Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Estimating the frequency of infectious diseases in heterogeneous populations: methodological overview and application to the COVID-19 outbreak in Belgium

Kristien Verdonck

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Epidemiology & Public Health Methodology

SUPERVISOR :

Prof. dr. Christel FAES

2019
2020

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics

Master's thesis

Estimating the frequency of infectious diseases in heterogeneous populations: methodological overview and application to the COVID-19 outbreak in Belgium

Kristien Verdonck

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Epidemiology & Public Health Methodology

SUPERVISOR :

Prof. dr. Christel FAES

In memoriam

Marleen Boelaert (21 October 1960 - 12 June 2020), professor of epidemiology and of life

Acknowledgements

In the first place, I want to thank my thesis supervisor, Prof. Christel Faes, for patiently allowing me to wander and at the same time keeping me on track. I also thank the other lecturers at UHasselt for generously sharing their most recent insights.

Ideas have many parents, some of whom I would like to name here. Koen Peeters, Larissa Otero, Katja Polman, and Bruno Marchal: many thanks for the motivating conversations of the past years. I really hope that there are many more conversations to come, in corridors and stairwells, bars and restaurants, stations and airports, over coffee, beer, pisco, ginger thee, or crodino.

I am very grateful to my colleagues in Belgium and abroad and to my head of department who generously put up with my recurrent UHasselt-related disappearances of the past four years. I also thank the coordination team and the students of the Master of Public Health at the Institute of Tropical Medicine in Antwerp for taking me along to a real world between global and local. In that world, heterogeneity is a lived experience.

Finally, I dedicate this work to the people behind the models and the dots, who suffer from infectious diseases and from control measures against infectious diseases.

Abstract

Background. Knowing the frequency of infectious diseases in human populations is essential for disease surveillance and for monitoring the impact of disease control interventions. The ongoing outbreak of coronavirus disease 2019 (COVID-19) illustrates that estimating disease frequency is not a simple task. Statistical and mathematical models play an important role in public health decision making. Challenges in the formulation of such models often relate to questions on how to deal with uncertainty and with heterogeneity (sometimes referred to as diversity or variability) in human populations.

Research questions. The core theme of this thesis project is the estimation of disease frequency in heterogeneous populations. This is explored from three angles: (i) How is heterogeneity addressed in recent literature about frequency of infectious diseases? (ii) How can statistical models describing the early growth of an outbreak take up heterogeneity? (iii) How can considerations about heterogeneity be taken up in growth models for the COVID-19 outbreak in Belgium?

Methods. This project contains a literature review and a data analysis. Two frameworks are used to clarify concepts throughout the entire project: one about variability and uncertainty and the other about mechanistic and phenomenological modelling approaches. The review starts with a broad overview of 15 papers published since August 2019 about infectious diseases in heterogeneous populations. The second part of the review is more technical and describes how one phenomenological approach, the generalised growth model for the early phase of an outbreak, can capture heterogeneity. This approach is then applied in a data analysis concerning the COVID-19 outbreak in Belgium in 2020.

Findings. Tropical infectious diseases are prominent in the literature about heterogeneity. Heterogeneity is addressed using diverse, often mechanistic, methods. Approaches to define relevant subpopulations can be classified as pragmatic, deductive, or inductive. Insight in what makes populations heterogeneous for specific infections helps to design interventions, to reach key individuals or subgroups, and to estimate the impact of interventions. A generalised growth model can accommodate heterogeneity by allowing overdispersion, sub-exponential growth, and different growth patterns across population subgroups. In the case of COVID-19 in Belgium, models allowing for overdispersion and sub-exponential growth fit the incidence data better than simpler models. Growth patterns are similar across subgroups based on age, sex, and province.

Discussion. Equidispersion and exponential growth are common assumptions for incidence data during outbreaks. In the case of COVID-19 in Belgium, these assumptions are not met. Assuming exponential growth when, in fact, growth is slower than exponential typically leads to an overestimation of epidemic size and an overestimation of the impact of control interventions. Models that account for heterogeneity typically contain more parameters than their simpler counterparts. However, taking up a few extra parameters can still be relatively straightforward, as illustrated in this work. Approaches incorporating heterogeneity also produce results surrounded by more uncertainty than simpler methods. This may look unattractive at first sight, but information will only be useful for public health if it reflects the true uncertainty and variability that is associated with the presence of infectious diseases in a population.

List of Tables

List of Figures

Table of Contents

1 Introduction

1.1 Estimating the frequency of infectious diseases

"*Infectious disease is one of the few genuine adventures left in the world.*

The dragons are all dead and the lance grows rusty in the chimney corner."

Hans Zinsser, in Rats, Lice and History (1935) (1).

Disease frequency is essential information in the field of public health; it is the first pillar of epidemiology (2). Especially in vertical health programmes, which focus on one or a few conditions, it is important to know how many people have that condition. To be able to properly organise health care services, policy makers need to have a good idea of which infrastructure, how many human resources, and what amount of diagnostic and treatment materials are required (3). Information about disease frequency also plays a role in the development of disease control interventions and in the evaluation of their impact. Indeed, disease frequency turns up in diverse public health scenarios such as detecting outbreaks, organising patient care, identifying determinants of disease, developing interventions, and eliminating diseases in entire populations.

Measuring disease frequency, even though this is expressed as a simple count, proportion (prevalence), or rate (incidence), is not a simple task. To directly measure disease frequency requires a representative study population, clear disease definitions, and accurate diagnostic procedures (2). In the field of infectious diseases, it is often so that only some of the people who are exposed to a microorganism get infected, and that only some of the infected get ill -after variable periods of time. In addition, diagnostic tests are never 100% accurate in real-life practice, and some assays only work in people who are already ill. Furthermore, for infectious diseases that are transmitted easily or are of short duration, prevalence and incidence figures can change rapidly. Finally, observations in health care establishments are often poor indicators of what happens in the larger community, because the way people seek care correlates with socioeconomic, cultural, and health system factors.

Statistical and mathematical modelling is commonly used to estimate disease frequency in a way that addresses both uncertainty and variability. The purpose of such models is usually (i) to predict the number of people who get infected or ill across populations and over time and (ii) to understand the factors that drive the dynamics of an epidemic or endemic disease, i.e. the determinants of disease frequency (4). Famous examples of mathematical models that underpinned disease control measures are, among many others, those for smallpox in Europe (18th century), plague in India (20th century), and measles, influenza, human papillomavirus, and coronavirus (nowadays) (4).

Modelling involves the translation of information into mathematical equations (4). An essential part of this modelling process is to make assumptions about microorganisms, vectors, hosts, and their interactions. Examples of things to consider are how and at what speed microorganisms reproduce, survive, and cause illness. From the human perspective, there are assumptions about the natural evolution of infection and disease, mechanisms of infection transmission, health seeking behaviour and access to health care, accuracy of diagnostic procedures, and effectiveness of and adherence to interventions. For many infectious diseases, environmental factors also play a role, such as temperature, humidity, vegetation, and abundance of mosquitos and other vectors. From an epidemiological point of view, the concept of heterogeneity can capture some of these assumptions about environment, biology, medicine, and human behaviour.

1.2 Heterogeneity

"*There is no consensus, there is no homogeneity, there is no truth.*"

Ward Churchill, in Glad Tidings, by Tade Reen (2011) (5)

The Cambridge online dictionary defines heterogeneity as "*the fact of consisting of parts or things that are very different from each other*" (6). In Wikipedia, this is extended as follows: "*A material or image that is homogeneous is uniform in composition or character (i.e. color, shape, size, weight, height, distribution, texture, language, income, disease, temperature, radioactivity, architectural design, etc.); one that is heterogeneous is distinctly nonuniform in one of these qualities*." (7) In some scientific domains, other terms that are related to heterogeneity are more commonly used and better known. Examples of such related terms are: diversity, variability, variance, interaction, modification, inconsistency, transferability, and generalisability. Applied to the context of the present thesis, a population of people is heterogeneous if it consists of subgroups that distinctly differ from each other in one or more variables that are relevant for the estimation of disease frequency.

Although it may seem obvious that models to estimate disease frequency should account for heterogeneity, this immediately raises a series of questions: which variables are relevant for the infectious disease at hand; how are subgroups in a population defined and what is their size; and how are the assumptions about heterogeneity translated in mathematical equations and statistical models? These questions are the topic of this thesis.

1.3 Perspective

I was trained as a medical doctor at KULeuven University in Belgium (1990-1997) and started clinical and epidemiological research while working at the *Instituto de Medicina Tropical Alexander von Humboldt* at the *Universidad Peruana Cayetano Heredia* in Lima, Peru (2000-2011) in the framework of a collaboration programme with the Institute of Tropical Medicine of Antwerp, Belgium. Since 2012, I have worked for the Department of Public Health at the Institute of Tropical Medicine in Antwerp. In 2016, I enrolled in the Master of Statistics programme at UHasselt as a distance learning student (specialisation Epidemiology and Public Health Methodology).

In my professional experience in the domain of infectious disease control, heterogeneity is a contested concept. On the one hand, there is a quest for straightforward explanations, universal solutions, and standardised interventions. With the first quote about the lance and the dragons, I intended to evoke the massive, global, and -according to many- heroic efforts that are ongoing to control and eliminate many infectious diseases. Where clear guidelines are needed, significant P-values, narrow confidence intervals, and reproducible results are pursued. For example, the GRADE framework for the formulation of evidence-based guidelines downgrades the quality of evidence if inconsistency (unexplained heterogeneity) is present (8). In such an environment, heterogeneity is nuisance, an unwelcome guest.

On the other hand, understanding the reasons behind heterogeneity can lead to breakthroughs and better tailored disease control interventions. But public health is about populations rather than individuals; how much diversity and disaggregation is manageable? This question is what I intended to evoke with the second ("*there is no truth*") quote. Studying heterogeneity opens the doors to complexity. How much heterogeneity can we rationally handle without opening a Pandora's box?

The original intention of this thesis project was to explore the issue of heterogeneity from a statistics perspective, based on a literature review and a concrete application. The pandemic due to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, further referred to as coronavirus) reached Belgium just when I started to write this thesis. Unprecedented measures were put in place to control the spread of the virus. Questions about the number of coronavirus disease 2019 (COVID-19) cases and deaths continue to be all over the news. How many people are infected and how many become ill? What are the effects of different disease control interventions on the incidence of infection and disease? Because these basic epidemiological figures cannot be measured directly, public health decisions heavily rely on model-based estimates. Varying assumptions about the heterogeneity of the populations in which coronavirus spreads may lead to different estimates of infection frequency and hence, different public health recommendations. It seemed logical to take the coronavirus outbreak as the application for the thesis project. As a result, this thesis document consists of three distinct parts: a brief exploration of the heterogeneity-in-infectious-diseases literature, a more technical discussion on the accommodation of heterogeneity in phenomenological growth models, and a study of the growth pattern of the coronavirus outbreak in Belgium.

2 Research questions

The nature of this thesis project is observational and reflexive. The core theme of heterogeneity in the estimation of the frequency of infectious diseases is explored from three different angles. The research questions are:

1. How is heterogeneity addressed in recent literature about frequency of infectious diseases?
2. How can statistical models describing the early growth of an outbreak take up heterogeneity?
3. How can considerations about heterogeneity be taken up in growth models for the COVID-19 outbreak in Belgium and what are the implications?

3 Methods

3.1 Conceptual frameworks

3.1.1 Variability and uncertainty

Because heterogeneity is central to this work, I looked for a framework in which heterogeneity is defined and distinguished from related concepts. Among several alternatives (9–12), I chose the approach by Krupnick et al., which was developed for the US Environmental Protection Agency by "Resources for the Future", a non-profit research institute in Washington (13). Their report is not peer-reviewed and it was written in a context of regulatory choices, far away from infectious diseases epidemiology. Yes I still chose this framework because it is based on an extensive literature review; it addresses both variability and uncertainty; and it contains a public health perspective.

The structure of the framework by Krupnick et al. is summarised in Figure 3.1. It starts with a distinction between variability and uncertainty. *Variability* is an inherent characteristic of the world; it is about how nature is. Variability is an ontological concept. It can be described using frequency distributions and is not reducible with additional research. As an example, if the weight of individuals in a complete, well-defined population is precisely measured, the variability of the weight (or variance of the weight distribution) will not decrease if further research is done. This variability is truly present in the population (13).

Uncertainty is different: it is about our limited understanding of the world around us. It is a characteristic of our knowledge, not of the world itself. Uncertainty is an epistemological concept. It can be modelled using probability distributions and is reducible with additional research. For example, if the weight of a population is estimated through a study in which the weight of a subpopulation is measured using an imperfectly calibrated scale, there is uncertainty about the weight of the overall population. Further research with better scales and more study participants can reduce that uncertainty. It is worth noting that this framework classifies "sampling variability" -as it is commonly used in statistics-, under uncertainty and not under variability (13).

Uncertainty is further subdivided into parameter, model, and decision uncertainty. *Parameter uncertainty* comes from the practical limitations of data. It is about the value of an empirical quantity such as the weight of individuals. Continuing with the example described above, the incomplete study population corresponds to extrapolation error on Figure 3.1 and the imperfect weighing scale corresponds to measurement error. *Model uncertainty* is not about an empirical quantity anymore, it is situated at a higher level. Model uncertainty results from limitations in our ability to make a model of a real-world system on the basis of data. It is related to our ignorance about the world, lack of imagination, or deliberate choice to understand real-world systems based on simple models. Finally, *decision uncertainty* plays at yet another level, where information (e.g. risk estimates) is translated into policy in an environment where societal values and political elements also interfere (13).

In real life, for example when looking at a given dataset, it is usually not possible to distinguish between variability and uncertainty. Nevertheless, Krupnick et al. suggest that keeping these different concepts in mind enriches the process of research and evidence-informed decision making (13). The way in which I have understood the term *heterogeneity* throughout this thesis comes close to the term *variability* in the framework. However, other authors and readers may have different understandings. For clarity, I will refer back to Figure 3.1 in different sections of this document.

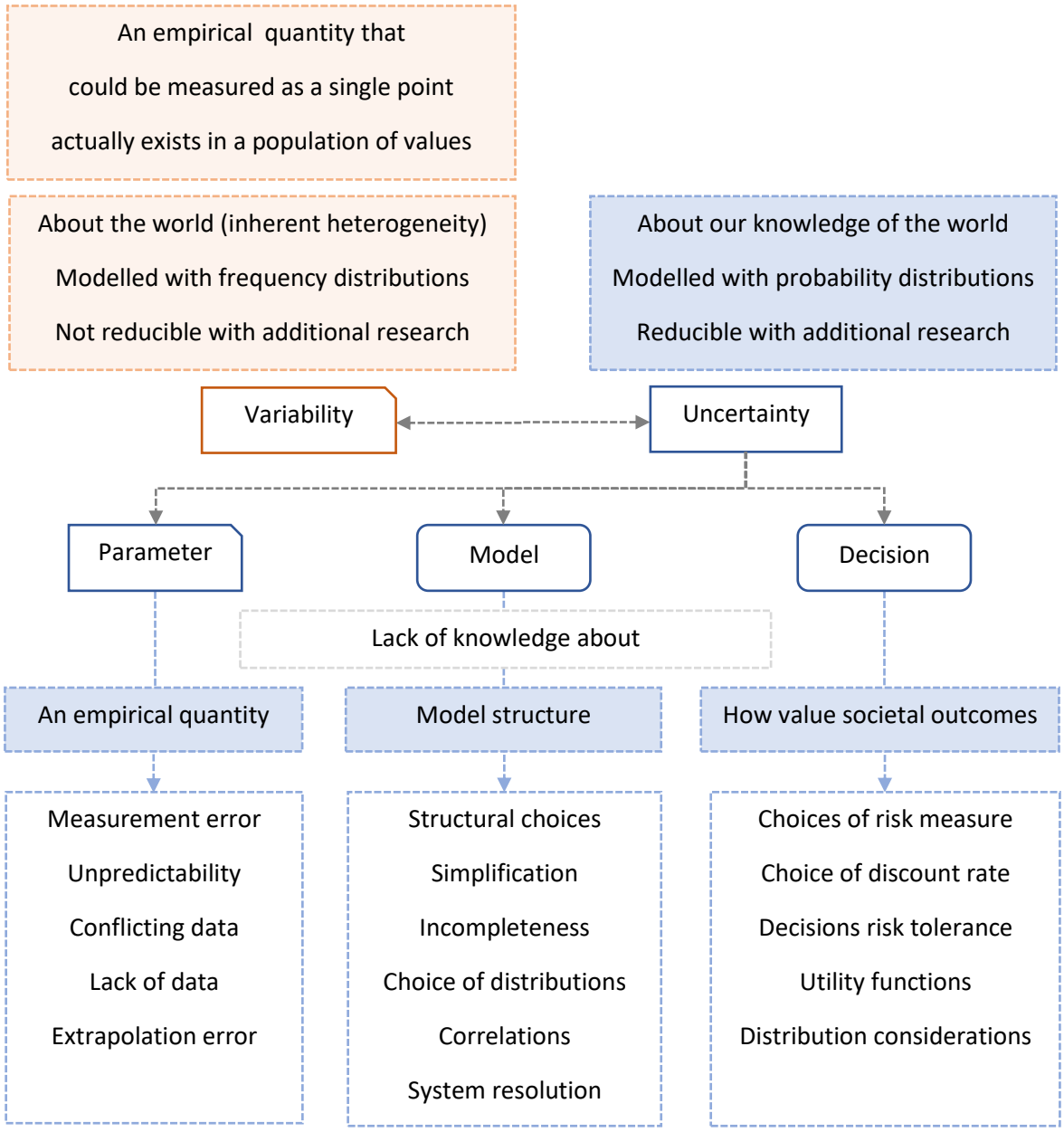

Figure 3.1 Conceptual framework about variability and uncertainty

Figure legend. Figure based on the text by Krupnick et al. 2006 (13)

3.1.2 Classification of models of transmission dynamics during outbreaks

Several statistical and mathematical approaches have been proposed to characterise the growth of infectious disease outbreaks. In this thesis project, I focus on one approach, i.e. a phenomenological generalised growth model. To position this method between alternatives, I refer to the classification used in a review paper by Chowell et al., which is summarised in Figure 3.2 (14). The main distinction is between phenomenological and mechanistic models.

Phenomenological models describe the empirical relationship between phenomena without immediately trying to explain why certain variables interact the way they do. The assumption is that relationships may be more complex than what is captured in an available dataset (15). Phenomenological models have an inductive flavour. Although they usually are in line with fundamental theory, they are not derived from detailed theory (15). In other words, we do not need

a detailed understanding (theory) of how exactly transmission takes place to be able to construct a phenomenological model to characterise an outbreak. In the domain of outbreak research, phenomenological models are sometimes called “statistical models”.

Mechanistic models, by contrast, have a deductive flavour. They are derived from an explicit theory about underlying causal mechanisms, for example theory about how infectious diseases spread through human populations. The assumed causal mechanisms are then translated into mathematical formulations (14). In the domain of outbreak research, mechanistic models are sometimes called “mathematical models”. The purpose of mechanistic models is to mimic real-life events and to assess whether the range of possible input and output behaviours predicted by the model is consistent with empirical observations (16). If they fit the empirical data well, mechanistic models are powerful tools, because they can immediately show the expected impact of interventions and they can be extrapolated to other settings, just by changing the input conditions (16). Mechanistic models are further classified into population-based and individual-based models (14).

