherwise.

In their article Bolt et al. (2021) analyze the modelling artefacts from the knowledge-management perspective. Based on the analysis of their case studies, they found two types of modelling processes: The first type aims to create new insights by communicating and sharing ideas among participants. In this case, the model acts as an *epistemic boundary object*. The second type aims to codify expert knowledge by creating a realistic representation of the real-world system. In this case, the model acts as a *technical representative object*. In the first case, one of the added values of the GMB approach over other approaches is that participants are taught the basics of SD modelling, which they can use as a common language (i.e., boundary object) to translate their knowledge for other participants. While this implies that in the GMB approach, the model development process is just as important as the resulting model itself, there is a far more relevant implication. In an interdisciplinary workshop, a boundary object can help to start the discussion.

3.3.3 Considered Guidelines for the Workshop

Despite defining workshops based on the scientific literature was a fairly easy step, no clear methodological guidelines were found, especially about the perspective of a workshop as a research tool. Freytag and Young (2018) describe that participatory design workshops use artefacts as boundary objects to help participants express their opinions. These artefacts enable non-designers to think more creatively and help participants share their tacit knowledge. However, they found that in the literature on action-oriented research, there are no specific guidelines about facilitating a workshop or effectively involving boundary objects in the process. They also propose a workshop framework Freytag and Young (2018, p. 164); however, it is still quite generic and did not help determine the exact methodological details.

Other scholars also found that the workshop literature is non-informative about methodological considerations. For example, Ørngreen and Levinsen (2017, p. 72) identified two goals of the workshop format as a research methodology: "to fulfil participants' expectations to achieve something related to their own interests"

and to "produce reliable and valid data about the domain in question". They also found that workshops often consist of roleplays or scenarios which are realistic and recognizable.

In the literature of group model building, it was also found that a workshop is hardly used on its own. There are either interviews before (Ibrahim Shire, Jun, & Robinson, 2020; Rouwette & Vennix, 2020) or successive workshops (Leerapan et al., 2020). There was also a study which used immediate follow-up surveys to evaluate the effectiveness of the workshop (Ibrahim Shire et al., 2020). After some consideration, we decided to limit this thesis to the workshops only, to limit the scope to a reasonable extent.

3.3.4 Workshop Description

To address the second sub-research question (*How are, or can healthcare resource models be used?*), the workshop was held along the following directions:

- Examine when the model is useful for the workshop participants.
- Present outputs of the model.
- Examine what outputs are useful for the workshop participants.

As the model consists of both a hospital and a public health resource side, three participants were invited from a hospital to represent the hospital side and three from RIVM to represent the public healthcare side. The exact participants got selected by following a snowballing approach. Also, in the absence of clear recommendations in the literature, a 90 minutes long workshop was planned. Unfortunately, due to limited availability and last-minute cancellations, we ended up having two separate sessions with two-two participants. Although methodologically, this resembles to discussion groups more than to workshops, to keep the linguistic consistency with the rest of this thesis, these will be kept referring to as workshops. In the first workshop, two participants were present from a hospital (further referred to as P1 and P2), and in the second workshop, two participants were present from RIVM (P3 and P4). As the reduced participant count made us expect 60-minute-long discussions, the remaining 30 minutes could be kept as a buffer time and for general discussion.

Building on the idea of 'roleplays' (Ørngreen & Levinsen, 2017), participants were asked to imagine a respiratory outbreak where they were in a decision-making position. Then they were told to examine: *How to support you in making your decisions?* To guide the conversation, different sets of scenarios were presented in different styles, and for each, the following practical questions were asked:

- How easy is it to understand this type of output? (1 to 9 Likert scale)
- What does this communicate to you? (discussion)
- What does it prompt you to do? (discussion)

We decided that the presentation of the different model outputs will be done via scenario discovery, where different scenarios will be created by defining interventions. Here under a *scenario*, a single model run with a fixed vector of parameters is meant. An *intervention* is defined as an opportunity to change the progression of the pandemic. For creating an intervention, several model runs are created using slightly different parameters. These differences in the parameters are referred as *uncertainties*. In an intervention, the differences in the scenarios are presented along important model variables. These are called Key Performance Indicators (KPIs) in this thesis. Furthermore, a specific scenario, based on the most probable parameter values, got the nickname *Baseline*, which remained the same across all interventions. For communicating the interventions, we decided to utilize a presentation.

3.3.5 Determining Interventions

Since a singular purpose for the model is not defined, automatic scenario discovery methods cannot be used. This is further elaborated in section A.5 in the appendix. To tackle this, a traditional scenario discovery was performed. Two things were needed to create the presentation of the interventions: uncertainties to create multiple scenarios and KPIs to communicate model outputs.

Determining Uncertainties

As part of the PANDEM-2 project, Beishuizen and et al. (2022a) examined which resources are important to include in the modelling and the operational planning of a pandemic via a Delphi study. The former is relevant because it shows where the attention of public health experts and clinicians lies. Translating his conceptual findings to potential scenarios happened in the following way: First, each concept was collected into a table alongside the model variable that models the concept most accurately or marked with 'N/A' in case it is not modelled. Then, for the modelled variables, it was identified which parameter is the best to influence by going upstream along the causal relations. Then for each parameter, the value in the model is collected. Due to the lack of better data, it was assumed that the uncertainty is the $[0.5 * parameter\_value, 2 * parameter\_value]$ range. The results of this approach can be found in the repository[6] accompanying this thesis.

From these uncertainties, using the insight gained into model behaviour during the verification and validation process, a few values were selected manually, with the goal in mind that the difference between the scenarios should be relatively easy to understand. The exact final parameters that were used to generate the interventions can be found in the repository[7] accompanying this thesis.

Determining KPIs

Due to the lack of a clear purpose for the model, there was no clear-cut problem definition, which also meant that defining the KPIs of the model was not straightforward. In exploratory modelling, identifying the model KPIs should be based on the stakeholder's values (Steinmann, Auping, & Kwakkel, 2020); however, this can become a relatively lengthy process. To cut some corners, we took an educated guess in the following way: First, a shortlist of possible KPIs was created based on the detailed knowledge of the refactored model. This can be found in the repository[8].

From this shortlist, relying on the tacit knowledge of two RIVM researchers, the items visible on Table 3.2 were selected (M. Steinand B. Beishuizen, personal communication, 14th September, 2022). Furthermore, to limit the expected mental workload on the workshop participants, it was also decided to show only the gap (or surplus) of the resources and only display a KPI when there is a difference between the scenarios.

Recommendation (conceptual)[9]	parameter equivalent in model or formula to calculate[10]
Epidemic progression	
Number of infected cases (per day)	infection
Number of infected cases that need hospitalization	symptomatic hospitalized
Number deceased	deceased
Hospital resources	
number of ward beds needed / gap	ward beds gap = ward - ward capacity supply
number of ICU beds needed / gap	ICU beds gap = ICU - ICU capacity supply
number of PPE needed / gap	PPE gap = PPE usage - PPE
*total ward beds occupied per day	ward
*total ICU beds occupied per day	ICU
Public health resources	
*Testing capacity per day needed	testing gap = combined testing demand - testing rate by age group
*Contact tracing capacity per day needed	contact tracing gap = tested unisolated symptomatics - capacity for contact tracing per day

Table 3.2: KPIs used to differentiate scenarios.

3.3.6 Description of Interventions

Given the 90 minutes planned workshop time, we decided to experiment using four interventions. Furthermore, to package the workshop nicely, it was decided that a presentation would be used. This is published in the

[6] /auxilary published material/Delphi to model lookup table.xlsx
[7] /src/model_setup/constant_scenario_definitions.py
[8] /auxilary published material/Possible model KPIs.xlsx
[9] KPIs marked with a star (*) got excluded or modified when a later revision of the workshop presentation focused more on the gaps.
[10] As most of these parameters are subscripted; therefore, a summation over the subscripts was performed.

repository[11]. The presented interventions in this presentation were the following:

Intervention 1

Intervention 1 is about an increase in PPE acquisition. In the model, this means that more PPE flow into the PPE stock each day. This intervention was chosen partly due to the high relevance to real life: There was public concern about the availability (NL Times, 2020), and some issues about the quality of PPE gears (National Institute for Public Health and the Environment, 2022b). This intervention is also relatively simple, therefore, partly acts as a warm-up exercise.

The effects of this intervention are visible on Figure 3.3, where a line plot was chosen for this output. The unplotted KPIs are the same, regardless of the scenario. The direct effect of the intervention happened on the PPE gap: Increased PPE acquisition meant sufficient PPE along the entire modelled time period (the orange line goes toward negative infinity). This led to a slightly less severe ward beds gap, as in the baseline scenario (blue line), the sudden jump around day 160 is caused by the PPE stock emptying out.

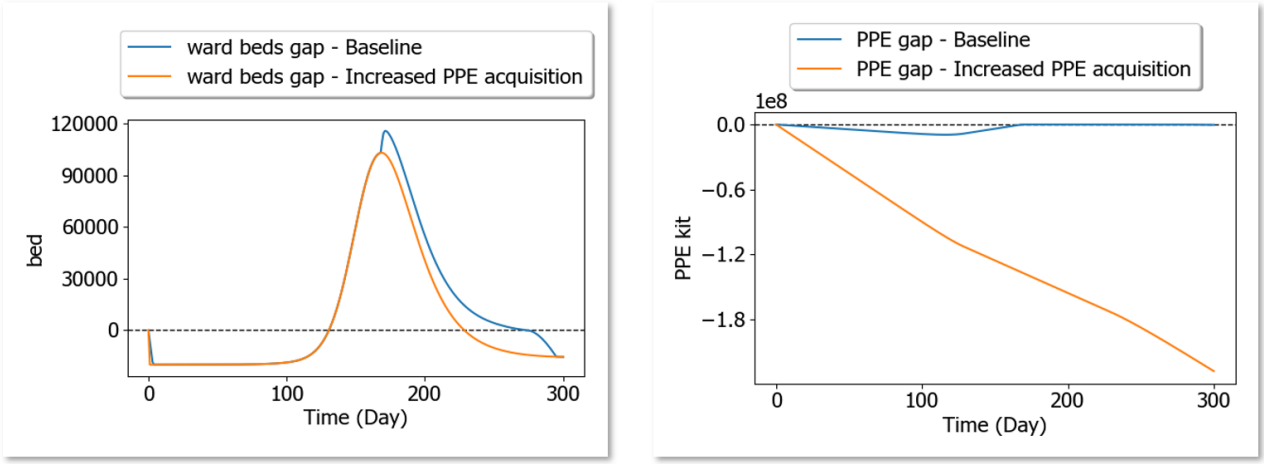


Figure 3.3: The presented graphs of intervention 1

Intervention 2

Intervention 2 is about an increase in test acquisition, meaning more tests flow into the tests stockpile daily. This is highly relevant because it shows an example that a pandemic needs an orchestrated public health and hospital capacity response.

The effects of this intervention are visible on Figure 3.4. This intervention resulted in more changes: First, the direct effect of increased test acquisition meant a less severe shortage of tests (not shown directly), which led to a smaller 'testing gap' (i.e., fewer tests were missing per day). This had a twofold effect: Since more of the people who got the disease got identified, contact tracing had more indices to start with, which led to an enormous gap compared to the baseline scenario (graph of 'contact tracing gap'). Also note that while the gap got bigger in the increased scenario, more people got contact traced in total. The other effect of the increased testing was that people who received confirmation about being infected with the disease were more likely to be isolated. This second effect was strengthened by the fact that more contact tracing also meant more people isolated in the model. More isolated people led to a smaller peak in the epidemic wave (graph of 'infected cases'), which directly affected 'hospitalization'.

Intervention 3

Intervention 3 is named 'Reduced ward length of stay', referring to the model parameter 'ward length of stay'. This part of the model originated from an idea at RUMC that during the peak of a wave, in an extremely desperate case, more hospital capacity could be created by releasing suitable patients early with a ration of oxygen to boost their recovery (Meeting between RIVM, NUIG, and RUMC, personal communication, 25th May, 2022). While there is an extra layer of uncertainty surrounding the topic, as it never got implemented, it is possible to explore the effects of this intervention in the model by decreasing the 'ward length of stay' parameter.

