PREDICTING THE OCCURRENCE AND SPREAD OF HARMFUL MICROORGANISMS WITH STATISTICAL MODELS

Johan Magnus van Niekerk

PREDICTING THE OCCURRENCE AND SPREAD OF HARMFUL MICROORGANISMS WITH STATISTICAL MODELS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. ir. A. Veldkamp,
on account of the decision of the Doctorate Board,
to be publicly defended
on 1 December 2021 at 16:45

by

Johan Magnus van Niekerk

born on the 22nd of August 1985
in Pretoria, South Africa

This dissertation has been approved by:
prof. dr. ir. A. Stein, supervisor
prof. dr. J.E.W.C. van Gemert-Pijnen, supervisor
dr. L.M.A. Braakman-Jansen, co-supervisor

Cover design: Job Duim
Printed by: University of Twente (ITC)
Lay-out: Johan Magnus van Niekerk
ISBN: 978-90-365-5294-3
DOI: https://doi.org/10.3990/1.9789036552943

Graduation Committee:

Chairman/Secretary

prof. dr. F.D. van der Meer

Supervisors

prof. dr. ir. A. Stein

prof. dr. J.E.W.C. van Gemert-Pijnen

Co-supervisor

dr. L.M.A. Braakman-Jansen

Committee Members

dr. N. Beerlage - de Jong

prof. dr. J.I. Blanford

prof. dr. A. Friedrich

prof. dr. A. Bekker

dr. N. al Naiemi

Contents

Contents ..... **i**

List of Figures ..... **iv**

List of Tables ..... **vi**

List of abbreviations ..... **vii**

List of symbols ..... **ix**

1. Introduction **1**

1.1 Motivation 1

1.2 Background research project (EurHealth-1Health) 3

1.3 Thesis outline 3

1.4 Identifying knowledge gaps in AMR research 4

1.5 Data-driven risk factors for surgical site infection 5

1.6 Transmission of harmful microorganisms through connected HCWs 6

1.7 Predicting the spread of AMR in a hospital 7

2. Antimicrobial resistance: Identifying knowledge gaps using semi-automated topic modelling **9**

Abstract 9

2.1 Background 10

2.2 Methods 11

2.2.1 Search string and data extraction 13
2.2.2 Corpus 13
2.2.3 Metadata 13
2.2.4 Structural topic modelling 13
2.2.5 Implementation 15
2.2.6 Thematic clusters 16
2.2.7 Thematic cluster focus 17
2.2.8 Knowledge gap identification 17
2.2.9 Data availability and interactive user interface 17
2.2.10 Software 18

2.3 Results 18

2.3.1 Topic modelling 18
2.3.2 Thematic clusters 19
2.3.3 Identifying knowledge gaps 20
2.4 Discussion 26
2.4.1 Thematic clusters 27
2.4.2 Knowledge gaps 27
2.4.3 Limitations 29
2.5 Conclusion 30

3. Risk factors for surgical site infections using a data-driven approach 31

Abstract 31
3.1 Background 32
3.2 Methods 33
3.2.1 Literature search 33
3.2.2 Setting and data collection 33
3.2.3 Statistical analysis 34
3.3 Results 35
3.3.1 Literature search 35
3.3.2 Risk factor identification 35
3.4 Discussion 43
3.4.1 Limitations 45
3.4.2 Future work 46
3.5 Conclusion 47

4. A spatiotemporal simulation study on the transmission of harmful microorganisms through connected healthcare workers in a hospital ward setting 48

Abstract 48
4.1 Background 49
4.2 Methods 51
4.3 Results 57
4.3.1 Simulation results 60
4.4 Discussion 65
4.4.1 Limitations and future work 67
4.5 Conclusion 67

5. Spatiotemporal prediction of the occurrence of vancomycin-resistant Enterococcus 69

Abstract 69
5.1 Background 70
5.2 Methods 72
5.2.1 Patient movement and antibiotic data 72
5.2.2 Spatiotemporal graph 72
5.2.3 PageRank algorithm 72
5.2.4 VRE screening data 72
5.2.5 Modelling 73
5.2.6 Decision trees 73
5.2.7 Random forest 74
5.2.8 Software 74
5.3 Results 74
5.3.1 Covariates 74
5.3.2 Decision tree 77
5.3.3 Random forest 79
5.3.4 Model performance 81
5.4 Discussion 81
5.4.1 Future work 82
5.4.2 Limitation 82
5.5 Conclusion 83
6. Synthesis 84
6.1 Findings 84
6.2 Significance and prospects 86
6.3 Limitations 89
6.4 Reflection 91
Bibliography 92
Summary 112
Samenvatting 115
Acknowledgements 119
Overview of publications 121

List of Figures

Figure 2-1: Workflow diagram illustrating the process followed to identify potential knowledge gaps in AMR research. 12
Figure 2-2: A graphical illustration of the STM generative process with the observed variables shaded. 15
Figure 2-3: Topic names weighted by the topic proportion correlations. 19
Figure 2-4: Thematic clusters and their foci. 20
Figure 2-5: A comparison between thematic groups and thematic clusters connected by topics. Thematic group = names assigned to similar topics [15]. Only topics with summed topic proportion of 2 000 are displayed for conciseness. 26
Figure 3-1**:** The occurrence of SSI for the time between the J01 antibiotics prescription start time and the start time of the surgery by decile. 46
Figure 4-1: Floorplan of the 32-bed general hospital ward, for stomach, gut and liver patients. **A** = Floorplan of the 32-bed general hospital ward, for stomach, gut and liver patients where sample data were collected using **B** = RFID tags worn by HCWs during data collection using **C** = RFID readers placed on the ceiling of the rooms inside the ward..... 52
Figure 4-2: The four-part of the simulation workflow. A and C depend upon the sampled data and B and D on initial assumptions from literature. The simulation ends when the HCW successfully performs hand hygiene in part B. 54
Figure 4-3: Transition probability matrix $\boldsymbol{P}$ for movement of *nurse* between ward rooms. The transmission probabilities are given as pij in the ith row and jth column for the movement of *nurse* between rooms. Each element is the estimated probability that a nurse will transition from the room i to room j after the next transition. 59
Figure 4-4: The expected number of transmissions expressed as a percentage of the worst-case scenario. For A, B and C: the highest number of expected transmissions (worst-case scenario) occur for the scenario where the transmission probability is 0.05 and hand hygiene compliance is 0.05. The expected number of transmissions is expressed as a percentage of the worst-case scenario. 64
Figure 5-1: VRE tests and the number of positive VRE test results during 2018 - 2019. 73
Figure 5-2: Number of patient and patients using antibiotics. pat_num_ant = the number of patients using antibiotics in each ward; pat_num = the number of patients in each ward. 75
Figure 5-3: Number of patient and patients using antibiotics in example general care ward. pat_num_ant = the number of patients using antibiotics in each ward; pat_num = the number of patients in each ward. 75

Figure 5-4: Average daily PageRank covariate and the number of VRE positive patients. PR_pat_num = PageRank of patient movements between wards; PR_pat_num_ant = PageRank of patient movements using antibiotics. 76
Figure 5-5: Average daily PageRank covariate and the number of VRE positive patients in example ward general care ward. PR_pat_num = PageRank of patient movements between wards; PR_pat_num_ant = PageRank of patient movements using antibiotics. 77
Figure 5-6: Decision tree for the daily VRE occurrence in a hospital ward using PageRank and traditional covariates. pat_num_ant = the number of patients using antibiotics in each ward; PR_pat_num_ant = PageRank of patient movements currently using antibiotics; PR_pat_num = PageRank of patient movements between wards. In each node, the percentage of ward with at least one VRE positive patient is shown above the sample distribution of the node. 78
Figure 5-7: Minimal depth for each covariate in the 500 random forest decision trees. pat_num_ant = the number of patients using antibiotics in each ward; PR_pat_num_ant = PageRank of patient movements currently using antibiotics; pat_num = the number of patients in each ward; PR_pat_num = PageRank of patient movements between wards. 79
Figure 5-8: The change in mean squared error when covariate values are replaced with random values. PR_pat_num = PageRank of patient movements between wards; PR_pat_num_ant = PageRank of patient movements currently using antibiotics; pat_num = the number of patients in each ward; pat_num_ant = the number of patients using antibiotics in each ward. 80
Figure 5-9: The change in residual sum of squares when covariate values are replaced with random values. PR_pat_num_ant = PageRank of patient movements currently using antibiotics; PR_pat_num = PageRank of patient movements between wards; pat_num_ant = the number of patients using antibiotics in each ward; pat_num = the number of patients in each ward. 80
Figure 5-10: Lorenz curves of the decision tree and random forest models. 81

List of Tables

Table 2-1: The ten most unrelated topics to each thematic cluster. 20
Table 2-2: Potential knowledge gaps in TC3. 23
Table 2-3: Potential knowledge gaps in TC5. 24
Table 3-1: Variable names and definitions used to investigate the occurrence of SSI in this study. 35
Table 3-2: Digestive system surgical procedures: univariate analysis of risk factors for the future occurrence of SSI. 37
Table 3-3: Multivariate analysis risk factors for the occurrence of SSI by group of surgeries using standard medical cut-offs. 41
Table 3-4: Multivariate analysis risk factors for the occurrence of SSI by group of surgeries using data-driven cut-offs. 41
Table 3-5: Statistical significance of risk factors and the source which lead them to be considered by surgical procedure. 42
Table 4-1: Example of data collected using the RFID sensors and readers. 53
Table 4-2: Agent-based model parameters (Thomas Hornbeck et al.). 55
Table 4-3: The number of contacts and duration of those contacts by occupation group. 58
Table 4-4: Number of HCWs or patients and the time they co-occurred in each room. 60
Table 4-5: Simulated HMO transmissions potential of a colonised nurse in a hospital ward under various assumed transmission rates and hand hygiene compliance levels. 61
Table 4-6: Simulated HMO transmissions potential of a colonised nurse in a hospital ward under various assumed transmission rates and hand hygiene compliance levels (between 7am-5pm on weekdays). 62
Table 4-7: Simulated HMO transmissions potential of a colonised nurse in a hospital ward under various assumed transmission rates and hand hygiene compliance levels (between 6pm and 6am or on weekends). 63

List of abbreviations

AIC	Akaike information criterion
AMR	Antimicrobial resistance
ASA	American Society of Anaesthesiologists
ASCII	American Standard Code for Information Interchange
AST	Antimicrobial susceptibility testing
ATC	Anatomical Therapeutic Chemical
BMI	Body mass index
CDC	Centres for Disease Control and Prevention
CI	Confidence interval
CoNS	Coagulase-negative staphylococci
CRP	C-reactive protein
CTM	Correlated Topic Model
DG	Dynamic directed spatiotemporal graph
DGNN	Dynamic graph neural networks
ECDC	European Centre for Disease Prevention and Control
ECDC	European centre for disease prevention and control
EHR	Electronic health records
ELBO	Evidence lower bound
ESBL	Extended spectrum beta-lactamase
ESKAPE	Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa or Enterobacter spp.
FAO	Food and Agriculture Organization
HAI	Healthcare associated infection
HCW	Healthcare worker
HHC	Hand hygiene
HMO	Harmful microorganism
IC	Intensive care
ICP	Infection control practitioners
IPM	Intrahospital patient movement
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MCMC	Markov chain Monte Carlo
MeSH	Medical subject heading
MIC	Minimum inhibitory concentration

MRSA	Methicillin-resistant Staphylococcus aureus
MSE	Mean square error
NA	Not applicable
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NII	Nosocomial infection index
OIE	World Organization for Animal Health
OR	Odds ratio
PKPD	Pharmacokinetic/Pharmacodynamic
PLSA	Probabilistic Latent Semantic Analysis
PR	PageRank
RF	Random forest
RFID	Radio Frequency Identification
RIVM	Netherlands National Institute for Public Health and the Environment
RO	Risk outcome
ROC	Receiver operating characteristic curve
RSS	Residual sum of squares
SARS	Severe Acute Respiratory Syndrome
SD	Standard deviation
SGNN	Streaming Graph Neural Networks
SSE	Super-spreading events
SSI	Surgical site infection
STM	Structural topic modelling
TC	Thematic cluster
UMCG	University Medical Center Groningen
VRE	Vancomycin resistant enterococci
WBP	Dutch Personal Data Protection Act
WHO	World Health Organization
WMO	Medical Research Involving Human Subjects Act

List of symbols

Y	Outcome variable
d	Document
$Docs$	Number of documents
$Topics$	Number of topics
Voc	Number of words in the vocabulary
p	The number of prevalence covariates
P_Vars	$Docs \times p$ matrix containing p document-level topic prevalence covariates
$\boldsymbol{x}_d$	Vector of containing topic prevalence covariates for document d
l	The number of topical content covariates
C_Vars	$Docs \times l$ matrix containing l document-level topical content variables
$\boldsymbol{y}_d$	Vector of containing topical content variables for document d
N_Words_d	Number of words in document d
n	Word index for each document
$topic_{d,n}$	Topic of each word with index n in document d
θ_d	Vector of topic proportions for document d
γ	$p \times 1$ coefficient vector
Σ	$(K-1) \times (K-1)$ covariance matrix
m	Baseline word distribution
$\kappa_k^{(t)}$	Topic specific covariates group deviations
	Topic specific covariates group deviations for document d
$\kappa_{y_d,k}^{(i)}$	Interaction term between $\kappa_k^{(t)}$ and $\kappa_{y_d}^{(c)}$
$w_{d,n}$	Word in document d with index n
$\beta_{d,k}$	Probability that each word $w_{d,n}$ is equal to $k = topic_{d,n}$
K	Number of topics
K^*	Optimal number of topics
C	Correlation matrix
$Dist$	Distance matrix
T	Number of thematic clusters
T^*	Optimal number of thematic clusters
m	Minutes
h	Hours
R_i	Room i

p_{ij}	Transition probability from R_i to R_j
$\boldsymbol{P}$	Transition probability matrix
ψ_{R_i}	Time spent in room R_i
η	The average number of HCWs or patients co-occurring
λ	Hand hygiene efficacy using alcohol rub
γ	Hand hygiene compliance level
P	Probability of transmission per 30 s of contact
ω_{R_i}	HCWs or patients co-occurring in room R_i with the infected HCW
pat_num	The number of patients in each ward
pat_num_ant	The number of patients using antibiotics in each ward
PR_pat_num	PageRank of patient movements between wards
PR_pat_num_ant	PageRank of patient movements currently using antibiotics
°C	Degrees Celsius

1. Introduction

1.1 Motivation

Antimicrobial resistance (AMR) is the most significant threat to modern healthcare. It is estimated that 10 million people will die due to AMR by 2050, more than the yearly death toll of cancer and road traffic accidents combined [1]. The discovery of antibiotics in the early 20th century significantly changed the course of modern healthcare. It enabled the treatment of previously deemed fatal infections and made numerous lifesaving surgical procedures possible. The use of antimicrobials leads to the natural evolution of microbes to become resistant to the newly discovered medicine. Due to the generous use of antimicrobials, selection pressure enabled the spread of AMR worldwide [2]. New antimicrobials were sought and manufactured, but AMR soon followed. Alternatives to antimicrobials have been explored but with limited success [3]. To date, there are no major antimicrobial classes for which no AMR has been found [4]. The world is facing an ever-growing number of AMR cases, and without significant scientific leaps treatment options, the main strategies are to detect and limit the occurrence and spread of AMR [3].