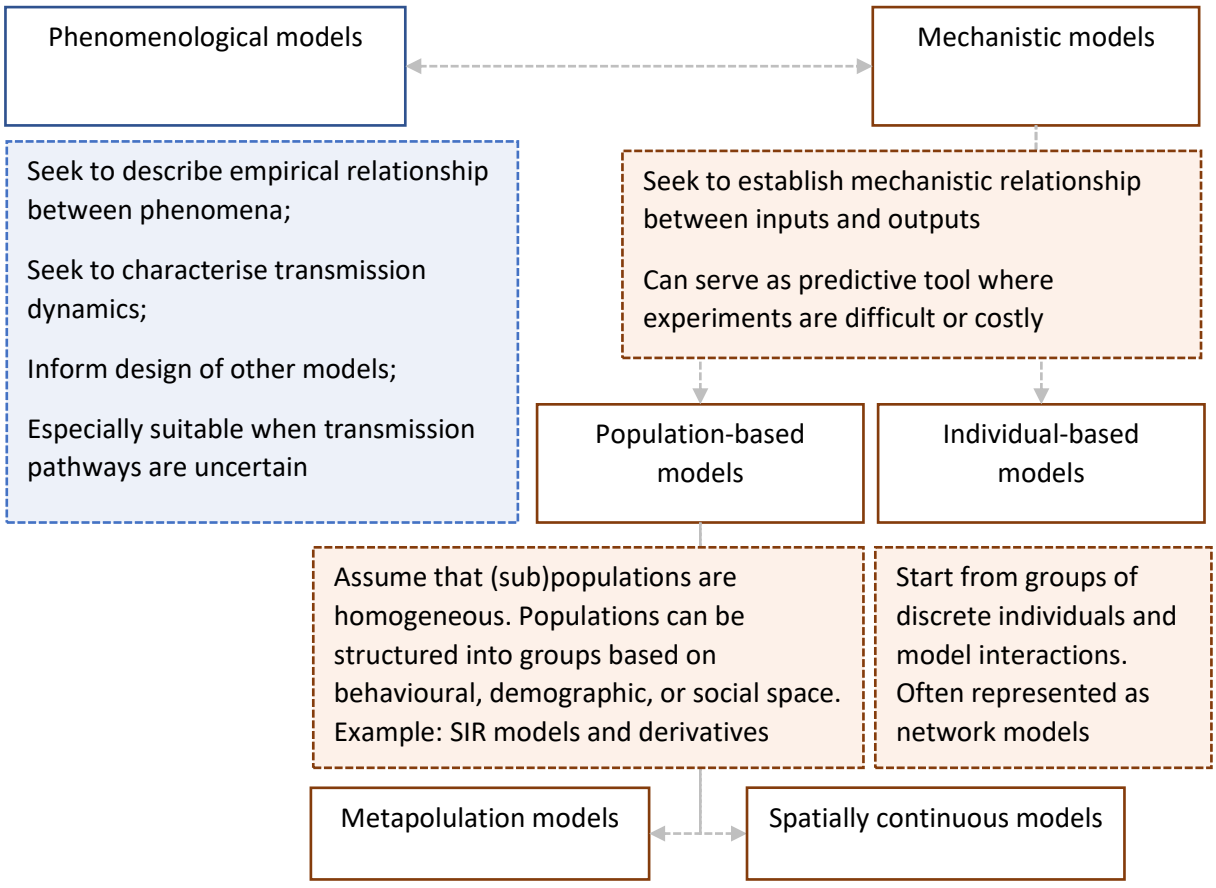

Figure 3.2 Overview of statistical and mathematical models used to capture the transmission dynamics of infectious diseases during outbreaks

Figure legend. Figure based on the review by Chowell et al. (14) and the opinion piece by Baker et al. (16). SIR model: susceptible, infectious, recovered model.

Population-based models start from the assumption that populations are homogeneous, which is useful in theory but unrealistic in practice. In metapopulation models, the overall population is divided into subgroups based on for example age, place of residence, sociocultural background, or behaviour. By disaggregating the population, predefined sources of heterogeneity become part of

the model structure. In a metapopulation model, it is the subgroups -and not the overall population- that are assumed to be homogeneous. In the structure of a metapopulation model, the subgroups are linked to one another with some sort of coupling mechanism, for example a matrix ("who acquires infection from whom" or WAIFW matrix) representing the interaction and transmission across subgroups. The dimensions of the population structure, i.e. the variables used to disaggregate the population, can be viewed as some kind of space in a broad sense, for example demographic, social, or behavioural space. In metapopulation models, the subgroups are discrete and identifiable. In spatially continuous models, the population is continuously distributed across space (14).

Whereas population-based models look at a population as one group or a set of linked subgroups, *individual-based models* explicitly consider each individual that is part of a population. Individual-level models are often analysed as contact network models, where the nodes are the individuals and the edges between the nodes are interactions that may lead to disease transmission. Individual-based models can capture heterogeneity at a very fine scale but require detailed information about for example the structure of people's contact networks, the types of contact that are relevant for disease transmission, and the individual-level variation in susceptibility and infectiousness (14).

3.2 Specific methods for research question 1: scoping review

To get a broad overview of which approaches are currently used to address heterogeneity in the infectious diseases literature, I did a rapid scoping review. The review methods are described below and are in line with the PRISMA guidelines for scoping reviews, although I did not formally register a protocol (17).

Records were eligible if they described studies about the frequency of infectious diseases in human populations and explicitly mentioned heterogeneity. The source of references was the MEDLINE database of citations for the biomedical literature. Because there were many eligible records and the timeframe of this thesis project was limited, I selected the 15 most recently published records. The search and selection of records took place in March 2020.

The search strategy combined four concepts: disease frequency, infectious diseases, statistical modelling, and heterogeneity. The exact search syntax for the PubMed interface is given in Annex 7.1. I ordered the retrieved records from the most recent to the oldest and included the first 15 eligible records in the present review.

During the process of data extraction (charting), I copied and pasted text extracts from the included papers to a Microsoft Excel datasheet. The data items of interest were: information about the record (type and year of publication, journal, authors), setting (infectious disease, country, public health perspective), heterogeneity (how introduced/justified, which sources of heterogeneity mentioned/evaluated), and statistical methods (which methods, advantages and disadvantages, data availability). Where possible, I classified the statistical methods as in Figure 3.2.

The review results are presented in a narrative synthesis accompanied by tables, organised along the following themes: (i) public health perspective (why do the authors of the selected papers consider that addressing heterogeneity is important?), (ii) sources of heterogeneity (which variables capture relevant heterogeneity?), and (iii) methodological approaches (how is heterogeneity accounted for, and what are assumptions and challenges?). Because the objective was to get a broad overview of methodological approaches, I did not conduct a formal critical appraisal of the individual records.

3.3 Specific methods for research question 2: narrative review

The focus now shifts to one setting, i.e. the early growth of outbreaks and one statistical approach, i.e. phenomenological growth models. To get a better understanding of how heterogeneity in human populations can be translated into statistical models, I reviewed a small set of purposefully selected publications. I started with a few key papers in the field and retrieved additional papers from the reference lists. In addition to the set of included papers, I also cited original seminal publications where needed. In the process of reading and synthesising the information, I focused on how heterogeneity can be ignored or addressed and on what the implications can be.

3.4 Specific methods for research question 3: data analysis

3.4.1 Study design

This is an observational study concerning the COVID-19 outbreak in Belgium in 2020. The epidemic trajectory is characterised using phenomenological growth models. For comparison, some of the analyses are also applied to COVID-19 outbreak data from other settings, i.e. the Princess Diamond Cruise ship and country-level data from Sweden, Spain, Peru, and Chile.

3.4.2 Setting

COVID-19 is an infectious disease caused by SARS-CoV-2, a betacoronavirus. Common symptoms of COVID-19 are fever, disorders of smell and taste, cough, and shortness of breath. (18). The virus is transmitted from human to human via close contact, infectious droplets, surfaces, and possibly also via the faecal-oral route (18).

In Belgium, local transmission of SARS-CoV-2 was first recognised in the beginning of March (19,20). Public health officials linked the start of the outbreak to Belgian travellers returning home from southern Europe after the spring holidays and to participants in carnival festivities in the bordering regions of Belgium, Germany, and the Netherlands (19–22). The outbreak rapidly spread over the national territory (23). Containment measures were gradually introduced. On 10 March, the government advised to cancel indoor events for more than 1000 people and recommended all companies to allow their personnel to work from home if possible. A few days later, Belgium moved towards a lockdown: on 13 March, bars and restaurants had to close and public gatherings for sporting, cultural or festive purposes were cancelled. On Monday 16 March, the schools closed. On 17 March, non-essential shops had to close as well, and non-essential travel was prohibited. On 20 March, Belgium closed its borders (20,24). The population density in Belgium is 378 people/km² (25). While I am writing this document, the pandemic is still ongoing, but it seems that Belgium will be among the world's most affected countries in terms of number of deaths per population (26). For this study, I focused on the early growth phase of the outbreak. I used secondary data which I obtained on 16 May 2020. Unless explained otherwise, I used the data from 10 March to 29 March for model fitting and up to 30 April for graphical display.

The settings outside Belgium were chosen purposefully as examples of very different settings in terms of virus spread, population vulnerability, testing conditions, control measures, and reporting. The Diamond Princess is a cruise ship where a large COVID-19 outbreak took place in February: the ship was put in quarantine off a port in Japan; and 19% of 3711 people on board got infected (27,28). Spain is one of the early and severely hit countries in Europe, where many people may have been

infected before formal control measures were put in place (29,30). The Spanish health system got overwhelmed and this may have interfered with testing and reporting. Spain has a moderate population density (94 people/km²) (25,30). The situation in Sweden is very different: the outbreak took off a few days later; there was extensive testing, control measures were less stringent (no formal lockdown); and there were less cases and deaths than in Spain (29,30). Sweden has a low population density (23 people/km²) (25).

In Peru and in Chile, the outbreak started about one week later than in Sweden. There are important differences between Peru and Chile in terms of cultural and ethnic background, climate conditions, socio-economic situation, and health system capacity. The outbreak response was also different: in Peru, there is a country-wide lockdown while in Chile, control measures are more localised (29). Both countries have a low population density overall (25 people/km² in Peru and 23/km² in Chile), but a very high density in the capitals (11000/km² in Lima and 9800/km² in Santiago) (25). Where countries are compared, the start of the time window for model fitting was defined in the same way for all countries and measures, i.e. the first day of the first three days of monotonic growth.

3.4.3 Variables

The variable to be modelled is disease frequency as a function of time. For Belgium, I used three measures of disease frequency: the incidence of confirmed COVID-19 cases, hospital admissions for COVID-19, and deaths due to COVID-19. Sciensano is the Belgian institute for health responsible for the epidemiological follow-up of outbreaks (31). The definitions used by Sciensano are available online (32). Of note, in Belgium, admissions and deaths due to COVID-19 combine probable and confirmed diagnoses. Confirmed diagnoses are mainly based on molecular testing. Incidence (counts) and time (in days) are treated as discrete variables. Three other variables are used to define subpopulations within Belgium: province of residence, sex, and age group (as defined by Sciensano). For the settings outside Belgium, I only used the incidence of cases and deaths over time.

3.4.4 Data sources

The data are publicly available. Data for the main analyses about Belgium were obtained via the Sciensano website (33). Data for the Diamond Princess come from the World Health Organization situation reports (34). When countries are directly compared, all data (including those for Belgium) were taken from the same source (European Centre for Disease Prevention and Control) (35).

3.4.5 Statistical methods

A generalised growth model (phenomenological approach) is used to model the incidence of COVID-19 cases, admissions, and deaths (36). This method is explained and discussed in detail in section 4.2. The analysis first focuses on ways to relax the assumptions of equidispersion and exponential growth. Second, there is an evaluation of the influence of specific modelling choices (time window for model fitting; incidence *versus* cumulative incidence; and incidence of cases, admissions, or deaths) on the characterisation of the growth pattern. Third, I explore whether there are differences in the growth pattern of the COVID-19 outbreak between (i) the Belgian population as a whole, (ii) different subpopulations within Belgium (defined based on age group, sex, and province of residence), and (iii) settings outside Belgium.

The analyses were done with R software (*gamlss*, *stats4*, *bbmle*, and *msm* packages); the core code is available in annex 7.2. Parameters are estimated in a maximum likelihood framework; and 95% confidence intervals are based on standard errors obtained during parameter estimation. Akaike's information criterion (AIC) is used for model comparison.

4 Results

4.1 How is heterogeneity addressed in recent literature about infectious diseases?

4.1.1 Overview of included records

The 15 records retrieved via the PubMed search were original research studies published between August 2019 and February 2020. Nine records were published in specific journals about infectious or tropical diseases: *PLoS Neglected Tropical Diseases* (n=3) (37–39), *Malaria Journal* (n=2) (40,41), *Parasites and vectors* (n=2) (42,43), *BMC Infectious Diseases* (n=1) (44) and the *Journal of the Brazilian Society of Tropical Medicine* (n=1) (45). The remaining six records appeared in more general medical or science journals: *BMC Medicine* (46), *BMC Public Health* (47), *Medicine (Baltimore)* (48), *PLoS ONE* (49,50), and *Nature Communications* (51). The studies were based on information from Africa (n=7) (37,38,41,43,44,49,51), South America (n=4) (39,40,45,50), and Asia (n=2) (47,48). The two remaining studies either used data from more than one continent (Africa and Asia) or were theoretical without localised information (42,46).

The infectious diseases studied in these records were malaria (n=5) (40,41,49–51), helminth infections (n=4) (37,38,42,43), HIV/AIDS (n=3) (44,45,48), leishmaniasis (n=1) (39), and hand foot and mouth disease (n=1) (47). One study presented several diseases: severe acute respiratory stress syndrome (SARS), plague, and Ebola virus disease (46). Seven of the 15 studies were about infectious diseases that are transmitted via insects (mosquitos, sand flies, and black flies) (37,39–41,49–51). More than half of the studies (9 out of 15) were set up in an environment in which there is an ambition to eliminate the infectious disease, e.g. via vector control or mass preventive chemotherapy (37,38,40–43,49–51).

Taken together, the field of tropical diseases is remarkably present in the recent literature about infectious diseases that explicitly mentions heterogeneity. This may be related to the complex transmission cycles that are common in this field as well as the enormous disease control and elimination efforts that are ongoing. This is illustrated in the following quote by Truscott et al.:

> *"As many countries with endemic soil-transmitted helminth (STH) burdens achieve high coverage levels of mass drug administration (MDA) to treat school-aged and pre-school-aged children, understanding the detailed effects of MDA on the epidemiology of STH infections is desirable in formulating future policies for morbidity and/or transmission control. (...) For a given value of infection prevalence in a population, parasite transmission intensity could vary considerably depending on the level of parasite aggregation within the human host population. Hence, prevalence alone may not be a reliable indicator of transmission intensity. (…) The high degree of parasite aggregation associated with low prevalence values after multiple rounds of MDA suggest that in the 'end game' of STH control once prevalence is low, different approaches to MDA distribution may be desirable. High levels of aggregation suggest that infection may be localized in small hotspots, possibly at the household level, or in groups who are consistently non compliant to control. As such, novel approaches to identifying, monitoring and treating such hotspots and or non-compliers in order to maintain low prevalence or achieve a break in transmission, are required to avoid unnecessary treatment of a largely uninfected population."* (43)

4.1.2 Public health perspective: why is addressing heterogeneity considered important?

The study justification was similar in most of the records and had the following structure: "Disease distribution (or exposure, transmission, susceptibility, intervention impact) is heterogeneous. Capturing this heterogeneity leads to better public health actions". What these actions are, is summarised in Table 4.1. This comes down to choosing the right interventions, reaching the right individuals or subgroups, and appropriately estimating the impact of different interventions.

The concept of heterogeneity is quite prominent in most of the studies: the term "heterogeneity" appears in four of the 15 titles (37,42,43,47); the term "spatiotemporal" appears in five titles (40,44,45,47,49). The following quote from Cooper et al. is one illustration of how heterogeneity is introduced:

"*Heterogeneity shapes infectious disease epidemiology and transmission. Understanding the causes and consequences of heterogeneity is important for analysis of infectious disease data and for determining target intervention coverage levels for control. A "Pareto rule" has been proposed for many infectious disease systems: 80% of infectious disease transmission is concentrated on 20% of hosts. Such heterogeneity, often called "super-spreading", has drawn interest because the efficiency of disease control could be dramatically improved if it were possible to identify and target individuals who account for most of transmission, who are sometimes called "super-spreaders", a topic on which mathematics and mechanistic models have provided some of the most important insights*." (51)

Table 4.1 Overview of the type of public health actions that can be informed by studies accounting for heterogeneity

Target subpopulations / Prioritise activities
Allocate resources / Rationally deploy / Improve programme performance and efficiency
Improve surveillance / Establish risk maps / Monitor, explore, plan
Serve (sub) population / Avoid unnecessary treatment of a largely uninfected population
Make decision-making more relevant / Formulate future policies / Revise control guidelines
Better understand role of drivers or determinants / Identify hotspots
Better understand why interventions do not work as predicted
Design or add more effective interventions / Predict the impact of control measures

4.1.3 Sources of heterogeneity: which variables capture relevant heterogeneity?

The variables that were meant to capture heterogeneity in the included studies are listed in Table 4.2. Some authors worked with relatively general variables that were routinely available, probably because that information was collected for other purposes, such as district of residence; population density; age, sex, and marital status; and meteorological data. Other authors collected information with specific hypotheses about (heterogeneous) disease transmission in mind such as data about mobility patterns, shoe wearing, mosquito biting frequency, immunity, and compliance with specific disease control interventions.

Classifying these variables in meaningful categories is not easy because they are often correlated among each other, and they correlate with unobserved information as well. Indeed, a general variable such as *location* probably correlates with wealth, natural environment, and disease control interventions, among many other known and unknown factors. Also more specific variables such as, for example, *occupation*, can be a proxy for other information, such as social class, ethnic background, proximity to forest areas, mobility, immunity, etcetera. Furthermore, simple variables such as age and sex can play a role in different aspects of the disease dynamic, such as exposure, transmission, and natural evolution of a disease.

Table 4.2 Overview of possible sources of heterogeneity mentioned in the included studies

Category	Examples of variables
Time	Year, season, day
Place of residence	Geocoordinates, township – municipality, distance to lake – river – forest edge
Weather, climate	Temperature, rainfall, humidity – drought, air pressure, wind speed, hours of sunshine – cloud cover, El Niño-related events
Natural and built environment	Elevation, water bodies (dams, irrigation canals, wetlands, ponds, puddles), land cover (cropland, tree cover, shrub land, grass land, urban area), biomass above ground, forest canopy height, species richness in mammals, soil sand content, human footprint, deforestation, level of urbanisation, sanitation conditions at school and in the community, infrastructure index, density of tracks and road network
Household environment	Household floor – wall – roof materials, sanitation conditions, shoe wearing, family size, screens in doors and windows
Demography	Sex, age, population density, population movement, mobility patterns
Socioeconomic and cultural factors	Human development index, ethnic background, education, marital status, occupation, human activity patterns, mining, income, poverty index, income equality, ownership of household assets, access to electricity, access to sanitation
Human behaviour	Shoe wearing, sleeping and wakeup time, compliance with interventions, sexual behaviour
Interventions	Mass drug administration, insecticide-treated bed net available in household, reported bed net use, indoor residual spraying
Microorganism and vector	Parasite species, parasite density – aggregation – sexual reproduction, vector species, mosquito distribution, mosquito biting rates
Human biology and immunology	Susceptibility, health index, infectiousness – gametocyte density – super-spreaders, duration of infectiousness, immunity – immunologically naïve immigrants, asymptomatic human reservoirs
Other	Accuracy of diagnostic procedures and reporting

To deal with complex sets of correlated variables, some authors emphasised that it is important to study several factors simultaneously, such as space & time (44,45,47), land use & time (40), population density & natural environment (39,40,43), disease control interventions & population movement & age (42), and sanitation conditions at home & elsewhere (38). Other research groups organised possible sources of heterogeneity in a comprehensive way: they combined variables in patterns or scenarios (40), or followed the biotic-abiotic-mobility (BAM) framework to choose variables (39). Such strategies imply reflection about how to approximate real-world systems with models and data. In the framework by Krupnick et al. (Figure 3.1), these issues (e.g. structural choices, simplification, correlation, and incompleteness) fall under model uncertainty (13).

The spatial scale or temporal resolution at which information was collected, varied from exact geocoordinates to large spatial areas, and from daily data to yearly averages. Few records included a reflection on what would be an appropriate scale for the purpose of the study (39,46). In the framework of Figure 3.1, considerations about system resolution are classified under model uncertainty (13).

Data was usually available at different levels: individual, household, and larger population subgroups. These population subgroups were defined in various ways, which I classified as *pragmatic* (subgroups

are municipalities or units of randomisation in cluster-randomised trials, *deductive* (subgroups are ecological units based on theory about disease transmission), or *inductive* (subgroups are transmission hotspots, based on the observation of heterogeneous disease frequency).

4.1.4 Methodological approaches: how is heterogeneity accommodated?

Statistical and mathematical models were formulated to explain or predict different types of outcome variables, the most frequent of which were count data (in 10 out of 15 included papers). The counted phenomena were disease episodes, cases, or mosquitos (39,40,43–47,49–51). The remaining outcome variables were disease prevalence (37,38,41–43), infection intensity (43), belonging to a hotspot (49), or proportion of transmission via a specific mode (48).

The models often had a hierarchical (multilevel) structure (n=10) (38,40,41,44,45,47–51). Levels of aggregation were households, schools, villages, and municipalities or other administrative levels. These levels were usually but not always nested. A Bayesian inference framework was used to develop the models in four of the 15 included studies (44,45,47,48). One approach, i.e. ecological niche modelling using maximum entropy was described as a type of machine learning (39).

Spatial heterogeneity was explicitly studied in nine papers, with findings shown on maps (39,40,43–45,47–50). Most models (n=13) contained sociodemographic, environmental, or intervention-related explanatory variables (37–45,47–50). In six studies, time and space were modelled together (40,44,45,47,49,51). In two of these studies, time was included as a random effect in a hierarchical model (45,51). One study used time series analysis (autoregressive integrated moving average) (40).