[11] /auxilary published material/workshop presentation.pptx

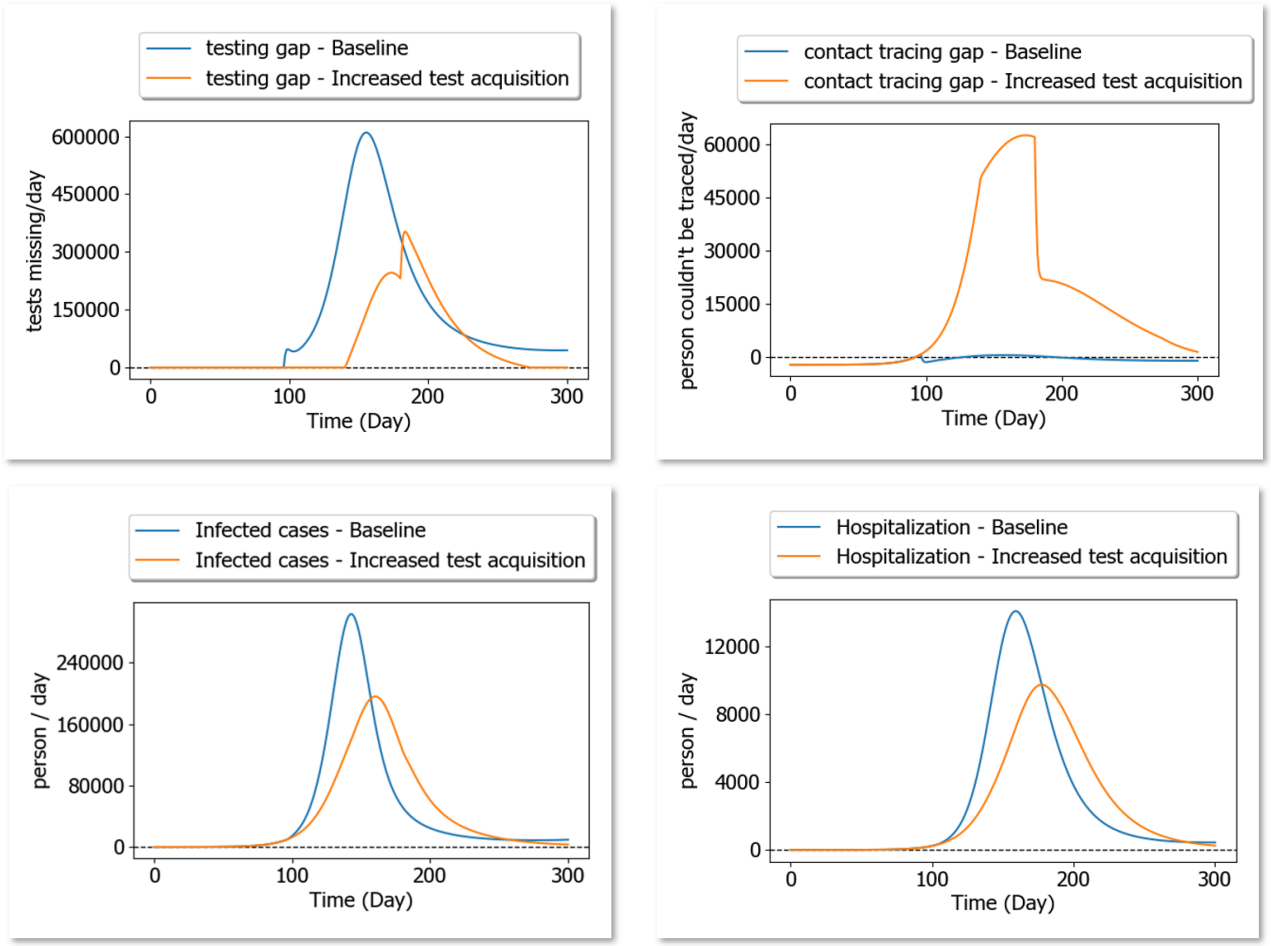


Figure 3.4: The presented graphs of intervention 2

While the naming of this parameter is slightly misleading, it is not decreasing the hospitalization time of all patients for the following reasons: In the model, the output flow from the hospital ward is split into multiple flows, where each flow is a first-order material delay. The 'ward length of stay' parameter is the coefficient for one of these material delays. This means that the 'ward length of stay' is the average time a patient stays in the ward compartment. Decreasing it means the average time decreases, and not every patient gets released earlier. While this is quite a coarse simplification of the real world, the theoretical foundations are well-established, for example, by J. Sterman (2000).

The effects of this intervention are visible on Figure 3.5, on the graphs. Firstly, the change in the 'Deseaed' is highly affected by a bug in the model (discussed in section A.3); therefore, it should not be interpreted. The second effect is the anticipated change in hospital capacity, which led to a smaller 'ward beds gap'. The second effect was that hospital staff interacted with patients less and used less PPE due to the early-release intervention. This slightly reduced usage meant that the PPE stockpile could grow bigger before the depletion started, resulting in the shortage happening later (in time). The effect is visible to the careful observers on the 'ward beds gap' graph: the sudden bump around day 170 happens slightly later in the 'reduced' scenario (it is more visible by zooming in or by using a ruler).

As we were interested in the participants' opinions about different types of data presentations too, a tabular representation of the data was created, also visible on Figure 3.5. The table was created by looking at the interesting part (i.e., the peak) of the graphs. This data representation is one step further processed version of the model outputs.

Parameter	Baseline	Decreased LoS
Ward bed capacity gap	Max: 115 000 Day 120 - 280	Max: 60 000 Day 125-230
Period of PPE gap	130 days	85 days
Deceased (total)	110 000	75 000
Other parameters	same	

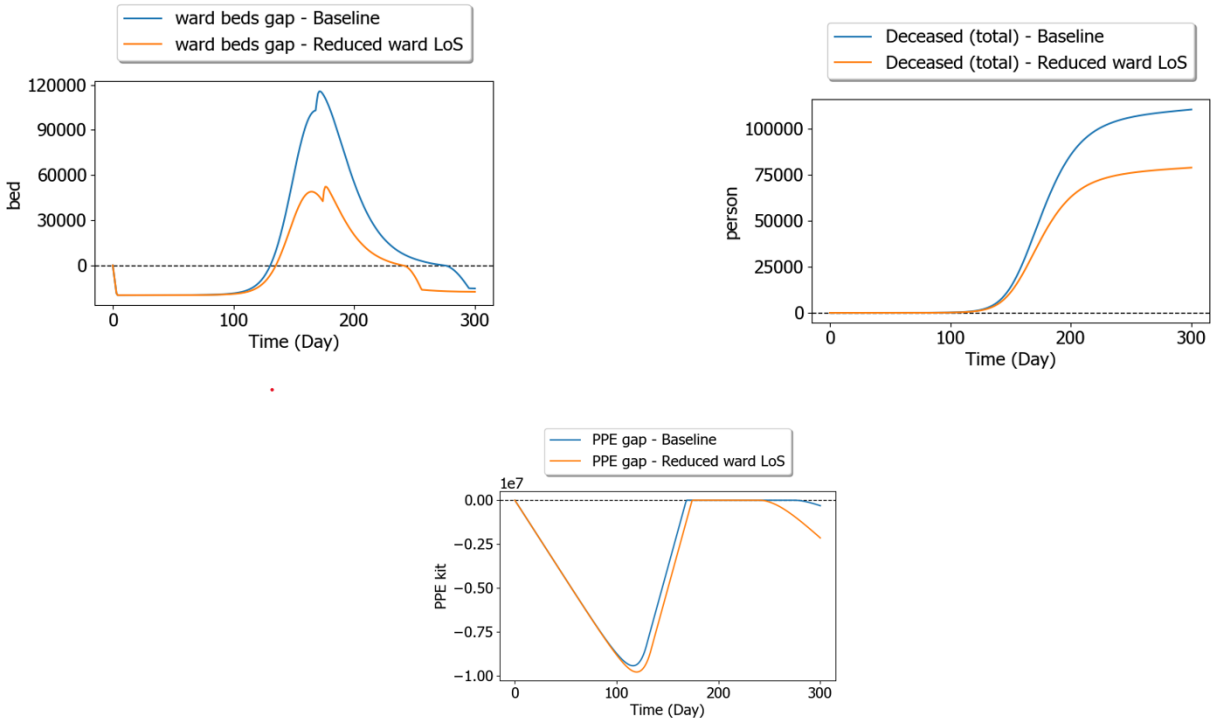


Figure 3.5: The presented table of intervention 3 and the graphs that were used to make the table

Intervention 4

Intervention 4 is related to hospital staffing. In one of the WP4 meetings, it was expressed in detail that the most pressing capacity limit was staffing, not tangible resources (Meeting between RIVM, NUIG, and RUMC, personal communication, 25th May, 2022); therefore, a related intervention was created, by modifying the 'visit per patient' model parameter. Reducing this parameter is equivalent to the idea of the staff visiting patients less frequently. This parameter reduction did not change the ward capacity; it only decreased PPE usage. After a minimal investigation, it was discovered that the lack of capacity change in the model was caused by the other limiting factors of delivering healthcare. Out of curiosity, these limits were removed from the ICU bed capacity ('bed count', 'ventilator count', and 'patient-to-staff ratio', the need for changing the last is related to another bug, discussed in section A.3). This resulted in three scenarios for this intervention: Baseline, limits lifted, limits lifted and visits reduced. We also wanted to experiment with bar plots; therefore, this intervention was presented in the way visible on Figure 3.6.

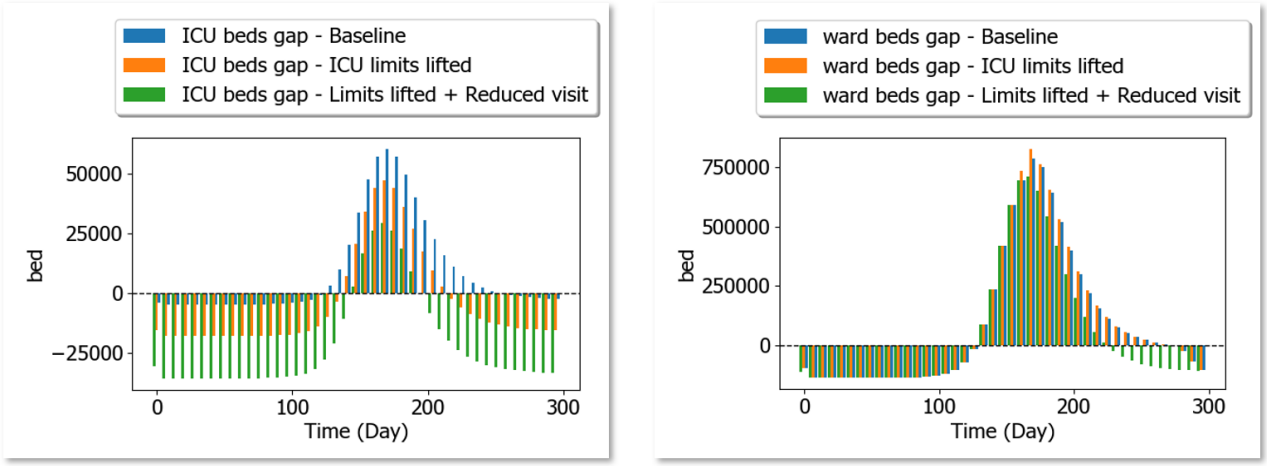


Figure 3.6: The presented graphs of intervention 4

3.4 Summary

First, a description was provided of how the practical work consists of verification, validation, and holding a workshop. After that, verification and validation were explained in detail, including that the model needs to be fit for the purpose of generating scenarios for the workshop. Next, the model was examined using different validation tests and was corrected and reparametrized in the process. While the results are still incredibly inaccurate, for the purpose of scenario generation, the model passes validation. After the validation, the details of the workshop were explained: Participants were asked to imagine themselves in a decision-making position, while the difference made by the interventions were presented to them along with KPIs.

Chapter 4

Results of the Workshop

The workshops were very insightful, as participants raised several critical questions. In addition, the data visualization greatly helped to foster discussion. The direct observations of the workshop can be classified into four topics: Participants' Reaction to the Presentation of the Data (section 4.1), Participants' Reaction to the Communicated Information (section 4.2), How Participants would Make Decisions Based on Data (section 4.3), and Limitations of Current Modelling Approach (section 4.4). *Also, at the risk of repeating myself, workshop participants are denoted in the following way: P1 and P2 are from the hospital-resource side, while P3 and P4 are from the public health resource side.*

4.1 Participants' Reaction to the Presentation of the Data

We have gained some insights regarding the practicalities of data presentation:

Data in Context

During the presentation of data, multiple questions were received about the context of the data, such as 'What does the graph mean ... What does it display?' (P1) or 'What is the difference between the orange and the blue line again?' (P1 referring to Figure 3.3, then P2 a few minutes later). This resulted in approximately half of the communication being about the data and the other half being about how that data got generated. Also, when presenting multiple scenarios via multiple line charts on the same canvas, such as in the case of interventions 1 and 2 (Figure 3.3 and Figure 3.4), the legend alone was not enough to communicate the difference between the different series.

Presentation Medium

Despite the accelerated digital transformation, we cannot assume that everyone will have a large enough screen to look at the graphs. This is especially the case when people look at data not because that is core to their work but because it is interesting or falls in the good-to-know category. In this case, they might commute or be on a lunch break while looking at the data, and more interruptions can happen. As P2 pointed it out: 'It is difficult to understand because I was not paying the attention that I intended to', while was travelling from one building to another. Furthermore, participants needed not only to understand the data but also needed subsequent time to understand the implications of the data for their work or agenda. 'It takes a while for a good question', as P2 put it, this time paying full attention to the presented data.

From another perspective, MS Teams was barely sufficient as an ad-hoc tool for holding the workshops. The digital environment caused some nuisances, such as: 'I can see it (the poll), there is just a little bit of switching' (P1), as on a mobile device, the presentation and the poll were on different views. In another case (on a computer), the polling system covered part of the presentation.