The World Health Organisation (WHO) has proposed a holistic action plan in response to the increasing AMR threat [5]. It was created using the consolidated objectives of existing action plans and best practices related to AMR and stretches across international sectors, industries and disciplines. A global strategic AMR research agenda was proposed to understand AMR at the global level. Knowledge gaps can be used to inform AMR research agendas [6–11]. To identify knowledge gaps in a research field, a thorough overview and understanding of the available knowledge in that research field are needed [12,13]. Knowledge gaps constantly need to be assessed for comprehensiveness and relevance. However, with the exponential increase in AMR research output, it is increasingly challenging to objectively organise and synthesise the current state of AMR research to stay informed about previous and most recent scientific contributions [14]. Scalable statistical models for text analysis can be used to determine underlying topics in large quantities of literature automatically and objectively [15]. In addition to understanding the AMR research field, statistical models are essential to address these knowledge gaps.

Predicting the occurrence and spread of microbes is important to support decision-making from the perspective of microbiology and epidemiology [16]. These models gradually evolved from deterministic models into stochastic models at the end of the previous century. Aggregate models neglect to use valuable individual-based data but can be less complicated and require less data and computational resources to build and maintain [16]. More recently, models started to use individual electronic

health records (EHR), pharmaceutical data and laboratory data rather than aggregated patient data [17]. The occurrence of surgical site infection (SSI) can be modelled using risk factors identified from covariates constructed using these data. The methodology used to construct these covariates is frequently determined by experts rather than the data, and the effects on the risk factors identified remain unclear. Using the standard medical cut-offs is convenient and has the advantage of easily comparing the results between studies [18]. Although, this is a form of confirmation bias and may lead to the statistical misclassification of risk factors [19].

The most effective precautionary measure to reduce the risk of transmitting harmful microorganisms in hospitals is adherence to well established and effective hand hygiene policies [20]. The lack of hand hygiene compliance (HHC) can result in the outbreak of harmful microorganisms in hospitals. The use of spatiotemporal data has become more prevalent when modelling the transmission and spread of microbes in hospitals. Innovative unobtrusive technology for tracking hospital assets, patients and healthcare workers is currently being explored [21–23]. Recently, radio frequency identification (RFID) has been progressively implemented in hospitals this effect [24]. The output of this technology can also be used to predict the spread of harmful microorganisms in hospitals [25]. Although, the spatiotemporal effects of varying levels of HHC on the transmission and spread of HMO in hospitals must still be quantified.

Spatiotemporal data in statistical models are essential to accurately model and predict the transmission dynamics of harmful microorganisms in hospitals. These models typically focus on single hospital wards, while interactions between wards and hospitals were later introduces [26]. RFID contact data are most desirable for detailed information, but they are typically unavailable for all hospitals [27]. Alternatively, a more commonly available source of spatiotemporal data to track patient movements in hospitals is the intrahospital movement data captured in the EHR. These data can predict the spread of microbes between hospital wards but are usually not considered risk factors for the contraction of AMR [28–31].

To address the current knowledge gaps in AMR research, this thesis investigates how statistical models and novel spatiotemporal data can enrich existing risk factors and identify new risk factors to predict the occurrence and spread of harmful microorganisms and the complication of AMR. Chapter 2 investigates how a bibliometric data-driven methodology can be used to identify knowledge gaps in AMR research. Traditional risk factor identification methodology evaluated and improved using statistical models in Chapter 3. Chapter 4 and 5 focus on enriching statistical models with spatiotemporal data to incorporate the spatiotemporal transmission dynamics of harmful microorganisms in hospitals. Chapter 4 describes

how the spatiotemporal movements of healthcare workers can identify potential a super-spreader occupation group in a hospital using spatiotemporal risk outcomes. Chapter 5 determines how the occurrence and spread of VRE can be explained using intrahospital patient movements (IPM) and their antibiotic use between hospital wards.

This thesis aims to answer the following research questions:

1. How can knowledge gaps in AMR research be identified objectively and automatically?
2. What are the risk factors for the occurrence of SSI when using data-driven cut-off values for continuous variables?
3. How can the spatiotemporal movements of healthcare workers identify potential a super-spreader occupation group of harmful microorganisms in a closed healthcare setting?
4. How can the occurrence of VRE in a hospital be predicted using intrahospital patient movements and antibiotic usage?

1.2 Background research project (EurHealth-1Health)

Due to the broad use of antimicrobials, AMR exists in multiple international sectors and across species. To this end, the "One-Health" approach was introduced by the Food and Agriculture Organization (FAO), World Organization for Animal Health (OIE), and World Health Organization (WHO) to study AMR across national borders and industries at a national and global level using the skillset of multiple disciplines [32,33]. The INTERREG VA project, EurHealth-1Health, was established to combat infections caused by the occurrence and spread of antimicrobial resistance (AMR) across the Dutch and German borders and between species. This all-encompassing approach was necessary as AMR does not adhere to the societal borders we set. Rhine-Westphalia and the Ministry for National and European Affairs and Regional Development of Lower Saxony.

This research was also supported by the INTERREG VA (202085) funded project EurHealth-1Health (http://www.eurhealth1health.eu), part of a Dutch-German cross-border network supported by the European Commission, the Dutch Ministry of Health, Welfare and Sport, the Ministry of Economy, Innovation, Digitalisation and Energy of the German Federal State of North.

1.3 Thesis outline

This thesis consists of six chapters. Chapters 3-4 are based on journal articles, and chapters 2 and 5 are based on articles under revision still to be published. The four research questions are addressed in Chapters 2-5, respectively.

Chapter 1 provides the background of the research setting and motivates the need for the research performed in this thesis. It sets the focus of the subsequent chapters and provides an overview of the thesis structure. Chapter 2 is the first research study performed in this thesis. Twenty years of AMR research is used to identify the main research areas in AMR research and obtain an objective data-driven view of the potential knowledge gaps. Examples of knowledge gaps are highlighted, and a complete list of potential knowledge gaps in AMR research is provided for the community to investigate further. This thesis challenges standard medical cut-off values used in modelling techniques to identify risk factors in healthcare in Chapter 3. Data-driven cut-off values are used instead of standard medical cut-offs to identify risk factors for surgical site infection from digestive, thoracic and orthopaedic system surgeries. Chapter 4 introduces empirical spatiotemporal data to model the spread of harmful microorganisms in an academic medical centre. It shows how indoor localisation data collected using RFID sensors can identify potential super spreading occupation groups and quantify the effects of varying levels of hand hygiene compliance in a healthcare setting. In the final study of this thesis (Chapter 5), common hospital data present in most electronic healthcare records are used to predict VRE occurrence at the hospital ward level. Patient movement and antibiotic use data are transformed into covariates through centrality measures that present how patients and antibiotics move to each ward. The results of this study are two models which can be used to calculate the probability that at least one patient has VRE in a particular ward on a specific day. In Chapter 6, the results of this thesis are synthesised and further discussed in terms of implications, limitations and future research opportunities.

In the following sections, more detail is provided about what can be expected in the subsequent chapters of this thesis.

1.4 Identifying knowledge gaps in AMR research

Knowledge gaps are used to inform research agendas in the AMR research field and are often identified using a manual expert method [6–11]. This method is prone to selection bias as the identified knowledge gaps and topics might have differed when experts from another discipline were involved [34]. It may also lead to nonreproducible and dated results [14]. A new scalable data-driven approach is needed to determine the knowledge gaps in AMR research. Chapter 2 identifies the potential research gaps in the AMR research field using a scalable and data-driven methodology.

Recent advances in scalable statistical models for text analysis made it possible to estimate the latent topics which generated the words of an observed body of text. Structural topic modelling (STM) is state of the art in unsupervised topic modelling

[35]. This thesis identifies the latent topics that generated the text in the AMR using STM. The PubMed database was queried using a broad search string related to AMR research published over the past 20 years.

To identify the knowledge gaps between the topics, the AMR topics are clustered into larger AMR research areas based on how they are studied together in the AMR literature and determine the strength of the relationship between them and the topic using Spearman's rank correlation coefficient based on the topic proportions [36]. Next, we highlight and discuss knowledge gaps identified using this semi-automated data-driven methodology.

Knowledge gaps are identified in the AMR research field, using a scalable data-driven, statistical approach, providing a repeatable and scalable way to identify potential knowledge gaps in AMR research. Examples of knowledge gaps are highlighted and discussed, and a complete list of potential knowledge gaps is provided for the community to investigate further.

1.5 Data-driven risk factors for surgical site infection

Surgical site infection (SSI) is the largest category of HAI [37]. The consequences of these infections are exacerbated when the infectious bacteria are resistant to the antibiotics administered to kill them [38]. Risk factors associated with SSI have received much attention in the scientific literature [18]. These factors typically include patient demographics and comorbidities and are based on the well-established categorical groupings and standard medical cut-offs for continuous variables rather than data-driven cut-offs, which may be suboptimal [39]. Using standard medical cut-offs is convenient and makes for easy comparisons between studies [18]. Although, this is a form of confirmation bias and may lead to the statistical misclassification of risk factors [19]. It should be determined if the risk factors for SSI are different when using data-driven cut-offs compared to if medical cut-offs are used. In Chapter 3, risk factors for the occurrence of SSI are identified using data-driven cut-off values for continuous variables.

This study was performed using data from the Erasmus MC University Medical Centre in Rotterdam are used, one of the largest university medical hospitals in the Netherlands with more than 1 320 beds [21]. A multivariate logistic regression model is built using a forward stepwise approach for each of the three groups of surgeries [41]. The data-driven cut-off values for the continuous variables are used determined using recursive partitioning [42]. Model performance is compared using the Gini coefficient and cross-validated using 5-fold cross-validation to estimate how the model would perform on new data [43]. The difference between the risk factors

identified using the standard medical cut-offs and the data-driven cut-offs are reported.

The results will inform better decision making when determining how to use continuous data when identifying risk factors for SSI.

1.6 Transmission of harmful microorganisms through connected HCWs

Hand transmission of harmful microorganisms (HMO) may lead to infections and poses a major threat to patients and HCWs in healthcare settings [44]. The most effective countermeasure against these transmissions is the adherence to spatiotemporal hand hygiene policies, but adherence rates are relatively low and vary over space and time [20]. Although the impact of HHC has been studied, the spatiotemporal determinants and potential consequences in healthcare settings remain undetermined. Potential super-spreaders occupation groups are identified in a closed healthcare setting and the risk of HMO transmission for different levels of HHC was quantified using empirical movement data in Chapter 4.

This study is based on empirical RFID contact data collected at the University Medical Center Groningen (UMCG), one of the largest university medical hospitals in the Netherlands with more than 10 000 employees almost 1 400 beds. Using the RFID contact data, a transition probability matrix $\boldsymbol{P}$ is constructed with p_{ij}, the probability of an HCW transitioning from room R_i to R_j room j (Formula 1.1-1.2).

$$p_{ij} = P\left(\text{Next room} = R_j \middle| \text{Current room} = R_i\right) \text{ for i, j} \in (1, \dots, \text{n}) \qquad \text{Formula 1.1}$$

$$\boldsymbol{P} = \begin{array}{c} \\ R_1 \\ \vdots \\ R_n \end{array} \begin{array}{c} \begin{array}{ccc} R_1 & \dots & R_n \end{array} \\ \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \end{array} \qquad \text{Formula 1.2}$$

A custom agent-based model is used to simulate the spread of harmful microorganisms in a hospital. Difference levels of assumption are used for microbe transmission, hand hygiene compliance and hand hygiene efficacy [45]. The sensitivity of the results to the change in the assumptions is quantified using four risk outcomes based on the number of minutes spent colonised, number of contacts, number of people in contact, number of transitions between rooms and the expected number of transmissions of an infection HCW.

In Chapter 4, a potential super-spreader occupation group is identified based on the potential risk of transmitting harmful microorganisms quantified using

spatiotemporal movements and social mixing patterns. These results will increase our insight into the consequences of varying levels of adherence to spatiotemporally specific healthcare policies such as hand hygiene compliance in a closed healthcare setting.

1.7 Predicting the spread of AMR in a hospital

Vancomycin-resistant enterococci (VRE) was first reported in Europe in 1986 and since then has been the cause of severe public health and monetary burdens [46,47]. These microorganisms can survive on inanimate surfaces for several months while spreading throughout hospital departments within days if the proper infection prevention strategies are not in place [48]. Studies have shown a significant relationship between intrahospital patient movements IPM and the occurrence of HAI infection [28,49]. The effects of IPM and antibiotic usage in hospitals are usually studied separately in AMR research. The use of antibiotics is usually included as a possible confounding effect to predict VRE in patients, but the antibiotics used by patients who may have frequented the same ward as the patient in question is often neglected. Since hospitals are dynamic systems with many moving objects that can serve as vectors for VRE, the occurrence of VRE should be studied using the spatiotemporal patterns of patients and antibiotics in the hospital. The occurrence of VRE is predicted at the ward level using conventional spatiotemporal patient and antibiotics data in Chapter 5.

Retrospective patient movement and antibiotic data are used from UMCG. The data are used to create a directed graph where the nodes are the wards. The patient movements are the directed edges between the nodes. The antibiotic data are used to create a binary indicator for each edge based on the patient's antibiotic use ($0 = not\ using\ antibiotics,\ 1 = using\ antibiotics$). Using this indicator, a second graph is created using only the edges with antibiotic use. Two daily centrality covariates are created using the PageRank algorithm to quantify the flow of patients and antibiotics at the ward level [50]. These daily centrality measures are based on the graph data over the past 30 days. In addition, two traditional covariates are calculated the daily number of patients present in each ward and how many of them are using antibiotics. In total, four covariates are calculated to explain the occurrence of VRE at the ward level.