Transmission dynamics and epidemic trajectories were studied in four papers (37,42,43,46). Following the classification of Figure 3.2, one study was phenomenological (46) and three were mechanistic (37,42,43). The three mechanistic studies were about helminth infections; two were stochastic and one deterministic. Two mechanistic models were used to assess the effect of specific conditions (mobility, exposure heterogeneity) on the outcome of mass drug administration for helminth control; one was individual-based (37) and the other used both individual-based and population-based (metapopulation) elements (42). The third study with a mechanistic approach focused on the estimation of model parameters at endemic equilibrium (43).

The issue of overdispersion was raised in five papers (37,42,43,50,51). In three studies, a negative binomial distribution (with an aggregation parameter) was used to model worm burden among human hosts (37,42,43). In two studies, a zero-inflated negative binomial distribution was used to represent malaria episodes or mosquito bites (50,51). In the study about mosquito biting, the term super-spreading was used in relation with overdispersion, and this was explained by means of the Pareto fraction (51). Overdispersion in malaria count data was explained by Corder et al. as follows:

"Statistical modeling of routinely collected malaria surveillance data can be particularly challenging. Poisson regression models are commonly used to analyze count-type data in epidemiology, but cannot adequately fit overdispersed malaria case distributions that are typically found in endemic settings. A variety of alternative models have been used instead, e.g. the negative binomial. However, as malaria rates decline, more subjects will remain uninfected over extended periods of time, increasing the proportion of zero counts in cohort studies. Zero-inflated statistical models, such as the zero-inflated negative binomial, usually provide a better fitting to malaria count data and household-level malaria vector densities with an excess of zero counts." (50)

Several papers discussed challenges in relation with the modelling of heterogeneous phenomena. For example, important parameters of mechanistic models vary across subpopulation and over time, which raises the question at which level and at which frequency they need to be (re)estimated

(42,43). Moreover, some key factors cannot be directly measured or estimated. Hamley et al. described this as follows:

"The phenomenological nature of density-dependent establishment of the parasite in the human host is an important limitation of this work. The use of population-level data to simultaneously estimate both density dependence and exposure heterogeneity allows the parameters involved to counteract, reducing identifiability and diminishing mechanistic interpretation. Since direct estimation of density-dependent parasite establishment within humans is not feasible (…), data collection on heterogeneity in exposure to fly bites will be an important step in better resolving density dependence. This has been discussed previously but available data remain limited. Since blackflies have very specific environmental requirements, with breeding sites varying in distance to human settlements, the estimation of individual-level variation in exposure to bites poses a substantial challenge." (37)

Finally, even if it is theoretically possible to estimate the parameters of complex (flexible) mathematical models, this is hindered in practice by the limitations in available epidemic data as illustrated in the following quote from Chowell et al.:

"Limited epidemic data limits the complexity of the mathematical models in terms of the number of mechanisms and parameters that can be estimated from data. These models often use a metapopulation framework to incorporate population heterogeneity by dividing the population into socio-demographic groups based on the susceptibility, infectivity, mobility patterns, or other individual characteristics related to the transmission dynamics. The individuals in the same group are assumed to be homogenous, and the heterogeneity of the population is limited by the number of groups. Even when the number of parameters that can be estimated from limited data is small, the model must include enough complexity to account for the underlying transmission dynamics." (46)

4.2 How can statistical models for the early growth of an outbreak capture heterogeneity?

4.2.1 Introduction

The present section contains a more technical discussion on how heterogeneity can be taken up in statistical models. After the broad overview of the previous section, the focus now goes to one particular case: phenomenological models for the characterisation of the early growth phase of disease outbreaks. In this setting, the purpose of statistical modelling is to understand how the number of cases grows before coordinated control measures or herd immunity alter the course of the outbreak. Early growth models have direct applications in public health: they are used to predict pattern, height, and time course of epidemic trajectories during the first generations of infection transmission. These phenomenological models can also be used to estimate parameters such as the force of infection which feed into mathematical, mechanistic models of disease transmission (14).

The synthesis below draws from seven publications in which a generalised growth model is used to examine different outbreaks (14,36,52–56). The authors work with well-known historical datasets, for example concerning plague in Bombay in 1905, influenza in San Francisco in 1918, and HIV/AIDS in the US and Japan in the 1980s. They also characterise more recent epidemics such as Zika virus disease in Colombia in 2015, and especially Ebola virus disease in West Africa in 2013 to 2016. In five of the seven publications, the authors produce simulated data to examine, contrast, and visualise the consequences of specific methodological choices (14,36,52,54). The synthesis below is organised from a perspective of heterogeneity.

4.2.2 Ways to ignore or address heterogeneity

Equidispersion and overdispersion

To characterise the early growth phase of an outbreak, we need a model that predicts the number of new illness cases (incidence) over time. As a starting point, we use a Poisson approach, in which we assume that the incidence is a random variable that follows a Poisson distribution (Box 4.3). This growth model has two parameters: the number of cases at the start of the outbreak (C_0) and the growth rate (r). The number of new illness cases at time t (y_t) is a Poisson random variable with mean μ_t. We assume that at the start of an outbreak, C_0 is close to zero and nearly the whole population can get the illness because there is no immunity yet. We also assume that in the early phase of an outbreak, there is exponential growth of the form $\mu_t = C_0 * e^{rt}$ (4). The counts predicted by this model depend on the time but are, apart from the time, independent of each other (conditional independence). The parameters C_0 an r can be estimated from observed data. The growth rate r is related to the basic reproduction number R_0 known from classical compartmental models as follows: $R_0 = 1 + r/\gamma$ where $1/\gamma$ is the mean infectious period (36,57).

An essential property of a Poisson distribution is equidispersion: at each time point, the variance of the incidence y_t is assumed to be equal to its mean (variance=mean=μ_t). In other words, the incidence y_t is allowed to fluctuate randomly around the mean but y_t is also expected to closely follow the qualitative behaviour of the mean trend μ_t. This implies that both the mean and the variance of the incidence are a consequence of time and nothing else. Hence, unobserved heterogeneity -which would lead to larger variance- is not considered (52).

The Poisson model is illustrated in Figure 4.1 (panel A), where C_0 is set to be 1 and r is set to be 0.2. At each point in time, 100 counts are simulated from a Poisson distribution with $\mu_t = 1 * e^{0.2*t}$. It is clear from Figure 4.1 (panel A) that the variability in the simulated counts follows the pattern of the mean.

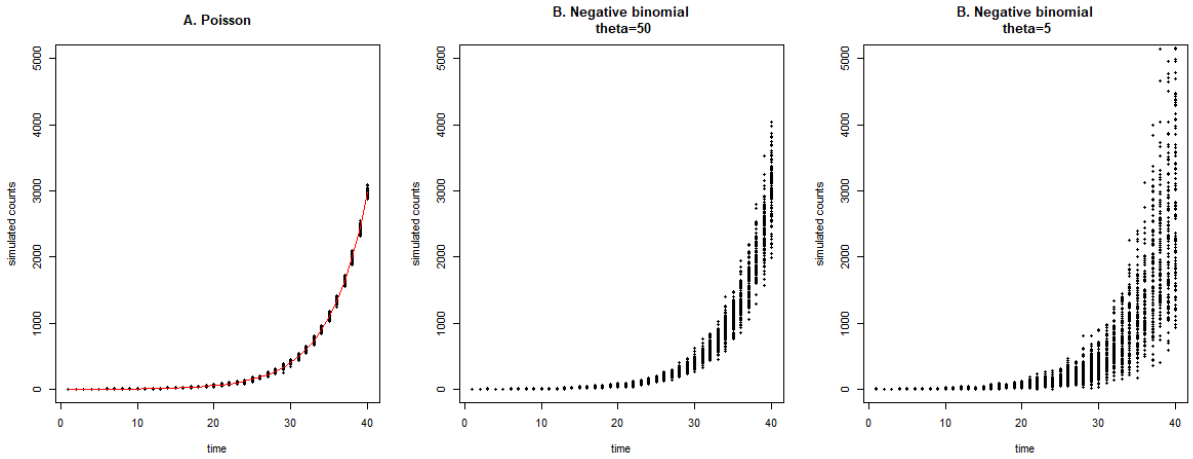

Figure 4.1 Illustration of the relation between variance and mean in Poisson and negative binomial exponential growth models

Figure legend. At each point in time, 100 counts are simulated from a Poisson model (panel A) and negative binomial models (panels B and C) with initial case number C_0=1 and growth rate r=0.2. The black dots represent the simulated counts. The red line represents the fixed mean count predicted by the Poisson model ($\mu_t = 1 * e^{0.2*t}$).

In real-life settings, observed data often display more variability than what is predicted by a Poisson model. This greater-than-expected variability is called overdispersion and can be viewed as a result of unobserved heterogeneity (52). The overdispersion reflects the effects of variables that are not in the model as well as other sources of randomness. One way of dealing with overdispersion is to assume that at a given time t, μ_t is not fixed (as in the Poisson approach) but that it comes from a distribution

of μ_t's. A gamma distribution is often chosen to represent the variability of μ_t's due to unknown reasons. Averaging over all possible values of μ_t gives the negative binomial model (Box 4.3) (52).

The negative binomial model includes three parameters that are to be estimated from the data: the same two parameters as in the Poisson model (initial number of cases C_0 and growth rate r) and the dispersion parameter θ. As explained in Box 4.3, if θ is small, there is considerable overdispersion. If θ goes to infinity, the negative binomial model reduces to the Poisson model. Figure 4.1 illustrates the negative binomial model with C_0=1, r=0.2, and two different values for the dispersion parameter θ. In panel B, θ is large (θ=50) and the model comes close to the Poisson model. In panel C, θ is small (θ=5) and the variance does not clearly follow the pattern of the mean anymore.

Exponential and sub-exponential growth

Exponential models are commonly used to describe growth in diverse domains such as demography, biology, and economy. The basis for these growth models was laid by Robert Malthus in the 18th century. The underlying idea is that populations tend to grow exponentially as long as the necessary resources abound (58). In classical outbreak studies, it is often assumed that the early growth phase is exponential, as long as there is no lack of susceptible individuals and before interventions are put in place (4,14,57). This assumption is called the mass action principle. It implies that the anticipated trajectory of an epidemic relates to the number of susceptible persons in a population, the number of infectious individuals, the incubation period, and the infection transmission parameter which is constant for a specific infectious disease (59). This mass action principle is related to the law of mass action in chemistry, which states that -in a solution with homogeneous mixing- the rate of a chemical reaction is directly proportional to the product of the concentrations of the reactants.

In practice, however, the observed growth of outbreaks can be less than exponential, even in very early phases. Sub-exponential growth has been documented for diverse infectious diseases, such as HIV/AIDS, Ebola, and hand foot and mouth disease (Figure 4.2) (14,36,52,53). One explanation for sub-exponential growth is that when an infection is transmitted via close contact, it does not spread smoothly through a population due to constraints in people's contact structures. The cubic polynomial trajectory of the HIV/AIDS epidemic in the United States in the 1980s has been explained in this way, i.e. via population mixing mechanisms (53,60). Second, behavioural changes may interfere with transmission, for example when people spontaneously reduce the number of contacts per unit time when they feel ill. A third explanation may be the heterogeneity in susceptibility and infectiousness within populations (36). However, for the majority of infectious disease outbreaks, the precise mechanisms underlying sub-exponential growth are poorly understood (14).

One way of visualising the difference between exponential and sub-exponential growth is by using epidemic trees. Figure 4.3 is taken from a review paper about mathematical models to characterize early epidemic growth (14). In the exponential tree (Panel A), the number of new cases increases smoothly from generation to generation, while the sub-exponential tree (Panel B) seems to falter.

Sub-exponential growth dynamics can be addressed by using a flexible generalised-growth model (Box 4.4) This model comes from the field of demography (61,62) and was introduced in outbreak research in 2016 (36). In this model, a tuning parameter p, called the growth scaling or growth deceleration parameter allows to reproduce a variety of growth profiles. The parameter p helps to quantify departure from the exponential situation: if p equals 1, growth is exponential; values of p between 0 and 1 correspond to sub-exponential growth; and if p equals 0, the incidence is constant, i.e. the number of new cases is the same at each point in time. In theory, p can also be larger than 1 (more-than-exponential growth), but because this situation has not been documented in real illness outbreaks, it will not be considered further.

4.3 Box: growth models using Poisson and negative binomial distributions

Note: the content of this box is adapted from Ganyani et al. 2020 (52)

The Poisson distribution has one parameter μ_t which represents both the mean and the variance of the distribution (equidispersion). Under a Poisson distribution, the probability that at time t, the incidence $y_t = m$ is given by:

$$P_P(y_t = m) = \left(\frac{\mu_t^m}{m!}\right) * e^{-\mu_t}$$

If we assume exponential growth, then

$$\mu_t = C_0 * e^{rt}$$

In the context of growth models, using a Poisson approach implies that we assume (i) that there is a fixed mean value μ_t at each time point t and (ii) that variation only occurs because the incidence is observed at different time points. There is no room for other sources of variation. However, in real-life settings, the variance in observed data is often larger than what is predicted by a Poisson model. This phenomenon is called overdispersion.

The assumption of equidispersion can be relaxed if we assume that, at a given time t, μ_t is not fixed but that it comes from a distribution of μ_t's. This new $\breve{\mu_t}$ is given by $\breve{\mu_t} = \varphi_t * \mu_t$, where φ_t is a random error term that represents the combined effects of all processes that are not captured by the variables included in the model. If φ_t is chosen to be a white-noise process, the conditional independence assumption is preserved. This corresponds to the introduction of a random effect into a Poisson distribution, which can then be written as a continuous mixture distribution:

$$P_M(x) = \int_0^\infty P_P(x|\breve{\mu_t})h(\breve{\mu_t}|\beta)d\breve{\mu_t}$$

Where:
$P_P(x|\mu_t)$ is a Poisson distribution with mean μ_t
h describes the variability of the mean μ_t
$P_M(x)$ is a marginal distribution of the counts (continuous mixture of a discrete distribution)

Several choices are possible for $h(\mu_t)$ some of which have an analytical solution. For example:
if $h(\mu_t)$ is a gamma distribution, $P_M(x)$ corresponds to a negative binomial distribution;
if $h(\mu_t)$ is an inverse Gaussian distribution, $P_M(x)$ is a Poisson-inverse Gaussian distribution;
if $h(\mu_t)$ is a generalised inverse-Gaussian distribution, $P_M(x)$ is a Sichel distribution.

A common choice for $h(\mu_t)$ is a gamma distribution, i.e. φ_t is assumed to be a gamma white-noise process. Using a gamma with scale parameter α_t and shape parameter β_t as a mixing distribution results in a negative binomial model with parameters $\mu_t = \alpha_t * \beta_t$ and $\theta = \beta_t$.

$$P_{NB}(y_t = m) = \frac{\Gamma(m + \theta)}{\Gamma(\theta)m!}\left(\frac{\mu_t}{\theta + \mu_t}\right)^m \left(\frac{\theta}{\theta + \mu_t}\right)^\theta$$

The mean of this negative binomial distribution is μ_t and the variance is $\mu_t + \frac{\mu_t^2}{\theta}$. The negative binomial model can be used to model count data with varying degrees of overdispersion. If the dispersion parameter θ goes to infinity, the negative binomial model is equivalent to the Poisson model because in that case, variance = mean = μ_t (equidispersion). If the dispersion parameter θ takes on low values, the variance is much larger than the mean; in other words, the deterministic solution is swamped by noise.

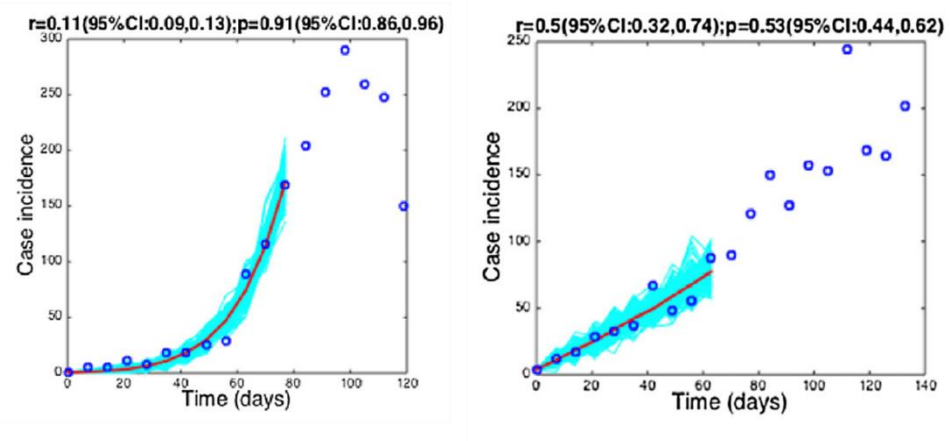

Figure 4.2 Two different growth patterns during the 2013-2016 Ebola epidemic in West Africa

Figure legend. This figure is reproduced from Viboud et al. 2016 (36). The left panel shows exponential growth observed in Montserrado, Liberia. The right panel shows sub-exponential growth observed in Western Area Urban, Sierra Leone. The blue dots are observed counts; the red fitted curve is based on a generalised growth model with two parameters: the growth rate r and the growth scaling factor p. If p is close to 1, the growth pattern is close to exponential (left panel); lower values of p correspond to less-than-exponential growth (right panel).

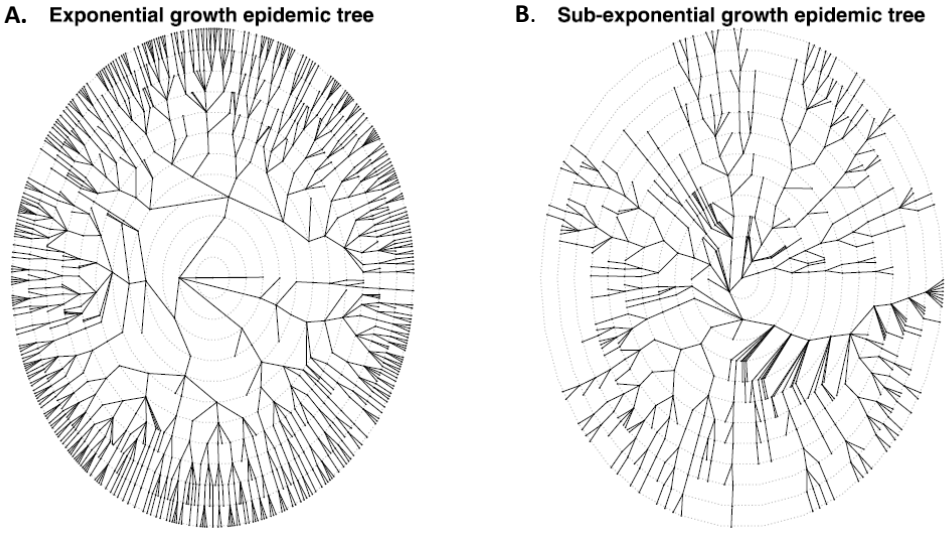

Figure 4.3 Schematic epidemic trees illustrating exponential and sub-exponential growth

Figure legend. This figure is reproduced from Chowell et al. 2016 (14). In both panels, the index case is located in the centre and there are 12 generations of disease transmission (concentric ellipses). In the exponential growth scenario, the numbers are generated stochastically based on a mean reproduction number of 1.5. In the sub-exponential growth scenario, the effective reproduction number decreases from 1.5 to 1 over subsequent generations.

If we use a Poisson model allowing for sub-exponential growth, we have to estimate three parameters from the data: the initial number of cases C_0, the growth rate r, and the growth scaling parameter p. Figures 4.4 and 4.5 illustrate what happens with the expected mean number of new cases over time if these three parameters vary. If we use a negative binomial model, a fourth parameter θ needs to be estimated as well.

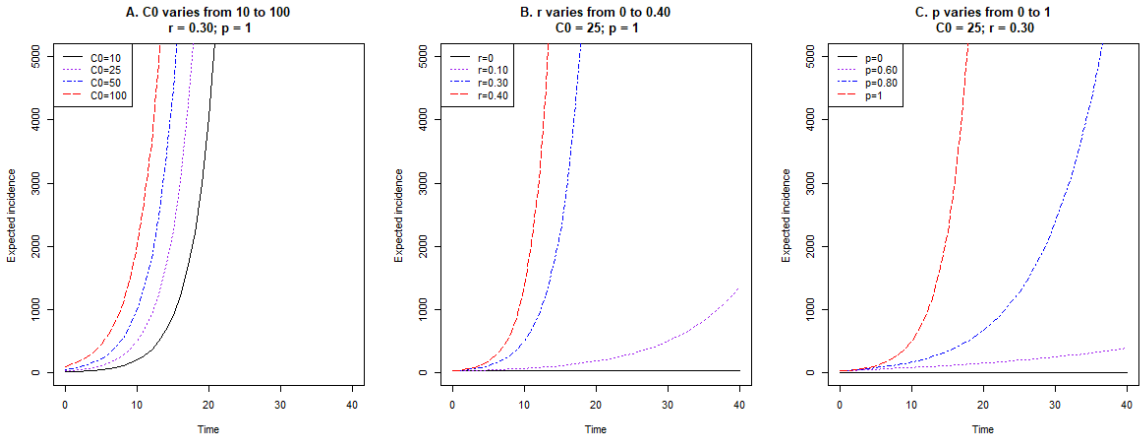

Figure 4.4 Illustration of the effect of varying the three parameters of a Poisson model allowing for sub-exponential growth on the expected mean number of new cases over time

Figure legend. In each panel, one parameters varies and the other two are fixed. Panel A illustrates the effect of the initial case number C_0, panel B the effect of the growth rate r, and panel C the effect of the growth scaling parameter p.