Discipline-specific Knowledge

Physicians generally seem to deal with fewer graphs than people in public health decision-making positions, according to (P3). This different level of familiarity means that different levels of explanation are required for different audiences. A related statement was received from P2 when it was expressed that the average professional cannot be expected to be able to transform the data from line plot format to table formats, such

as done at Figure 3.5. However, for operational decision-making, the tabular style of presentation of the data is needed.

We also perceived a lack of common jargon over the entire field, leading to abbreviations and definitions being non-trivial. For example, 'Maybe a stupid question, but what is a PPE?' (P2), and another participant misunderstood what reducing patient visits stands for in intervention 4. However, these might be related to the fact that the participants work in Dutch while the workshop was held in English.

Timing

When asked what the displayed data prompts participants, P4 responded with: 'It depends also on the timeline, when I am on day 100 and we get a model like this. And you see that there is a enormous gap in the available beds, it is alarming. But when you are on the day 175, ...' While P4 did not finish his sentence, it is clear that it was a reference to the timing of the model results, which seems to be a factor to count in.

Templatized Visualization

The people who look at the data tend to learn the visualization style quickly. As P3 put it: 'Since this was the first picture that was shown it was good to get a little bit of explanation, ... But I can imagine when you have given the first kind of explanation, like the next pictures will be easier to read'. This 'kind of explanation' referred to the templatization of visuals, such as the same graph type, colours, and general layout.

Scoring of Outputs

As part of the workshop, we asked participants to score how easy it is to understand the outputs on a Likert scale of one to nine. The aggregated results are visible on Table 4.1.

Intervention	average score	median score
Intervention 1	5	5
Intervention 2	7.5	7.5
Intervention 3	7.75	8
Intervention 4	5.25	5

Table 4.1: Average and median scores received to the question: *How easy is it to understand this type of output?*

4.2 Participants' Reaction to the Communicated Information

The following are our findings about the participant's reactions to the presented data:

Practical Perspective

It was noted that participants had a slightly more heated response about people not getting care in a simulation model. Participants tended to focus on the practical meaning of the graph more than the analyst, probably due to this approach being superior in day-to-day work. As part of this practice-oriented perspective, participants expressed the need to 'look for other options' when the presented intervention did not lead to a sizeable reduction: 'The difference between increased PPE acquisition or just the baseline doesn't seem that high, ..., but it's not that great to me, so, I'd look perhaps to other options' (P1). However, they also noted that when there is a gap, 'every bit counts'. They also expressed a 'need to react' (P4).

Data Augmented with Experience

When asked *What does it prompts you to do?*, P2 responded with the following: 'When I am looking back at the situation of the spring of 2020 this slight decrease would not have affect any decision that we have taken, because the difference was to increase the PPE was so minor to the lack of capacity we had.' Referring to mostly the lack of personnel (especially nurses), but in the first wave, also to the lack of equipment. This shows that participants tended to augment the shown data with their past experience. During the discussion with the hospital side, this 'lack of nurses' came up several times before intervention 4, which is the intervention related to the lack of personnel. P3 also expressed that from experience, they know that reducing visits per patient will not work, despite the model outputs suggesting this conclusion.

Probably this experience augmentation also resulted in the observation that the different disciplines thought along different types of possible actions: Hospital participants expressed that knowing how many beds are needed in the baseline and how many beds are needed in case the intervention is chosen would be helpful. This also implies the need for predictive models. Meanwhile, RIVM participants expressed their desire to share the knowledge with relevant partners or to 'notify the minister to get oxygen for the country' (P4) because that could help with the current situation, or to 'speak with somebody or organize something' (something referring to a meeting, workshop, or conference here). 'Especially for the PPE, if they see a gap, they have to act quickly to decide if we have enough, where do we get some extra from' (P3). At one point, P3 referred to executing preparedness plans, which are plans made to enable a fast yet organized response to a crisis situation.

Communication of the Modelling Paradigm

Due to recommendations during peer-checking of the workshop presentation, the explanation of the basics of the SD paradigm got removed. However, this seems to be a mistake due to the following reason while discussing intervention 3: Using the jargon of SD modelling: intervention 3 is about changing the parameter of a first-order material delay between the ward and the compartments after. This means changing the average time a patient stays in the ward compartment (this was more elaborately explained in Intervention 3). While presenting this output (without the SD jargon), we got feedback from both hospital participants that it is a case-by-case decision whether patients can be sent home early. What potentially happened was that due to the lack of communication, they assumed that every patient gets sent back home earlier and pointed out that this is highly unrealistic. In retrospect, this minor misunderstanding could have been by rephrasing the intervention to 'reducing the length of stay *statistically*', but this whole class of these misunderstandings can be avoided by communicating the modelling paradigm. Relatedly, P2 posed an interesting question: Do we want to present the model or the effect of the interventions? P2 said that he perceived differently when the graphs were presented and when only the table was presented: 'There are difference between the graphs to show the model, and the table to understand the decision'.

4.3 How Participants would Make Decisions Based on Data

Participants were searching for actionable insight when they imagined themselves in a decision-making position, which was not immediately obtainable from the graph type of output. However, it was noted that there is no single decision-maker.

It was also observed that participants sought for full solutions. In cases where the model outputs suggest that the intervention only leads to a partial solution (i.e., having a smaller resource gap), P2 expressed that 'it is hard to make a good decision about what should be done because either way, the gap is still there'.

While discussing the utility of model-based scenarios, P2 recalled a problem their hospital experienced over the summer: due to a few days of festivity, many nurses caught COVID-19, resulting in an unexpected level of nurse absenteeism. Unfortunately, they did not have preparedness plans for this kind of absency. In P2's opinion, a model could have been useful to generate a scenario like this and induce discussions about the topic by showing the effects of staff absency.

4.4 Limitations of Current Modelling Approach

There are quite some previously unknown or neglected limitations with the model, which have been raised during the discussions in the workshops.

Unknown and Fast Changing Dynamics

When faced with a fundamentally new situation, such as the first COVID-19 wave, the tendency to implement ad-hoc solutions increases. For example, at the hospital, 'Many anesthesiologists went to work at the ICU, so, therefore, we could increase it (the ICU personnel) in a way we could not increase anymore' (P2, referring to the fact that anesthesiologists went back to help with surgeries). In the same way, some operation rooms have also been used as ICUs, but not anymore.

It is interesting that the real-world system can change at a speed comparable to the speed of model building: P3 expressed that while the model accounts for contact tracing, the current (14th October 2022) approach is to not contact trace infections. It is unnecessary to contact trace the virus, as people already know what to do in case of an infection, and there is a fair amount of knowledge about how the strain behaves. Of course, if a new strain emerges, they would restart contact tracing for that strain.

Focus on Hospital

RIVM participants pointed out that the current interventions focus on the hospital part. They provided a few ideas that would be interesting to research more in-depth: selective testing and contact tracing. Selective testing in case of a more contagious disease than the coronavirus works in the following way: the tested community needs to be identified, for example, a nursing home, daycare, or school. A few (e.g. three) people need to be tested, and if all three are positive, it can be assumed that the entire community is infected with the disease (P4). Selective contact tracing works by only doing contact tracing for specific groups (e.g. for people over 50) (P3). P3 also suggested looking into modelling the effect of self-testing. It was expressed that predicting the need for an increased testing capacity is relatively easy, but not for contact tracing. Therefore, a model that could forecast the need for CT would be interesting.

PR versus Modelling

P3 recalled a case where the action taken was counterintuitive compared to what their (mental) models suggested: Contact tracing a public KPI in the eyes of the government, so for PR-related reasons, they went beyond the otherwise reasonable limits.

4.5 Summary of the Workshop Results

We gained substantial practical knowledge about how to present data. A key takeaway was that more effort should have been invested into communicating the surrounding context of the data, especially the model. Interdisciplinary communication (i.e., communication between the analyst and participants) was non-trivial, as an under-communicated context led to misunderstandings and confusion.

We have also seen that participants tend to augment their thinking with their recent experiences. In this workshop's case, this was the experience accumulated during earlier waves of the COVID-19 pandemic. This became helpful, as the underlying model also simulated a COVID-like illness. However, this augmentation is a potential source of difficulty when interpreting the data if the past experience does not match the simulation's base assumptions.

The hospital participants indicated that in intervention 3, making the inference from the graphs to the table cannot be expected from every professional. From an analytical perspective, making this inference is quite a trivial step. Therefore, to provide valuable insight, the work does not need to stop at visualizing the model outputs; those could be further analyzed. This leads naturally to the next point: Data-based decision-making.

The end goal of the participants was quite clear: to get rid of the perceived gap. While the presentation included data that supported their reasoning, they were not searching for the data but for means to eliminate the resource gap. The ideal insight would be to provide a set of interventions and their expected effects or provide which interventions are needed to eliminate the resource gap. Furthermore, it was observed that hospital-side participants thought along 'what options can we take', while RIVM participants thought along 'whom do we need to notify'. This likely indicates that different things count as an intervention in different organizations.

Lastly, some limitations of the model were discovered. First, RIVM participants mentioned an example: When contact tracing became a publicized KPI of the ministry, for PR reasons, it is evident to put more effort into that activity than the model outputs would suggest. Secondly, they also suggested that contact tracing is underresearched in the model, as it is possible to do group-specific tracing and testing. It was pointed out, too, that the model completely ignores the effects of self-testing. Finally, a need for contact tracing demand forecast was also expressed, as currently, it is a challenge to employ enough people without risking over-employment.

Chapter 5

Discussion and Recommendation

The discussion of this study is presented along four themes: First, uncertainty and validation are discussed in Modelling and Epistemic Uncertainty (section 5.1) to answer the first sub-research question. After this the practicalities of How to Present Model Outputs (section 5.2) are discussed. Thirdly, in Needed Data for Situational Awareness (section 5.3), it is discussed how can be healthcare resource models used, and sub-research question two is answered. The last theme is Implications of the Need for Consolidative Models (section 5.4), where the main research question is answered. After this, the Limitations and Future Research (section 5.5) of this study are briefly mentioned.

5.1 Modelling and Epistemic Uncertainty

During the workshop, we received the question of whether *we want to present the model or present the effect of the interventions* (section 4.2). This question is interesting for the following reason: an argument can be made that presenting the model is needed to present the effects of the intervention, as the intervention's outcome is directly dependent on the model structure and parameter values. The option of only presenting the effects of the intervention is possible if we assume that the modellers, analysts and audience all share the same knowledge of the ground truth. However, when the exploratory approach is beneficial, parts of the ground truth are not shared due to uncertainties (Bankes, 1993); therefore, presenting the model alongside the interventions should be chosen in our case.

On the other hand, an interesting idea about consolidative models is that shared ground truth has accuracy. An out-of-domain example is that using the Newtonian model to calculate the gravitational forces, it was possible to get to the Moon. However, the Newtonian model is incorrect compared to the theory of general relativity. The reason why the Newtonian model is still can be used because, within the spacecraft's operational parameters, the two theories result only in negligible differences. This implies that not all uncertainty is important to eliminate, even from a consolidative model.

Based on this idea, I propose that the consolidative and exploratory approach should not be viewed as exclusive techniques, as Bankes (1993) and section A.2 presents. Instead, I propose that these techniques should be viewed as two ranges on the spectrum of epistemic uncertainty. This idea of a spectrum fits into the conclusion of a chapter by Edmonds (2017, p. 56), where he discusses the different purposes of modelling: "There is a natural progression in terms of purpose attempted as understanding develops: from illustration to description or theoretical exposition, from description to explanations and from explanations to prediction". The key part is that as 'the understanding develops', epistemic uncertainty naturally decreases, leading to increased model accuracy (both in terms of representing the system and in the predictive sense).

For a moment, let us return to sub-research question one: *How can healthcare resource models be validated?* We actually already answered this question implicitly in the Methods chapter, but let it make explicit: Healthcare resource models should be validated as any other models: by examining whether the model is fit for purpose (Auping et al., 2022). At the detailedness of the SD models we were working with, due to the uncertainties, these models should not be validated by accuracy alone. Instead, these models should pass a set of tests on the three levels of validity (Barlas, 1996). An appropriate set of tests should be defined and executed based on the available resources, expertise, and software support. Placing the idea of validation into the just proposed spectrum perspective is also possible. As various validation tests compare the model with the real-world system, data about the real-world system is produced, which reduces epistemic uncertainty. Therefore in the spectrum perspective, validation means reducing epistemic uncertainty.

5.2 How to Present Model Outputs

We had an interesting idea while reading the literature: Consolidative and exploratory modelling (Bankes, 1993) fits together well with the model as an artefact view of Bolt et al. (2021). A consolidative model can be described as a technical representative object, given that it collects, unifies, and codifies the knowledge of multiple domain experts. On the other hand, an exploratory model is used to create new insights, classifying it as an epistemic object. While the differentiation between boundary and representative objects is not that straightforward, given the tendency of exploratory modelling to include domain experts from different disciplines to share their ideas about the inner mechanism of the system, an argument can be made that it is a boundary object. This especially holds if deep uncertainty is present, where it is, by definition, guaranteed that different ideas about the system mechanisms exist.