The binary outcome variable Y is defined such that

$$Y = \begin{cases} 1, & \text{number of VRE positive patients in ward} > 0 \\ 0, & \text{otherwise} \end{cases}$$

The outcome variable is modelled using decision trees and random forest statistical models [39,51]. Decision trees are based on systematically splitting the outcome

variables according to covariates considered. The result is a set of simple rules based on the covariate values that are easy to implement in practice. The random forest model is an ensemble of decision trees. The random forest model will perform better than the decision tree but will not result in a simple set of rules like the decision tree. The difference in model results is compared using the Gini coefficient, which summarises all levels of model sensitivity and specificity [52].

In Chapter 5, two daily centrality measures were proposed to summarise the flow of patients and antibiotics at the ward level using data present in most electronic healthcare records. A simple set of rules was produced that can be used to monitor VRE risk in hospital wards. An ensemble model was proposed to improve the prediction performance at the cost of simplicity. An early warning system for VRE can be developed to test and further develop infection prevention plans and outbreak strategies using these results.

2. Antimicrobial resistance: Identifying knowledge gaps using semi-automated topic modelling

Abstract

Antimicrobial resistance is a multifaceted global problem and a significant threat to sustainable modern healthcare. Strategic action plans to tackle the increasing international threat of AMR are based upon research agendas that can be informed using knowledge gaps in the AMR research field. Currently, these knowledge gaps are identified manually and are often subjective. The first objective was to use a data-driven methodology to identify knowledge gaps in AMR research. To this end, the twenty years of AMR related articles were extracted from the PubMed database. We identified the topics comprising the AMR research field with structural topic modelling, while topic clusters were created using hierarchical clustering on the topic proportions. Potential AMR knowledge gaps were obtained using Spearman's correlation between topic clusters and topics and between individual topics. A total of 88 topics and seven topic clusters were identified from 158 616 scientific AMR research articles. In total, 421 potential knowledge gaps were identified between the topic clusters and topics and 2 663 between individual topics. Key knowledge gaps between molecular and laboratory AMR research were highlighted. The knowledge gaps between AMR research regarding water and the environment and both institutional and international surveillance topics were highlighted at the topic level. These results provide an innovative, data-driven way to identify knowledge gaps in AMR research. Technical advisory groups across sectors and industries can use these results to guide future AMR research agendas.

This chapter is partially based on Luz C, van Niekerk JM, Keizer J, Beerlage-de Jong N, Braakman-Jansen A, Stein A, Sinha B, van Gemert-Pijnen L, Glasner C. Mapping twenty years of antimicrobial resistance research trends. Available at SSRN 3792901. 2021 Jan 1. This article was submitted to Artificial Intelligence In Medicine and is undergoing minor revision.

2.1 Background

Antimicrobial resistance (AMR) is a multifaceted global problem and a significant threat to sustainable modern healthcare [53]. Current research estimates that 700 000 people die due to AMR annually and that this number may increase to 10 million by 2050, even though global estimates remain difficult to determine [54,55]. Since AMR microorganisms can occur in and spread between humans, animals and the environment, it is also studied as a *One Health* problem [56].

To tackle this problem, the World Health Organisation (WHO) has proposed a holistic action plan in response to the increasing AMR threat [5]. It was created using the consolidated objectives of existing action plans and best practices related to AMR and stretches across international sectors, industries, and disciplines. Governments are urged to ensure long-term investment for research and development to counter AMR. Healthcare professionals are encouraged to prescribe antimicrobials when necessary, practice proper hygiene measures to prevent new infections and communicate the dangers of misuse of antimicrobials to their patients. Agriculture is encouraged to limit the use of antimicrobials on animals and take preventative measures to reduce the need for antimicrobials. In parallel, the pharmaceutical sector should develop new antimicrobials and find alternatives. But even with the increase in multidisciplinary collaboration, these actions can be disrupted by knowledge gaps in and between industries and disciplines [5]. To this end, a global strategic AMR research agenda was proposed to understand AMR at the global level [5].

Knowledge gaps can be used to inform AMR research agendas [6–11]. To identify knowledge gaps in a research field, a thorough overview and understanding of the available knowledge in that research field are needed [12,13]. Knowledge gaps constantly need to be assessed for comprehensiveness and relevance. However, with the exponential increase in AMR research output, it is increasingly challenging to organise and synthesise the current state of AMR research to stay informed about previous and most recent scientific contributions [14].

A scoping review presented a thematic overview of knowledge gaps research covering research related to preventing antibiotic and antimicrobial resistance [3]. A total of 431 225 references from the initial search results was reduced to 622 unique references using a two-step process. Potential knowledge gaps were manually identified by an expert specialising in epidemiology and global health. However, this manual expert method has a possible selection bias as the identified knowledge gaps and topics might have been different when experts from another discipline would have been involved.

To overcome this possible bias, data-driven computational techniques for text analysis can be used to determine underlying topics in large quantities of literature automatically [15]. Recent advances in statistical models for text analysis made it possible to estimate the latent topics which generated the words of an observed body of text. Structural topic modelling (STM) is state of the art in unsupervised topic modelling [35]. STM is preferred to Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) as it results in a probabilistic instead of a rigid classification [35]. While Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) offer probabilistic results similar to STM, both use simplistic prior distributions, while STM takes full advantage of the covariates at the document level [57].

To create thematic groups of topics in a systematic way, statistical clustering can be used to automatically group the latent topics thematically [58,59]. The relationships between thematic clusters and other topics can then be quantified automatically using the correlation between topic proportions. In this way, less human interaction is required to understand their structure.

This study aims to identify potential knowledge gaps in the AMR research field using a data-driven, statistical approach. As a sub-objective, we determine the underlying research topic and groups of topics in the AMR research field. It provides a repeatable and scalable way to identify potential knowledge gaps in AMR research.

2.2 Methods

An overview of the steps followed to arrive at the potential knowledge gaps in AMR research is presented in Figure 2-1. The PubMed database was queried using an AMR related search string to extract the scientific research related to AMR. A corpus with its associated metadata was extracted from these data. These data were used as input for the STM algorithm to identify topics in the corpus. The topics were first grouped using expert judgement to obtain thematic groups and then quantitatively to obtain thematic clusters. Potential knowledge gaps were identified by assessing the correlation between the topics and the thematic clusters and between the topics themselves.

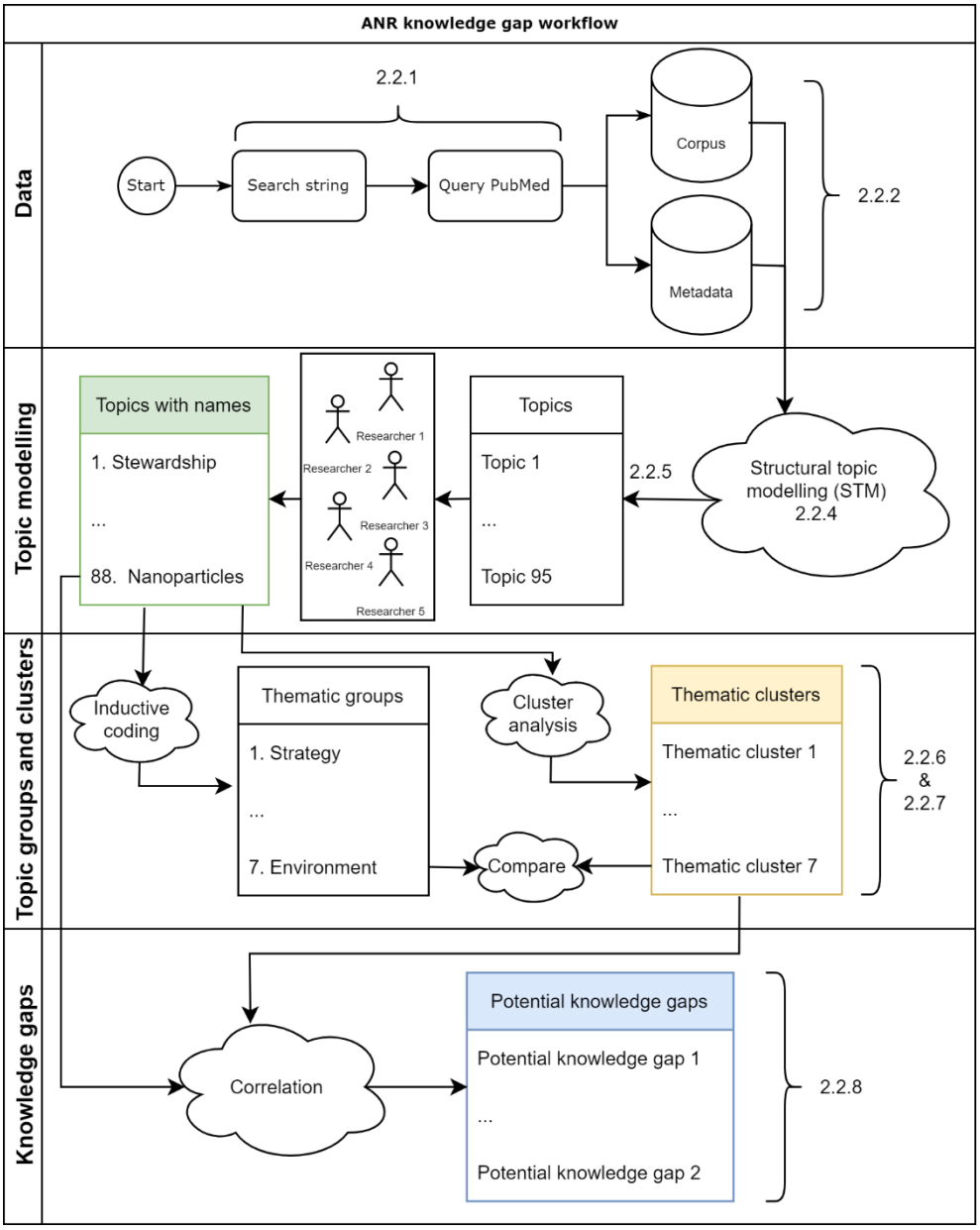

Figure 2-1: Workflow diagram illustrating the process followed to identify potential knowledge gaps in AMR research.

2.2.1 Search string and data extraction

The PubMed database was queried using a combination of free text terms (tiab) and medical subject headings (MeSH) in a search string consisting of two parts [60]. The following search string was used: 1) ("Anti-Bacterial Agents"[MeSH] OR Anti-Bacterial* [tiab] OR antibacterial* [tiab] OR antibiotic* [tiab] OR antimicrobial* [tiab] OR antimycobacterial* [tiab] OR "Antifungal Agents"[MeSH] OR Antifungal* [tiab] or anti-fungal* [tiab]); 2) ("Drug Resistance"[MeSH] OR resistan* [tiab] OR "Microbial Sensitivity Tests"[MeSH]). The first part of the search string covers the broad research field of antimicrobials, while the second part narrows the search results down to antimicrobial resistance. The inclusion criteria were such that all results were journal articles with a title and abstract and were published between January 1, 1999 to December 31, 2018. PubMed identification number, author affiliations, title, abstract, year of publication and citations were extracted from the NCBI Entrez database.

2.2.2 Corpus

Title and abstract were merged to create a text variable for each article. The text was cleaned from non-words and short (≤ 2 character) character strings and parsed to American Standard Code for Information Interchange (ASCII) encoded characters. Using the snowball language for stemming algorithms, a list of generic stopwords was created. Those, together with a list of domain-specific stopwords, were excluded [61]. The final text variable was created by stemming the resultant article text using snowball stemming.

2.2.3 Metadata

The article affiliation data were used to determine the affiliation of the first author of each article. This country was then used as the country variable, while the Google Maps API was used to determine the country of the first author if no country was listed but an affiliation was given [62]. The PageRank was added as a centrality measure based on the citations between the literature extracted to indicate the relative importance of each publication [63]. The metadata variables used for STM were: 1) year of publication; 2) country; 3) the total number of citations; 4) PageRank.

2.2.4 Structural topic modelling

STM assumes that a probabilistic generative process generated the text in the corpus according to a specified process structure, associated parameters and document-level covariates. Once the process is defined, the observed data are used to estimate the parameters defined in the generative process using Bayesian inference [64].

The probabilistic generative process is defined by letting each document d in corpus with size $Docs$ be generated from $Topics$ distinct topics, consisting of a possible vocabulary of size Voc. Let P_Vars be a $Docs \times p$ matrix containing p document-level topic prevalence covariates as rows $\boldsymbol{x}_d$. Similarly, let C_Vars be a $Docs \times l$ matrix containing l document-level topical content variables as rows $\boldsymbol{y}_d$.

Let each document have N_Words_d words indexed by n such that $n \in \{1, \ldots, N\_Words_d\}$. Each n is assigned a topic $topic_{d,n}$ according to its assumed distribution:

$topic_{d,n}|\theta_d \sim$Multinomial$(\theta_d)$ (Formula 2.1)

where $0 \leq \theta_d \geq 1$ is defined for each document d as

$\theta_d|P_Vars_d\gamma, \Sigma \sim$LogisticNormal$(\mu = P_Vars_d\gamma, \Sigma)$ (Formula 2.2)

where γ is the $p \times 1$ coefficient vector such that $P\_Vars_d\gamma$ is the covariate specific prior and Σ the $(K-1) \times (K-1)$ global topic covariance matrix.

The probability distribution of the words in vocabulary Voc for each topic $topic$ and document d with its associated covariates P_Vars_d is given by

$\beta_{d,k} \propto$exp$(m + \kappa_k^{(t)} + \kappa_{y_d}^{(c)} + \kappa_{y_d,topic}^{(i)})$ (2.3)

where exp is the exponential distribution, m is the baseline word distribution and $\kappa_k^{(t)}$ and $\kappa_{y_d}^{(c)}$ are the topic-specific and covariate group deviations respectively and $\kappa_{y_d,topic}^{(i)}$ is their interaction.

For each word $w_{d,n}$ in document d we sample a word from the multinomial distribution with parameter $\beta_{d,k}$ where $k = topic_{d,n}$ defined as

$w_{d,n}|topic_{d,n}, \beta_{d,k=topic_{d,n}} \sim$Multinomial$(\beta_{d,k=topic_{d,n}})$ (2.4)

The complete generative process is presented graphically in Figure 2-2.