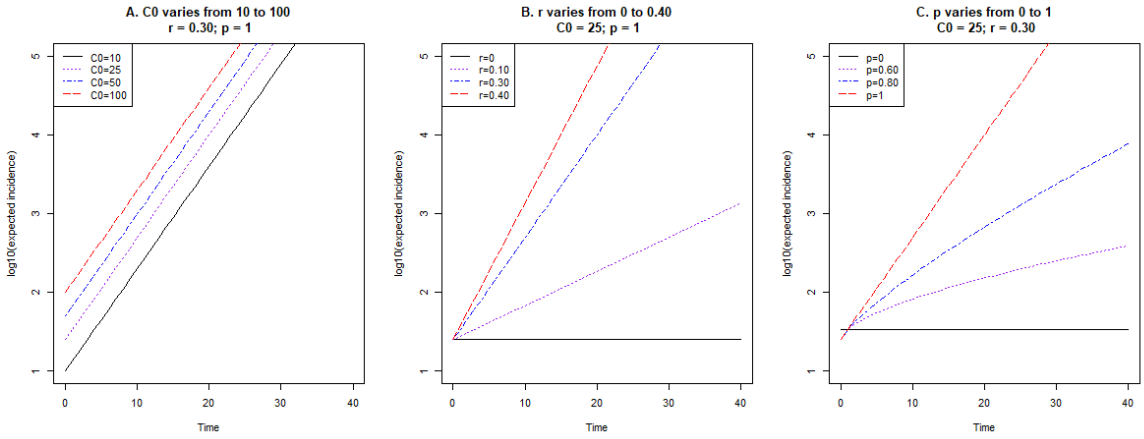

Figure 4.5 Illustration of the effect of varying the three parameters of a Poisson model allowing for sub-exponential growth on the expected mean number of new cases over time, on a semilogarithmic scale

Figure legend. In each panel, one of the three parameters varies. Panel A illustrates the effect of the initial case number C_0, panel B the effect of the growth rate r, and panel C the effect of the growth scaling parameter p. On a semilogarithmic scale, straight lines indicate exponential growth and lines with downward curvature indicate sub-exponential growth.

Populations and sub-populations

The phenomena of overdispersion and sub-exponential growth described above can be assumed and described in a statistical analysis without knowing the underlying reasons. However, as discussed in section 4.1, if considerable heterogeneity is observed, understanding where this comes from is of public health importance. If data are available for variables that are potentially relevant, the population can be disaggregated and growth models can be fitted for sub-populations. For example, an outbreak can be described at the level of a country, its provinces, or districts, which may highlight differences both within and between levels (*heterogeneity in space*). A population can also be disaggregated based on personal characteristics such as age, socio-economic conditions, or health-related behaviour (*heterogeneity in person*). Finally, one can also imagine *heterogeneity in time*.

4.4 Box: generalised growth model

Note: the content of this box is adapted from Viboud et al. 2016 and Ganyani et al. 2020 (36,52)

The generalised growth model is based on the differential equation:

$$\frac{dC_t}{dt} = C'_t = rC_t^p$$

Where:
C_t represents the cumulative number of cases at time t (cumulative incidence).
The derivative of the function $\frac{dC_t}{dt}$ can also be written as C'_t. It describes the incidence over time.
The parameter r is the growth rate; r can take positive values.
The parameter p is the growth scaling parameter; p can take values from 0 to 1.

Interpretation of p:
- If p = 0, the incidence is constant and the cumulative incidence grows linearly.
- If p = 0.5, incidence grows linearly and cumulative incidence follows a quadratic polynomial.
- If p = 2/3: incidence grows quadratically and cumulative incidence follows a cubic polynomial.
- If p equals 1: growth is exponential (Malthus equation).
- If p > 1: growth is faster than exponential; this is not considered here.
- Values of p between 0 and 1 indicate sub-exponential growth.

In practice, outbreak data are observed in discrete time intervals. A discrete approximation is:

$$\frac{C_{(t+h)} - C_t}{h} = C'_t = rC_t^p$$

Where:
h denotes the length of the discrete time step between consecutive reporting periods
$C_{(t+h)} - C_t$ is the number of individuals who become infected during the time interval [t, t+h)

The generalised growth model can be solved via a polynomial equation of degree m:

$$C_t = \left(\frac{r}{m}t + A\right)^m$$

Where:
The parameter m is a positive integer, and the growth scaling parameter $p = 1 - \frac{1}{m}$.
A is related to the number of cases at the start of the outbreak C_0 as follows: $A = \sqrt[m]{C_0}$

The generalised growth model is formulated as a generalised nonlinear model, which consists of three elements: (i) a probability function of the exponential family; (ii) a nonlinear predictor; and (iii) a link function. Two candidate probability functions (Poisson and negative binomial) are discussed in Box 4.3. Let y_t denote the observed number of individuals who become infected in the interval [t, t+h), then $y_t \sim Poisson\ (\mu_t)$ or $y_t \sim Negative\ binomial\ (\mu_t, \theta)$, where $\mu_t = r * C_t^p * h$ is the nonlinear predictor with identity link function. This implies that given C_t and model parameters, the incidence counts y_t are independent (conditional independence assumption).

For sub-exponential growth dynamics, the relative growth rate decreases with time:

$$\frac{dC_t/dt}{C_t} \propto \frac{m}{t}$$

Exponential growth is characterised by constant doubling times. However, when growth dynamics are sub-exponential, the doubling time T_d increases with time:

$$T_d \propto \frac{t(ln2)}{m}$$

For example, the spread of Ebola in West Africa in 2013 to 2016 appeared exponential at country level over short time intervals. However, that turned out to be a composition of smaller epidemics at district or county level. These epidemics at a lower level of aggregation had varying growth patterns (from sub-exponential to exponential) and occurred one after the other (Figure 4.2) (53,55,63).

4.2.3 Consequences of ignoring heterogeneity

Biased parameter estimates and low coverage of the 95% confidence intervals

Using a Poisson distribution in a generalised growth model when, in fact, there is overdispersion, can lead to misleading results. Ganyani et al. used a simulation experiment to evaluate parameter estimation procedures in different circumstances (52). Three problems with the Poisson model were identified in the presence of overdispersion, regardless of whether the true growth pattern was exponential or sub-exponential. First, the Poisson model overestimates the growth rate r and underestimates the growth scaling factor p (the point estimates are biased). Second, the sample-to-sample variability of the parameter estimates is high (sample-to-sample estimates are not well concentrated around their average). Third, the coverage, calculated as the proportion of times that the 95% confidence interval contains the true value, is low (the 95% confidence intervals are narrow but this impression of precision is false). As expected, the negative binomial model gives better results in the presence of overdispersion: bias and sample-to-sample variability are smaller and coverage is higher. On the other hand, the estimate of the dispersion parameter θ is variable and tends to underestimate the degree of dispersion. In absence of overdispersion, the Poisson and negative binomial models yield similar results (52). The problems related with overdispersion and model misspecification are especially prominent if time series are short (few data points) (52,56).

Mismatch between predictions and observed data

Different types of models are being used to predict the trajectory and size of an epidemic, including phenomenological (e.g. logistic growth) and mechanistic (e.g. compartmental) models. The classical standard versions of these models assume homogeneous mixing and exponential growth during the early phase of an epidemic (57). In outbreaks with an exponential growth pattern such as the 1905 outbreak of bubonic plague in Bombay and the 1918 influenza epidemic in San Francisco, compartmental models provide an excellent fit (64,65). On the other hand, if growth is sub-exponential, classical compartmental models may produce poor forecasts. More elaborated mechanistic models such as metapopulation and individual-based models can account for sub-exponential growth if the sources of heterogeneity are known, identifiable, and quantifiable (14). When this is not the case, the generalised growth model is expected to fare better for short-term forecasting of epidemic growth (36,52). Indeed, as illustrated in Figure 4.4, epidemic size is highly sensitive to small variations in the growth scaling parameter p (36).

Predictions of the expected number of Ebola cases in West Africa based on models assuming exponential growth greatly overestimated final epidemic size (14). Some researchers forecasted hundreds of thousands and up to 1.4 million Ebola cases (66). However, according to the US Centres of Disease Control, the outbreak ended with approximately 28600 cases and 11325 deaths (67). This led to criticism in academic as well as in newspaper publications (14,68,69). For example, one journalist described it as "*a habit of willfully ignoring the complexities of disease outbreaks, resulting in estimates that overdramatize how bad an outbreak could get — estimates that may be skewed by politics. They say (…) also overestimate how much vaccine is needed and how beneficial it has been*" (69). It is important to note that if model assumptions are not met, not only the forecasted case numbers are wrong, but also the predicted and reported impact of disease control measures.

4.2.4 Consequences of addressing heterogeneity

Interpretation of generalised growth model parameters

In a generalised growth model, the interpretation of the parameters changes because they are now correlated (Box 4.4, p is in the exponent of the expression). The estimated initial case number C_0 and the growth rate r do not have an easy interpretation anymore. In addition, the parameter r depends on the scaling of time in the equation (53). As a consequence, a generalised growth model does not give a direct equivalent to the basic reproduction number R_0 which has become a well-known parameter in the domain of public health and beyond (36).

On the other hand, the growth scaling parameter p is an essential parameter of the generalised growth model. It helps to assess whether the assumption of exponential growth is appropriate. This information is important for the development of further disease transmission models, whether they are phenomenological or mechanistic (54). Moreover, the parameter p by itself can be interpreted as a signature feature of the growth kinetics of an outbreak: it helps to understand the transmission dynamics of an infectious disease and may shed light on the type and intensity of interventions that are needed (36,53). For example, the parameter p could be used to assess the risk of observing a major outbreak: a value close to 1 indicates a high threat level.

In the Ebola outbreak in West Africa in 2013 to 2016, there was a clear correlation between the estimated parameter p before the epidemic peak and the observed epidemic size later on. For 24 subnational outbreaks in the West African epidemic, Spearman's rank correlation coefficient for the correlation between the parameter p estimated three weeks before the peak and the final observed epidemic size was 0.67 (95% confidence interval 0.33-0.87) (53). It was also possible -by defining a negative binomial regression model- to predict the epidemic size of other Ebola outbreaks (in DRC in 1976 and in Uganda in 2000) based on the parameter p (53).

Identification of growth patterns across contexts

Fitting growth models for relevant subgroups of a population (e.g. at fine spatial scales) can lead to insights that cannot be obtained via the study of a larger population (55). Such insights could be used to target control measures to areas with close-to-exponential growth. Fine-scale information is also immediately useful if response decisions are taken at sub-national level (53,55,63). Such an approach can also be used to assess shifts in epidemic growth patterns at lower levels. In this way, the characterisation of growth patterns across contexts (in terms of space, person, and time) can improve the reliability of forecasts and the assessment interventions (36). On the other hand, systematically characterising spatial (and other) heterogeneity increases complexity, which, in times of an emergency, may not be what politicians and the general public want to hear (69).

4.3 Modelling early growth of COVID-19 in Belgium with considerations about heterogeneity

4.3.1 Introduction

The considerations about heterogeneity in phenomenological models describing the early growth of epidemics that were discussed in the previous section are now applied to the COVID-19 epidemic using publicly available incidence data. The specific research questions in this section are:

- Are the COVID-19 incidence data for Belgium compatible with equidispersion or overdispersion?
- Is the early growth pattern exponential or sub-exponential?

- Does the time window on which the model is based influence key parameter estimates?
- Do models based on incidence *versus* cumulative incidence counts lead to similar conclusions?
- Do different measures of disease frequency (incidence of COVID-19 cases, hospital admissions, or deaths) lead to similar parameter estimates?
- Does an analysis of the data at provincial level lead to similar parameter estimates?
- Does an analysis in subgroups based on age and sex affect the parameter estimates?
- Do data from other countries lead to similar conclusions regarding overdispersion and sub-exponential growth?

4.3.2 Equidispersion or overdispersion

One consideration in the statistical analysis is the choice of a probability distribution. In the conceptual framework by Krupnick et al. (Figure 3.1), this consideration is classified under *model uncertainty* (13). As this project is about modelling count data (incidence of admissions), I consider the following probability distributions as candidates: a Poisson distribution with two parameters (initial case number C_0 and growth rate r) *versus* more flexible distributions with three parameters (initial case number C_0, growth rate r, and dispersion parameter θ) such as the negative binomial or the Poisson inverse Gaussian distribution.

The incidence of hospital admissions in Belgium was approximated using Poisson, negative binomial, and Poisson inverse Gaussian distributions (Table 4.5). The AIC values of the more flexible models are clearly better than that of the simpler Poisson model, suggesting that there is considerable overdispersion. The low point estimates of the dispersion parameter θ lead to the same conclusion, because θ is parameterised such that lower values of θ indicate larger deviations from a Poisson process. Overdispersion can be interpreted as unobserved heterogeneity, i.e. there is more variability in the data than what can be explained by a simple model including time and no other covariates.

Table 4.5 Comparison of Poisson, negative binomial, and Poisson inverse Gaussian distributions to model hospital admissions in Belgium: parameter estimates and Akaike's information criterion

	Poisson	Negative binomial	Poisson Inverse Gaussian
C_0 [95% CI]	44 [41-48]	24 [12-36]	20 [10-29]
r [95% CI]	0.150 [0.145- 0.156]	0.204 [0.156- 0.251]	0.223 [0.179- 0.266]
θ [95% CI]	assumed ∞	3.8 [1.8-8.2]	3.3 [1.3-8.1]
AIC	549	240	245

AIC: Akaike's information criterion; CI: confidence interval; C_0: initial number of cases; r: growth rate; θ: dispersion parameter. The time window for model fitting is 20 days (10 to 29 March).

Because the negative binomial model gives the best AIC values, I used this as the standard approach for the remainder of the results section. However, to illustrate the impact of this decision, I repeated key analyses using the alternative approach (Poisson model) and present the results as supplementary material. Expected consequences of the choice for the negative binomial model are (i) wider confidence intervals (Table 4.5), and (ii) the possibility of computational difficulties due to the extra parameter θ that has to be estimated from limited data points. Of note, the decision on the probability distribution does not only affect the confidence intervals; it also leads to different estimates of parameters C_0 and r. This is illustrated in Figure 4.6, where negative binomial and Poisson models lead to different fitted curves.

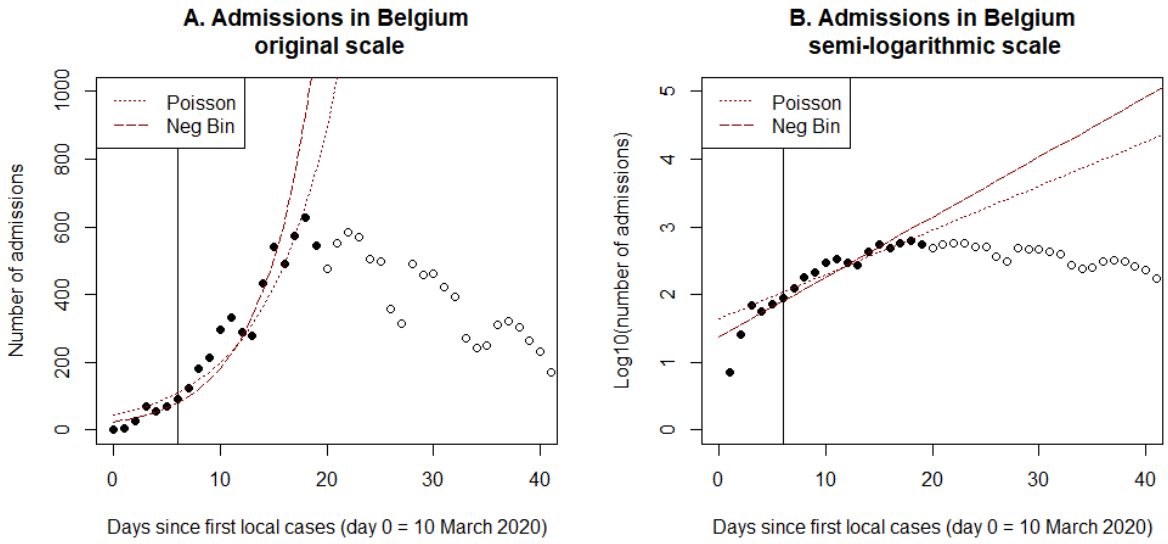

Figure 4.6 Incidence of hospital admissions in Belgium and comparison of Poisson and negative binomial model approximations (equidispersion versus overdispersion)

Figure legend. The Poisson and negative binomial model approaches both assume exponential growth. The Poisson model (red dotted line) also assumes equidispersion. The negative binomial model (red dashed line) allows for overdispersion. Based on Akaike's information criterion, the negative binomial model fits the data better than the Poisson model. The black solid vertical line indicates the lockdown date (16 March). The dots are the observed admission counts; those filled in black indicate the time window used for model fitting (10 to 29 March). Panel A: original scale. Panel B: semi-logarithmic scale.

4.3.3 Exponential or sub-exponential growth

Another consideration is whether it is appropriate to assume that the incidence data follow an exponential growth pattern. In the framework of Krupnick et al., such a consideration is classified under *simplification* and is viewed as a type of *model uncertainty* (13). The assumption of exponential growth can be relaxed through the formulation of a generalised growth model that incorporates a growth scaling parameter (Box 4.4). This growth scaling parameter p ranges from 0 to 1 and allows for less-than-exponential growth: p=0 indicates that the incidence count does not change over time; p=0.5 is equivalent to linearly growing counts; and p=1 is equivalent to exponentially growing counts.

The incidence of hospital admissions in Belgium was approximated using negative binomial models with and without a growth scaling parameter. The findings are summarised in Table 4.6 and Figure 4.7. The model allowing for sub-exponential growth appears to be the best choice because the growth scaling parameter p is smaller than 1 (confidence interval does not include 1). Moreover, despite the extra parameter p, this more complex model has a better AIC value (Table 4.6). The point estimate of p (0.59) indicates a growth pattern that lies between linear and exponential growth. On a semi-logarithmic scale, this results in a line with downward curvature (Figure 4.7). These findings indicate that the assumption of homogeneous mixing leading to exponential growth is unlikely.

The approaches with and without growth scaling parameter p lead to different estimates of the growth rate r and the dispersion parameter θ. The point estimate of the growth rate r is higher in the model allowing for sub-exponential growth (4.120) than in the model assuming exponential growth (0.240). This can be interpreted as follows. In the scenario of sub-exponential growth, the outbreak has, in the early phase, the potential to grow very fast, but due to factors such as inhomogeneous population mixing and reactive behaviour changes (reflected in the exponent p smaller than 1), the effective growth rapidly decays. This decay is illustrated by the downward curvature of the blue dashed line in Figure 4.7. The point estimate for the dispersion parameter θ is also higher in the

model that allows sub-exponential growth (43.7) than in the exponential model (3.8). The sub-exponential model comes closer to the Poisson situation (as higher values of θ correspond to less overdispersion as illustrated in Figure 4.1). In other words, a part of the overdispersion described before was due to model misspecification and can be reduced by allowing sub-exponential growth.

Table 4.6 Comparison of a negative binomial model with and without growth scaling parameter p: parameter estimates and Akaike's information criterion

	Negative binomial model without growth scaling parameter (assuming exponential growth)	Negative binomial model with growth scaling parameter (allowing for sub-exponential growth)
C_0 [95% CI]	24 [12-36]	0*
r [95% CI]	0.204 [0.156-0.251]	4.120 [3.652-4.649]
θ [95% CI]	3.8 [1.8-8.2]	43.7 [5.3-82.1]
p [95% CI]	assumed 1	0.59 [0.57-0.61]
AIC	240	194

AIC: Akaike's information criterion; CI: confidence interval; C_0: initial number of cases; p: growth scaling parameter; r: growth rate; θ: dispersion parameter. The time window for model fitting is 20 days (10 to 29 March). *The 95% confidence interval is not given as computation is not straightforward anymore (C_0 indirectly parameterised). The interpretation of the parameters C_0 and r is not the same in the two approaches: the estimates in the right column come from an expression with nonlinearity in the parameters (p is in the exponent of the expression), hence, the parameters are correlated.