From this perspective: an added value of GMB is the process of creating a boundary object (i.e., the model), which enables cross-disciplinary communication (Bolt et al., 2021; Scott, 2018). Just like a model, the PANDEM-2 dashboard can also be classified as a boundary object. The users can interact with the dashboard to define scenarios, which can be shared with professionals from other disciplines. These scenarios maintain enough integrity because the underlying mathematics is immutable (i.e., the numbers do not change depending on which discipline looks at the graph). However, how the numerical data is interpreted depends on the discipline; therefore, in this section, some recommendations are presented for avoiding ambiguity while communicating healthcare resource models.

Based on Participants' Reaction to the Presentation of the Data, we think extra care should be taken to explain the difference between the different scenarios and respective time series, and model parts, as insufficient or bad communication can result in misunderstanding and wasted effort. People without an extensive data science background will have questions they lack the hands-on knowledge to answer. The root is not the often blamed STEM/non-STEM division but the discipline-specific knowledge in generating and analyzing data. In such situations, it is helpful if these questions can be answered immediately. The analyst who made the visualization is the best person to have around for this. Wrongly understood concepts can also be caught and corrected before moving on with the analysis. In an ideal case, the analyst can also 'on the fly' fine-tune the presentation of the data by explaining the required parts more in-depth. Furthermore, given some familiarity with the audience, it is possible to expect some of these questions. The answer to these questions could be prepared in advance and put as 'backup slides' (when presenting) or as 'frequently asked questions' (on the dashboard). However, it should be noted that documenting the correct part of the analyst's tacit knowledge is non-trivial, as it is highly audience-dependent which part of the analyst's knowledge needs to be documented. As there are potentially too many questions to answer, without senior-level domain experience, making these prepared answers should be expected to become an iterative process.

Moreover, we noticed that the findings deemed interesting by the analyst can be trivial for the decision-makers. For example, this happened in intervention 2 in the workshop (where increased testing causes quite a big contact tracing demand), but this was also observed at other workshops, where a model's results were communicated (W. Auping, personal communication, 18th October, 2022). This can be addressed by asking subject-matter experts beforehand if the finding is worth presenting. It could be the case that a 'novel' finding is, in fact, trivial to domain experts or exists as passive knowledge (i.e., it can only be recalled when reminded of it by a related topic). Despite this, even when the novelty of the insights gained from the pandemic model is negligible (such as in our case), the discovered scenarios were still good discussion starters in the workshop. This is likely the mechanism of the scenarios acting as a reminder for passive knowledge.

During the workshop, we evaluated the different presentation styles via a Likert scale (Table 4.1). Interestingly, interventions 1 and 4 scored substantially lower than interventions 2 and 3. The low score of intervention 4 is probably related to the misunderstanding about the bar graphs and the presence of 3 scenarios. However, interventions 1 and 2 are very alike, yet they received different scores. This is probably the result of ensuring participants understand intervention 1 in detail to be able to answer the following two guiding questions about it. At intervention 2, they benefitted from the understanding of intervention 1 as it was built on the same template. Based on these results, two presentation modes are easier to process: The first option is to analyze the graphs of the KPIs further and present the key insights in a tabular format. Alternatively, the second option is to build all visualization on the same template and explain that template on the first occurrence in detail. On the subsequent occurrences point out the interesting points, then give time to participants to process the information. However, it should be kept in mind that translation from the graphs to the tabular format could be cumbersome for professionals (section 4.1).

5.3 Needed Data for Situational Awareness

The end goal of participants during the workshop was pretty clear: completely eliminate the presented resource gap (section 4.3). They used the presented data to asses what actions they should take (and they concluded several times that actions not presented in the intervention are needed because the resource gap is not eliminated by the intervention alone). To achieve this, they expressed the need for data which could be used for operational planning purposes (section 4.2, and section 4.4). Therefore this thesis argues that this type of data is the best for participants to achieve situational awareness. However, this also requires an experimentally validated consolidative model that can be used as a surrogate of the modelled phenomenon.

Given this, sub-research question two can be answered as well: *How are, or can healthcare resource models be used?* Based on the workshop findings, when in possession of a healthcare resource model of sufficient predictive power, it gives the most utility, as participants actively look for data that they can use for operational planning purposes. This could be why in the second literature review, we have seen that a quarter of the models are autoregression models. In essence, an autoregressive model is a typical black-box model, preferring accuracy over explainability. However, this is not the only way to utilize healthcare resource models. Exploratory models can be created to help to address specific questions or problems, and subsequent analysis of these models could provide insight on how to develop risk-averse strategies or discover plausible worst cases (Bankes, 1993). Furthermore, a third way of utilizing models has been identified: to use them as a means of communication, either to codify knowledge or to be used as boundary objects. However, under uncertainty, the model building should happen with a specific problem or purpose in mind, as having multipurpose models, essentially encompassing the entire system, does not provide much utility (Bankes, 1993; J. Sterman, 2000).

5.4 Implications of the Need for Consolidative Models

We have seen that consolidative models would add quite a value. However, to achieve such models, datasets of high reliability are needed to validate the predictive power of the model, and as far as we know, no such datasets exist (M. Stein, personal communication, 16th October, 2022); therefore, these should be created. This idea is not far from the PANDEM-2 project, as WP2 was partially aimed to "aggregate surveillance data from multiple sources to provide useful surveillance indicators" to be presented on the dashboard (European Research Executive Agency, 2020, Annex 1 - p.14). However, these plans are aimed at data related to disease spread and not at healthcare resources. There is a possibility that data related to resources are already being collected for operational purposes; however, the collection and aggregation of this data is not happening, despite, in theory, hospitals' enterprise resource management systems could be queried relatively automatically.

Our best idea why this does not happen is that widespread data aggregation would need quite a significant upfront investment (S. Hinrichs-Krapels, personal communication, 29th November, 2022). Furthermore, collecting these data has far-reaching security and privacy implications, which need to be addressed. Fundamentally, there is a value traded-off between preparedness via data collection and privacy, indicating that it is not only an engineering problem. Another possible problem with data aggregation is that different organizations are likely using different methods to measure the same resources. A simple example would be that one hospital measures PPE in kits while the other measures it in gloves, masks, and so forth. Nevertheless, better situational awareness would address some limitations of both the refactored and the NUIG model, as it would significantly reduce the uncertainty related to healthcare resource management.

Now that both sub-research questions are answered, we can revisit the main research question of this study: *How to support healthcare resource managers in acquiring situational awareness via an SD model?*. This thesis argues that, by far, the biggest utility could be achieved by strengthening data collection and aggregation, as it enables the possibility to develop surrogate models. However, as this requires a significant upfront investment, question-driven exploratory models are an alternative way to address uncertainties.

5.5 Limitations and Future Research

We have encountered various setbacks in this study, and not all of these could be solved elegantly. Most of these impose some limitations on this study, which are collected in this section. Also, some ideas that are worth future examination are presented.

5.5.1 Limitations of the Literature Reviews

The literature review was not fully comprehensive, as is often the case on the timescale of a thesis. Multiple models in the grey literature are known to be relevant but were not returned by the search process and therefore

got, excluded. Also, the review was conducted without including relevant scientific frameworks (such as Barlas (1996)), which led to the interaction between these frameworks and the literature remaining unexplored.

Furthermore, the literature search might have a hidden bias towards hospital resources, as in the search terms we defined *healthcare resources* instead of *public health resources*, which seem to have a slightly different meaning. This difference might be connected to the fact that only 20% of the returned articles passed the abstract screening. However, this difference in the meaning was discovered months after performing the review; therefore, its effect remains an open question.

5.5.2 Limitations of the Refactored Model

We could not correct all of the original model's errors in the refactored model. This happened because we discovered far more errors in the original model than we initially expected. This also meant that the structure-oriented behaviour validity was not done thoroughly; the set of tests performed was not extensive by any measure. Furthermore, no well-established methodology was found regarding how to evaluate the validity of transient model behaviour, such as one-time resource depletion. Therefore the refactored model could still contain many undiscovered errors, which means that the model's utility is minimal. For example, it should not be viewed as an approximation or a description of the real-world system.

It was also discovered during the workshops that the model completely ignores the effect of self-testing, which is considered to be an important factor. In light of this finding, the chosen set of validation tests is questionable. An alternative approach would have been using boundary-adequacy tests to examine whether the important phenomenons have been all modelled (Auping, 2018) and to examine what could have been done to simplify the model.

There are also some practical limitations: Due to limited knowledge about the model, we decided to start with face-checking to understand it better. However, the lack of detailed knowledge about the real-world system transformed this into less of a test and more into learning based on the model. While the conceptual level of both the original and the refactored models are thought to represent reality, there are some structural errors resulting from either formalization errors or uncertainty. Furthermore, using unit testing instead of extreme-condition testing was made to speed up the validation process at the cost of having less thorough testing. Lastly, not all of the discovered errors and shortcomings were fixed.

What raises further questions is that Pace (2004) found that qualitative validation and verification assessments are not thought to be credible and repeatable but also noted that there is a lack of any large study backing up this notion. It was also noted that there is always an impression of room for improvement, but this improvement did not seem to happen over time. Though this article was published 18 years ago, it might not be longer relevant.

There is also the problem that many interesting ideas and scenarios would need extensive structural modification of the model to explore. For example, implementing selective testing and contact tracing would mean that the entire testing and contact tracing parts would need to be changed. Another example is that the effects of hospital capacity overflow are not modelled. Also, the disclaimer applies that comes with every exploratory model: the model should not be used for numerical predictive purposes.

In conclusion, due to these errors, neither the original nor the refactored model should be used anymore. Since the start of this thesis, a refined model was developed by NUIG, and the RIVM team working within the PANDEM-2 project does not have the capacity to maintain the model used in this thesis.

5.5.3 Repurposed Workshop

At the beginning of this study, we expected that another study would be conducted about how to present the data to end-users. Therefore, our first ideas about the workshop were very different, as the other study would have covered exploring the topic of data presentation. However, we decided to repurpose the workshop when we realized this would not happen. There were two problems with this approach. First, given the relatively late redesign, the literature about workshops could not be looked into in detail. Secondly, we could not organize the workshop in a way that both resource sides are represented, resulting in the fact that interaction between the participants from different backgrounds could not be observed. In retrospect, if we knew at the beginning of this study that the workshop would be about presenting model data, much of the effort used for understanding exploratory modelling in detail could have been directed to understanding participatory design techniques instead.

5.5.4 Faster Change than Model Development Speed

An interesting finding of the workshop is that the contact-tracing policy of the dutch healthcare system changed on the conceptual level since the development of the original model (section 4.4). This indicates that a model with the current detailedness could soon become outdated. However, there are no plans for continuous development to address these structural changes. Interestingly, this is also a limitation of the advocated data collection approach: nothing guarantees that information currently relevant for COVID will remain relevant for another strain or another disease. Also, at the resolution of these models, every disease has a specific structure (including patient pathway, spread mechanism, resources, and possible interventions). Therefore, the refactored model is only applicable for COVID-like illnesses, which effectively limits it to the different COVID strains.

There is one last limitation: In the SD paradigm, the dynamic of the system has to be explicitly modelled (for example, in agent-based modelling, the system behaviour is the result of the interaction between the agents, which is not explicitly coded into the model). However, in the workshop, it was indicated that the discovery of unknow-unknowns would be beneficial. These mechanics are easy to miss and would rarely be uncovered by validation since these *unknown* unknowns cannot be explicitly checked for. An example of this happened with the refactored model, as the lack of relationship between the number of infectious people and staff absenteeism went unnoticed for most of the validation process, while in reality, these are very related.

5.5.5 Future Research

Some ideas are presented in this thesis, which could be used as starting points for further research. Firstly, it should be explored whether the spectrum idea is worth further discussion by examining how other (non-SD) paradigms deal with uncertainty. Furthermore, the idea of the spectrum could be refined by looking at the model boundary in detail, for example, through a bull's-eye diagram.

Secondly, as situational awareness is a central part of the planned dashboard, it might be worth examining when healthcare resource managers perceive a resource gap or when it becomes significant. For example, a single missing mask is unlikely to cause a pandemic wave, but can 1000 missing masks do?

From another angle, during the workshop, RIVM participants expressed that they would organize 'something' to communicate with relevant stakeholders. It could be interesting to see how these communication channels will be affected by the opening of the new National Functionality for Infectious Disease Control (LFI) division of RIVM, which will also be tasked with future large-scale crisis response (National Institute for Public Health and the Environment, 2022a).

5.6 Chapter Summary

By discussing how to address uncertainty, we have concluded that SD models should not be validated by accuracy alone, but other levels of validity should be considered too. It was also discussed how to present the outputs of such models and that cross-disciplinary communication is one of the utilities provided by the dashboard. By examining the workshop results, we also conclude that model outputs are more easily understood when visualizations are based on the same template or when outputs are further processed and presented in a tabular format. When examining how can be healthcare resource models used, we saw that workshop participants would appreciate models that can be used for operational planning; therefore, the biggest utility could be provided by consolidative models. This comes with the implication that better data collection systems should be developed; however, as this requires quite an investment, exploratory models could be used as an alternative way to address uncertainties. Lastly, the limitations and ideas for future research are collected.