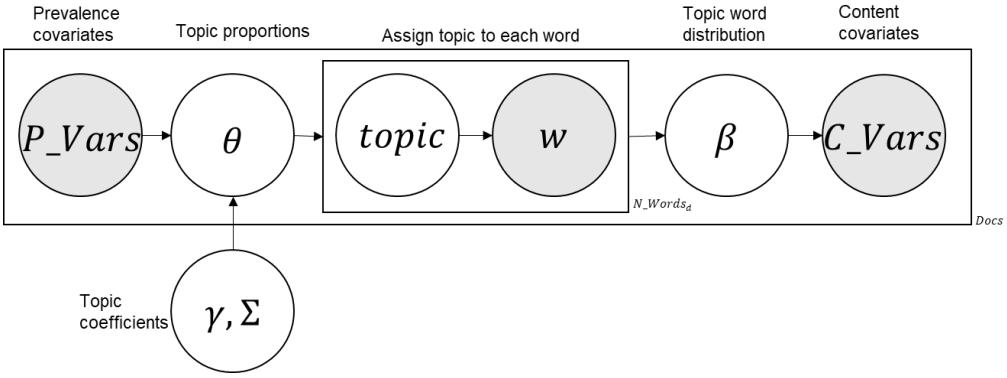

Figure 2-2: A graphical illustration of the STM generative process with the observed variables shaded.

The parameters are estimated using semi-collapsed variational expectation maximisation [65]. The joint optimum for topic proportions θ_d and word-level topic assignment $topic_{d,n}$ are obtained in the E-step by iterating through each document and updating the variational posteriors followed by the M-step where the evidence lower bound (ELBO) is maximised with respect to the global parameters (κ, y and Σ) [64–70]. This process is repeated until the ELBO convergences. The result is the estimated multinomial posterior distribution θ_d over the latent topics for each document (Formula 2.1) and the multinomial distribution over the vocabulary for each of the identified topics (Formula 2.4).

2.2.5 Implementation

The problem of obtaining the optimal number of topics assumed to have generated the literature is NP-hard [71]. To proceed, we used a grid analysis to identify possible optimal values for (K) [72]. Mixed-membership topic models have non-convex posteriors, which may result in the convergence to a local optimum when applying an expectation maximisation optimisation [69]. Spectral learning can be used to find the global maximum in multi-modal models consistently. In the case of STM, it can be used to determine a value of K which is in a region where the optimal value of K can be found [69,73]. Topic models based on different values of K can be compared quantitatively in terms of semantic coherence, i.e. how frequently cooccur high probability words for a topic, and the exclusivity the topics, i.e. how unique are the topics in terms of words with high probability. This study used spectral learning to find the initial value for K^* and further investigated the semantic coherence and exclusivity for topics based on values of K in the region of K^* to find the optimal number for K quantitatively.

The interpretability of the topics determined the final decision of the number. A panel of healthcare professionals compared the interpretability of the topic models, which produced the best qualitative fit [72]. They assigned topic names by evaluating the ten words most highly associated with each topic and reviewing the five documents with the largest associated topic proportions (θ_d) [74]. Topic names were further refined by scanning titles and abstracts of five highly associated documents per topic and five important documents per topic by PageRank. No topic name was assigned if this process did not converge to a meaningful name. Consensus was reached when both researchers differed in their generated topic names. Five independent AMR researchers reviewed this process and verified the generated topic names. The final model was chosen based on the highest number of topics with an assigned topic name. Each document was assigned the topic name of the topic comprising the highest proportion of the document's text. Topics in the final model were inductively coded into thematic groups to navigate the results.

2.2.6 Thematic clusters

The thematic groups were created using expert judgement and may be prone to bias. This possible limitation is overcome by an optimal grouping of topics quantitatively as thematic clusters.

The topic proportions (θ_d) of each of the articles were used to create a $Docs \times K$ topic proportion matrix, where the columns contain the K topic proportions. We use the topic proportions across the articles to identify larger research areas in AMR research. We cluster the identified topics using the degree to which the topic proportions are related across the articles. Using the topic proportion matrix, an $K \times K$ correlation matrix C was obtained with elements c_{ij}, the correlation between topic proportions of topics i and j. The degree to which the topics are related was quantified using the distance matrix $Dist = 1 - C$.

Topics were clustered into T thematic clusters (TC) using $Dist$ and applying the complete-linkage algorithm [75]. This algorithm sequentially combines topics nearest to each according to the distance matrix $Dist$ until the optimal value T^* was reached. Using the *silhouette* width, T^* was determined by minimizing the average distance between topics in the same cluster and maximizing distance to the topics in the nearest other cluster [76]. From manual topic modelling performed on AMR literature, we know that at least five TCs exist in the literature [3]. Thus, we apply the additional condition that $T^* \geq 5$ and T^* is the largest value before a significant decrease in the *silhouette* width occurs.

2.2.7 Thematic cluster focus

The topics proportion was used to determine the most prominent topics for each TC and describe the focus of the content concisely. The number of focus topics for each TC was limited to five to make the results tractable. Another criterion of the focus topics was that their topic proportion must be at least 15% of the largest topic proportion in that TC. For example, if 70% of the topic proportion in TC was assigned to the *surveillance* topic and the second-largest topic proportion was assigned to the *surgical site infections* topic with 10%, then the focus of the TC would be *surveillance*. Since $\frac{10\%}{70\%} = 14.23\% \leq 15\%$, the TCs only has one topic as the focus. If *surgical site infections* had a topic proportion of 25%, the focus would be *surveillance* as well as *surgical site infections*, since $\frac{25\%}{70\%} = 35.71\% > 15\%$.

2.2.8 Knowledge gap identification

Knowledge gaps are areas in scientific research where the information to support answers to the questions asked in those areas is either insufficient or non-existent [77]. We assume that the number of knowledge gaps between two research areas is negatively correlated with the strength of the relationship between them. Should there exist a strong relationship between the two research areas, we assume fewer knowledge gaps exist. If the relationship between them is weak, then we assume there is less evidence to support the questions asked between them.

In this context, we use the AMR topics and TCs as research areas. The degree of relatedness between these topics and TCs was determined using Spearman's rank correlation coefficient based on the topic proportions [36]. Potential knowledge gaps can be identified in two ways: 1) between a TC and a topic, which indicates a lesser representation of the topic in the larger research area; 2) between individual topics, which shows that the two topics are typically not studied together. We first identify the potential knowledge gaps between TCs and a topic and then describe the knowledge gaps in more detail by considering the potential knowledge gaps between topics. Potential knowledge gaps were identified as pairs of TCs and topics or between two topics with statically significant negative correlations at the 0.05 significance level.

2.2.9 Data availability and interactive user interface

This study generated substantial amounts of data that enable detailed analyses. The results in this manuscript present only selected highlights from these data. An interactive web-based application was developed to repeat this study's analyses and enable further analyses (https://topicsinamr.shinyapps.io/amr_topics/). Additionally, individual articles can be searched and assessed and the topic model

can be leveraged to evaluate texts from new articles not included in this study. Moreover, the data used and generated in this study are openly available under (https://osf.io/j3d65/).

2.2.10 Software

The R statistical programming language was used to perform the analyses in this study [78]. The RISmed and easyPubMed packages were used to extract the data from the PubMed database [79,80]. In addition, the tidyverse R packages was used to clean and structure the data [81]. Structural topic modelling was performed using the R stm package [70].

2.3 Results

In total, 158 616 articles were included, showing a steady increase over the past 20 years (8.5% nominal annual increase). In 2018, 14 547 articles were published, an increase of 450% compared to 1999.

2.3.1 Topic modelling

The optimal number of topics for the structural topic model was determined to be equal to $K^* = 75$. To investigate the sensitivity around K^*, we determined the semantic coherence and exclusivity measures for K from 15 to 205 in steps of 10. Semantic coherences suggested that the topics are more coherent as the number of topics increase beyond 205. The exclusivity measure indicated that the topic fits the corpus best between $K = 95$ and $K = 155$. The healthcare experts determined that $K = 95$ was the best qualitative fit based on the interpretability of the topics identified.

Topic names were manually assigned to each topic as well as one of seven thematic group names: 1) Strategy; 2) Methods; 3) Clinical; 4) Pathogen; 5) Compound; 6) Structure; 7) Environment. Correlations of the topic proportions and the thematic group names are shown in Figure 2-3.

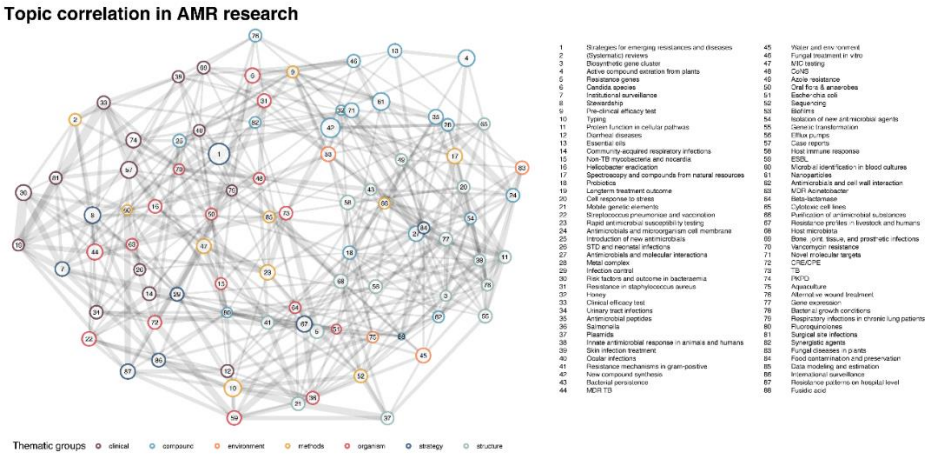

Figure 2-3: Topic names weighted by the topic proportion correlations.

2.3.2 Thematic clusters

The silhouette width indicated that the optimal number of thematic clusters is $T^* = 2$. The first significant decrease in the average silhouette after $T = 5$ was observed from $T = 7$ to $T = 8$. Thus, $T^* = 7$ was used for the optimal number of TCs.

The seven TCs with their respective foci are shown in Figure 2-4. The sub-figures are radar plots with evenly spaces axis leading to the perimeter of the graphs. A radar plot of a TC with similarly proportioned foci resembles a regular pentagon, e.g., TC2. This resemblance indicates that there is no single strong focus and that the main topics in the TC have a similar representation in the article text. In comparison, the radar plot of TC1 resembles a sharp triangle as it is the only TC with less than five foci with concentrated topic proportions in *emerging resistances and diseases* and *stewardship*. The focus of TC2 consists of topics related to patient outcomes and institutional surveillance. TC3 pertains to laboratory research with a focus on resistance testing and new antimicrobial compounds. Topics related to genomic resistance patterns in healthcare and agriculture and international surveillance topics are the focus of TC4. Treatment outcomes and antimicrobial efficiency is the focus of TC5. TC6 is centred around sequencing the human microbiome, and the environment. TC7 has its focus in new compounds and novel molecular targets.

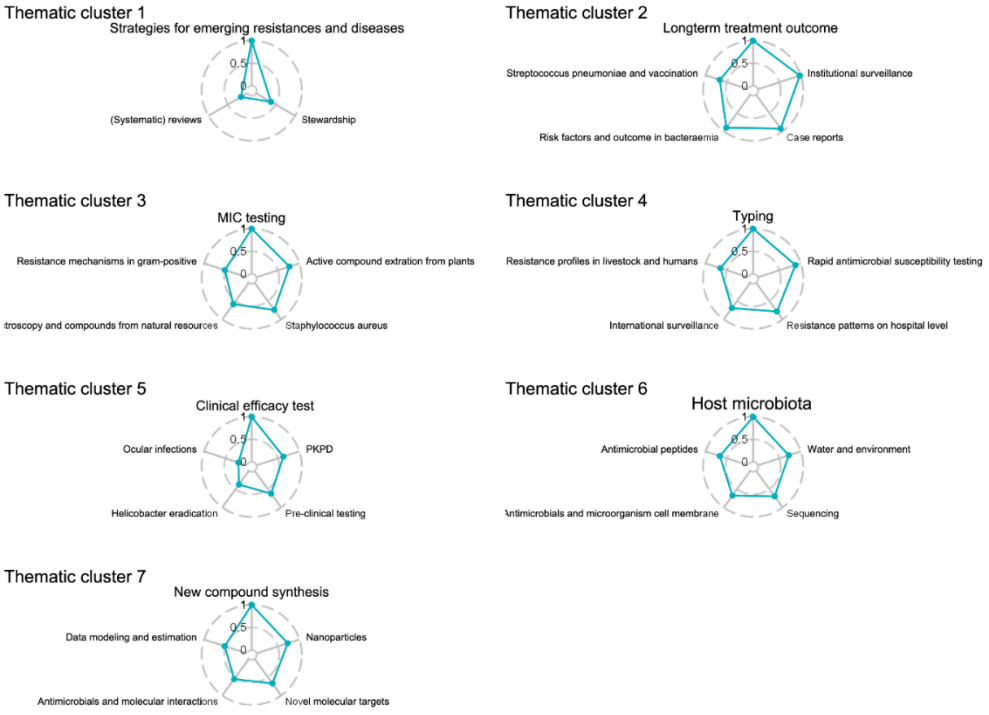

Figure 2-4: Thematic clusters and their foci.

2.3.3 Identifying knowledge gaps

In total, 421 potential knowledge gaps were identified between TCs and topics and 2 663 between individual topics. We highlight some key results in this section and invite the reader to identify more knowledge gaps from the complete set of results. The ten topics with the largest negative correlation for each TC are shown in Table 2-1. All the correlations shown in this table are statistically significant with p-values of < 0.001. We highlight potential knowledge gaps in TC3, TC5 and two potential knowledge gaps at the topic level.

Table 2-1: The ten most unrelated topics to each thematic cluster.