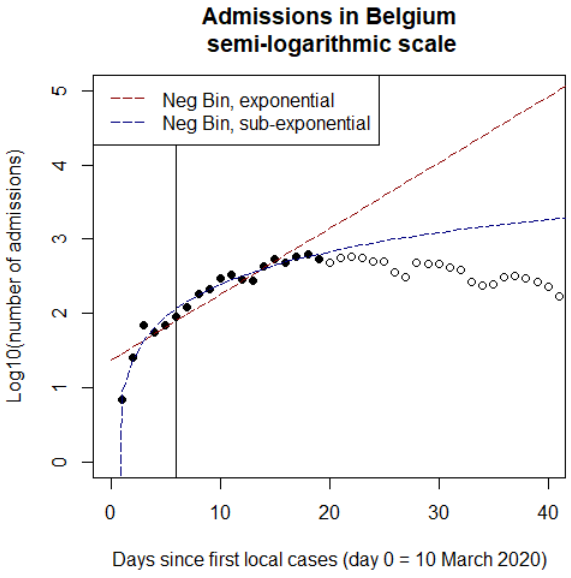

Figure 4.7 Incidence of hospital admissions in Belgium and comparison of negative binomial models with and without a growth scaling parameter (exponential versus sub-exponential growth)

Figure legend. Two negative binomial models are compared. The red dashed line is based on a model assuming exponential growth, which is equivalent to fixing the growth scaling parameter p to 1. On a semi-logarithmic scale, exponential growth results in a straight line. The blue dashed line is based on a model allowing for sub-exponential growth, where the growth scaling parameter p is estimated from the data. The point estimate for p is 0.59 (95% confidence interval: 0.57 - 0.61). The sub-exponential model provides the best fit to the data. On a semi-logarithmic scale, sub-exponential growth results in a line with downward curvature. The black solid vertical line indicates the lockdown date (16 March). The dots are the observed admission counts; those filled in black indicate the time window used for model fitting (10 to 29 March).

Because the data are in line with a pattern of sub-exponential growth, I used the generalised growth model with a growth scaling parameter p as the standard approach for the remainder of the results section. To illustrate how this decision affects the findings, I repeated key analyses under the alternative assumption of exponential growth and summarise them in the supplementary material. The decision to continue with the generalised growth model has several implications: (i) there is yet another parameter (p) to estimate from the data; (ii) problems with parameter identification may arise if there are various local minima; and (iii) the interpretation of the parameters C_0 and r is not straightforward anymore because the different parameters are now correlated (nonlinearity in parameters: p appears in the exponent of expression Y). Therefore, from now onwards, the focus will be on characterising the early growth pattern of the incidence data by estimating the growth scaling parameter p, with less emphasis on the parameters C_0 and r.

4.3.4 Variability in time

In this project, I try to characterise the pattern of the early ascending phase of the outbreak. The data used to estimate the model parameters should therefore represent the beginning of the outbreak, before formal interventions change its course. On the other hand, a sufficient amount of data points is needed to fit models with reasonable precision. To model hospital admissions, I took 10 to 29 March as the basic time window. In this subsection, I assess if different choices would lead to different findings. In the framework by Krupnick et al., this can be viewed as a consideration of *variability*, i.e. the growth pattern may really change over time (true heterogeneity). There may also be a component of *parameter uncertainty* (and in particular *extrapolation errors*) if the way cases and admissions are diagnosed, registered, and reported changes over time (13).

In Belgium, the first COVID-19 cases started to appear on 1 March 2020, but these people probably contracted the infection elsewhere as many of them had travelled in the days before symptom onset (19–22). Regarding hospital admissions, from 10 to 14 March, there is data available from newspapers and other situation reports. From 15 March onwards, systematically collected admission data is available via Sciensano (33). Decisions about control measures were taken in the week of 9 to 15 March, and key elements of the lockdown were in place by Monday 16 March (20,24). The lockdown was expected to have a clear impact on the incidence of hospital admissions by the end of March, taking into account the incubation time (average of five days between exposure and symptom onset) and the additional time lag between symptom onset and hospital admission.

Changing the length and the starting day of the time window used to estimate the model parameters does not affect the finding of sub-exponential growth: the growth scaling parameter p is always lower than 1 (Figure 4.8). Extending the time window to 25 days (Figure 4.8, panel C) results in a slightly lower point estimate of p (0.57) and a slightly narrower 95% confidence interval (0.56-0.60). Shifting the starting day of the time window has a larger impact. Taking 18 March as a starting day (Figure 4.8, panel F) gives a low estimate of the growth scaling parameter p (0.35; 95% confidence interval: 0.24-0.53). This last scenario seems to capture a different phase of the outbreak, where growth deviates even further from the exponential pattern.

4.3.5 Incidence or cumulative incidence

Model approximations for the ascending phase of an outbreak can be fitted using incidence or cumulative incidence data. As illustrated in Figure 4.9, the curve of the observed cumulative incidence counts (panel B) is smoother than that of the incidence counts (panel A). This is also reflected in the dispersion parameter θ: there is far less overdispersion with the cumulative incidence (high estimate of θ: 167.6) than with the incidence (low estimate of θ: 43.7).

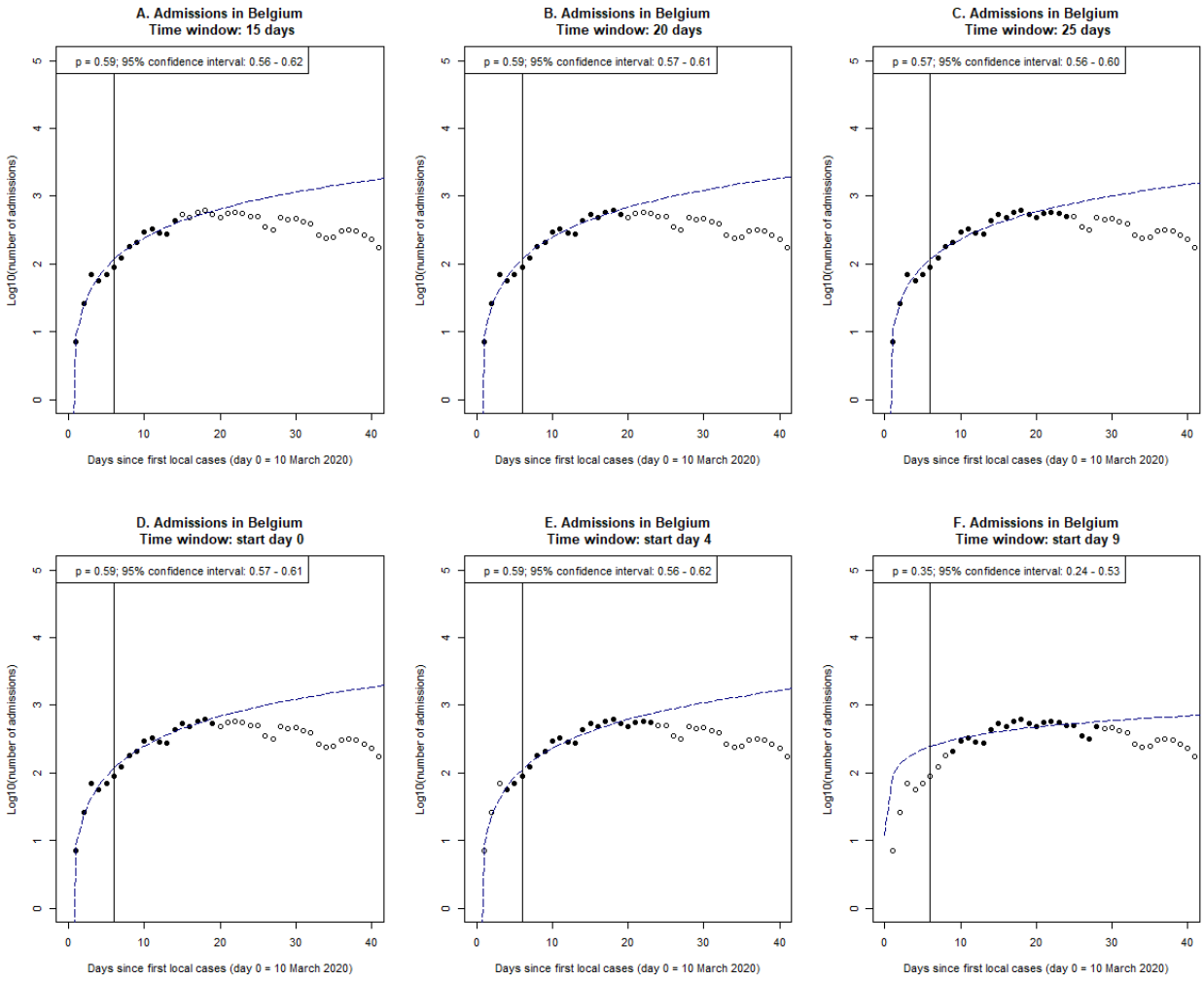

Figure 4.8 Incidence of hospital admissions in Belgium: influence of time window on fitted curves using a negative binomial model with a growth scaling parameter

Figure legend. The dots represent the observed admission counts and those filled in black indicate the time window used for model fitting. In panels A to C, the length of the time window varies. In panels D to F, the starting day of the time window varies. The models indicate sub-exponential growth regardless of the time window. The black solid vertical line indicates the lockdown date (16 March).

Because the cumulative incidence somehow contains variability (e.g. noise in the data) that is carried forward from previous moments, the sequential measurement error is not independent. This would be problematic if model parameters were estimated via a least squares approach (68). However, in this study, maximum likelihood estimation was used, which allows for heteroscedasticity. As shown in Figure 4.9, modelling incidence or cumulative incidence of admissions leads to similar point estimates for the growth scaling parameter p (0.59 versus 0.60).

Although the authors of the first publications on generalised growth models for outbreaks used cumulative incidence data (14,36), we chose to fit the models based on incidence data. This is a decision that concerns the way in which an empirical quantity is measured, which can be classified under *extrapolation errors*, a type of *parameter uncertainty* in the framework by Krupnick et al. (Figure 3.1) (13).

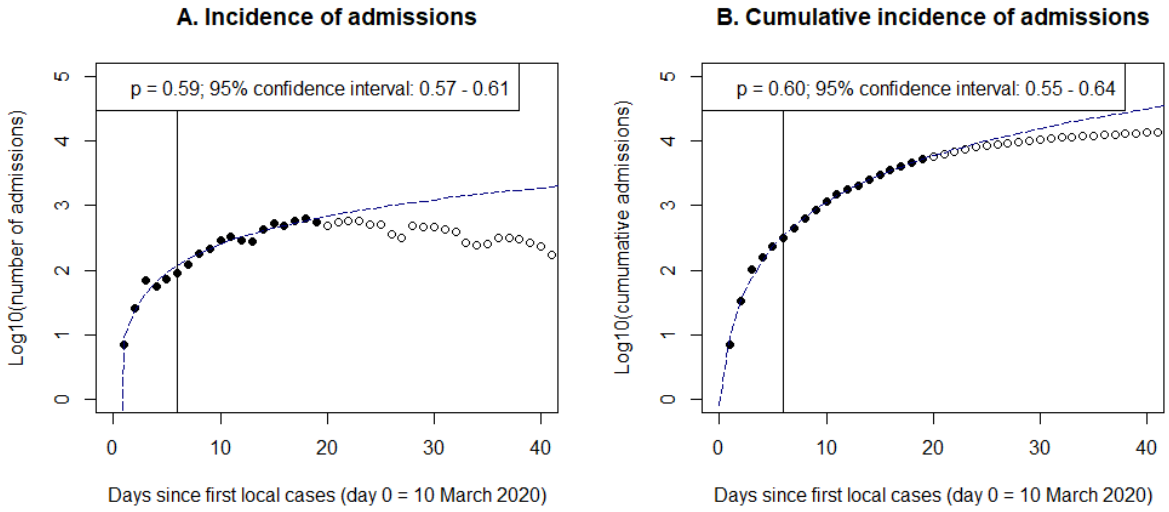

Figure 4.9 Incidence and cumulative incidence of hospital admissions in Belgium and fitted curves using a negative binomial model with a growth scaling parameter

Figure legend. The dots represent the observed incidence counts (panel A) and cumulative incidence counts (panel B). The observed cumulative incidence (panel B) appears smoother than the incidence (panel A). The two approaches lead to similar estimates of the growth scaling parameter p. The black solid vertical line indicates the lockdown date (16 March). The dots filled in black indicate the time window used for model fitting (10 to 29 March).

4.3.6 Incidence of registered cases, hospital admissions, or deaths

In the Sciensano database, there are three measures of epidemic growth: the incidence of registered COVID-19 cases, hospital admissions, and deaths. None of these measures is perfect. The number of registered cases per day depended on the availability and accuracy of diagnostic tests, health seeking behaviour, testing processes in the health (and care) sector, and accuracy of the reporting, all of which changed over time (20,23,33). The incidence of hospital admissions could be considered a more stable measure than the case incidence, because -at least in Belgium- the admissions depended less on patient preferences and testing algorithms (nearly all admitted patients with COVID-19 suspicion were tested). On the other hand, many COVID-19 patients living in homes for the elderly were never referred to a hospital, so they are underrepresented in the admission data. Finally, the incidence of deceased people may represent the severe subcomponent of the outbreak but does not necessarily reflect the spread of the virus in the whole population. In addition, the case definition of COVID-19-related death includes both confirmed and unconfirmed cases (32). The choice for one of the three measures implies considerations of *parameter uncertainty* (which measure is less prone to error) and of *decision uncertainty* (which measure reflects what is needed for policy; Figure 3.1) (13).

The impact of the measure choice on the characterisation of the growth pattern is shown in Figure 4.10. The three measures all lead to the conclusion of sub-exponential growth, but there are differences as well. First, although the point estimate of the growth scaling parameter p for the cases is similar to that of the admissions, the confidence interval is considerably wider, suggesting more uncertainty or variability for the cases. Second, using the incidence of deaths as a measure leads to a higher value of p (0.77) compared to the incidence of cases (0.60) or admissions (0.59). In other words, the deviation from exponential growth is more pronounced for the admissions than for the deaths.

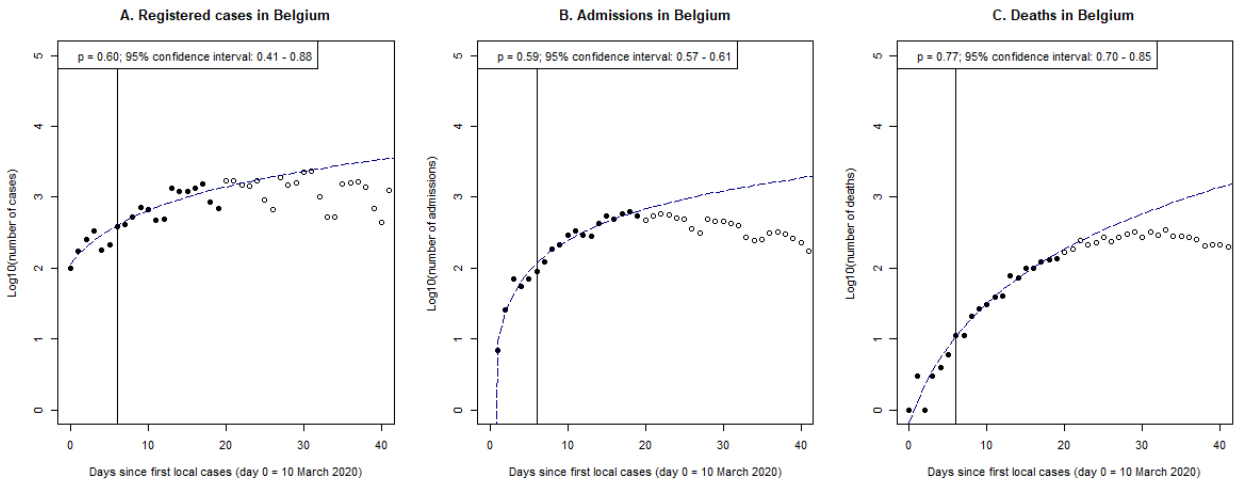

Figure 4.10 Incidence of registered cases, hospital admissions, and deaths in Belgium with fitted curves from a negative binomial model approximation with a growth scaling parameter

Figure legend. The black solid vertical line indicates the lockdown date (16 March). The dots represent the observed incidence counts; the dots in black were used to fit the models (time window from 10 to 29 March).

4.3.7 Variability across provinces

The next question is whether the growth patterns per province resemble that of the aggregated national level. This is can be viewed as a matter of *variability* (spatial heterogeneity) and of *system resolution* (classified under *model uncertainty* in Figure 3.1) (13).

The estimates of the growth scaling parameter p for the 10 Belgian provinces and for the capital Brussels are summarised in Figure 4.11. The estimate at Belgian level is included in the figure for comparison. These estimates are based on the incidence of hospital admissions from 15 to 29 March. The province of Hainaut has the highest estimate for p and the confidence interval includes 1 (p=0.84; 95% confidence interval: 0.66-1.07). Hence, the growth pattern in Hainaut cannot be significantly distinguished from exponential growth. In all other provinces, the growth pattern is sub-exponential. The province with the lowest p value is Brabant wallon, but the confidence interval is very wide (p=0.37; 95% confidence interval: 0.14-0.98). Taken together, apart from Hainaut, there are no clear differences in growth patterns across provinces or between provincial and national level. In addition, there is no clear correlation between the growth scaling parameter p and outbreak size (as of 16 May 2020) per province: Spearman's rank correlation coefficient ρ is 0.42 (P-value 0.20).

4.3.8 Variability according to age and sex

Within one large population, there may be subpopulations displaying different patterns of epidemic growth. Indeed, characteristics of persons are yet another possible source of *variability* or true heterogeneity (Figure 3.1) (13). The Sciensano database contains information about age and sex for the incidence of registered cases and for the incidence of deaths (but not for the admissions).

The findings according to sex are summarised in Table 4.7. When the incidence of registered cases is used to fit the models, the growth pattern is not well defined: for men, the confidence interval for the growth scaling parameter p is very wide; for women, the confidence interval could not be determined using the methods described in this project. When the incidence of deaths is used to fit the models, the point estimate of p is somewhat higher in women (0.80) than in men (0.74), but the confidence intervals overlap considerably.

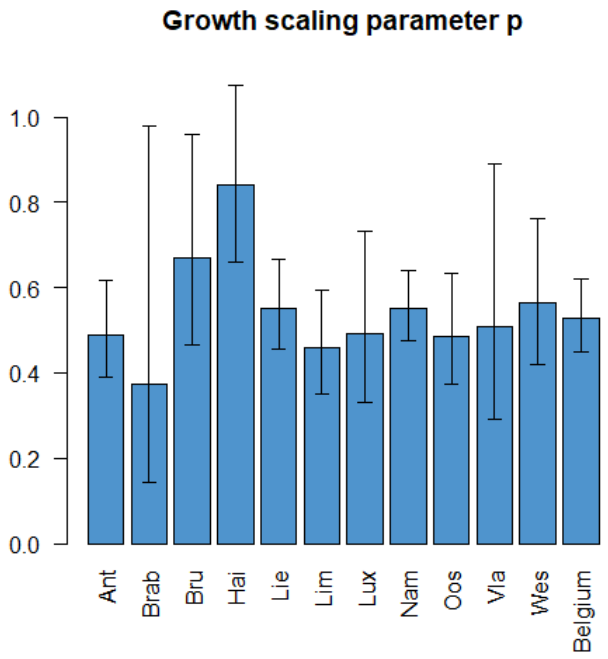

Figure 4.11 Point estimates and 95% confidence intervals for the growth scaling parameter p per province in Belgium

Figure legend. Negative binomial models allowing for sub-exponential growth were fitted using the incidence of admissions in each province and in the capital Brussels. The values at national level are also given for comparison. The time window for model fitting was from 15 to 29 March (admission data are not available per province before 15 March).

Table 4.7 Comparison of the epidemic growth patterns in Belgian men and women using models based on incidence of cases and incidence of deaths

	Men, p [95% CI]	Women, p [95% CI]
Registered cases	0.60 [0.41 - 0.87]	0.32*
Deaths	0.74 [0.64 - 0.85]	0.80 [0.67 - 0.96]

CI: confidence interval; p: growth scaling parameter. The time window for model fitting is 20 days (10 to 29 March). *The 95% confidence interval is not given because of computational difficulties. The estimations are based on a generalised growth model using a negative binomial probability distribution.

With regard to age, the findings for registered cases and deaths are summarised in Figure 4.12. As in previous comparisons, confidence intervals are wider if the growth parameter p is estimated using cases compared to deaths, particularly for younger age categories. As illustrated in Figure 4.12, there is no clear trend in growth pattern according to age. Although it is clear from the clinical literature that the case fatality rate of COVID-19 increases with age (70), the pattern of epidemic growth does not appear to depend on age alone.