Chapter 6

Conclusion

Through this thesis, we have discovered many aspects of pandemic resource modelling. We started by understanding the problem that significant healthcare resource shortages make pandemics worse and that the importance of preparedness has long been recognized within the European Union. One of the several innovation projects in this domain is PANDEM-2, aiming to improve pandemic preparedness by creating cutting-edge digital tools for cross-border resource management and sharing. Therefore, we have explored how to support healthcare resource managers in acquiring situational awareness via an SD model.

To answer this question, two literature searches were conducted to examine the existing health resource models and how these are validated. It was found that the models found in the scientific literature are usually less detailed than the model examined in this thesis and that no common approach is followed for the validation process. After this, a snapshot of the literature about model building and validation was examined and subsequently used to inform the methodology. Following a framework proposed by the validation literature, it was decided to examine the model via multiple validation tests, and the model was validated for the purpose of producing outputs for the workshop. Using the model, multiple interventions were created, which were presented in different styles. In the workshop, participants evaluated how much the presented interventions were understandable and how these could support them in a decision-making position.

During the workshop, we discovered that the interventions should be communicated with more context: furthermore, presenting the model can be used for communicating assumptions. We have also seen that participants tend to augment the data with their past experiences. Furthermore, it was also discovered that participants are not satisfied with partial solutions (i.e., a smaller resource gap). Instead, they expressed interest in seeking out a solution where the resource gap is completely eliminated.

We have discussed that the type of situational awareness that would benefit workshop participants the most needs a consolidative model and that developing such models need better datasets. However, that requires a significant upfront investment and needs to be continuously maintained to address the changes of the real-world system. While healthcare resource modelling needs more data to be improved substantially, in the meantime, exploratory modelling can help by offering a way to address uncertainties.

References

Abdin, A. F., Fang, Y.-P., Caunhye, A., Alem, D., Barros, A., & Zio, E. (2021). An optimization model for planning testing and control strategies to limit the spread of a pandemic–the case of covid-19. *European journal of operational research*.

Abdolhamid, M. A., Pishvaee, M. S., Aalikhani, R., & Parsanejad, M. (n.d.). A system dynamics approach to covid-19 pandemic control: a case study of iran. *Kybernetes*. doi: 10.1108/K-01-2021-0038

Abramovich, M. N., Hershey, J. C., Callies, B., Adalja, A. A., Tosh, P. K., & Toner, E. S. (2017). Hospital influenza pandemic stockpiling needs: a computer simulation. *American journal of infection control*, *45*(3), 272–277.

Araz, O. M., Bentley, D., & Muelleman, R. L. (2014). Using google flu trends data in forecasting influenza-like–illness related ed visits in omaha, nebraska. *The American journal of emergency medicine*, *32*(9), 1016–1023.

Auping, W. (2018). *Modelling uncertainty: Developing and using simulation models for exploring the consequences of deep uncertainty in complex problems* (Doctoral dissertation). TU Delft, Faculty of Technology, Policy and Management.

Auping, W., d'Hont, F., van Daalen, E., Pruyt, E., & Thissen, W. (2022). *System dynamics - the Delft method.* (Distributed as lecture supporting material during the course: EPA1324 Introduction to TPM Modelling (2021/22 Q2))

Bankes, S. (1993). Exploratory Modeling for Policy Analysis. *Source: Operations Research*, *41*(3), 435–449. Retrieved from https://about.jstor.org/terms doi: https://doi.org/10.1287/opre.41.3.435

Barlas, Y. (1989a). Multiple tests for validation of system dynamics type of simulation models. *European journal of operational research*, *42*(1), 59–87.

Barlas, Y. (1989b). Theory and Methodology Multiple tests for validation of system dynamics type of simulation models *. *European Journal of Operational Research*, *42*, 59–87.

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, *12*(3), 183–210. doi: 10.1002/(sici)1099-1727(199623)12:3<183::aid-sdr103>3.0.co;2-4

Beishuizen, B., & et al. (2022a). *Identifying essential resources for pandemic response: an international delphi study within the eu-pandem-2 project.* (Manuscript in writing. I could not find the list of all authors for this work)

Beishuizen, B., & et al. (2022b). *A systematic literature review on health-care resources that can be included in pandemic response modelling.* (Manuscript in writing. I could not find the list of all authors for this work)

Berta, P., Paruolo, P., Verzillo, S., & Lovaglio, P. G. (2020). A bivariate prediction approach for adapting the health care system response to the spread of covid-19. *Plos one*, *15*(10), e0240150.

Bolt, T., Bayer, S., Kapsali, M., & Brailsford, S. (2021). An analytical framework for group simulation model building. *Health Systems*, *10*. Retrieved from https://doi.org/10.1080/20476965.2020.1740613 doi: 10.1080/20476965.2020.1740613

Brauer, F. (2008). Compartmental models in epidemiology. In F. Brauer, P. van den Driessche, & J. Wu (Eds.), *Mathematical epidemiology* (pp. 19–79). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-540-78911-6_2 doi: 10.1007/978-3-540-78911-6_2

Cakan, S. (2020). Dynamic analysis of a mathematical model with health care capacity for covid-19 pandemic. *Chaos Solitons & Fractals*, *139*. doi: 10.1016/j.chaos.2020.110033

Campillo-Funollet, E., Van Yperen, J., Allman, P., Bell, M., Beresford, W., Clay, J., ... others (2021). Predicting and forecasting the impact of local outbreaks of covid-19: use of seir-d quantitative epidemiological modelling for healthcare demand and capacity. *International journal of epidemiology*, *50*(4), 1103–1113.

Canese, K., & Weis, S. (2013). Pubmed: the bibliographic database. *The NCBI handbook*, *2*(1).

Clarivate. (n.d.). *Web of science core collection.* Retrieved from https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/

confidential presentation, U. (2018). *TDD bowling game.*

Council of the European Union. (2013). 1082/2013/EU of the european parliament and of the council of 22 october 2013 on serious cross-border threats to health and repealing decision no 2119/98/EC. *Official Journal of the European Union, L 293/1*. Retrieved from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:293:0001:0015:EN:PDF

Cui, Q., Qiu, Z., Liu, W., & Hu, Z. (2017). Complex dynamics of an sir epidemic model with nonlinear saturate incidence and recovery rate. *Entropy*, *19*(7). doi: 10.3390/e19070305

de Schipper, L. (2022a). *Pandemic resource modelling: facing the dutch omicron winter and summer of 2022.*

de Schipper, L. (2022b). *Technical documentation.*

Du, Z., Fox, S. J., Ingle, T., Pignone, M. P., & Meyers, L. A. (2022). Projecting the combined health care burden of seasonal influenza and covid-19 in the 2020–2021 season. *MDM policy & practice*, *7*(1), 23814683221084631.

Earnest, A., Chen, M. I., Ng, D., & Sin, L. Y. (2005). Using autoregressive integrated moving average (arima) models to predict and monitor the number of beds occupied during a sars outbreak in a tertiary hospital in singapore. *BMC Health Services Research*, *5*(1), 1–8.

Edmonds, B. (2017). Different modelling purposes. *Understanding Complex Systems*(9783319669472), 39–58. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-66948-9_4 doi: 10.1007/978-3-319-66948-9_4/TABLES/1

Ejigu, B. A., Asfaw, M. D., Cavalerie, L., Abebaw, T., Nanyingi, M., & Baylis, M. (2021). Assessing the impact of non-pharmaceutical interventions (npi) on the dynamics of covid-19: A mathematical modelling study of the case of ethiopia. *PLOS ONE*, *16*(11). doi: 10.1371/journal.pone.0259874

European commission. (n.d.). *Horizon 2020 - pandemic preparedness and response.* Retrieved from https://cordis.europa.eu/project/id/883285

European Research Executive Agency. (2020). *Grant agreement - number 883285 — PANDEM-2.* (Unpublished confidential document)

Forrester, J. W., & Senge, P. M. (1980). Tests for building confidence in system dynamics models. *TIMS studies in the management sciences*. Retrieved from https://www.albany.edu/faculty/gpr/PAD724/724WebArticles/ForresterSengeValidation.pdf

Freytag, P. V., & Young, L. (2018). *Collaborative research design.* Springer.

Galbraith, E., Li, J., Rio-Vilas, V. J. D., & Convertino, M. (2022). In. to. covid-19 socio-epidemiological co-causality. *Scientific reports*, *12*(1), 1–25.

Garcia-Vicuna, D., Esparza, L., & Mallor, F. (2022). Hospital preparedness during epidemics using simulation: the case of covid-19. *Central European Journal Of Operations Research*, *30*(1), 213-249. doi: 10.1007/s10100-021-00779-w

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. Retrieved from http://arxiv.org/abs/1512.03385

Ibarra-Vega, D. (2020). Lockdown, one, two, none, or smart. modeling containing covid-19 infection. a conceptual model. *Science Of The Total Environment*, *730*. doi: 10.1016/j.scitotenv.2020.138917

Ibrahim Shire, M., Jun, G. T., & Robinson, S. (2020). Healthcare workers' perspectives on participatory system dynamics modelling and simulation: designing safe and efficient hospital pharmacy dispensing systems together. *Ergonomics*, 1044–1056. doi: 10.1080/00140139.2020.1783459

Joulaei, H., Honarvar, B., Zamiri, N., Moghadami, M., & Lankarani, K. B. (2010). Introduction of a pyramidal model based on primary health care: A paradigm for management of 2009 h1n1 flu pandemic. *Iranian Red Cresent Medical Journal*, *12*(3), 224-230.

Kamerlin, S. C., & Kasson, P. M. (2020). Managing coronavirus disease 2019 spread with voluntary public health measures: Sweden as a case study for pandemic control. *Clinical Infectious Diseases*, *71*(12), 3174–3181.

Keeling, M. J., Hill, E. M., Gorsich, E. E., Penman, B., Guyver-Fletcher, G., Holmes, A., ... Tildesley, M. J. (2021). Predictions of covid-19 dynamics in the uk: Short-term forecasting and analysis of potential exit strategies. *PLOS COMPUTATIONAL BIOLOGY*, *17*(1). doi: 10.1371/journal.pcbi.1008619

KPI.ORG. (2022). *What is a key performance indicator (KPI)?* Retrieved from https://www.kpi.org/KPI-Basics/

Kuzdeuov, A., Baimukashev, D., Karabay, A., Ibragimov, B., Mirzakhmetov, A., Nurpeiissov, M., ... Varol, H. A. (2020). A network-based stochastic epidemic simulator: Controlling covid-19 with region-specific policies. *IEEE journal of biomedical and health informatics*, *24*(10), 2743–2754.

Leerapan, B., Teekasap, P., Urwannachotima, N., Jaichuen, W., Chiangchaisakulthai, K., Udomaksorn, K., ... Sawaengdee, K. (2020). System dynamics modelling of health workforce planning to address future challenges of Thailand's Universal Health Coverage. *Human Resources for Health*. Retrieved from