Thematic Cluster	Unrelated topic	Correlation	p-value
Thematic Cluster 1	MIC testing	-17.2%	< 0.001
	Typing	-13.0%	< 0.001
	Bacterial growth conditions	-12.3%	< 0.001
	Mobile genetic elements	-12.2%	< 0.001
	Fusidic acid	-12.1%	< 0.001

	New compound synthesis	-11.9%	< 0.001
	Escherichia coli	-11.8%	< 0.001
	Resistance mechanisms in gram-positive	-10.8%	< 0.001
	Protein function in cellular pathways	-10.4%	< 0.001
	Resistance genes	-9.9%	< 0.001
Thematic Cluster 2	Novel molecular targets	-18.0%	< 0.001
	New compound synthesis	-14.3%	< 0.001
	Antimicrobials and microorganism cell membrane	-14.0%	< 0.001
	Gene expression	-14.0%	< 0.001
	Antimicrobials and molecular interactions	-13.4%	< 0.001
	Bacterial growth conditions	-13.1%	< 0.001
	Fusidic acid	-12.8%	< 0.001
	Protein function in cellular pathways	-12.1%	< 0.001
	Cell response to stress	-12.1%	< 0.001
	Pre-clinical testing	-11.9%	< 0.001
Thematic Cluster 3	Strategies for emerging resistances and diseases	-20.7%	< 0.001
	Long-term treatment outcome	-19.1%	< 0.001
	Stewardship	-15.2%	< 0.001
	Clinical efficacy test	-14.3%	< 0.001
	Risk factors and outcome in bacteraemia	-13.0%	< 0.001
	Institutional surveillance	-12.9%	< 0.001
	MDR TB	-12.8%	< 0.001
	Host microbiota	-12.6%	< 0.001
	Data modelling and estimation	-10.8%	< 0.001
	Case reports	-10.7%	< 0.001
Thematic Cluster 4	Novel molecular targets	-18.7%	< 0.001
	Pre-clinical testing	-17.9%	< 0.001
	New compound synthesis	-17.4%	< 0.001
	Antimicrobials and microorganism cell membrane	-15.3%	< 0.001
	Cytotoxicity	-14.4%	< 0.001
	Nanoparticles	-14.3%	< 0.001

	Strategies for emerging resistances and diseases	-14.3%	< 0.001
	Cell response to stress	-12.8%	< 0.001
	Antimicrobial peptides	-12.6%	< 0.001
	Active compound extraction from plants	-12.2%	< 0.001
Thematic Cluster 5	Typing	-14.4%	< 0.001
	Mobile genetic elements	-11.7%	< 0.001
	Sequencing	-10.7%	< 0.001
	Resistance patterns on hospital level	-9.5%	< 0.001
	ESBL	-8.5%	< 0.001
	Antimicrobials and molecular interactions	-8.4%	< 0.001
	Plasmids	-8.4%	< 0.001
	International surveillance	-8.3%	< 0.001
	New compound synthesis	-8.1%	< 0.001
	Escherichia coli	-8.1%	< 0.001
Thematic Cluster 6	Long-term treatment outcome	-26.3%	< 0.001
	Institutional surveillance	-20.8%	< 0.001
	Clinical efficacy test	-20.0%	< 0.001
	Resistance patterns on hospital level	-19.4%	< 0.001
	Risk factors and outcome in bacteraemia	-17.5%	< 0.001
	MIC testing	-17.3%	< 0.001
	Microbial identification in blood cultures	-16.8%	< 0.001
	International surveillance	-16.3%	< 0.001
	Stewardship	-15.6%	< 0.001
	Case reports	-13.2%	< 0.001
Thematic Cluster 7	Long term treatment outcome	-19.2%	< 0.001
	Institutional surveillance	-16.7%	< 0.001
	Resistance patterns on hospital level	-16.4%	< 0.001
	Typing	-15.2%	< 0.001
	Risk factors and outcome in bacteraemia	-14.3%	< 0.001
	Clinical efficacy test	-14.1%	< 0.001
	International surveillance	-13.7%	< 0.001

Microbial identification in blood cultures	-13.3%	< 0.001
Infection control	-13.2%	< 0.001
Resistance profiles in livestock and humans	-13.0%	< 0.001

The knowledge gap analysis revealed that the TC3 is unrelated to *Risk factors and outcome in bacteraemia* (-13.0%), *Long term treatment outcome* (-19.1%), *Stewardship* (-15.2%) *and clinical efficacy test* (-14.3%). Comparing the topics that make up TC3, *Bacterial growth conditions* has a relatively large negative correlation (-6.5%, -10.0%, -5.8% and -8.0%) while *Staphylococcus aureus* and *Vancomycin resistance* are positively correlated with the highlighted, unrelated topics except for *Stewardship* (5.2%, 7.2%, -3.4% and 8.8%) (Table 2-2). This result highlights the potential knowledge gaps in the research area related to AMR laboratory research (TC3) and four topics related to clinical practice.

Table 2-2: Potential knowledge gaps in TC3.

Thematic Cluster 3 topic	Institutional surveillance	Risk factors and outcome in bacteraemia	Long term treatment outcome	Stewardship	Clinical efficacy test
Active compound extraction from plants	-6.4%	-5.4%	-8.6%	-4.4%	-6.7%
Bacterial growth conditions	**-7.7%**	**-6.5%**	**-10.0%**	**-5.8%**	**-8.0%**
CoNS	0.0%*	-1.1%	-0.5%	-2.7%	-2.5%
Essential oils	-4.2%	-3.5%	-5.7%	-3.0%	-4.4%
Food contamination and preservation	-5.1%	-4.6%	-7.5%	-2.9%	-5.5%
Fusidic acid	-7.4%	-7.2%	-10.0%	-6.2%	-7.1%
Honey	-2.4%	-2.3%	-3.4%	-1.8%	-2.5%
Introduction of new antimicrobials	1.3%	-2.1%	0.7%	-3.2%	8.2%
Isolation of new antimicrobial agents	-6.2%	-5.4%	-8.5%	-4.6%	-7.0%
MIC testing	-3.8%	-6.6%	-7.9%	-8.0%	-6.7%
Oral flora & anaerobes	-3.4%	-3.1%	2.4%	-1.7%	2.5%
Probiotics	-4.3%	-3.6%	-5.6%	-3.1%	-3.8%

Purification of antimicrobial substances	-5.4%	-4.4%	-7.2%	-3.7%	-4.9%
Resistance mechanisms in gram-positive	-1.4%	-3.9%	-5.7%	-5.4%	-6.0%
Spectroscopy and compounds from natural resources	-5.3%	-4.4%	-7.2%	-3.8%	-6.0%
Staphylococcus aureus	**6.2%**	**5.3%**	**2.8%**	**-4.7%**	**-3.6%**
Vancomycin resistance	**4.6%**	**5.2%**	**7.2%**	**-3.4%**	**8.8%**

*Correlation is not statistically significant at a 0.05 level of significance. Bold formatting indicates the results highlighted. CoNS= Aoagulase-negative staphylococci; MIC=Minimum inhibitory concentration.

Next, we highlight the potential knowledge gap between TC5 and *Typing* (-14.4%), *Mobile genetic elements* (-11.7%), *Sequencing* (-10.7%), *Resistance patterns on hospital level* (-9.5%) and *ESBL* (-8.5%). The unrelated topics have high negative correlations with *Clinical efficacy test* (-10.0%, -9.0%, -8.1%, -5.4% and -5.6%) and *Pre-clinical testing* (-11.0%, -7.9%, -6.6%, -11.0% and -6.1%). This result highlights potential knowledge gaps between the clinical AMR research area (TC5) and topics related to molecular AMR research.

Table 2-3: Potential knowledge gaps in TC5.

Thematic Cluster 5 topic	Typing	Mobile genetic elements	Sequencing	Resistance patterns on hospital level	ESBL
Clinical efficacy test	**-10.0%**	**-9.0%**	**-8.1%**	**-5.4%**	**-5.6%**
Helicobacter eradication	-2.7%	-2.5%	-2.4%	-0.8%	-2.5%
Ocular infections	-2.8%	-2.5%	-2.3%	1.3%	-2.1%
PKPD	-6.6%	-5.2%	-5.0%	-5.1%	-3.5%
Pre-clinical testing	**-11.0%**	**-7.9%**	**-6.6%**	**-11.0%**	**-6.1%**

Bold formatting indicates the results highlighted. PKPD=Pharmacokinetic/Pharmacodynamic.

At the topic level, there exist potential knowledge gaps between *Institutional surveillance* and *international surveillance* and *Water, and environment* with a

correlation of -3.9% and -2.0%, respectively. Another potential knowledge gap at the topic level is *Resistance patterns on hospital level* and *Data modelling and estimation* with a correlation of -8.1%.

We compare the TCs derived using the semi-automated framework introduced in this study with the thematic groups that manually group the topics in [15]. The comparison shows that most TCs are combinations of two to three thematic groups. TC1 only relates to topics associated with the *Strategy* thematic group, whereas TC3 and TC4 relate to three thematic groups.

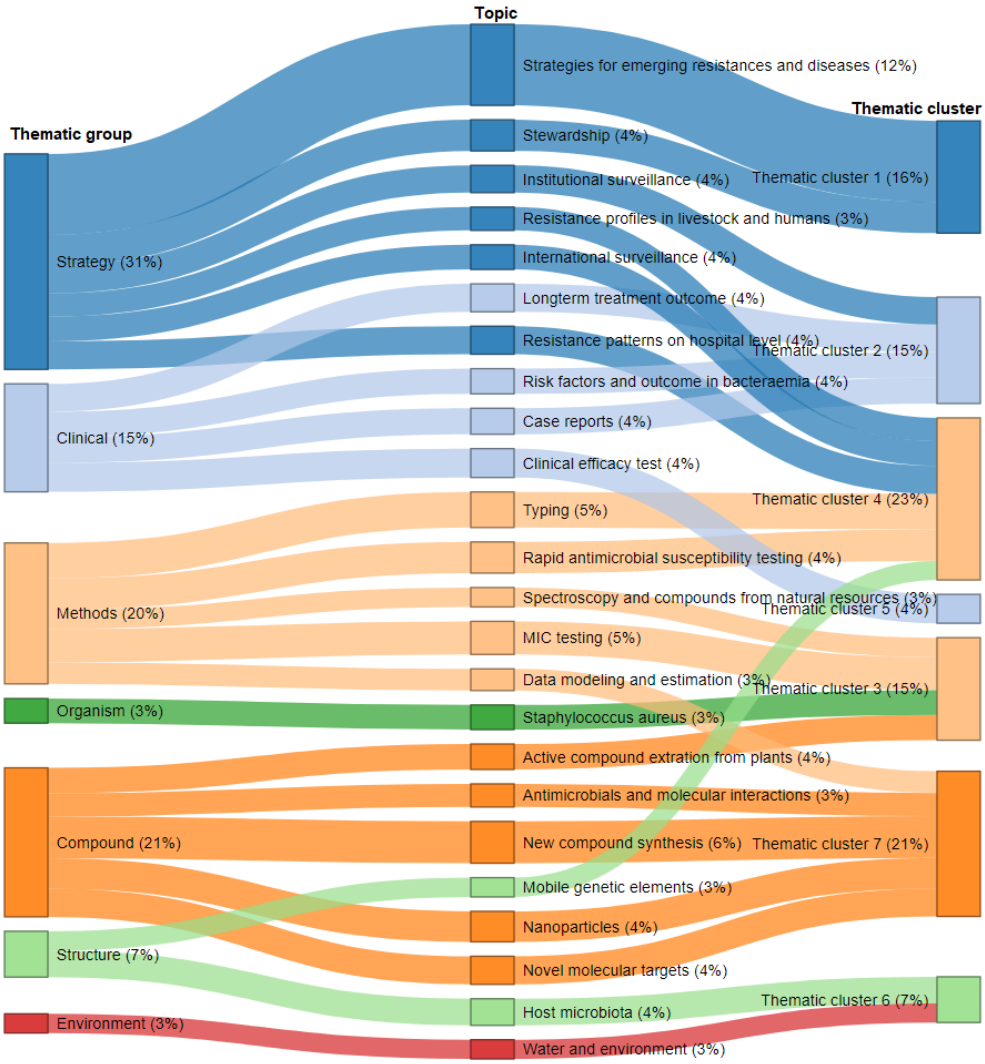

Figure 2-5: A comparison between thematic groups and thematic clusters connected by topics. Thematic group = names assigned to similar topics [15]. Only topics with a summed topic proportion of 2 000 are displayed for conciseness.

2.4 Discussion

We identified potential knowledge gaps in published research in the field of antimicrobial resistance (AMR) from the past 20 years using a semi-automated framework. Seven research areas were identified in the form of thematic clusters (TCs) using the 88 identified AMR topics. In total, 421 potential knowledge gaps were identified between the TCs and topics and 2 663 between individual topics. We

highlighted potential knowledge gaps in two TCs and between two topics and provided the complete results for future research.

2.4.1 Thematic clusters

The optimal number of clusters explaining the variance in the 88 AMR topics was determined as seven. This result is similar to the manual approaches used even though the content of the thematic clusters differs [3,15]. Comparing the TCs and the intuitive, thematic groups suggested that some of the TCs and thematic groups are very similar, such as TC1 and the *Strategy* thematic group and TC2 and the *Clinical* thematic group. Other TCs were a clear combination of thematic groups, e.g., TC6 contains topics from the *Structure* and *Environment* thematic groups. Using STM, we were able to model the complicated nature of how research topics are studied together, which would not be possible to do manually. This comparison illustrated the difference between manually grouping research based on the perceived topic and using a data-driven framework to understand the complexities of how those topics are studied together in scientific literature.

2.4.2 Knowledge gaps

Thematic cluster 3

Risk factors and outcome in bacteraemia, *Long term treatment outcome*, S*tewardship and clinical efficacy test* were identified as knowledge gaps in TC3. These topics are closely related to each other and are typically studied in clinical AMR research. Since the foci of TC3 are five topics related to AMR laboratory research, we conclude that these are the knowledge gaps related to a combination of clinical and laboratory AMR research.

The topics *Risk factors and outcome in bacteraemia, Long term treatment outcome,* and *Clinical efficacy test* are based on clinical outcomes under existing or potential AMR conditions. A limited number of studies have shown how AMR research related to *Bacterial growth conditions* can be studied with these clinical-outcome topics to significant effect. One study showed that favourable conditions for the germination of clinical *Clostridium difficile* spores are correlated with disease severity and adverse treatment outcomes [82]. Another study found that high levels of heavy metals (lead and cadmium) in patient blood is a risk factor for AMR [83]. These studies are an example of how technical laboratory analysis can be applied in clinical research.

Antimicrobial stewardship programmes, defined as a coherent set of actions that promote using antimicrobials responsibly, have become mandatory parts of institutional healthcare in many countries [84]. Over the last years, the stewardship term is also broadly used to incorporate antimicrobial use and other essential aspects

in diagnostics and infection prevention (e.g., diagnostic stewardship) [85]. This study identifies significant gaps with *Staphylococcus aureus, vancomycin resistance, isolation of new antimicrobial agents* and *CoNS* that can raise awareness in these unique aspects of stewardship.

Thematic cluster 5

The top five topic unrelated topics to TC5 are closely linked to each other. *Sequencing* is slowly replacing the role of *Typing,* although *Typing* is still widely used as a faster and cheaper alternative. These two topics are naturally related to the study of *Mobile genetic elements*, *Resistance patterns on hospital level* and *Extended spectrum beta-lactamase* (ESBL).