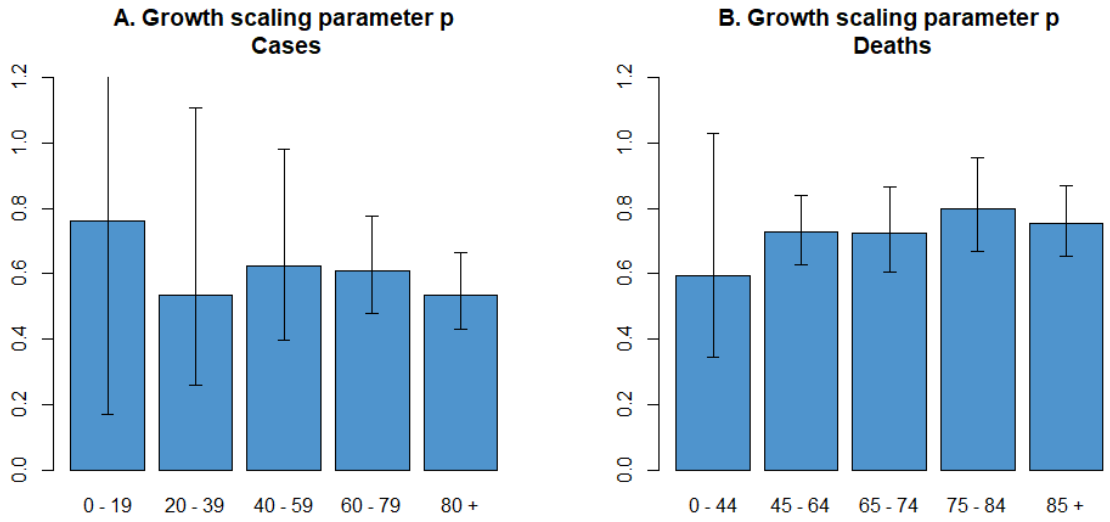

Figure 4.12 Point estimates and 95% confidence intervals for the growth scaling parameter p per age category, calculated for the incidence of registered cases and deaths

Figure legend. A negative binomial model allowing for sub-exponential growth was fitted for each age category. The time window used for model fitting was from 10 to 29 March. The growth scaling parameter p was estimated in models approximating the incidence of registered cases (panel A) and deaths (panel B). Cases and deaths are reported in Belgium using different age categorisations.

4.3.9 Comparison across countries

In this last subsection, I explore whether incidence data from diverse settings outside Belgium lead to similar conclusions in terms of overdispersion and sub-exponential growth. This can be viewed as another dimension of *variability* (true heterogeneity); as *extrapolation* errors (a type of *parameter uncertainty* or heterogeneity due to methods); or both (Figure 3.1) (13).

The findings regarding the issue of overdispersion are summarised in Table 4.8. As expected, the 95% confidence intervals obtained with the Poisson model are narrower than those obtained with the negative binomial model. However, the negative binomial distribution results in a better fit than the Poisson distribution in nearly all comparisons, across settings and measures (incidence of registered cases or deaths). The only exception is the model for the incidence of deaths in Chile, where the AIC value from the Poisson model is better than that from the negative binomial model. In this setting, the two models give nearly the same point estimate and 95% confidence interval for the growth scaling parameter p. The difference in AIC is 2, which corresponds to the penalisation for the extra parameter (dispersion parameter θ) in the negative binomial model. Taken together, assuming overdispersion appears to be the best option.

Regarding the growth pattern, two settings are compatible with exponential growth: the incidence of cases at the Diamond Princess cruise ship and the cases in Belgium (Table 4.8 and Figure 4.13). All other settings indicate sub-exponential growth (Table 4.8 and Figure 4.14). For the incidence of cases, Belgium is the country with the highest point estimate for the growth scaling parameter p (0.88; closest to exponential growth) and Sweden has the lowest estimate (0.50). For the incidence of deaths, Belgium has the highest estimate for p (0.77) and Chile the lowest (0.19) (Figure 4.14). It is important to bear in mind that in this subsection, the data for Belgium come from the European Centre for Disease Prevention and Control (instead of Sciensano in previous subsections) and the time window for model fitting starts the first day of three consecutive days of monotonic growth

(instead of 10 March in previous subsections). For the incidence of cases, this leads to a higher point estimate of the growth scaling parameter p here (0.88 in Table 4.8) than before (0.60 in Figure 4.10). For the incidence of deaths, the estimates are similar (0.77 in Table 4.8 and 0.77 in Figure 4.10).

Table 4.8 Comparison of Akaike's information criterion and growth scaling parameter p using Poisson and negative distributions for the incidence of cases and deaths in different countries

	Poisson AIC	Poisson p [95% CI]	Negative binomial AIC	Negative binomial p [95% CI]
A. Cases				
Diamond Princess	356	1.00*	152	0.92 [0.52 - 1.64]
Belgium	713	0.93 [0.92 - 0.94]	277	0.88 [0.74 - 1.04]
Spain	680	0.81 [0.80 - 0.82]	275	0.86 [0.83 - 0.89]
Sweden	536	0.49 [0.47 - 0.51]	250	0.50 [0.44 - 0.57]
Peru	614	0.68 [0.61 - 0.76]	223	0.63 [0.52 - 0.77]
Chile	318	0.66 [0.64 - 0 .69]	231	0.70 [0.64 - 0.76]
B. Deaths				
Belgium	328	0.79 [0.72 - 0.87]	189	0.77 [0.60 - 0.99]
Spain	459	0.71 [0.69 - 0.73]	233	0.73 [0.62 - 0.86]
Sweden	213	0.62 [0.54 - 0.73]	163	0.56 [0.45 - 0.70]
Peru	161	0.45 [0.35 - 0.57]	130	0.45 [0.29 - 0.69]
Chile	98	0.19 [0.08 - 0.47]	100	0.19 [0.08 - 0.48]

CI: confidence interval; p: growth scaling parameter. For the Diamond Princess setting, all available data are used for model fitting. For the country data, the time window for model fitting is 25 days for the incidence of cases and 20 days for the incidence of deaths. The starting day of the time windows is chosen per indicator and per country: it is the first day of the first series of three days of monotonic growth. *The 95% confidence interval is not given because of computational difficulties. The estimations are based on a generalised growth model using a Poisson or a negative binomial probability distribution.

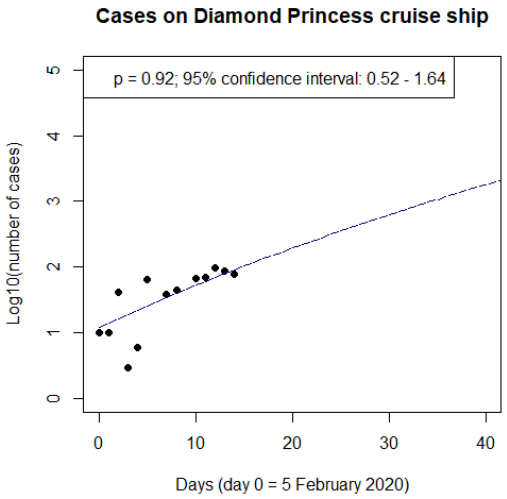

Figure 4.13 Incidence of cases on board of the Diamond Princess cruise ship and fitted curve based on a negative binomial model with a growth scaling parameter p

Figure legend. All available data (a series of 15 days) were used for model fitting. The growth pattern is close to exponential.

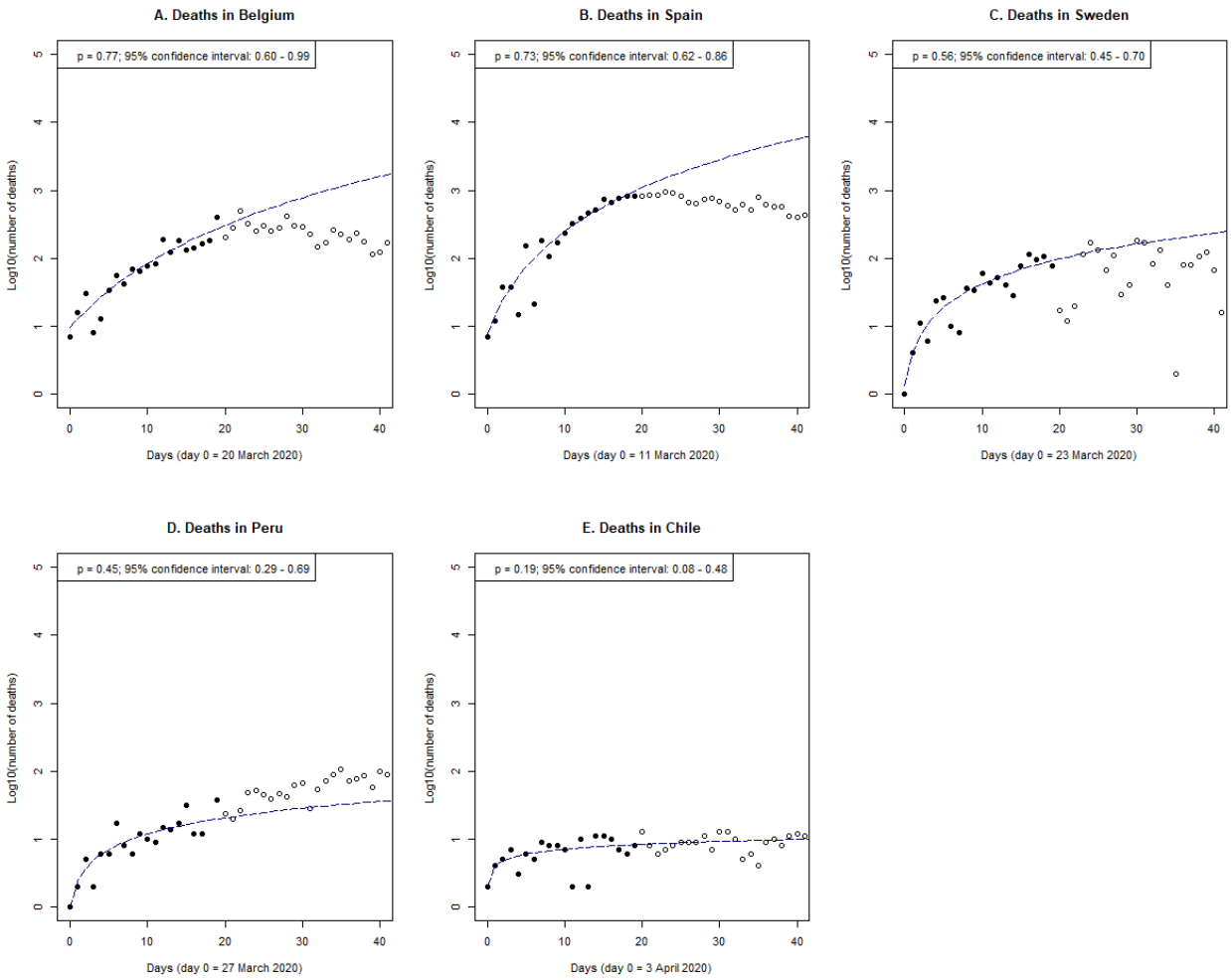

Figure 4.14 Incidence of deaths and estimation of growth scaling parameter p in different countries

Figure legend. Negative binomial models with a growth scaling parameter p were used to approximate the incidence of deaths in various countries in Europe (Panels A to C) and Latin America (Panels D and E). In each country, day 0 is defined as the first day of the first three days of monotonic growth in the incidence of deaths. The dots represent the observed incidence of deaths; the dots filled in black are the data points that were used to fit the models.

5 Discussion

5.1 Main findings

Heterogeneity is addressed in the recent literature on the epidemiology of infectious diseases using diverse methodological approaches. Usually, there are attempts to pin down the variables that underly population heterogeneity (mechanistic). Sometimes, statistical techniques just allow the heterogeneity to be there (phenomenological). Approaches to define subpopulations are pragmatic (working with data at hand), deductive (based on theory about transmission and pathogenesis), or inductive (data-driven disaggregation of populations). Insight in what makes populations heterogeneous for specific infections helps to design interventions, reach key individuals or subgroups, and properly estimate the impact of interventions.

In the case of COVID-19, I used a phenomenological generalised growth model to characterise the early phase of the outbreak and explored three ways to account for heterogeneity. First, there was overdispersion in the incidence data in nearly all the analyses, regardless of the level (province, Belgium, other countries), incidence measure (cases, admissions, deaths), and time window for model fitting. Second, the growth pattern was sub-exponential in most of the analyses, and the deviation from exponential growth varied depending on incidence measure and the starting date of the time widow. Third, growth patterns were similar across subgroups in Belgium defined based on age, sex, and province of residence. Repeating the analyses with data from other countries and comparing the findings revealed more variation in early growth patterns.

Accounting for overdispersion and sub-exponential growth in the model structure has a marked influence on the expected course of the COVID-19 outbreak: the number of expected cases and deaths is smaller and the uncertainty is larger in comparison with simpler models.

5.2 Limitations

The findings of this work have to be seen in light of some limitations. First, the study of the early growth of COVID-19 in Belgium is limited by the availability of data. I worked with variables included in public datasets. In other words, I took the pragmatic approach which I criticise in other parts of this document. I also compared growth patterns across countries as an exercise, without a thorough discussion on differences in context, case definitions, and data collection. This study did not include a formal validation strategy, apart from repeating key analyses for different countries and visualising expected together with observed epidemic trajectories on the graphs.

Second, the COVID-19 outbreak is ongoing and the sheer amount of information about it is daunting. Papers are being published at an unprecedented rate. As Nature Index put it: "*COVID-19 is not just a global pandemic but also a research publishing phenomenon*" (71). On 11 June 2020, about six months after the discovery of the virus, a PubMed search with "COVID-19" as a search term retrieved 21275 records (not including papers on preprint platforms). I most likely missed research papers relevant for the discussion of the COVID-19-related findings, as many came out while I was trying to round up the writing. Moreover, the thesis project grew somewhat organically from theoretical considerations into a practical exercise. This report is, therefore, inevitably incomplete.

Third, I proposed a broad framework to reflect about variability and uncertainty (Figure 3.1) but did not explore all its dimensions in this project. For example, I did not discuss the following elements of parameter uncertainty: unpredictability, conflicting data, and misclassification (viewed in the framework as one specific type of extrapolation error). The latter element could open a line of discussion by itself, i.e. about the accuracy of diagnostic procedures and case definitions of COVID-19 and differences in time-dependent misclassification across countries (72). I also disregarded decision uncertainty. An example of such a decision in the case of COVID-19 in Belgium is the emphasis that was put on hospital and intensive care unit admissions as a measure of disease frequency relevant for policy. Now, a few months later, some people argue *post hoc* that a focus on the number of cases in the community or in specific risk groups as a measure of disease frequency (instead of hospital admissions) could have led to different disease control interventions and a different outcome.

Fourth, I started this thesis project from the impression that in many epidemiological studies, heterogeneity is insufficiently accounted for. Although I set out to explore -from a statistical perspective- how heterogeneity can be addressed, my personal view on this may have influenced the way in which I approached the literature and presented this report.

5.3 Interpretation in the context of the COVID-19 pandemic

COVID-19 modelling studies play a prominent role in decision making. Many modellers used mechanistic approaches and estimated the effects of disease control measures by comparing the observed situation with a hypothetical counterfactual scenario (without control interventions) (72–76). In the context of this thesis, it is interesting to see how the modellers imagined these counterfactual scenarios. Some assumed that without centrally organised disease control interventions, the effective reproduction number R_t would remain constant, which is equivalent to assuming exponential growth (72,76). Others expressed R_t as a function of viral elements, season, and cross-immunity (75).

Diverse sources of heterogeneity that may influence epidemic trajectory are proposed in the COVID-19 literature. They overlap with the type of variables studied in other infectious diseases (Table 4.2). Examples are time and weather (season and temperature) (74,75); place of residence; built and household environment (population density) (74); demographics (age, sex, ethnic background) (77); socioeconomic and cultural factors (74); human behaviour (74); interventions and response (72–74,76); virus (transmission routes) (74); human biology and immunology (comorbidities and degree of cross-immunity with other coronaviruses) (75–77); and accuracy of diagnosis and reporting (77).

The COVID-19 modelling studies apply a broad range of methods that capture specific aspects of heterogeneity. In the family of mechanistic approaches, there are, among others, individual-based models (73), population-based models with many compartments (75), metapopulation models with age and social activity structure (78), and Bayesian perspectives allowing parameters to vary (28,72). Some studies highlighted overdispersion in COVID-19 transmission (also called super-spreading) and dealt with the phenomenon by using negative binomial distributions (79) or by creating a special compartment for super-spreaders (80). The idea of super-spreading is not specific for COVID; it has been studied in the past for infectious diseases with various transmission routes such as malaria, tuberculosis, SARS, measles, smallpox, monkeypox, and pneumonic plague (51,81,82).

Regarding the family of phenomenological approaches, I found three studies in which a generalised growth model was used (similar to what I did in this project) (83–85). In two of these studies, the authors reported estimates of the growth scaling parameter: 0.99 in Hubei (exponential pattern), 0.69 (sub-exponential pattern) in other Chinese provinces; and 0.9 (exponential pattern) in Lima, Peru (83). The fact that growth patterns vary across settings is in line with what I found. However, I reported a lower value for the growth scaling parameter p in Peru (0.63 compared to 0.9). Of note, there were at least five differences between the pre-published study (85) and my approach: they focused on the capital Lima while I looked at country level; they took the data from Peruvian sources (*Centro Nacional de Epidemiología, Prevención y Control de Enferemedades* and *Instituto Nacional de Salud*) while I took them from the European Centre for Disease Prevention and Control (30); they calibrated the model using an earlier time window (starting 29 February compared to 10 March in this study); they did not include the initial number of cases as a parameter in the model; and they used least squares while I used maximum likelihood estimation. One additional study using another type of growth model reported early sub-exponential growth in various countries (86). Taken together, such varying and sometimes sub-exponential growth patterns indicate that population factors may decelerate the spread of COVID-19 before official control measures have an effect.

Many COVID-19 models did not incorporate the possibility of subnational variation and of gradual or informal behaviour changes (72,75,76). Another strong assumption was that the efficacy of the interventions is constant over time and similar across countries (72). In addition, authors described concerns about parameter uncertainty, related to for example the number of imported cases, the amount of underreporting, and the effects of different, almost simultaneously organised interventions (72,75). The more authoritative studies mentioned such issues as study limitations (72,75). These studies usually contained a sensitivity analysis, which tended to focus more on parameter uncertainty than on model uncertainty (72,73,75).

One may wonder whether capturing fine-scale variation is relevant in times of a public health crisis where the positive impact of large standardised control measures appears to be clear (69,72,73,75). However, if specific sources of heterogeneity are structurally excluded from modelling studies, it becomes of course impossible to see their relevance. In a different domain, that of malaria elimination strategies, Peeters-Grietens et al. described such processes as follows: "*A crucial component of this misdirection process is the global standardization of intervention methodologies operating independent of local social contexts and the perceived impossibility to 'localize' such interventions. This conviction requires – and is simultaneously supported by- the production of decontextualized evidence through the application of methodologies aiming at generalizability, in detriment of social context and variability. This process produces pseudo measurements and conclusions that are at the same time real in their adherence to paradigmatically valid methodologies and fake as they either remain empty of empirical significance or whose validity cannot be assessed as we have lost sight of the (local, social, cultural) variation it has decided to ignore.*" (87)

It is worth mentioning that there is a clear and growing opposition against generalised control measures to reduce the spread of COVID-19. Arguments in this discussion overlap with those described in Table 4.1 about other infectious diseases. The motivation for the opposition is sometimes political, sometimes academic, and is fuelled by concerns about the negative impact of general measures on health (health problems other than COVID-19), wellbeing, and the economy, especially in low-and middle-income countries (88). It is worth noting that some countries have chosen, from the beginning, for a localised rather than a national lockdown (29); and some appeal to voluntary rather than enforced behaviour changes (29).

5.4 Challenges

While I was reviewing papers for this project, I came across recurring challenges that appear to be triggered by attempts to accommodate heterogeneity in studies of disease frequency. I summarised these challenges in five categories:

Confusion. There is confusion about the meaning of heterogeneity and its relation with uncertainty. In the literature, there are several classification schemes but no consensus. In the document introducing the framework of Figure 3.1, this is explained as follows: "*For example, take the circularity of the sentence, "There are two kinds of uncertainty: variability and uncertainty." Depending on the context, uncertainty might refer to the overall degree of imprecision or unpredictability or it might refer only to that not due to the inherent heterogeneity across individuals, space, or time*" (13). Even when there is clarity about the concept of true heterogeneity, it is usually impossible, in practice, to disentangle it from parameter, model, and decision uncertainty.

Data sparsity. Taking up heterogeneity in statistical models implies that model flexibility increases. More parameters need to be estimated and this estimation requires more data. In addition, more flexible statistical methods tend to be more variable. As models get more flexible, assessing bias-variance trade-offs is increasingly important, and the validation procedures to do this require more data (89). The issue of data sparsity in relation with model complexity is known in some fields as the curse of dimensionality. More flexible models may also pose computational challenges.