https://doi.org/10.1186/s12960-021-00572-5 doi: 10.1186/s12960-021-00572-5
Lempert, R. J., Popper, S. W., & Bankes, S. C. (2003). *Shaping the next one hundred years: New methods for quantitative, long-term policy analysis.* Santa Monica, CA: RAND Corporation. doi: 10.7249/MR1626
Liu, M., Cao, J., Liang, J., & Chen, M. (2020). A novel fpea model for medical resources allocation in an epidemic control. In *Epidemic-logistics modeling: A new perspective on operations research* (p. 143-166). doi: 10.1007/978-981-13-9353-2_8
Liu, M., & Zhang, D. (2016). A dynamic logistics model for medical resources allocation in an epidemic control with demand forecast updating. *Journal Of The Operational Research Society*, *67*(6), 841-852. doi: 10.1057/jors.2015.105
Mu, R., Wei, A., & Yang, Y. (2019). Global dynamics and sliding motion in a(h7n9) epidemic models with limited resources and filippov control [Article]. *Journal Of Mathematical Analysis And Applications*, *477*(2), 1296-1317. doi: 10.1016/j.jmaa.2019.05.013
Mu, R., & Yang, Y. (2018). Global dynamics of an avian influenza a(h7n9) epidemic model with latent period and nonlinear recovery rate. *Computational And Mathematical Methods In Medicine*, *2018*. doi: 10.1155/2018/7321694
Mugisha, J. Y. T., Ssebuliba, J., Nakakawa, J. N., Kikawa, C. R., & Ssematimba, A. (2021). Mathematical modeling of covid-19 transmission dynamics in uganda: Implications of complacency and early easing of lockdown. *PLOS ONE*, *16*(2). doi: 10.1371/journal.pone.0247456
National Institute for Public Health and the Environment. (2022a). *New crisis response organisation at rivm to control future pandemics.* Retrieved from https://www.rivm.nl/en/news/new-crisis-response-organisation-at-rivm-to-control-future-pandemics
National Institute for Public Health and the Environment. (2022b). *Quality of protective equipment insufficient during covid-19 crisis.* Retrieved from https://www.rivm.nl/en/news/quality-of-protective-equipment-insufficient-during-covid-19-crisis
National Library of Medicine. (2022a). *Mesh database - delivery of health care.* Retrieved from https://www.ncbi.nlm.nih.gov/mesh/68003695
National Library of Medicine. (2022b). *Pubmed user guide.* Retrieved from https://pubmed.ncbi.nlm.nih.gov/help/
Newell, S., Morton, J., Marabelli, M., & Galliers, R. (2019). *Managing digital innovation: A knowledge perspective.* Bloomsbury Publishing.
Nguyen, H. M., Turk, P. J., & McWilliams, A. D. (2021). Forecasting covid-19 hospital census: A multivariate time-series model based on local infection incidence. *JMIR Public Health and Surveillance*, 7(8), e28195.
Nikolic, I. (2021). *I. Nikolic PhD Thesis (Ch 3 and Appendix).* Distributed as lecture supporting material during the course: SEN1211 Agent-based Modelling (2021/22 Q2). Retrieved from https://brightspace.tudelft.nl/d2l/le/content/401504/viewContent/2239611/View
NL Times. (2020). *Hotline launched for healthcare workers with insufficient protective gear.* Retrieved from https://nltimes.nl/2020/10/18/hotline-launched-healthcare-workers-insufficient-protective-gear
Olivieri, A., Palu, G., & Sebastiani, G. (2021). Covid-19 cumulative incidence, intensive care, and mortality in italian regions compared to selected european countries. *International Journal Of Infectious Diseases*, *102*, 363-368. doi: 10.1016/j.ijid.2020.10.070
Operatív törzs. (2021). *Kórházi főigazgatóság: egészségügyi tanuló és végzettségü önkéntesek jelentkezését várják.* Retrieved from https://koronavirus.gov.hu/cikkek/korhazi-foigazgatosag-egeszsegugyi-tanulo-es-vegzettsegu-onkentesek-jelentkezeset-varjak (Hungarian government covid response team)
Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641–646.
Ørngreen, R., & Levinsen, K. (2017). Workshops as a research methodology. *Electronic Journal of E-learning*, *15*(1), 70–81.
Pace, D. K. (2004). Modeling and simulation verification and validation challenges. *Johns Hopkins APL technical digest*, *25*(2), 163–172.
PANDEM-2. (n.d.-a). *Pandem-2: Pandemic preparedness and response.* Retrieved from https://pandem-2.eu/
PANDEM-2. (n.d.-b). *Pandem-2 partners.* Retrieved from https://pandem-2.eu/rivm/
Pei, Z., Yuan, Y., Yu, T., & Li, N. (n.d.). Dynamic allocation of medical resources during the outbreak of epidemics. *ISEE Transactions On Automation Science And Engineering*. doi: 10.1109/TASE.2021.3102491
Picchiotti, N., Salvioli, M., Zanardini, E., & Missale, F. (2020). Covid-19 pandemic: a mobility-dependent seir

model with undetected cases in italy, europe and us. *arXiv preprint arXiv:2005.08882*.
Pierce, K. A., Ho, E., Wang, X., Pasco, R., Du, Z., Zynda, G., ... Meyers, L. A. (2020). Early covid-19 pandemic modeling: Three compartmental model case studies from texas, usa. *Computing in science & engineering*, *23*(1), 25–34.
Pierce, K. A., Ho, E., Wang, X., Pasco, R., Du, Z., Zynda, G., ... Meyers, L. A. (2021). Early covid-19 pandemic modeling: Three compartmental model case studies from texas, usa. *Computing In Science & Engineering*, *23*(1), 25-34. doi: 10.1109/MCSE.2020.3037033
Rijksoverheid. (2020). *Nog steeds te veel testaanvragen door mensen zonder klachten.* Retrieved from https://www.rijksoverheid.nl/actueel/nieuws/2020/08/26/nog-steeds-teveel-testaanvragen-door-mensen-zonder-klachten (Dutch government)
Rocha, R., Atun, R., Massuda, A., Rache, B., Spinola, P., Nunes, L., ... Castro, M. C. (2021). Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to covid-19 in brazil: a comprehensive analysis. *Lancet Global Health*, *9*(6), E782-E792. doi: 10.1016/S2214-109X(21)00081-4
Rouwette, E. A. J. A., & Vennix, J. A. M. (2020). Group Model Building. *System Dynamics*. Retrieved from https://doi.org/10.1007/978-1-4939-8790-0_264 doi: 10.1007/978-1-4939-8790-0_264
Roy, S., Dutta, R., & Ghosh, P. (2021). Towards dynamic lockdown strategies controlling pandemic spread under healthcare resource budget. *Applied Network Science*, *6*(1), 1–15.
Sarkar, S., Pramanik, A., Maiti, J., & Reniers, G. (2021). Covid-19 outbreak: A data-driven optimization model for allocation of patients. *Computers & Industrial Engineering*, *161*, 107675.
Scott, R. (2018). *Group Model Building Using Systems Dynamics to Achieve Enduring Agreement.* Springer. Retrieved from http://www.springer.com/series/11467
Senge, P. M., & Forrester, J. W. (1980). Tests for building confidence in system dynamics models. *System dynamics, TIMS studies in management sciences*, *14*, 209–228.
Shariatmadar, K., Wang, K., Hubbard, C. R., Hallez, H., & Moens, D. (2022). An introduction to optimization under uncertainty–a short survey. *arXiv preprint arXiv:2212.00862*.
Smith, B. A., Bancej, C., Fazil, A., Mullah, M., Yan, P., & Zhang, S. (2021). The performance of phenomenological models in providing near-term canadian case projections in the midst of the covid-19 pandemic: March–april, 2020. *Epidemics*, *35*, 100457.
Stein, M. L., Rudge, J. W., Coker, R., Van Der Weijden, C., Krumkamp, R., Hanvoravongchai, P., ... others (2012). Development of a resource modelling tool to support decision makers in pandemic influenza preparedness: The asiaflucap simulator. *BMC public health*, *12*(1), 1–14.
Steinmann, P., Auping, W. L., & Kwakkel, J. H. (2020, jul). Behavior-based scenario discovery using time series clustering. *Technological Forecasting and Social Change*, *156*. doi: 10.1016/J.TECHFORE.2020.120052
Sterman, J. (2000). *Business dynamics: Systems thinking and modeling for a complex world.* New York: McGraw.
Sterman, J. D. (1984). Appropriate summary statistics for evaluating the historical fit of system dynamics models. *Dynamica*, *10*(2), 51–66.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
Tembine, H. (2020). Covid-19: Data-driven mean-field-type game perspective. *Games*, *11*(4). doi: 10.3390/g11040051
The Local. (2020). *Geneva hospitals call for volunteers as covid-19 virus surges.* Retrieved from https://www.thelocal.com/20201025/geneva-hospitals-call-for-volunteers-as-virus-surges/
Thoring, K., Mueller, R., & Badke-Schaub, P. (2020). Workshops as a research method: Guidelines for designing and evaluating artifacts through workshops. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.
Tran, T. N.-A., Wikle, N. B., Albert, E., Inam, H., Strong, E., Brinda, K., ... others (2021). Optimal sars-cov-2 vaccine allocation using real-time attack-rate estimates in rhode island and massachusetts. *BMC medicine*, *19*(1), 1–14.
Tsioptsias, N., Tako, A., & Robinson, S. (2016). Model validation and testing in simulation: a literature review. In *5th student conference on operational research (scor 2016)*.
TU Delft. (2020). *Typologie van modellen.* Retrieved from https://sysmod.tbm.tudelft.nl/wiki/index.php/Typologie_van_modellen#Black_box_versus_white_box
van der Wal, N., & Nikolic, I. (2022, 02 12). *SEN1211: Agent-based Modelling (2021/22 Q2): ABM Formalisation.* Distributed through Brigthspace (Electronic Learning Envrionment).
Vecchi, V., Cusumano, N., & Boyer, E. J. (2020). Medical supply acquisition in italy and the united states in the era of covid-19: The case for strategic procurement and public–private partnerships. *The American*

Review of Public Administration, *50*(6-7), 642–649.

VENTANA systems inc. (2022a). *Vensim help - checking for model syntax and units errors.* Retrieved from https://www.vensim.com/documentation/20570.html

VENTANA systems inc. (2022b). *Vensim help - reality check.* Retrieved from https://www.vensim.com/documentation/usr14.html

Verma, V. R., Saini, A., Gandhi, S., Dash, U., & Koya, S. F. (2020). Capacity-need gap in hospital resources for varying mitigation and containment strategies in india in the face of covid-19 pandemic. *Infectious Disease Modelling*, *5*, 608-621. doi: 10.1016/j.idm.2020.08.011

Vierlboeck, M., Nilchiani, R. R., & Edwards, C. M. (2020). Systems approach to localize tipping points for the emergency services in face of the covid-19 pandemic. In *2020 6th IEEE international symposium on systems engineering (IEEE isse 2020).* (6th IEEE International Symposium on Systems Engineering (IEEE ISSE), Electr Network, Conference Proceedings)

V&VN. (2020). *Peiling v&vn: tekorten maskers houden aan, psychische druk hoog.* Retrieved from https://www.nu.nl/coronavirus/6164661/is-de-nederlandse-ic-capaciteit-zoveel-beperkter-dan-die-elders-in-europa.html

Walker, W. E. (1982). Models in the Policy Process: Past, Present, and Future. *Interfaces*, *12*(5), 91–100.

Wang, A., Xiao, Y., & Zhu, H. (2018). Dynamics of a filippov epidemic model with limited hospital beds. *Mathematical Biosciences and Engineering*, *15*(3), 739-764. doi: 10.3934/mbe.2018033

Wang, X., Li, Q., Sun, X., He, S., Xia, F., Song, P., ... others (2021). Effects of medical resource capacities and intensities of public mitigation measures on outcomes of covid-19 outbreaks. *BMC public health*, *21*(1), 1–11.

Watson, O. J., Alhaffar, M., Mehchy, Z., Whittaker, C., Akil, Z., Brazeau, N. F., ... others (2021). Leveraging community mortality indicators to infer covid-19 mortality and transmission dynamics in damascus, syria. *Nature communications*, *12*(1), 1–10.

Weissman, G. E., Crane-Droesch, A., Chivers, C., Luong, T., Hanish, A., Levy, M. Z., ... Halpern, S. D. (2020). Locally informed simulation to predict hospital capacity needs during the covid-19 pandemic. *Annals of Internal Medicine*, *173*(1), 21+. doi: 10.7326/M20-1260

Wikle, N. B., Tran, T. N.-A., Gentilesco, B., Leighow, S. M., Albert, E., Strong, E. R., ... others (2022). Sars-cov-2 epidemic after social and economic reopening in three us states reveals shifts in age structure and clinical characteristics. *Science advances*, *8*(4), eabf9868.

Wood, R. M., McWilliams, C. J., Thomas, M. J., Bourdeaux, C. P., & Vasilakis, C. (2020). Covid-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive care. *Health Care Management Science*, *23*(3), 315-324. doi: 10.1007/s10729-020-09511-7

Yanez, A., Duggan, J., Hayes, C., Jilani, M., & Connolly, M. (2017). Pandemcap: Decision support tool for epidemic management. In *2017 ieee workshop on visual analytics in healthcare (vahc)* (pp. 24–30).

Yin, L., Zhang, H., Li, Y., Liu, K., Chen, T., Luo, W., ... others (2021). A data driven agent-based model that recommends non-pharmaceutical interventions to suppress coronavirus disease 2019 resurgence in megacities. *Journal of the Royal Society Interface*, *18*(181), 20210112.

Zhao, X., Li, W., Wang, Y., & Jiang, L. (2021). Evaluation of the number of visits to chinese medical institutions using a logistic differential equation model. *Complexity*, *2021*. doi: 10.1155/2021/7943651

Appendices

Appendix A

Additional methodology

A.1 Overview of the Scientific Theories Presented in this Thesis

Given that many ideas from many topics of the scientific literature was merged together in this thesis, it could be helpful to have a visualization to organize these. This visualization is given in Figure A.1. The map is created by exmaining the relationship between the practical perspective of the tasks of this thesis, the steps of the modelling cycle, and where the relevant theories fit into this.

The colour coding differentiates this thesis's planned tasks from the steps already done. The prussian blue coloured bocks are the tasks already done, while the red ryb blocks are the tasks planned accomplished during this thesis. The continuous arrows represent the next step in the processes, while the dashed arrows represent an association between the diferent topics, though some of these associations are presented only in the discusion chapter.