ESBL can be produced by certain bacteria, making them more resistant to beta-lactam antibiotics [86]. Even though the treatment outcome of patients with ESBL-producing bacteria are well represented in the scientific literature [87,88], this study found a research gap between ESBL and C*linical efficacy test, Pre-clinical testing* and *PKPD.* More septically, we found that the *Clinical efficacy test* and *Sequencing* are still unrelated in the AMR research field. AMR genetic determinants such as the production of efflux pumps and enzymes that reduce the effectiveness of antimicrobials can be identified using sequencing and typing technologies [89]. These technologies remain unrelated to topics that study the efficacy of new antibiotics. This knowledge gap could be addressed by research where the efficacy of new antibiotics is studied for different strains of bacteria identified through typing or sequencing. The results may lead to improved personalised patient treatments where the antibiotics prescribed depend on the strain of the bacteria [90].

Since the foci of TC5 contain topics related to the clinical study of AMR, we describe these knowledge gaps as the gap existing between the clinical and molecular AMR research.

Water and environment and surveillance

The results showed that *Water and environment* is unrelated to both international and institutional surveillance-related topics. Local, regional and international surveillance systems have been set up and used for extensive research, mainly focussing on hospitalised patients [91]. AMR can occur in healthy patients, but sampling this healthy population remains challenging. The water and environment surrounding the healthy population can be used to estimate the current state of AMR across the entire population.

Recent studies have shown that sewage and wastewater can be sampled to monitor the level of AMR at the community level [91,92]. At the institutional level, it was shown how AMR fluctuates on the personal mobile devices of nurses in intensive care units [93]. Together, these measurements can enrich surveillance reports to provide a holistic view of the institutional and international levels of the AMR situation. Yet, these topics are still unrelated in AMR research.

Data modelling and estimation and resistance patterns on hospital level

The knowledge gap between *Data modelling and estimation* and *Resistance patterns on hospital level* was highlighted in AMR research. The scientific literature available on data modelling and estimation of AMR patterns at the hospital level is limited and was confirmed by the results of this study. Combining epidemiological and microbiological data is essential to understand the resistance dynamics at the genetic, cellular, patient, and population levels. The lack of these data leads to the exclusion of essential mechanisms contributing to the uncertainty when modelling AMR [94].

The connectivity between patients, hospital wards and healthcare workers is how transmission of AMR occurs inside hospitals [28,30,95]. Few models aim to predict the occurrence and spread of AMR using empirical spatiotemporal data [96]. Typically, these movements are not tracked due to the lack of technologies, financial cost and the sheer complexity of system implementation. In some instances, these networks structures are estimated through expert knowledge and anecdotal case studies rather than empirical data. The mathematical models in the AMR research area remain predominantly deterministic. Some research recognises the critical role played by these data in viral epidemiology. For example, Brockmann et al. showed how complex spatiotemporal patterns extracted from global air traffic data could explain the spread of SARS [97]. No equivalent study could be found for AMR, and the gap still exists.

2.4.3 Limitations

This study assumed that a negative correlation between two topic proportions is an indication that there may exist a gap in the current AMR research. Without human intervention, it is not possible to determine if those potential knowledge gaps are worth exploring. Some topics may be unrelated for obvious reasons and are not worth exploring together. For example, the relationship between the topic *cytotoxic cell lines* and a TC focussing on surveillance may be weak. Even though some questions can be formulated in this area, they may not be sensible, relevant, or highly important, but they can be considered because of this study. A future research

opportunity is to automate and generalise this framework even further by introducing methods that understand the topic's contextual nature and their relationship to each other.

Another assumption made in this study was to determine the TC foci. The TC foci give the most prominent topics for each TC, but the next or sixth most prominent topic could also be essential to determine the context of the TC. Similarly, the 15% rule to determine the TC foci was used to create concise results and not to maximise the interpretability of the TC. Future research can improve this by using text analysis across the topics to extract a coherent contextual summary of each TC.

2.5 Conclusion

A semi-automated data-driven approach was used to identify the potential knowledge gaps based on 88 topics identified in AMR research published over the past 20 years. Examples of knowledge gaps were highlighted, and a complete list of potential knowledge gaps was provided for the community to investigate further. Technical advisory groups across industries and sectors can use these results to guide future AMR research agendas. Future research can use the applied methodology to other research fields to enumerate the potential knowledge gaps.

3. Risk factors for surgical site infections using a data-driven approach

Abstract

Surgical site infections make up 19.6% of healthcare-associated infections in Europe. Risk factor identification studies do not usually specify how continuous variable cut-offs are determined. In most cases, they are not determined by the data. The second objective was to identify risk factors for surgical site infection from digestive, thoracic and orthopaedic system surgeries using clinical and data-driven cut-off values. Retrospective surgery data were used from a tertiary care hospital in The Netherlands. Risk factors were identified using a multivariate forward-step logistic regression model. Standard medical cut-off values were compared with cut-offs determined from the data. For digestive, orthopaedic and thoracic system surgical procedures, the risk factors identified were preoperative temperature of 38 ^{o}C and antibiotics used at the time of surgery. C-reactive protein and the duration of the surgery were identified as risk factors for digestive surgical procedures. Being an adult (age $\geq$ 18) was identified as a protective effect for thoracic surgical procedures. Data-driven cut-off values identified for temperature, age, and CRP, explained the SSI outcome up to 19.5% better than standard medical cut-off values. Future studies should investigate if data-driven cut-offs can add value to explaining the modelled outcome and not solely rely on standard medical cut-off values to identify risk factors.

This chapter was published as van Niekerk JM, Vos MC, Stein A, Braakman-Jansen LM, Voor in 't holt AF, van Gemert-Pijnen JE. Risk factors for surgical site infections using a data-driven approach. PloS one. 2020 Oct 28;15(10):e0240995.

3.1 Background

Surgical site infections (SSI), as defined by the European Centre for Disease Prevention and Control (ECDC), make up 19.6% of the total number of healthcare-associated infections (HAIs) in Europe [1]. With an estimated 81 089 patients in Europe having an HAI on any given day, almost 16 000 people in Europe are suffering from some form of SSI at any given time [2]. The burden of SSI can be measured in terms of increased length of stay in hospital, additional (surgical) procedures required, increased morbidity and mortality, as well as in economic terms [100].

Risk factors relating to the patient, procedure and the environment alter the odds of an SSI occurring. Research has been done to identify risk factors for SSI with the aim to identify preventative actions to reduce the incidence rate of SSI [101–107]. Patient-related risk factors for SSI, such as obesity, diabetes, surgery duration and the American Society of Anaesthesiologists (ASA) score are risk factors for digestive system, thoracic and orthopaedic surgical procedures [15], [18–28]. In low-income countries, risk factors in low-income countries also include unemployment and level of education due to the disparity in socioeconomic status [117]. Risk factors can be modifiable or non-modifiable [109]. Modifiable risk factors are the most interesting of the two since they can be changed preoperatively to reduce the risk of SSI.

The Segmentation of surgical procedures into homogenous groups makes it possible to find useful and relevant risk factors unique to each segment. Digestive system surgical procedures are more prone to SSI as they are generally clean-contaminated or dirty surgeries, making deep space SSI more likely. The occurrence of SSI after thoracic and orthopaedic surgeries are both relatively low because they are both typically clean surgeries, but the probability of attracting a deep space SSI after thoracic surgery is much higher compared to orthopaedic surgeries [118]. Because of these differences, we focus on digestive system, thoracic and orthopaedic surgical procedures for this study.

Multivariate logistic regression is the most common statistical model used to identify risk factors in longitudinal study design data [119]. Not all studies report the discriminatory power of the multivariate logistic regression model fitted. Risk factor identification studies do not usually specify how continuous variable cut-offs are determined. Cut-off values for variables such as age (≥ 18) or patient temperature ($37\ ^{o}C$) may seem intuitive or standard for clinical practice, but they may not statistically be the best cut-offs values determined by the data [39].

The objective of this study is to identify risk factors for SSI from digestive, thoracic and orthopaedic system surgeries using clinical and data-driven cut-off values. A

second objective is to compare the identified risk factors in this study to risk factors identified in the literature.

3.2 Methods

3.2.1 Literature search

A literature search was performed to identify known risk factors for SSI associated with digestive system surgical procedures, thoracic surgery and orthopaedic procedures using the corresponding medical subject headings (MeSH) linked data representation and the MEDLINE database.

Search strings used for MEDLINE literature search:

1. "Surgical Wound Infection"[Mesh] AND "Risk Factors"[Mesh] AND "Digestive System Surgical Procedures"[Mesh]
2. "Surgical Wound Infection"[Mesh] AND "Risk Factors"[Mesh] AND "Orthopaedic Procedures"[Mesh]
3. "Surgical Wound Infection"[Mesh] AND "Risk Factors"[Mesh] AND "Thoracic Surgery"[Mesh]

The search results were sorted, using the *Best Match* algorithm developed by PubMed [120]. Search results were deemed relevant using title and abstract screening. Risk factors were extracted if they were significant in a multivariable analysis until data saturation was achieved [121]. Risk factors identified, which were common to all three groups of surgeries, were defined as "general risk factors" in this study.

3.2.2 Setting and data collection

The Erasmus MC University Medical Centre in Rotterdam is one of the largest university medical hospitals in the Netherlands with more than 1 320 beds [21]. The data used for this study were anonymised in accordance with the Dutch Personal Data Protection Act (WBP). Approval from the Medical Ethical Research Committee was obtained (MEC-2018-1185).

A weekly prevalence survey was performed by infection control practitioners (ICP) from January 2013 until December 2013 and two-weekly until June 2014 using a semi-automated algorithm proposed by Streefkerk et al. [40,122]. This algorithm was used to calculate a nosocomial infection index (NII) which was then verified by ICP in case of a positive outcome to determine whenever an HAI was present or not. An ICP verified all patients with an NII > 7, and a definite SSI outcome was concluded by the ICP using the electronic patient data system. This outcome was used in this study as the occurrence of SSI outcome variable.

Data were extracted from a centralised database, containing cross-departmental data, clinical synopsis reports, infectious disease consultation reports, laboratory results and imaging reports. Surgeries were included if they were part of the three groups of surgeries under investigation in this study and had a point prevalence measurement within 30 days after the surgery took place. If a second surgery took place within 30 days after an included surgery, then the recent surgery was excluded. All emergency surgeries were excluded to avoid possible undesirable confounding effects relating to the urgency and necessity of the surgeries.

3.2.3 Statistical analysis

The differences in the averages of variables with missing values and those without were evaluated using t-tests and were found statistically significant. Little's missing completely at random (MCAR) test was used to determine if the missing values were dependent on the data values themselves or not. These tests convinced us that the missing values were not completely randomly missing and that we could not make use of more simple imputation methods. Therefore, we chose to use conditional Markov chain Monte Carlo (MCMC) with multiple imputations for the imputation process [23,24].

Two methods were used to discretise continuous measurement variables: 1) standard medical cut-offs as used by Erasmus MC and 2) recursive partitioning [39]. Recursive partitioning is a data-driven, supervised discretisation method, used to group continuous values with similar outcomes optimally. The data-driven method was used to test and confirm if the standard medical cut-offs were the best way to explain the outcome variable for the groups of surgical procedures considered.

To build a prognostic prediction model for SSI, Hosmer et al. suggest fitting a univariate logistic regression model to each variable separately and if the p-value is less than a specific p-value, 0.1 is this case, then consider the variable good enough to include in the multivariate logistic regression model [125]. A univariate analysis was performed for each of the three groups of surgeries using the variables identified from the literature search. Significant variables ($p<0.1$) in the univariate analysis were added to the list of variables associated with each group of surgery, together with the variables identified from the literature search. This resulted in an extended list of general risk factors as more risk factors were common across the three groups of surgeries.

A multivariate logistic regression model was built using a forward stepwise approach for each of the three groups of surgeries [41]. The general risk factors were first added to the model and then the risk factors unique to each surgery group in the order of the Akaike information criterion (AIC) until convergence was reached. In this

case, we chose the conversion of the model to imply that there are no additional variables which can be added which will be statistically significant with a p-value of less than 0.05 or an AIC of 3.8415. Model performance was determined using the Gini coefficient after each step of the multivariate model, and the difference is reported as the marginal contribution of surgery group-specific risk factors for this study [52,121]. Model performance was cross-validated using 5-fold cross-validation to estimate how the model would perform on new data [43]. R [126] was used in this study together with packages mice (multiple imputation) [127], smbinning (recursive partitioning) [128], dplyr (data wrangling) [129], finalfit (formatting of tables) [130] and scorecard (cross-validation) [131].

Approval was obtained from the Medical Ethical Committee of Erasmus MC (MEC-2018-1185) to perform this study. Data were analysed anonymously, and thus no further consent was obtained.

3.3 Results

3.3.1 Literature search

The literature search resulted in 1 422 research papers (as at 5 March 2020) using the MeSH headings in the PubMed search engine. We identified 24 research papers, published from 2008 until 2019, which contained statistically significant results from a multivariate analysis. A total of 79 risk factors were identified for the three groups of surgical procedures [108–116,119,132–145]. Age, ASA class, body mass index (BMI), preoperative length of stay and diabetes were identified as general risk factors from the literature search. In total, 29 risk factors for digestive system surgical procedures, 31 for orthopaedic procedures and 19 for thoracic surgeries were identified. This amounted to 59 unique risk factors, of which 15 were present in more than one group of surgeries.

3.3.2 Risk factor identification

A total of 21 of the 59 unique risk factors could be replicated using our own data. The variable describing the type of surgery was used to create three homogenous groups of surgical procedures. The emergency classification variable was used to exclude emergency surgeries from the study such that 19 risk factors remained (Table 1). We observed 3 250 surgeries over the study period and excluded 526 (16.2%) emergency surgeries to be left with 2 724 surgical observations. CRP and temperature data were available for 52.55% (60.47% for in-patients) and 96.88% of all surgeries, respectively.

Table 3-1: Variable names and definitions used to investigate the occurrence of SSI in this study.