Changing methods. The field of modelling infectious diseases is evolving rapidly. Diverse methodological approaches have been proposed, many of which address population heterogeneity in one way or another. Keeping track of the strengths and weaknesses of new methods has become a challenge in itself.

Mismatch with what people want to hear. Considerations about heterogeneity and uncertainty may be unwelcome in situations where quick and clear answers are demanded by politicians and the general public, for example at the beginning of an outbreak (69). Mechanistic models can provide direct quantitative arguments for decision-making but are heavily dependent on model structure and assumptions. Sometimes, key parameters are inherently unmeasurable (e.g. density-dependent helminth establishment in humans) (37) or very difficult to capture (e.g. reactive behaviour changes during an outbreak). Phenomenological models contain less risky assumptions. They do not require detailed knowledge about the mechanics of disease transmission, which is a valuable characteristic in case of newly emerging infectious diseases such as COVID-19. However, although phenomenological models allow us to imagine sources of heterogeneity, they do not prove what the drivers of an outbreak are and they cannot directly show the expected impact of disease control interventions.

Questions about level of decision-making. Considerations about heterogeneity may lead to politically sensitive questions about the appropriate level of autonomous decision-making in public health. Some have argued that it is convenient to model an outbreak at the level at which decisions are taken (resolution of the model determined by administrative entities) (63). However, one can also argue that the way in which an infectious disease spreads and clusters should determine the administrative level at which disease control is organised. The "political dance" between the president of Colombia and the mayor of Bogotá regarding coronavirus control measures is one of many examples in the world where the coronavirus pandemic acutely sharpens the tension between local and national levels (90). Similar tensions abound -albeit in a more chronic way- in the field of global elimination initiatives against neglected tropical diseases.

5.5 Recommendations

Based on this project, I have formulated recommendations for studies about the frequency of infectious diseases in heterogeneous populations in general and for studies about COVID-19 in particular.

Concerning heterogeneity in general:

- Reflect about the possibility of heterogeneity and what it implies for research and policy. In this reflection, avoid confusion between heterogeneity and uncertainty. In practice, variance is composed of different elements some of which are reducible while others are not. Recognise the tension between the pursuit of reproducibility, consistency, and standardised solutions on the one hand and the identification of heterogeneity on the other.
- Learn from conceptual and technical developments in fields where heterogeneity is studied intensively. The views from academic disciplines with an experienced eye for human behaviour and local context are enriching in this respect. Also within the biomedical environment, there are fields (e.g. ecohealth and tropical infectious diseases) where considerations about heterogeneity come in naturally, for example when there are biological reasons (e.g. in the domains of helminth infections and vector-borne diseases) or when universal interventions such as mass drug administration or generalised lockdowns do not work as expected.
- Plan routine and research data collection and statistical analyses in a way that allows the possibility of heterogeneity. Which variables define relevant population subgroups and how can these variables be measured? Skipping this step means opting for the pragmatic attitude, i.e. working with the data at hand. Investing in this step implies that the data that are to be collected may differ across pathogens and contexts, which leads to less standardised and seemingly less generalisable findings.

Concerning COVID-19 in particular:

- Consider aspects of model uncertainty throughout the entire process of producing, reporting, interpreting, and using evidence. Model uncertainty is analysed less frequently than parameter uncertainty but may have a larger impact on the findings (13). Allowing the possibility of overdispersion and sub-exponential growth affects the findings of a simple COVID-19 growth model. This raises questions about the accuracy of models used for decision-making that assume equidispersion or exponential growth. The incorporation of dispersion and growth scaling parameters in other families of COVID-19 models is worth considering. The technicalities of the incorporation as well the interpretation of these parameters are areas of further research. Examples of research questions in this direction are: is there a better option than the gamma distribution to account for overdispersion in the case of COVID-19? How does the growth scaling parameter correlate with final epidemic size? Which of several extra parameters proposed in phenomenological growth models (dispersion, growth scaling, sub-epidemics) is most useful in the case of COVID-19 in different contexts?
- Plan data collection for a possible next wave of COVID-19 (and for outbreaks of other pathogens) based on lessons learned in the past. The growth pattern of the COVID-19 outbreak may differ across subpopulations within a country. Such differences could not be identified in this thesis project because the relevant information was not (publicly) available. The possibility of heterogeneous populations was also underexplored in some key COVID-19 publications. Collecting and reporting data on a few additional, carefully chosen variables would facilitate analyses that can generate potentially relevant insights. Examples of research questions along

this line are: Is the growth pattern of the COVID-19 outbreak closer to exponential in constrained or vulnerable populations (e.g. people living in homes for the elderly in Belgium) than in other population groups? Does the level of overdispersion vary across population subgroups (urban *versus* rural, commuters *versus* noncommuters, people with large *versus* small social networks)? Assuming that phenomenological models allowing for sub-epidemic waves will fit the COVID-19 incidence data well, can these sub-epidemic waves then be linked to population subgroups?

5.6 Conclusions

When heterogeneity is ignored, mathematical and statistical models can produce misleading results such as biased estimates, narrow confidence intervals that do not contain the true value, and inappropriate generalisations. Such mistakes are difficult to identify once papers are published, but can be discovered when specific assumptions are investigated. In the case of the COVID-19 outbreak, incidence data are often assumed to present equidispersion and exponential growth. In Belgium and in diverse other settings, these two assumptions are not met. Assuming an exponential growth pattern when, in fact, the growth is slower than exponential typically leads to an overestimation of the epidemic size and an overestimation of the impact of disease control interventions.

Alternative methodological approaches that account for heterogeneity do exist. Two examples of research fields where such approaches are commonly used are the epidemiology of helminth infections (longstanding tradition) and the characterisation of the Ebola virus outbreak in West Africa (more recent tradition). Some of these approaches require knowledge of the sources of heterogeneity so that the population can be disaggregated in relevant subgroups (e.g. mechanistic models). Other approaches are more generic: they allow the existence of heterogeneity without knowing its sources (e.g. phenomenological models).

Methodological approaches that account for heterogeneity are typically more complex (contain more parameters) than their simpler counterparts assuming homogeneous populations. However, taking up one or two extra parameters is still relatively straightforward, as illustrated in this thesis project. Approaches incorporating heterogeneity also produce results surrounded by more uncertainty than simpler methods. Although this finding of more uncertainty may not look attractive at first sight, it is better than the false impression of certainty produced by overly simple approaches. In the end, information will only be useful for public health if it reflects the true uncertainty and variability that is associated with the presence of infectious diseases in a population.

6 References

1. Zinsser H. Rats, Lice and History. Boston: Little, Brown and Company; 1935. 301 p.
2. Coggon D, Rose G, Barker D. Epidemiology for the uninitiated. London: BMJ Books; 2003. Chapter 1. What is epidemiology? Available from: https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/1-what-epidemiology (accessed 14 June 2020).
3. Cairncross S, Periès H, Cutts F. Vertical health programmes. Lancet. 1997; 349: S20-S21.
4. Hens N, Shkedy Z, Aerts M, Faes C, Van Damme P, Beutels P. Modeling infectious disease parameters based on serological and social contact data. A modern statistical perspective. New York: Springer; 2012. 298p.
5. Reen T. Glad tidings. Lulu.com; 2011. 192 p.
6. Cambridge Online Dictionary [Internet]. Heterogeneity. Available from: https://dictionary.cambridge.org/dictionary/english/heterogeneity (accessed 14 June 2020).
7. Wikipedia [Internet]. Homogeneity and heterogeneity. Available from: https://en.wikipedia.org/wiki/Homogeneity_and_heterogeneity (accessed 14 June 2020).
8. Schünemann H, Brożek J, Guyatt G, Oxman A, editors. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach (updated October 2013). GRADE Working Group; 2013. Chapter 5.2 Factors that can reduce the quality of the evidence. Available from: https://gdt.gradepro.org/app/handbook/handbook.html (accessed 14 June 2020).
9. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science. 1981; 211: 453-458.
10. van Asselt MBA. Perspectives on uncertainty and risk. The PRIMA approach to decision support. Springer Netherlands; 2000. 436 p.
11. Wynne B. Uncertainty and environmental learning. Reconceiving science and policy in the preventive paradigm. Global Environmental Change. 1992; 2: 111-127.
12. Funtowicz SO, Ravetz JR. Uncertainty and quality in science for policy. Springer Netherlands; 1990. 231p.
13. Krupnick A, Morgenstern R, Batz M, Nelson P, Burtraw D, Shih JS, et al. Not a sure thing: Making regulatory choices under uncertainty. Washington: Resources for the future; 2006. 233p.
14. Chowell G, Sattenspiel L, Bansal S, Viboud C. Mathematical models to characterize early epidemic growth: a review. Phys Life Rev. 2016; 18: 66-97.
15. Stanford encyclopedia of philosophy, fall 2012 edition [Internet]. Models in science. Available from: https://plato.stanford.edu/archives/fall2012/entries/models-science/ (accessed 14 June 2020)
16. Baker RE, Peña JM, Jayamohan J, Jérusalem A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol Lett. 2018; 14: 20170660.
17. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018; 169: 467-473.
18. Sciensano. Fact sheet COVID-19 disease (SARS-CoV-2 virus). Sciensano; 2020. Available from: https://covid-19.sciensano.be/sites/default/files/Covid19/COVID-19_fact_sheet_ENG.pdf (accessed 14 June 2020).
19. Belgian federal public service; health, food chain safety and environment. Six new cases of Covid-19 by the end of the spring holidays [Internet]. Available from: https://www.info-coronavirus.be/en/news/6-new-cases-of-covid-19-by-the-end-of-the-spring-holidays/ (accessed 14 June 2020).
20. Wikipedia. COVID-19 pandemic in Belgium [Internet]. Available from: https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Belgium (accessed 14 June 2020).
21. Advanced Solutions Nederland. Covid-19: Netherlands – a disastrous combination of events [Internet]. Available from: https://www.advsolned.com/covid-19-netherlands-a-disastrous-

combination-of-events/ (accessed 14 June 2020).
22. Oltermann P, Davidson H, Laughland O, Ratcliffe R, Walters J, Willsher K, et al. The cluster effect: how social gatherings were rocket fuel for coronavirus. The Gardian [Internet]. 2020 Apr 9. Available from: https://www.theguardian.com/world/2020/apr/09/the-cluster-effect-how-social-gatherings-were-rocket-fuel-for-coronavirus (accessed 14 June 2020).
23. Sciensano. COVID-19 weekly epidemiological update of 26 March 2020. Sciensano; 2020. Available from: https://covid-19.sciensano.be/nl/covid-19-epidemiologische-situatie (accessed 14 June 2020).
24. Official information and services [Internet]. Coronavirus: reinforced measures. Beligum.be; 2020. Available from: https://www.belgium.be/en/news/2020/coronavirus_reinforced_measures (accessed 14 June 2020).
25. Food and Agriculture Organization and World Bank population estimates [Internet]. Population density (people per sq. km of land area). Available from: https://data.worldbank.org/indicator/EN.POP.DNST (accessed 14 June 2020).
26. Aron J, Muellbauer J. Measuring excess mortality: England is the European outlier in the Covid-19 pandemic. VOX CEPR Policy Portal [Internet]. 2020 May 18. Available from: https://voxeu.org/article/excess-mortality-england-european-outlier-covid-19-pandemic (accessed 14 June 2020).
27. Moriarty LF, Plucinski MM, Marston BJ, Kurbatova E V., Knust B, Murray EL, et al. Public Health Responses to COVID-19 Outbreaks on Cruise Ships - Worldwide, February-March 2020. MMWR Morb Mortal Wkly Rep. 2020; 69: 347-352.
28. Mizumoto K, Chowell G. Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020. Infect Dis Model. 2020; 5: 264-270.
29. Dunford D, Dale B, Stylianou N, Lowther E, Ahmed M, de la Torre Arenas I.Coronavirus: The world in lockdown in maps and charts. BBC News [Internet]. 2020 Apr 7. Available from: https://www.bbc.com/news/world-52103747 (accessed 14 June 2020).
30. European Centre for Disease Prevention and Control. COVID-19 situation update worldwide, as of 4 June 2020 [Internet]. Available from: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases (accessed 5 June 2020).
31. Sciensano [Internet]. Sciensano & Coronavirus. What role does Sciensano play? Available from: https://www.sciensano.be/en/health-topics/coronavirus/role (accessed 14 June 2020).
32. Sciensano [Internet]. COVID19BE open data codebook. Available from: https://epistat.sciensano.be/COVID19BE_codebook.pdf (accessed 14 June 2020).
33. Sciensano [Internet]. COVID-19 datasets. Available from: https://epistat.wiv-isp.be/Covid/ (accessed 14 June 2020).
34. World Health Organization [Internet]. Coronavirus disease (COVID-2019) situation reports. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/ (accessed 14 June 2020).
35. European Centre for Disease Prevention and Control [Internet]. Download today's data on the geographic distribution of COVID-19 cases worldwide. Available from: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
36. Viboud C, Simonsen L, Chowell G. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics. 2016; 15: 27-37.
37. Hamley JID, Milton P, Walker M, Basáñez MG. Modelling exposure heterogeneity and density dependence in onchocerciasis using a novel individual-based transmission model, EPIONCHO-IBM: Implications for elimination and data needs. PLoS Negl Trop Dis. 2019; 13: e0007557.
38. Oswald WE, Halliday KE, McHaro C, Witek-McManus S, Kepha S, Gichuki PM, et al. Domains of transmission and association of community, school, and household sanitation with soil-transmitted helminth infections among children in coastal Kenya. PLoS Negl Trop Dis. 2019; 13: e0007488.
39. Chavy A, Nava AFD, Luz SLB, Ramírez JD, Herrera G, Dos Santos TV, et al. Ecological niche

modelling for predicting the risk of cutaneous leishmaniasis in the Neotropical moist forest biome. PLoS Negl Trop Dis. 2019; 13: e0007629.
40. de Oliveira Padilha MA, de Oliveira Melo J, Romano G, Malveira de Lima MV, Alonso WJ, Mureb Sallum MA, et al. Comparison of malaria incidence rates and socioeconomic-environmental factors between the states of Acre and Rondônia: A spatio-temporal modelling study. Malar J. 2019; 18: 306.
41. Brunner NC, Chacky F, Mandike R, Mohamed A, Runge M, Thawer SG, et al. The potential of pregnant women as a sentinel population for malaria surveillance. Malar J. 2019; 18:370.
42. Vegvari C, Truscott JE, Kura K, Anderson RM. Human population movement can impede the elimination of soil-transmitted helminth transmission in regions with heterogeneity in mass drug administration coverage and transmission potential between villages: a metapopulation analysis. Parasit Vectors. 2019; 12: 438.
43. Truscott JE, Ower AK, Werkman M, Halliday K, Oswald WE, Gichuki PM, et al. Heterogeneity in transmission parameters of hookworm infection within the baseline data from the TUMIKIA study in Kenya. Parasit Vectors. 2019; 12: 442.
44. Otiende V, Achia T, Mwambi H. Bayesian modeling of spatiotemporal patterns of TB-HIV co-infection risk in Kenya. BMC Infect Dis. 2019; 19: 902.
45. Martinez EZ, Zucoloto ML, Galdino G, Nunes AA, da Silva Lizzi EA. Spatiotemporal distribution of acquired immunodeficiency syndrome incidence in Brazil between 2012 and 2016. Rev Soc Bras Med Trop. 2019; 53: e20190086.
46. Chowell G, Tariq A, Hyman JM. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Med. 2019; 17: 164.
47. Li J, Zhang X, Wang L, Xu C, Xiao G, Wang R, et al. Spatial-temporal heterogeneity of hand, foot and mouth disease and impact of meteorological factors in arid/ semi-arid regions: a case study in Ningxia, China. BMC Public Health. 2019; 19: 1482.
48. Xiao C, Jike C, Liu D, Jia P, Xu X, Xiao L, et al. The changing modes of human immunodeficiency virus transmission and spatial variations among women in a minority prefecture in southwest China: an exploratory study. Medicine (Baltimore). 2020; 99: e18776.
49. Solomon T, Loha E, Deressa W, Gari T, Lindtjørn B. Spatiotemporal clustering of malaria in southern-central Ethiopia: a community-based cohort study. PLoS One. 2019; 14: e0222986.
50. Corder RM, Paula GA, Pincelli A, Ferreira MU. Statistical modeling of surveillance data to identify correlates of urban malaria risk: a population-based study in the Amazon Basin. PLoS One. 2019; 14(8): e0220980.
51. Cooper L, Kang SY, Bisanzio D, Maxwell K, Rodriguez-Barraquer I, Greenhouse B, et al. Pareto rules for malaria super-spreaders and super-spreading. Nat Commun. 2019; 10: 3939.
52. Ganyani T, Faes C, Hens N. Inference of the generalized-growth model via maximum likelihood estimation: a reflection on the impact of overdispersion. J Theor Biol. 2020; 484: 110029.
53. Ganyani T, Roosa K, Faes C, Hens N, Chowell G. Assessing the relationship between epidemic growth scaling and epidemic size: the 2014-16 Ebola epidemic in West Africa. Epidemiol Infect. 2018; 147: 1-6.
54. Ganyani T, Faes C, Chowell G, Hens N. Assessing inference of the basic reproduction number in an SIR model incorporating a growth-scaling parameter. Stat Med. 2018; 37: 4490-4506.
55. Chowell G, Viboud C, Hyman JM, Simonsen L. The Western Africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. PLoS Curr. 2015; 7: ecurrents.outbreaks.8b55f4bad99ac5c5db3663e916803261.
56. Roosa K, Luo R, Chowell G. Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study. Math Biosci Eng. 2019; 16: 4299-4313.
57. Anderson RM, May RM. Infectious Diseases of Humans: dynamics and control. Oxford: Oxford University Press; 1991. 757p.
58. Bacaër N. A short history of mathematical population dynamics. London: Springer; 2011. 160 p.
59. Oxford Reference [Internet]. Mass action principle. Available from: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100138688 (accessed 14 June 2020).

60. Colgate SA, Stanley EA, Hyman JM, Layne SP, Qualls C. Risk behavior-based model of the cubic growth of acquired immunodeficiency syndrome in the United States. Proc Natl Acad Sci U S A. 1989; 86: 4793-4797.
61. Tolle J. Can growth be faster than exponential, and just how slow is the logarithm? The Mathematical Gazette. 2003; 87: 522-525.
62. Reppell M, Boehnke M, Zöllner S. The impact of accelerating faster than exponential population growth on genetic variation. Genetics. 2014; 196: 819-828.
63. Santermans E, Robesyn E, Ganyani T, Sudre B, Faes C, Quinten C, et al. Spatiotemporal evolution of Ebola Virus Disease at sub-national level during the 2014 West Africa epidemic: model scrutiny and data meagreness. PLoS One. 2016; 11: e0147172.
64. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc A. 1927; 115: 700–721.
65. Chowell G, Nishiura H, Bettencourt LM. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. J R Soc Interface. 2007; 4: 155-166.
66. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, et al. Estimating the future number of cases in the Ebola epidemic--Liberia and Sierra Leone, 2014-2015. MMWR Suppl. 2014; 63: 1-14.
67. US Centers for Disease Control and Prevention [Internet]. 2014-2016 Ebola Outbreak in West Africa. Available from: https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html (accessed 14 June 2020).
68. King AA, Domenech de Cellès M, Magpantay FM, Rohani P. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. Proc Biol Sci. 2015; 282: 20150347.
69. Stobbe M. CDC's overblown estimate of Ebola outbreak draws criticism. The Seattle Times [Internet]. 2015 Aug 1. Available from: https://www.seattletimes.com/nation-world/cdcs-overblown-estimate-of-ebola-outbreak-draws-criticism/ (accessed 14 June 2020).
70. Petrilli C, Jones S, Yang J, Rajagopalan H, O'Donnell L, Chernyak Y, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. BMJ. 2020; 369: m1966.
71. Nature index [Internet]. The most influential coronavirus research articles. Available from: https://www.natureindex.com/news-blog/the-top-coronavirus-research-articles-by-metrics (accessed 14 June 2020).
72. Flaxman S, Mishra S, Gandy A, Unwin H, Mellan T, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe [published online ahead of print, 2020 Jun 8]. Nature. 2020;10.1038/s41586-020-2405-7.
73. Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand [Report]. Imperial College London; 2020. https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-03-16-COVID19-Report-9.pdf (accessed 14 June 2020).
74. Eubank S, Eckstrand I, Lewis B, Venkatramanan S, Marathe M, Barrett CL. Commentary on Ferguson, et al., "Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand". Bull Math Biol. 2020; 82: 52.
75. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science. 2020; 368: 860-868.
76. Hens N, Vranck P, Molenberghs G. The COVID-19 epidemic, its mortality, and the role of non-pharmaceutical interventions. Eur Heart J Acute Cardiovasc Care. 2020; 9: 204-208.
77. Sominsky L, Walker DW, Spencer SJ. One size does not fit all - patterns of vulnerability and resilience in the COVID-19 pandemic and why heterogeneity of disease matters [published online ahead of print, 2020 Mar 20]. Brain Behav Immun. 2020; S0889-1591(20)30366-4.
78. Britton T, Ball F, Trapman P. The disease-induced herd immunity level for Covid-19 is substantially lower than the classical herd immunity level [e-print]. arXiv. 2020. Available from: https://arxiv.org/abs/2005.03085 (accessed 14 June 2020).