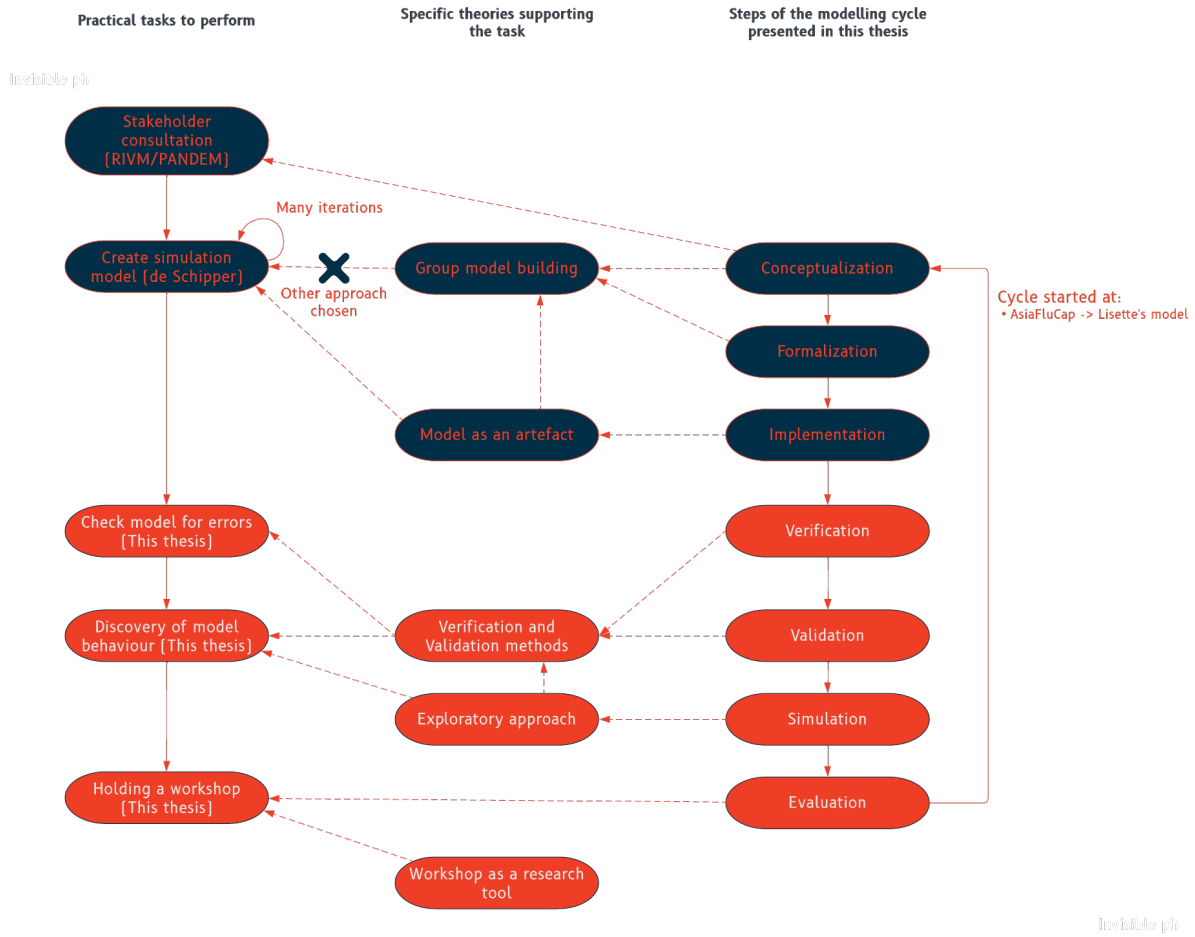


Figure A.1: Visulization of the workflow this thesis is part of.

One thing that is not evident, is the 'x' between the GMB approach, and the creation of the simulation model: While there are some recommendations in literature to use GMB to inform concleptualization and formalization, it remains slightly detached, as the approach taken by de Schipper (2022a) is more of a 'freestyle' approach. As she developed the original model, she showed parts of it in the periodical internal WP meetings. In these meetings, she asked for the opinion of the present domain experts and stakeholders of the project. She also consulted the scientific literature to get inspiration for further directions regarding model development (M. Stein, personal communication, 30th June, 2022). While this approach consists of similar elements, it differs from the GMB approach. Furthermore, Rouwette and Vennix (2020) specifically warns against having the dual role of the facilitator and the expert, which in my opinion, is a somewhat similar situation to the one that happened in the original model's freestyle approach.

A.2 Difference Between Consolidative and Exploratory Modelling

Uncertainty is a fundamental problem in modelling, and there are two well-established approaches, depending on the level of epistemic uncertainty. This two different approaches are: exploratory-, and consolidative modelling. To start, *exploratory modelling* is described as:

> When insufficient knowledge or unresolvable uncertainties preclude building a surrogate for the target system, modellers must make guesses at details and mechanisms. While the resulting model cannot be taken as a reliable image of the target system, it does provide a computational experiment that reveals how the world would behave if the various guesses were correct. (Bankes, 1993, p. 435)

On the other hand, under *consolidative modelling*, Bankes (1993) referred to models that can predict behaviour reliably enough to be used as a surrogate for the modelled system itself. These models are possible to build when enough knowledge about system characteristics is accumulated. This implies that there are no conflicts between the different theories about the system, and these can be combined into a single *internally consistent* theory. These models consolidate a large amount of knowledge and information into a single formal model, which can be implemented as a computer program. After validation, the program can be dispensed, including the information used in its construction.

The key concept to differentiate the two approaches is the presence of unresolvable uncertainties within the model boundaries. These uncertainties could refer to the fact that the modellers did not have the resources to resolve some of the key uncertainties. However, in more challenging cases, this could also mean the presence of *deep uncertainty*. This latter case refers to the situation where the stakeholders within the modelled system do not know or do not agree on how the system works (Lempert et al., 2003).

The consolidative approach encompasses simple models, like Newton's laws of motion, or complex models, like a finite element method, in engineering. What is common in these consolidative models is that both were constructed by combining a vast amount of past research with the latest 'computational' capabilities. In particular, Newton was a pioneer in calculus, and the engineers behind a finite element method pioneered harnessing the computer's power to solve mechanical calculations. However, all consolidative models require extensive experimental validation after their construction, which is not always possible due to ethical, legal, or financial considerations.

There are possible trade-offs. For example, uncertainties can be battled by spending extreme amounts of money. A case like this is the Standard Model (in particle physics) and the CERN facility, where the underlying uncertainties emerge from the theory using probability functions to describe the real-world phenomenon, and CERN battles that by running a vast amount of experiments required to determine p values over 99.99%. Another example is found in gravitational waves and the LIGO facility. In this case, the measurement instrument (interferometer) is so sensitive that it picks up minimal vibrations, such as ones caused by a truck breaking nearby the facility. This means that from a single measurement alone, it is impossible to determine whether it was a gravitational wave or the truck nearby; therefore, an active vibration-dampening system was built along the kilometres of the vacuum system. However, in both cases, the detector's reliability was improved, not the complexity of the tested model, which is the difference between these examples and Bankes's "ultimate combat simulator". Also, the important idea behind all models is to augment human knowledge and never is to replace it.

In Bankes's view, when extensive experimental validation of the model is not possible, an exploratory approach should be taken. This approach provides value by improving the modelled system's insight, guiding future analysis plus data collection and generating hypotheses to test. Further uses were identified by (Rouwette & Vennix, 2020): An exploratory model can be used to help decide in situations when the human mind is simply incapable of processing the vast amount of information about the system. In this case, while the model is unlikely to be true in every detail, the output could still be used to provide better decisions than guessing alone, or the output could aid risk aversion by pointing out the worst possible scenarios. The model could also be used to search for strategies where a little investment could lead to large returns.

One common pattern along all these exploratory uses is that the model is never used to generate explicit answers or predictions of the real-world system, but it is used to uncover new pieces of information, which help to make an informed decision. In other words, these models can be used to discover implications of what is known or to examine hypothetical ('what if') scenarios to improve our insight into the problem or decision.

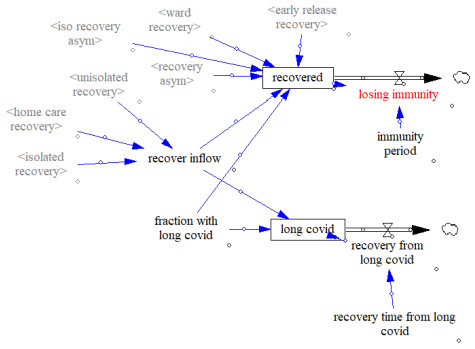
A.3 Errors in the original model

The following errors were identified during verification:

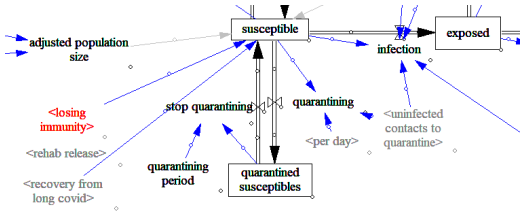
- de Schipper (2022b) conducted a sensitivity analysis of the original model along a few variables. While not mentioned explicitly in her report, there are two parameters which exhibit a strange behaviour: both 'willingness to quarantine' and 'vaccination per staff per day', albeit negligibly but positively correlated with the variables used to describe the seriousness of the disease ('deceased', 'total infections', 'admissions to ICU'). This is counter-intuitive; therefore, further investigation should be conducted. Unfortunately, there was no time to re-do the sensitivity analysis on the refactored model.
- Parts of the model were still under construction, and it was not indicated by anything. For example, the resource usage of long-covid was not modelled at all. Furthermore, there were several parts of the model where key parameters were set to placeholder values. This was the case for all of the home care and rehabilitation resources. These 'aftercare resources' were deleted from the model. Placeholder values also caused a vaccination capacity of 100 people/day, which is unrealistic in light of the modelled population size of more than 10 million people. In some cases, the placeholder value was zero, effectively turning off that part of the model. This was the case in modules related to quarantining, early release, rehab, PPE acquisition, medication, and noise testing demand. Besides the medication module, these modules got turned on by searching for a suitable parameter value.
- After turning the aforementioned modules on, people started to disappear over time in the model. After a thorough investigation, an error was found in the implementation of the rehabilitation stock, which was corrected.
- On the conceptual level, the case-fatality ratio determines the percentage of people surviving the disease despite hospitalization. This part got formalized using two first-order material delay out-flows from the ICU stock, with different delay times (surviving and deceased). When the coefficient for the material delay is different, the case-fatality ratio needs to be modified on the formalization level to account for these different delay times. However, this was not noticed in the original model, which resulted in the mortality ratio during simulation not being equal to the case-fatality ratio. An effort was made to correct this, but it was not completed due to the lack of time.
- For the epi-compartments, the original model utilized a very unorthodox method. Instead of a direct compartment-to-compartment flow, the model passes the value of the flow through an auxiliary variable, and the flow itself goes outside of the modelling boundary into the cloud symbol in Vensim. This is demonstrated on Figure A.2, where it is highlighted in red for better visibility. This unorthodox method was one of the main arguments for refactoring the model.
- The original model has a specific, directional hospital structure, but an alternative structure was identified during the NUIG modelling efforts. The alternative structure models a bi-directional flow between the ward and ICU, which was deemed to be more realistic than the structure in the original model (Caroline Green, personal communication, 15th June, 2022). This is one manifestation of the deep uncertainty which surrounds the model. While this specific structural uncertainty can be simplified into parameter uncertainty, it well demonstrates the level of uncertainty regarding the model.
- Throughout the NUIG modelling efforts, it was found that the whole hospital system behaves like a CAS. Multiple layers of fall-back strategies can be activated in case of a severe resource shortage. In essence, the hospital can decrease the quality of care in exchange for increased capacity. (Caroline Green, personal communication, 15th June, 2022). This indicated that many assumptions of the original model were not documented, if they were known at all. While the NUIG model made quite some progress towards this direction, the lack of assumptions was not addressed in the refactored model.
- The calculation of the R number seems to be wrong. There might be a mathematical equivalency overlooked, but this was not investigated thorougly during this thesis.
- The '"patient-to-bed ratio"' parameter should be only used to model the situation when only a percentage of the hospital is dedicated to covid care. This was a minor incorrectly documented part in the original model's report.

There were some concerns identified related to deviating from the SD modelling paradigm. These are the following:

- Auxilary variables were directly affecting a stock. An example of this is visible on Figure A.2a, where the arrows from the auxiliary variables point directly to a stock instead of pointing to a flow.
- Two variables were incompatible with the SD paradigm: the 'stock to flow converter' and the 'per day'. The former was used to ensure that more resources cannot be used in a single timestep than the entire stock. However, this was implemented erroneously (Andrade Ortiz, Jair Albert, personal communication, 23rd March, 2022). In the refactored model, there is an alternative approach: There are material delays with very small coefficients (i.e. a few hours), limiting the out-flow from these stocks. These coefficients follow the naming template: '[resource_name] emergency emptying time'. The 'per day' variable was deleted because it was not affecting the results numerically and was only used to ensure the unit check tests were passed. The refactored model passes the unit check without this shenanigan.
- Given that the original model was reported to have compatibility with python, the following should also be classified as an error: There were parameters that in Vensim acted like variables because there was a constant equation to calculate their value instead of a single numerical value. This caused some unhandled exceptions while importing the model through the EMA Workbench and had to be corrected.
- There was another implementation error where the ICU capacity depended on the ICU demand. This error was not causing numerical differences but decreased model understandability, which was especially painful considering the low quality of the documentation.
- There is a shenanigan introduced by the refactoring process: There was a falsely indicated circular equation resulting from Vensim's numerical approach to subscripts. As there is no circular equation in the symbolic equations, this got a quick fix via the 'symptomatic nonhospitalized manual sum shadow' variable. This could be resolved more elegantly, but that would take quite some time.



(a) Recovered compartment in the original model



(b) Susceptible compartment in the original model

Figure A.2: Demonstration of the unconventional pass-on auxiliary variable method.