Variable	Surgery group	Definition
Demographic		
Gender	D,O	Gender of patient (Male/Female)
Age	D,O,T	Age of patient on the day of surgery (Years)
ASA class	D,O,T	ASA class of patient (I-V)
BMI	D,O,T	BMI of patient at the time of surgery.
Behavioural		
Alcohol use	O	Alcohol use of patient at the time of surgery (Current/Never/Past).
Smoking	D,O	Smoking status of patient at the time of surgery (Current/Never/Past).
Comorbidities		
Heart disease	O,T	Patient has a history of heart disease at the time of surgery (Yes/No).
Liver disease	D	Patient has a history of liver disease at the time of surgery (Yes/No).
Hypertension	O	Patient has a history of hypertension (Yes/No).
Diabetes	D,O,T	Patient has diabetes Type I or II at the time of surgery (Yes/No).
Measurement		
Temperature	D	Highest temperature of patient in the past 7 days before surgery.
CRP	O	Highest CRP of patient in the 7 days before surgery.
Leukocyte	D	Highest leukocyte level of patient in the 7 days before surgery.
Serum total protein	D	Highest serum total protein of patient in the 7 days before surgery.
Glucose	D	Highest glucose level of patient in the 7 days before surgery.
Haemoglobin	D	Highest haemoglobin level of patient in the 7 days before surgery.
Operative		
Preoperative length of stay	D,O,T	Preoperative length of hospital stay of patient at the time of surgery (Days).

Antibiotic use	T	Antibiotic (WHO ATC code J01) use of patient at the time of surgery (Yes/No).
Duration of surgery	D,O	Duration of the surgical procedure (Minutes).

D=Digestive system surgical procedures; O=Orthopaedic system surgical procedures; T=Thoracic system surgical procedures; ASA=American Society of Anaesthesiologists; CRP=C-reactive protein; BMI=Body Mass Index; SSI=Surgical Site Infection; ATC=Anatomical Therapeutic Chemical; WHO=World Health Organization.

The significant univariate results of digestive system, orthopaedic and thoracic surgical procedures are shown in Table 2. Antibiotic use, CRP and temperature were added to the list of general risk factors after being found statistically significant in the univariate analysis – increasing the number of general risk factors to 8. Diabetes was identified as a general risk factor from our literature search but was not found significant in any of the three univariate analyses in our own study. For digestive system surgical procedure and thoracic procedures, the data-driven cut-off for age was obtained as 23 years and both the standard cut-off (18 years) and the data-driven cut-off were statistically significant with p-values of less than 0.001 which resulted in rejecting the null hypothesis that the coefficient associated with the age of the patient is zero. For orthopaedic procedures, the data-driven cut-off for the temperature (39 degrees) was found statistically significant, but the standard medical cut-off not. A data-driven CRP cut-off of 8.1 was identified for orthopaedic surgical procedures as opposed to a standard medical CRP cut-off of 10; both cut-offs are statistically significant.

Table 3-2: Digestive system surgical procedures: univariate analysis of risk factors for the future occurrence of SSI.

Variable		SSI = No (2 600)	SSI = Yes (124)	Univariate OR (95%CI, P-value)
Digestive System Surgical Procedures				
Gender	Female	359 (43.9)[2]	24 (33.8)	Reference
	Male	458 (56.1)	47 (66.2)	1.54 (0.93-2.60, p=0.099)
Age[1]	≤18	246 (30.1)	8 (11.3)	Reference
	>18	571 (69.9)	63 (88.7)	3.39 (1.70-7.77, p<0.001)
Age (data-driven)	≤23	258 (31.6)	8 (11.3)	Reference

		559		3.63 (1.82-8.32,
	>23	(68.4)	63 (88.7)	p<0.001)
		496		
Antibiotic use	No	(60.7)	17 (23.9)	Reference
		321		4.91 (2.85-8.86,
	Yes	(39.3)	54 (76.1)	p<0.001)
Temperature[1]	≤36.5	0 (0.0)	0 (0.0)	NA
		98		
	(36.5,37.5]	(12.0)	2 (2.8)	Reference
		719		4.70 (1.44-28.91,
	>37.5	(88.0)	69 (97.2)	p=0.033)
Temperature (data-driven)		535		
	≤38	(65.5)	20 (28.2)	Reference
		187		3.58 (1.95-6.66,
	(38,39]	(22.9)	25 (35.2)	p<0.001)
		95		7.32 (3.94-13.79,
	>39	(11.6)	26 (36.6)	p<0.001)
		397		
CRP[1]	≤10	(48.6)	21 (29.6)	Reference
		420		2.25 (1.35-3.89,
	>10	(51.4)	50 (70.4)	p=0.003)
		365		
CRP (data-driven)	≤8.1	(44.7)	18 (25.4)	Reference
		452		2.38 (1.39-4.24,
	>8.1	(55.3)	53 (74.6)	p=0.002)
Preoperative length of stay (Days)	Mean Days (SD)	6.6 (24.1)	12.1 (37.3)	1.01 (1.00-1.01, p=0.092)
Duration of surgery	Mean Minutes (SD)	243.6 (143)	330.4 (190.8)	1.00 (1.00-1.01, p<0.001)
Orthopaedic Procedures				
		196		
ASA class	ASA CLASS I	(26.8)	6 (33.3)	
		339		0.58 (0.18-1.87,
	ASA CLASS II	(46.4)	6 (33.3)	p=0.348)
		182		0.72 (0.18-2.55,
	ASA CLASS III	(24.9)	4 (22.2)	p=0.612)
	ASA CLASS ≥ IV	13 (1.8)	2 (11.1)	5.03 (0.69-24.47, p=0.062)
		327		
Alcohol use	Current	(44.8)	6 (33.3)	Reference

	Never	339 (46.4)	8 (44.4)	1.29 (0.44-3.94, p=0.645)
	Past	64 (8.8)	4 (22.2)	3.41 (0.85-12.26, p=0.063)
Antibiotic use	No	591 (81.0)	8 (44.4)	Reference
	Yes	139 (19.0)	10 (55.6)	5.31 (2.06-14.16, p<0.001)
Temperature (data-driven)	≤39	695 (95.2)	14 (77.8)	Reference
	>39	35 (4.8)	4 (22.2)	5.67 (1.55-16.79, p=0.003)
Thoracic Surgery				
Age[1]	≤18	232 (22.0)	16 (45.7)	Reference
	>18	821 (78.0)	19 (54.3)	0.34 (0.17-0.67, p=0.002)
Age (data-driven)	≤23	226 (21.5)	16 (45.7)	Reference
	>23	827 (78.5)	19 (54.3)	0.32 (0.16-0.65, p=0.001)
BMI	Mean (SD)	24.5 (5.3)	22.1 (4.2)	0.91 (0.85-0.98, p=0.010)
Alcohol use	Current	534 (50.7)	11 (31.4)	Reference
	Never	422 (40.1)	18 (51.4)	2.07 (0.98-4.57, p=0.061)
	Past	97 (9.2)	6 (17.1)	3.00 (1.01-8.09, p=0.034)
Antibiotic use	No	705 (67.0)	18 (51.4)	Reference
	Yes	348 (33.0)	17 (48.6)	1.91 (0.97-3.77, p=0.060)
Temperature[1]	≤36.5	0 (0.0)	0 (0.0)	NA
	(36.5,37.5]	302 (28.7)	3 (8.6)	Reference
	>37.5	751 (71.3)	32 (91.4)	4.29 (1.52-17.94, p=0.017)
Temperature (data-driven)	≤38	882 (83.8)	20 (57.1)	Reference

	>38	171 (16.2)	15 (42.9)	3.87 (1.91-7.67, p<0.001)
CRP[1]	≤10	684 (65.0)	17 (48.6)	Reference
	>10	369 (35.0)	18 (51.4)	1.96 (1.00-3.88, p=0.050)
Haemoglobin[1]	≤8.6	665 (63.2)	21 (60.0)	Reference
	(8.6,10.5]	358 (34.0)	11 (31.4)	0.97 (0.45-2.00, p=0.942)
	>10.5	30 (2.8)	3 (8.6)	3.17 (0.72-9.85, p=0.074)

CRP=C-reactive protein; OR=Odds Ratio; BMI=Body Mass Index; NA=Not Applicable; CI=Confidence Interval; SSI=Surgical Site Infection; Data-driven=cut-off values determined using recursive partitioning.

[1]Standard Erasmus MC clinical cut-offs.

[2]The percentage distribution of the SSI outcome is provided in brackets next to the frequency for each variable.

The multivariate results using standard medical cut-offs and data-driven cut-offs are shown in Table 3 and Table 4, respectively. The temperature variable was statistically significant in the multivariate analysis using the data-driven cut-offs for all three groups of surgeries, but not in one of the multivariate analyses using the medical standard cut-offs. The duration of the surgery was the only statistically significant variable in the multivariate analyses which was not identified as a general risk factor to increase the odds of SSI by approximately 6% for every 30 minutes spent in surgery. For digestive surgical procedures, the addition of duration of surgery to the multivariate model increased the Gini coefficient from 0.46 to 0.52 based on standard medical cut-offs and from 0.57 to 0.62 for the multivariate model based on the data-driven cut-offs. This increase translates into a 12.5% and 8.8% increase in the Gini coefficient, respectively. Neither the orthopaedic nor the thoracic group of surgical procedures had any statistically significant risk factors which are not part of the general risk factors group of surgeries. The Gini coefficient of the data-driven multivariate model is 19.5% (0.62 vs 0.52) higher than the multivariate model based on the standard medical cut-offs. The 5-fold cross-validated 95% confidence intervals for the Gini coefficients based on the validation samples of the data-driven models are (0.49, 0.72) for digestive procedures, (0.21, 0.86) for orthopaedic procedures and (0.21,0.70) for thoracic procedures.

Table 3-3: Multivariate analysis risk factors for the occurrence of SSI by group of surgeries using standard medical cut-offs.

Risk factor by surgery group[1]	Coefficient	Multivariate OR (95%CI)	P-value
Digestive System Surgical Procedures			
Antibiotic use	1.240	3.455 (1.951-6.384)	<0.001
Duration of surgery (Minutes)	0.003	1.003 (1.001-1.004)	<0.001
CRP >10	0.803	2.232 (1.302-3.951)	0.004
Orthopaedic Surgical Procedures			
Antibiotic use	1.670	5.315 (2.059-14.158)	<0.001
Thoracic Surgical Procedures			
Age >18	-4.195	0.146 (0.058-0.351)	<0.001
Antibiotic use	1.311	4.849 (2.035-12.266)	<0.001

CRP=C-reactive protein; CI=Confidence Interval; OR=Odds ratio.

[1]The multivariate analysis was performed using Erasmus MC clinical cut-offs.

Table 3-4: Multivariate analysis risk factors for the occurrence of SSI by group of surgeries using data-driven cut-offs.

Risk factor by surgery group[1]	Coefficient	Multivariate OR (95%CI)	P-value
Digestive System Surgical Procedures			
Temperature (38,39]	1.067	2.907 (1.556-5.497)	<0.001
Temperature >39	1.732	5.650 (2.952-10.947)	<0.001
Antibiotic use	1.201	3.322 (1.856-6.200)	<0.001
Duration of surgery (Minutes)	0.002	1.002 (1.001-1.004)	0.003
CRP >8.1	0.639	1.894 (1.062-3.510)	0.035
Orthopaedic Surgical Procedures			
Antibiotic use	1.552	3.665 (1.370-10.006)	0.009

Temperature >39	1.224	5.120 (1.316-16.387)	0.009
Thoracic Surgical Procedures			
Age >17	-1.847	0.158 (0.055-0.426)	<0.001
Antibiotic use	1.597	4.939 (1.896-14.043)	0.002
Temperature >38	0.824	2.280 (1.098-4.653)	0.024

Data-driven=cut-off values determined using recursive partitioning; CRP=C-reactive protein; CI=Confidence Interval; OR=Odds ratio.

[1]The multivariate analysis was performed using data-driven cut-offs.

An overview of the study results (Table 5) shows that 10 of the 19 risk factors, identified during the literature search, were not statistically significant in the univariate or multivariate analysis for any of the surgery groups. BMI and diabetes were identified across all three groups of surgeries and multiple studies as risk factors for SSI but were not statistically significant in this study. Temperature and the duration of the surgery were confirmed as risk factors for digestive system surgeries, and similarly, antibiotic use and age were confirmed as risk factors for thoracic surgeries. Antibiotic use and CRP were identified as risk factors for digestive surgeries from the multivariate analysis, which were identified during the literature search for thoracic and orthopaedic surgeries, respectively. Antibiotic use and temperature were statistically significant for all three groups of surgeries and were included because of two studies regarding thoracic and digestive system surgeries, respectively [114,146].

Table 3-5: Statistical significance of risk factors and the source which lead them to be considered by surgical procedure.

Risk Factor	Significance[1]	Digestive System[2]	Orthopaedic[2]	Thoracic[2]
Age	D_U,T_M	[111,112,134,138]	[119]	[115]
Alcohol use	O_U,T_U		[142]	
Antibiotic use	- D_M,O_M,T_M			[114]
ASA Class	O_U	[110,113,132,134,145]	[119,142,144]	[119]
BMI	None	[135]	[142–144]	[133]
CRP	D_M		[119]	
Diabetes	None	[111,138,141]	[119,136,142,144]	[116]

Duration of surgery	D_M	[108,111,132,134,135, 140,145]	[119,136,142 ,144]	
Gender	D_U	[111,112,134]	[119,142]	
Glucose	None	[138]		
Haemoglobin	None	[112,135,145]		
Heart Disease	None		[142]	[115]
Hypertension	None		[142]	
Leukocyte	None	[146]		
Liver disease	None	[145]		
Preoperative length of stay	D_U	[132,141]	[119,143]	[114–116]
Serum total protein	None	[108,140]		
Smoking	None	[140]	[142–144]	
Temperature	- D_M,O_M,T_M	[146]		

D=Digestive system surgical procedures; O=Orthopaedic system surgical procedures; $_U$=Significant in univariate analysis; $_M$=Significant in multivariate analysis; T=Thoracic system surgical procedures; ASA=American Society of Anaesthesiologists; CRP=C-reactive protein; SSI=Surgical Site Infection; BMI=Body Mass Index.

[1]During which part of the analysis the risk factor was found statistically significant.

[2]References to the literature which had the risk factor as a multivariate result for each group of surgeries.

3.4 Discussion

We identified temperature and antibiotics used at the time of surgery as risk factors for digestive, orthopaedic and thoracic system surgical procedures in this study. The duration of the surgery was identified as a risk factor for digestive surgical procedures. Being an adult (age ≥ 18) was identified as a protective effect for thoracic surgical procedures. Data-driven cut-offs were identified for temperature, CRP and age, which differ from the standard medical cut-offs. Temperature would not have been identified as a risk factor if only standard medical cut-offs were considered. From our literature search, we identified age, ASA class, BMI, preoperative length of stay and diabetes as general risk factors, while CRP, temperature and antibiotic use were identified as general risk factors because of this study.