79. Endo A, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott S, Kucharski AJ, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 1; peer review: 1 approved, 1 approved with reservations]. Wellcome Open Res. 2020; 5: 67.
80. Ndaïrou F, Area I, Nieto JJ, Torres DFM. Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan [published online ahead of print, 2020 Apr 27]. Chaos Solitons Fractals. 2020; 135: 109846.
81. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005; 438: 355-359.
82. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Sci Rep. 2018; 8: 5382.
83. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect Dis Model. 2020; 5: 256-263.
84. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13-23, 2020. J Clin Med. 2020; 9: 596.
85. Munayco C, Tariq A, Rothenberg R, Soto-Cabezas G, Reyes M, Valle A, et al. Early transmission dynamics of COVID-19 in a southern hemisphere setting: Lima-Peru: February 29th-March 30th, 2020 [Preprint]. medRxiv. 2020; 2020.04.30.20077594.
86. Manchein C, Brugnago EL, da Silva RM, Mendes CFO, Beims MW. Strong correlations between power-law growth of COVID-19 in four continents and the inefficiency of soft quarantine strategies. Chaos. 2020; 30: 041102.
87. Peeters Grietens K, Gryseels C, Verschraegen G. Misdirection in the margins of malaria elimination methods. Crit Public Health. 2019; 29: 390-400.
88. Hodgins S, Saad A. Will the higher-income country blueprint for COVID-19 work in low- and lower middle-income countries? [published online ahead of print, 2020 Jun 10]. Glob Health Sci Pract. 2020; GHSP-D-20-00217.
89. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statiscal Learning with Applications in R. New York: Springer; 2013. 426 p.
90. Alsema A. Colombia's two presidents combating coronavirus. Colombia Reports [Internet]. 2020 May 6. Available from: https://colombiareports.com/colombia-combating-coronavirus-with-two-presidents/ (accessed 14 June 2020).

7 Annexes

7.1 MEDLINE search strategy and syntax

MEDLINE was searched through PubMed on 14 March 2020. The search terms combined four concepts: disease frequency, infectious disease, statistical modelling, and heterogeneity. There were no restrictions regarding language, publication type, or study design.

Concept	Search terms
Disease frequency	prevalence [MeSH Terms] OR Incidence [MeSH Terms] OR epidemiology [MeSH Terms] OR Prevalence OR Incidence
AND	
Infectious disease	infections [MeSH Terms] OR disease outbreak, infectious[MeSH Terms]
AND	
Statistical modelling	models, statistical[MeSH Terms] OR modeling OR modelling OR Bayes*
AND	
Heterogeneity	heterog*

The search ((((prevalence [MeSH Terms] OR Incidence [MeSH Terms] OR epidemiology [MeSH Terms] OR Prevalence OR Incidence)) AND (infections [MeSH Terms] OR disease outbreak, infectious[MeSH Terms])) AND heterog*) AND (models, statistical[MeSH Terms] OR modeling OR modelling OR Bayes*) retrieved 1206 hits. These were sorted anti-chronologically; and he first 15 eligible records were included in the review.

7.2 Core R code

Key elements of the code are given below; the complete and detailed code is available via figshare: https://figshare.com/articles/R_code_for_thesis_project_entitled_Estimating_the_frequency_of_infectious_diseases_in_heterogeneous_populations_methodological_overview_and_application_to_the_COVID-19_outbreak_in_Belgium_/12479303.

Figure 4.1. Illustration of the relation between variance and mean in Poisson and negative binomial exponential growth models

```
Time<-seq(0,40,by=1)
C0=1
r=0.2
CountsP <- matrix(nrow = length(Time), ncol = 100)
for(i in 1:length(Time))
{CountsP[i,]<-rPO(100,mu=C0*exp(r*Time[i]))}
CountsP<-as.vector(CountsP)
CountsPTime<-cbind(Time, CountsP)
CountsPTime<-as.data.frame(CountsPTime)
plot(CountsPTime$Time,CountsPTime$CountsP, main="A. Poisson", xlab="time",
ylab="simulated counts", ylim=c(0,5000), pch=20)
lines(Time,C0*exp(r*Time),col="red")
```

```
theta=50
sigma=1/theta
CountsNB10 <- matrix(nrow = length(Time), ncol = 100)
for(i in 1:length(Time))
{CountsNB10[i,]<-rNBI(100,mu=C0*exp(r*Time[i]),sigma)}
CountsNB10<-as.vector(CountsNB10)
CountsNB10Time<-cbind(Time, CountsNB10)
CountsNB10Time<-as.data.frame(CountsNB10Time)
plot(CountsNB10Time$Time,CountsNB10Time$CountsNB10, main="B. Negative binomial\n
theta=50", xlab="time", ylab="simulated counts", ylim=c(0,5000), pch=20)

theta=5
sigma=1/theta
CountsNB10 <- matrix(nrow = length(Time), ncol = 100)
for(i in 1:length(Time))
{CountsNB10[i,]<-rNBI(100,mu=C0*exp(r*Time[i]),sigma)}
CountsNB10<-as.vector(CountsNB10)
CountsNB10Time<-cbind(Time, CountsNB10)
CountsNB10Time<-as.data.frame(CountsNB10Time)
plot(CountsNB10Time$Time,CountsNB10Time$CountsNB10, main="B. Negative binomial\n
theta=5", xlab="time", ylab="simulated counts", ylim=c(0,5000), pch=20)
```

Table 4.5. Comparison of Poisson, negative binomial, and Poisson inverse Gaussian distributions to model hospital admissions in Belgium: parameter estimates and Akaike's information

Figure 4.6. Incidence of hospital admissions in Belgium and comparison of Poisson and negative binomial model approximations (equidispersion versus overdispersion)

```
#Data Belgium country; from 10 March to 30 April 2020 - updated 16 May 2020
Cases<-c(99, 172, 252, 338, 179, 213, 386, 417, 534, 713, 663, 470, 487,
1329, 1196, 1198, 1363, 1519, 844, 683, 1739, 1682, 1503, 1460, 1691, 925,
665, 1922, 1501, 1585, 2221, 2321, 1023, 517, 532, 1561, 1619, 1658, 1379,
701, 448, 1266, 1257, 768, 965, 793, 387, 203, 744, 565, 512, 579)
Hosp<-c(0, 7, 26, 70, 56, 71, 90, 123, 183, 212, 295, 332, 290, 278, 434, 540, 490,
575, 629, 545, 478, 553, 584, 568, 504, 499, 358, 314, 490, 459, 462, 421, 393,
270, 242, 250, 310, 320, 303, 265, 232, 172, 263, 211, 210, 217, 202, 127, 123,
174, 178, 152)
Deaths<-c(1, 3, 1, 3, 4, 6, 11, 11, 21, 27, 31, 39, 41, 78, 74, 98, 101, 123, 133,
138, 168, 185, 247, 210, 232, 274, 233, 275, 304, 319, 273, 320, 290, 344, 283,
277, 275, 254, 203, 211, 214, 199, 202, 191, 173, 147, 155, 176, 125, 114, 98, 84)
Days<-seq(0,length(Cases)-1,by=1)
Dates<-seq(as.Date("2020-03-10"),as.Date("2020-04-30"), by="days")
Belgium=data.frame(Cases, Deaths, Hosp, Dates, Days)

#Subselect time period used to build model
x1<-1
x2<-20
BelIni=Belgium[x1:x2,]

PoiH<-function(par1,par2){-sum(
dPO(BelIni$Hosp,par1*exp(par2*BelIni$Days),log=TRUE))}
fit.PoiH<-mle(PoiH,start= list(par1=100, par2=0.10),method = "Nelder-Mead")

NBH<-function(par1,par2,par3){-sum(
dNBI(BelIni$Hosp,m=par1*exp(par2*BelIni$Days),sigma=exp(par3),log=TRUE)}
fit.NBH<-mle(NBH,start= list(par1=100, par2=0.10,par3=-2.06),method = "Nelder-
Mead")

PIGH<-function(par1,par2,par3){-sum(
dPIG(BelIni$Hosp,m=par1*exp(par2*BelIni$Days),sigma=exp(par3),log=TRUE))}
```

```
fit.PIGH<-mle(PIGH,start= list(par1=100, par2=0.10,par3=-2.06),method = "Nelder-
Mead")

window<-Belgium$Days
window[window<20]<-16
window[window>19]<-1
plot(Belgium$Days,Belgium$Hosp, xlab="Days since first local cases (day 0 =
      10 March 2020)", ylab="Number of admissions", ylim=c(0,1000), xlim=c(0,40),
      main="A. Admissions in Belgium\n original scale", pch=window)
lines(Belgium$Days, coef(fit.PoiH)[1]*exp(coef(fit.PoiH)[2]*Belgium$Days),
      lty=3,col="darkred")
lines(Belgium$Days, coef(fit.NBH)[1]*exp(coef(fit.NBH)[2]*Belgium$Days),
      lty=5,col="darkred")
abline(v=6,lty=1,col="black")
legend("topleft", c("Poisson", "Neg Bin"), col=c("darkred", "darkred"),
      lty=c(3,5),bg="white")

plot(Belgium$Days,log10(Belgium$Hosp), xlab="Days since first local cases
      (day 0 = 10 March 2020)", ylab="Log10(number of admissions)",
      ylim = c(0,5), xlim=c(0,40), main="B. Admissions in Belgium\n semi-
      logarithmic scale", pch=window)
lines(Belgium$Days,
      log10(coef(fit.PoiH)[1]*exp(coef(fit.PoiH)[2]*Belgium$Days)),
      lty=3,col="darkred")
lines(Belgium$Days,
      log10(coef(fit.NBH)[1]*exp(coef(fit.NBH)[2]*Belgium$Days)),
      lty=5,col="darkred")
abline(v=6,lty=1,col="black")
legend("topleft", c("Poisson", "Neg Bin"), col=c("darkred", "darkred"),
      lty=c(3,5),bg="white")

fit.PoiH
LLcH<-coef(fit.PoiH)[1]-qnorm(0.975)*sqrt(diag(fit.PoiH@vcov)[1])
ULcH<-coef(fit.PoiH)[1]+qnorm(0.975)*sqrt(diag(fit.PoiH@vcov)[1])
CIcH<-round(c(LLcH,ULcH),0)
LLrH<-coef(fit.PoiH)[2]-qnorm(0.975)*sqrt(diag(fit.PoiH@vcov)[2])
ULrH<-coef(fit.PoiH)[2]+qnorm(0.975)*sqrt(diag(fit.PoiH@vcov)[2])
CIrH<-round(c(LLrH,ULrH),3)
round(AIC(fit.PoiH),0)

fit.NBH
round(1/exp(coef(fit.NBH)[3]),1)
LLcHNB<-coef(fit.NBH)[1]-qnorm(0.975)*sqrt(diag(fit.NBH@vcov)[1])
ULcHNB<-coef(fit.NBH)[1]+qnorm(0.975)*sqrt(diag(fit.NBH@vcov)[1])
CIcHNB<-round(c(LLcHNB,ULcHNB),0)
LLrHNB<-coef(fit.NBH)[2]-qnorm(0.975)*sqrt(diag(fit.NBH@vcov)[2])
ULrHNB<-coef(fit.NBH)[2]+qnorm(0.975)*sqrt(diag(fit.NBH@vcov)[2])
CIrHNB<-round(c(LLrHNB,ULrHNB),3)
LLthHNB<-coef(fit.NBH)[3]-qnorm(0.975)*sqrt(diag(fit.NBH@vcov)[3])
ULthHNB<-coef(fit.NBH)[3]+qnorm(0.975)*sqrt(diag(fit.NBH@vcov)[3])
CIthHNB<-round(1/exp(c(LLthHNB,ULthHNB)),1)
round(AIC(fit.NBH),0)

fit.PIGH
1/exp(coef(fit.PIGH)[3])
round(AIC(fit.PIGH),0)
LLcHPIG<-coef(fit.PIGH)[1]-qnorm(0.975)*sqrt(diag(fit.PIGH@vcov)[1])
ULcHPIG<-coef(fit.PIGH)[1]+qnorm(0.975)*sqrt(diag(fit.PIGH@vcov)[1])
CIcHPIG<-round(c(LLcHPIG,ULcHPIG),0)
LLrHPIG<-coef(fit.PIGH)[2]-qnorm(0.975)*sqrt(diag(fit.PIGH@vcov)[2])
ULrHPIG<-coef(fit.PIGH)[2]+qnorm(0.975)*sqrt(diag(fit.PIGH@vcov)[2])
CIrHPIG<-round(c(LLrHPIG,ULrHPIG),3)
LLthHPIG<-coef(fit.PIGH)[3]-qnorm(0.975)*sqrt(diag(fit.PIGH@vcov)[3])
```

```
ULthHPIG<-coef(fit.PIGH)[3]+qnorm(0.975)*sqrt(diag(fit.PIGH@vcov)[3])
CIthHPIG<-round(1/exp(c(LLthHPIG,ULthHPIG)),1)
```

Figure 4.7. Incidence of hospital admissions in Belgium and comparison of negative binomial models with and without a growth scaling parameter (exponential versus sub-exponential growth)

```
SENBHb<-function(A,r,p,theta){-sum(
dnbinom(BelIni$Hosp,mu=exp(r)*(exp(A)+exp(r)*(1-exp(p))*BelIni$Days)^(exp(p)/(1-
exp(p))),size=theta,log=TRUE))}
fit.SENBHb<-mle(SENBHb,start= list(A=1, r=-0.10,p=-0.8,theta=1),method = "BFGS")
LLp<-coef(fit.SENBHb)[3]-qnorm(0.975)*sqrt(diag(fit.SENBHb@vcov)[3])
ULp<-coef(fit.SENBHb)[3]+qnorm(0.975)*sqrt(diag(fit.SENBHb@vcov)[3])
CInewp<-round(c(exp(LLp),exp(ULp)),2)
newp<-exp(coef(fit.SENBHb)[3])

plot(Belgium$Days,log10(Belgium$Hosp), xlab="Days since first local cases
       (day 0 = 10 March 2020)", ylab="Log10(number of admissions)",
       ylim=c(0,5), xlim=c(0,40), main="Admissions in Belgium\n semi-
       logarithmic scale", pch=window)
lines(Belgium$Days,
       log10(coef(fit.NBH)[1]*exp(coef(fit.NBH)[2]*Belgium$Days)),
       lty=5,col="darkred")
lines(Belgium$Days,
       log10(exp(coef(fit.SENBHb)[2])*(exp(coef(fit.SENBHb)[1])+
       exp(coef(fit.SENBHb)[2])*(1-exp(coef(fit.SENBHb)[3]))*Belgium$Days)^
       (exp(coef(fit.SENBHb)[3])/(1-exp(coef(fit.SENBHb)[3])))), lty=5,col="navy")
abline(v=6,lty=1,col="black")
legend("topleft", c("Neg Bin, exponential", "Neg Bin, sub-exponential"),
     col=c("darkred", "navy"), lty=c(5,5), bg="white")
```

Figure 4.9. Incidence and cumulative incidence of hospital admissions in Belgium and fitted curves using a negative binomial model with a growth scaling parameter

```
SENBHb<-function(A,r,p,theta){-sum(
dnbinom(BelIni$Hosp,mu=exp(r)*(exp(A)+exp(r)*(1-exp(p))*BelIni$Days)^(exp(p)/(1-
exp(p))),size=theta,log=TRUE))}
fit.SENBHb<-mle(SENBHb,start= list(A=1, r=-0.10,p=-0.8,theta=1),method = "BFGS")
LLp<-coef(fit.SENBHb)[3]-qnorm(0.975)*sqrt(diag(fit.SENBHb@vcov)[3])
ULp<-coef(fit.SENBHb)[3]+qnorm(0.975)*sqrt(diag(fit.SENBHb@vcov)[3])
CInewp<-c(exp(LLp),exp(ULp))
newp<-exp(coef(fit.SENBHb)[3])
summary(fit.SENBHb)

SENBCumHb<-function(A,r,p,theta){-
sum(dnbinom(cumsum(BelIni$Hosp),mu=(exp(A)+exp(r)*(1-exp(p))*BelIni$Days)^(1/(1-
exp(p))),size=theta,log=TRUE))}
fit.SENBCumHb<-mle(SENBCumHb,start= list(A=1, r=-0.10,p=-0.8,theta=1),method =
"BFGS")
LLpCumH<-coef(fit.SENBCumHb)[3]-qnorm(0.975)*sqrt(diag(fit.SENBCumHb@vcov)[3])
ULpCumH<-coef(fit.SENBCumHb)[3]+ qnorm(0.975)*sqrt(diag(fit.SENBCumHb@vcov)[3])
CInewpCumH<-c(exp(LLpCumH),exp(ULpCumH))
newpCumH<-exp(coef(fit.SENBCumHb)[3])
summary(fit.SENBCumHb)
```

7.3 Supplementary tables and figures

Comparison of a Poisson model with and without growth scaling parameter p: parameter estimates and Akaike's information criterion

	Poisson model without growth scaling parameter (assuming exponential growth)	**Poisson model with growth scaling parameter (allowing for sub-exponential growth)**
C_0 [95% CI]	44 [41-48]	0*
r [95% CI]	0.150 [0.145-0.156]	4.038 [3.686-4.425]
p [95% CI]	assumed 1	0.60 [0.58-0.61]
AIC	549	250

AIC: Akaike's information criterion; CI: confidence interval; C_0: initial number of cases; p: growth scaling parameter; r: growth rate. The time window for model fitting is 20 days (10 to 29 March). *The 95% confidence interval is not given as computation is not straightforward anymore (C_0 indirectly parameterised). The interpretation of the parameters C_0 and r is not the same in the two approaches: the estimates in the right column come from an expression with nonlinearity in the parameters (p is in the exponent of the expression), hence, the parameters are correlated.

Incidence of hospital admissions in Belgium and comparison of a Poisson model with and without a growth scaling parameter (exponential versus sub-exponential growth).

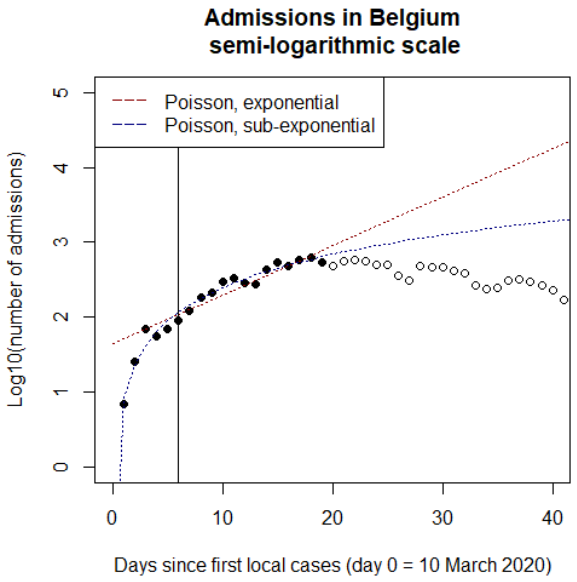

Figure legend. Two Poisson models are compared. The red dotted line is based on a model assuming exponential growth, which is equivalent to fixing the growth scaling parameter p to 1. On a semi-logarithmic scale, exponential growth results in a straight line. The blue dotted line is based on a model allowing for sub-exponential growth, where the growth scaling parameter p is estimated from the data. The point estimate for p is 0.60 (95% confidence interval: 0.58-0.61). The sub-exponential model provides the best fit to the data. On a semi-logarithmic scale, sub-exponential growth results in a line with downward curvature. The black solid vertical line indicates the lockdown date (16 March). The dots are the observed admission counts; those filled in black indicate the time window used for model fitting (10 to 29 March).

Estimates of the growth scaling parameter p using a Poisson approach and comparing incidence and cumulative incidence of admissions

Incidence measure	Estimates of growth scaling parameter p [95% CI] Poisson model
Incidence of admissions	0.60 [0.58-0.61]
Cumulative incidence of admissions	0.61 [0.60-0.63]

Estimates of the growth scaling parameter p using a Poisson approach and comparing the incidence of registered cases, hospital admissions, and deaths

Incidence measure	Estimates of growth scaling parameter p [95% CI] Poisson model
Cases	0.55 [0.53-0.58]
Admissions	0.60 [0.58-0.61]
Deaths	0.77 [0.75-0.81]