During the face-checking, many implicit assumptions were discovered. Despite the assumptions already documented in the original report, the following assumptions were discovered. However, as none of the V&V steps was directly aimed at discovering hidden assumptions, this list is certainly incomplete.

- When administering multiple jabs of vaccination, the time delay between the jabs is not modelled. This simplifies the modelling of the initial two covid jabs but falls short in modelling the subsequent booster shots.
- A 100% test specificity and a 100% test sensitivity are assumed.
- The 'time onset of symptoms' cannot be smaller than the 'look-back period of ct'. If that happens, the contact tracing will find people who are already infectious, which is not accounted for in the model. It is assumed that every contact found is either in the susceptible or the exposed compartment.
- The model was built with the January 2022 Dutch policies in mind. Furthermore, no other hospital was consulted during the development process despite RUMC.
- One staff visit of a patient in the ward or the ICU uses up one kit of PPE equipment.

- Staff absenteeism is not affected by the number of infected people.
- Two parameters: 'staff visits per patient per day' and 'ICU patient-to-staff ratio' are not independent. This dependence is not accounted for in the model.
- There was an effort to account for different 'staff-to-patient ratio' for each shift. This was disadvantageous for the model clarity, as demonstrated by Figure A.3. The correct approach would have been to subscript the entire hospital resource model with the type of shift. As this was deleted, the refactored model assumes that every shift is the same.
- There is no clear documentation of who and when needs to be tested in the model. An attempt was made to reverse-engineer the logic based on the model equations. However, this was not completed due to the time limit constraints of this thesis work.
- It was documented in the original model "It is assumed susceptibles do not meet multiple infectious people per day" de Schipper (2022b). However, it should be added that when this assumption does not hold (e.g. during a peak), the contact tracing module produces an invalid output, leading to over-isolating people. There is no quick fix or indicator when this problem happens; that part should be corrected.

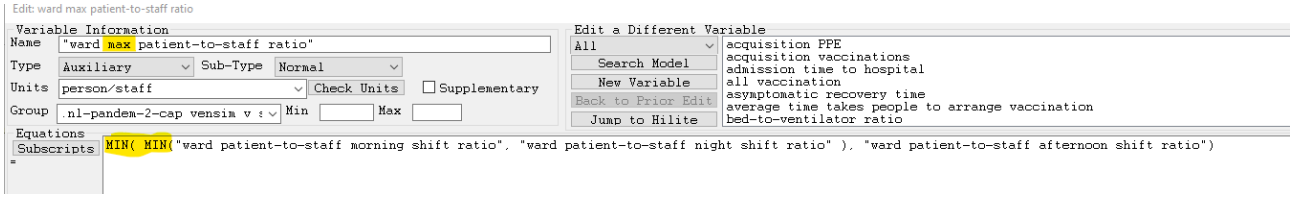


Figure A.3: Equation behind the '"ward max patient-to-staff ratio"' variable.

There were several concerns that could be best described as technical debt or bad codebase health. The size of this technical debt was so large that if we had been fully aware of it during the initial thesis planning, we would undoubtedly have reconsidered the plan's feasibility. This technical debt was manifesting itself in the following forms:

- The stock-flow diagrams were unorganized to the extent that it was time-consuming to find variables.
- While the meaning of the parameters is documented, this is not the case for the variables. This resulted in the need to reverse-engineer many parts, to understand the ideas behind the relationships of the variables. Furthermore, the naming style of variables is inconsistent within the model itself. This issue, paired with the lack of proper documentation, made working with the model extremely difficult. As a result, quite some effort went into understanding and renaming the variables, but this process is far from complete. A full documentation would be needed, but that is not realistic to include in this thesis.
- There were variables in the model which were not used to calculate the output and were not documented anywhere. These variables are deleted in the refactored model.
- Sometimes, basic built-in Vensim functions were implemented manually (e.g. XIDZ, ZIDZ), obscuring the model relationships.
- Assumptions were only documented in the original model's report. Equipped with some retrospective enlightenment, indicating these assumptions in the stock-flow diagrams would have accelerated refactoring.
- Some variables are not calculated on the same stock-flow diagram (view in Vensim) as similar variables. These were mostly cleared up; however, calculating the vaccination resources still happens in the compartment sub-model.
- As already mentioned above, there was no notice of which parameters had a placeholder value.
- Access was not obtained automatically for all of the project files related to the work of the original model. When access was finally obtained, it was found that the original model also used an excel file to store the parameter list. Unsurprisingly, this list was slightly different from the parameter values found within the model and the parameter values found in the technical documentation.

In brief, the documentation of the whole project was abysmal. Both the conceptual and formal levels of the documentation left a lot to be desired, which resembles the hypothetical situation described as the *ultimate combat model* by Bankes (1993). This lack of reporting made building on the original research outputs incredibly difficult.

With the same scrutiny as above, there are two notable moments where the refactoring process fell short.

- The code used to attach the model to the Vensim DLL is full of technical debt. There were no attempts to clear this, as the direction of automated scenario discovery was abandoned.
- The folder structure of the project has some unintuitive decisions. Furthermore, the files were stored in a way which carries the risk of not granting automatic access to all project files.

A.4 Description of Group Model Building

A fundamental problem needs to be addressed in any modelling task: How to measure and quantify the real-world phenomenon the model is about? The solution to this question is highly discipline-specific. This section will present the approach that seems to work best for recent SD modelling tasks.

In SD modelling, the 'measurements' of the real-world system happen in the conceptualization and formalization steps. It is also important to recognize that SD usually deals with problems in *Large Scale Socio-Technical Systems*. These systems have many properties, and corresponding descriptions (Nikolic, 2021). However, there are two problems in particular to overcome: There is no central control, as the individual agents' decisions collectively determine system behaviour in a non-trivial manner. The other problem is that every agent only knows part of the system in detail (i.e., the part they are actively engaged with) and, with less precision, some neighbouring parts.

While collecting data artefacts produced by agents might be possible in some settings (e.g. the logs of an enterprise resource planning system), these are usually confidential and partial. Therefore using a purely data-oriented approach is not plausible to overcome these problems. In this case, the best instrument to measure the real-world system is a qualitative approach: asking agents to share their tacit knowledge while examining the accessible artefacts of the system. Furthermore, SD is rarely used only for the sake of building a model; it is usually used as a method to solve a problem, usually at the request of the agents within the system.

Analyzing the gap between the expectations and results of models applied for aiding the policy-making process, Walker (1982) found that modellers learned more about the system during modelling than the policy-makers, who were supposed to learn about the system. Furthermore, the confidence of decision-makers in the models' results was relatively low, as the policymakers were excluded from most parts of the modelling process.

Rouwette and Vennix (2020) described that the SD modeller community reacted to these problems by experimenting with involving clients beyond the problem definition. Many of these efforts resulted in the approach called Group Model Building (GMB), which emphasizes client involvement more than earlier approaches: GMB utilizes client involvement in the entire conceptualization step and in parts of the formalization step too. When done right, the participants included in these steps should feel ownership over the developed model, leading to higher confidence in model outputs, as well as a better understanding of the model limitations, leading to fewer unused or misused models. In practice, this client involvement is usually achieved through interviews and workshops. While some workshops reportedly had around 40 participants (Leerapan et al., 2020), GMB generally utilizes a group size of 5-10 people (Bolt et al., 2021).

A.5 Scenario discovery

There was a research direction we made substantial progress towards but did not contribute towards the final output: Behaviour-Based Scenario Discovery (BBSD). It is a relatively novel way within the topic of scenario discovery (Steinmann et al., 2020). To understand the novelty of BBSD, we must first understand *conventional scenario discovery*, which is a method aiming to derive decision-relevant future scenarios from exploratory model behaviours. Conventional scenario discovery consists of 3 sequential steps. First, conducting experiments over the uncertainty space, then reducing each model output to a single value, and finally finding the regions in the uncertainty space where the model output is within an 'interesting region'. While outside the jargon of the SD modelling, in operational management, the process of reducing complex system behaviour into quantifiable indicators, as an analytical basis for decision-making is called *defining Key Performance Indicators (KPIs)* (KPI.ORG, 2022). Therefore the reduced model outputs will be referred to as KPIs in this thesis. BBSD introduces a variation into the second step of conventional scenario discovery. Instead of classifying model outputs into two classes (i.e. interesting and not interesting), it classifies those into n classes, using a user-defined KPI and time-series clustering.

As a KPI is a quantified property, combining multiple KPIs into a single KPI is also possible by defining a (mathematical) metric. In practice, this means defining how much an increase in one KPI can offset the decrease in another. For example, in the case of an epi-model, deciding how much an increase in ward-bed capacity cancels the effect of a decrease in a single testing capacity is an act of defining a metric. As demonstrated by this example, defining such a metric is a non-trivial process. While defining KPIs is possible for the model (as elaborated in section 7), defining the metric would have taken extra time due to the surrounding deep uncertainty and likely ethical value clashes.

It is now evident that in scenario discovery, reducing the model output into a single KPI is essential for the automated comparison of runs. However, this would have taken extraordinary effort; therefore, the BBSD approach was abandoned alongside every *automated* scenario discovery approach. As our initial plan was to perform BBSD on the model, a code was written to connect the model with the EMA Workbench python library. This resulted in a minor contribution towards the library, by finding the line that caused a runtime error: `https://github.com/quaquel/EMAworkbench/issues/155`. Furthermore a plan was created to perform BBSD. This can be seen in Figure A.4

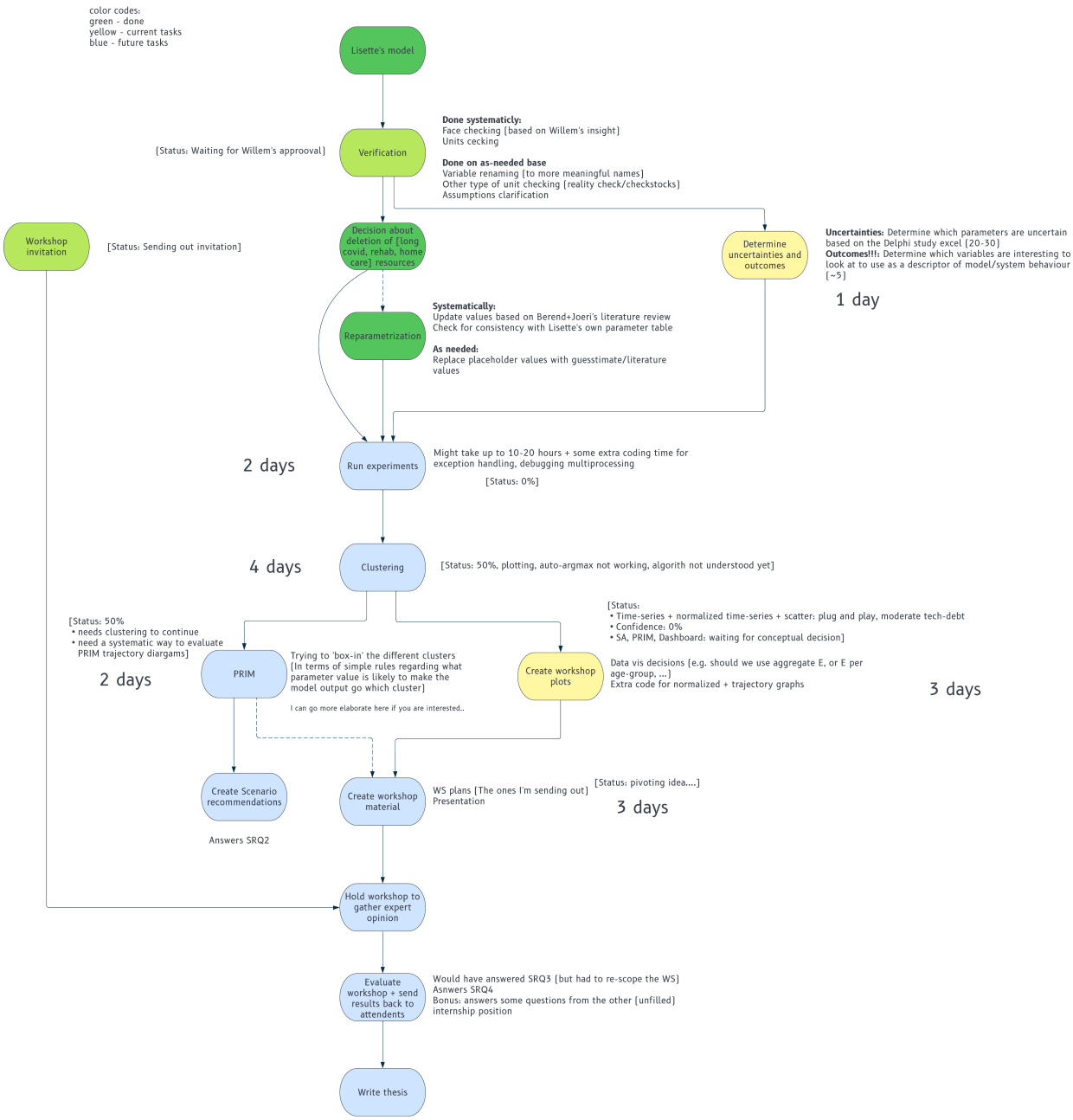


Figure A.4: Abandoned BBSD planning