The identified risk factors may be classified as modifiable or non-modifiable, depending upon the circumstances of the patient like the complexity of his condition. For instance, the temperature of a patient may be high because of an existing infection, which is why the surgery is needed in the first place and may not be

modifiable before surgery. Age, on the other hand, may be a modifiable risk factor if the surgery can be postponed for several years, e.g., due to a heart defect. This study revealed that children are more likely to be diagnosed with an SSI after thoracic surgery than adults. There are studies which identify risk factors for children after thoracic surgeries, but none found that being a child is a risk factor for SSI after undergoing thoracic surgery [133,139]. We segmented the thoracic surgeries between adults and children and obtained multivariate results for children and adults separately. The multivariate model based only on children (age $\leq$ 18) did not reveal any significant results, contrary to the results of the thoracic study which found age to be a risk factor for children [115]. This absence could be partly due to the small study population size of 248. Antibiotic usage was the only significant factor in the multivariate analysis of thoracic surgeries based on adults. The other two groups of surgical procedures were consistent in terms of their statistical significance of risk factors based on adults.

The data-driven cut-offs confirmed the existing standard medical cut-offs. On average the clinical cut-off for temperature was one degree Celsius lower, while for digestive system surgical procedures, the clinical cut-off for CRP (10) was just less than two units more than the data-driven cut-off of 8.1. This means that there is a greater difference between the occurrence of SSI for patients with a CRP below and above 8.1 than below and above 10. The data-driven cut-offs improved the ability of the statistical model to explain the occurrence of SSI. The performance of the digestive system surgical procedure prediction model increased by 19.5% due to using data-driven cut-offs rather than the standard medical cut-offs. Using data-driven cut-offs, we were able to identify temperature as a risk factor for all three groups of surgical procedures. If standard clinical cut-offs were used, temperature would not have been significant from the multivariate analysis. This potential oversight illustrates the importance of evaluating the cut-offs used for continuous variables against the data before identifying risk factors.

Antibiotic use, temperature and CRP were added to the list of general risk factors by incorporating the statistically significant results of the univariate analysis. These risk factors might have been overlooked when the focus was on only one type of surgery. Temperature was identified as a risk factor in the multivariate results for all three groups of surgical procedures, whereas the literature search identified it only for digestive surgeries. Antibiotic use was not found during our literature search for digestive or orthopaedic surgical procedures but was found significant for both groups of surgeries in the multivariate analysis of our study.

The Centres for Disease Control and Prevention (CDC), the European centre for disease prevention and control (ECDC), World Health Organisation (WHO) and

Netherlands National Institute for Public Health and the Environment (RIVM) suggest maintaining normothermia intraoperatively to prevent undesirable hypothermia (during some thoracic and neurosurgeries, hypothermia may be desirable). [147–149] A lower intraoperative bound for temperature of 35.5 °C to 36 °C is explicitly mentioned, and only the RIVM mention an upper bound of 38 °C which is consistent with the risk factors identified in our study. An upper limit for preoperative temperature should, therefore, be investigated instead of only the lower limit. The four health organisations refer to the proper administration and timing of surgical antimicrobial prophylaxis, but not to the proper preoperative use of standard prescription antibiotics. The proper preoperative use of antibiotics should be well defined, and the reason why antibiotic-use was identified as a risk factor for SSI should be further investigated.

3.4.1 Limitations

This is a retrospective, single-centre study, and therefore the data were not collected for the purpose of this study. Even though cross-validation was performed to estimate model performance on new data, the models were not externally validated. Surgeries were aggregated into three broad groups of surgical procedures which serve as a proxy for the reason for surgery but leads to the loss of information regarding possible comorbidities. Some measurements, like temperature and CRP, were not always present and was partly overcome using imputation. Patient information concerning smoking and drinking habits may be understated due to incomplete medical records. The literature search used for this study was not exhaustive but rather based on the principal on data saturation. We used a 30-day outcome period in which we observe if an SSI was present or not, but according to the CDC definition, this outcome period should be one year for surgical implantation procedures. Since our data only spans over 18 months, it was not possible to use a 12-month outcome window for all surgical implantation procedures, which is a limitation of this study. The administration of prophylaxis and the optimal timing thereof is an important risk factor for the occurrence of SSI. However, these data were not available. The definition of antimicrobials was limited to the J01 class of the Anatomical Therapeutic Chemical (ATC) classification system, which corresponds to anti-infectives for systemic use. The occurrence of SSI varies for different times between antibiotics prescription and surgery (Figure 3-1), but hypotheses regarding this relationship should be evaluated using case-control studies with specific data regarding the antimicrobials and the timing of the administration thereof.

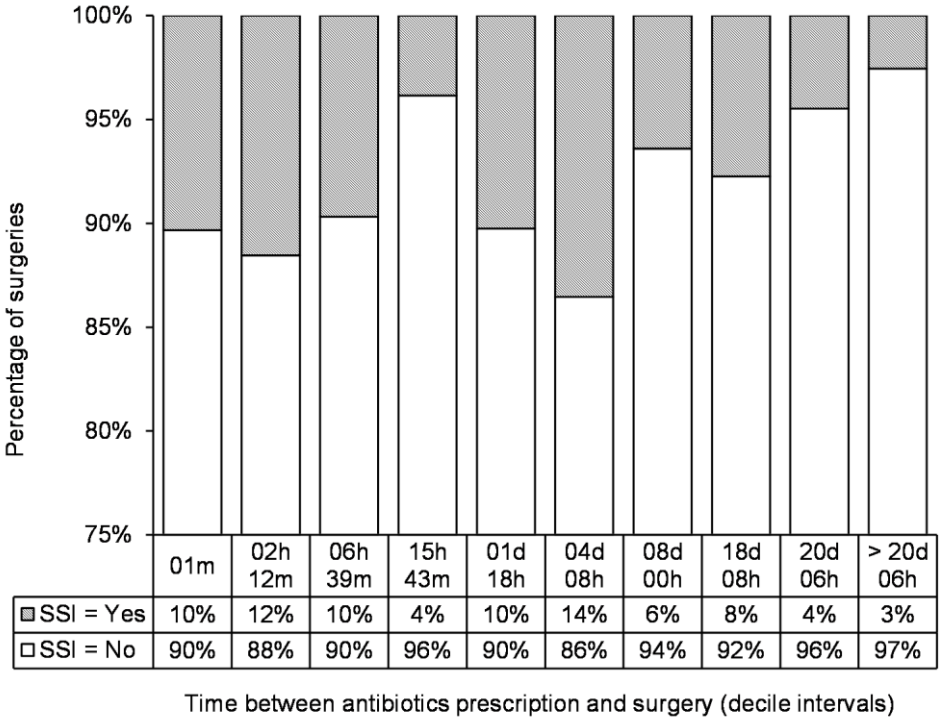

Figure 3-1**:** The occurrence of SSI for the time between the J01 antibiotics prescription start time and the start time of the surgery by decile.

m=minutes; h=hours; d=days.

[1]The vertical axis starts at 75% to increase visibility.

[2]The horizontal axis stipulates the endpoints of the respective deciles of the distribution of the time between prescription and surgery.

3.4.2 Future work

Future work will investigate the modifiability of the risk factors identified in this study in more detail, as the circumstances under which this occurs are hitherto unclear. The exact purpose of the use of antibiotics over the time of surgery was not investigated in depth, which can be done in future studies. Future research can also investigate differences between adults and children, which lead to the occurrence of SSI among children. Another opportunity for future research is to investigate which risk factors are predictive for the occurrence of SSI over different periods. Doing this will enable healthcare workers to identify which risk factors explain the occurrence of SSI soon after surgery, towards the end of the 30 days and even later for

implantation surgeries. These insights can help set guidelines to determine the vigilance necessary to mitigate the risk of SSI on a patient level.

3.5 Conclusion

This study shows that data-driven cut-offs can be used to identify risk factors that would not have been identified by only using standard medical cut-offs. Preoperative temperature and antibiotic use were identified as risk factors for digestive, orthopaedic, thoracic system surgeries, while the duration of surgery and age were identified as risk factors for orthopaedic and thoracic system surgeries, respectively. In contrast with literature, this study found that an SSI is more likely to occur in children (age < 18) than in adults after thoracic system surgeries. Statistical modelling has been important to quantify important risk factors and indicate their significance. Clinical studies using retrospective data are important to carry out, despite limitations in the data sets. To this end, future studies should use both standard medical cut-offs and data-driven cut-offs to investigate risk factors.

4. A spatiotemporal simulation study on the transmission of harmful microorganisms through connected healthcare workers in a hospital ward setting

Abstract

Hand transmission of harmful microorganisms (HMO) poses a major threat to patients and healthcare workers in healthcare settings. The most effective countermeasure against these transmissions is the adherence to spatiotemporal hand hygiene policies, but adherence rates are relatively low and vary over space and time. The third objective aimed to identify a healthcare worker occupation group of potential super-spreaders and quantify spatiotemporal effects on the hand transmission of HMO for varying levels of hand hygiene compliance caused by this group. Spatiotemporal data were collected in a hospital ward of a tertiary hospital using radio frequency identification technology. The effects of five probability distributions of HHC and three harmful microorganism transmission rates were simulated using a dynamic agent-based simulation model. The effects of initial simulation assumptions on the simulation results were quantified using five risk outcomes. Nurses were identified as the potential super-spreader healthcare worker occupation group. During lack of HHC (5%) and high transmission rates (5% per contact moment), a colonised nurse can transfer microbes to three of the 17 healthcare workers or patients encountered during the 98.4 minutes of visiting 23 rooms while colonised. The HMO transmission potential for nurses is higher during weeknights (5 pm – 7 am) and weekends as compared to weekdays (7 am – 5 pm). Spatiotemporal behaviour and social mixing patterns of healthcare can change the expected number of hand transmissions and spread HMO by super-spreaders in a closed healthcare setting. These insights can be used to evaluate spatiotemporal safety behaviours and develop infection prevention and control strategies.

This chapter was published as Van Niekerk JM, Stein A, Doting MH, Lokate M, Braakman-Jansen LM, van Gemert-Pijnen JE. A spatiotemporal simulation study on the transmission of harmful microorganisms through connected healthcare workers in a hospital ward setting. BMC infectious diseases. 2021 Dec;21(1):1-4.

4.1 Background

The majority of healthcare-associated infections are caused by direct or indirect transmission of *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa* or *Enterobacter spp.* (ESKAPE) [150]. These harmful microorganisms (HMO) can survive on human skin and hospital surfaces for extended periods and lead to high cross-transmission rates between healthcare workers (HCW) and patients [151–153]. The ease of transmission of HMO depends upon the features of the microorganism, patient characteristics and the behaviour of healthcare workers (HCW), whereas the damage caused by the infection that follows ranges from none to potentially fatal [154].

The most effective precautionary measure to combat hand transmission and spread of harmful microorganisms in closed healthcare settings is the adherence to well established and effective hand hygiene policies also known as hand hygiene compliance (HHC) [20]. Unfortunately, HHC is often unsatisfactory with highly variable levels within and between hospitals. Rates of hand hygiene compliance range from 5% to 81%, with average compliance of approximately 40% [44]. With a level of 80% adherence seen as high levels of HHC and 95% as very high, it is not surprising that the spread of HMO in closed healthcare settings remains a major dilemma [44]. Reasons for hand hygiene non-compliance include increased work intensity, lack of education and ineffective placement or defective alcohol dispensers. For instance, one hour of overtime worked by an HCW can lead to a 3% decrease in the level of HHC [155,156]. The result is a highly variable level of HHC within closed healthcare settings. Compounding the non-adherence to hand hygiene policies is that the medium and method used for hand hygiene are not 100% efficient. The efficacy of hand rubbing using alcoholic rub was compared with handwashing using antibacterial soap during routine patient care [157]. Some of the patients had methicillin-resistant Staphylococcus aureus (MRSA). The study estimated an efficacy rate of 83% (interquartile range 78% - 92%) for alcoholic rub compared to 58% (interquartile range -58% - 74%) for antibacterial soap. Even though alcoholic rub significantly outperforms antibacterial soap, some HMO may remain on the hands of the HCWs and lead to further transmissions.

The combination of colonised and uncolonized HCWs or patients, who are potentially immunocompromised and in a confined space, makes healthcare facilities a high-risk environment for the spread of HMO. The term super-spreader is used to categorise an individual with a disproportionately high potential to spread HMO. Super-spreaders were the cause of several super-spreading events (SSE) in the past with devastating consequences [158]. Highly connected HCWs can increase the risk of SSE in closed healthcare environments. The amount of contact between HCWs and

patients and HHC while performing regular duties are critical factors that contribute to the extent and severity of an SSE [159].

For these reasons, the SSE is affected by the joint spatiotemporal behaviour, i.e., the where and when, by the social mixing patterns, i.e., with whom of the HCW or patients inside a hospital ward, and by the level of HHC, including its variability. Therefore, it is necessary to understand the spatiotemporal effect on the hand transmission and spread of HMO for varying levels of HHC for potential super-spreaders in a closed healthcare setting.

Healthcare institutes are now adopting automatic contact tracking methods like Radio Frequency Identification (RFID) technology by tagging healthcare equipment, HCWs and patients to improve logistics and patient safety. There is still a reluctance to fully adopt this technology, mainly driven by security and privacy concerns [27]. Real contact data between patient and HCWs became more prevalent since 2002 when data were collected using shadowing. Medical records, surveys and sensors became more important for data and contact detection. Assab et al. (2017) showed that studies using empirical contact data within closed healthcare settings led to a better understanding of the transmission and spread of HMO. Such data can result in the development of improved control interventions. Using real-time RFID tracking data, it is possible to model the spread of HMO at an individual level rather than using a compartmental-based model [154,160]. RFID data have been used to model the spread of HMO in different closed healthcare settings and at different proximities using a temporal proximity network at schools [161,162], conferences [163], households [164], hospitals [165–171] and other healthcare facilities [172]. In addition to recent research data collection and modelling innovations are needed to implement better control strategies [173]. Studies based upon contact data only are unable to determine the effect of spatiotemporal healthcare policies like HHC. A few hours of RFID tracking can be sufficient to develop and calibrate a statistical model that shows the heterogeneity of spatiotemporal social contact patterns, representing how people socially interact in space and time [174].

The spatiotemporal effects of varying levels HHC on the transmission and spread of hand-transmitted HMO in a closed healthcare setting must still be quantified, based upon empirical spatiotemporal tracking data. HCWs and policymakers may benefit from understanding the impact of spatiotemporal infection control interventions and healthcare policies on the transmission and spread of HMO.

This study's objectives were to (1) identify an HCW occupancy group of potential super-spreaders and (2) quantify the spatiotemporal effects on the transmission and spread of HMO for varying levels of hand hygiene compliance caused by this group.

4.2 Methods

We used spatiotemporal data from the University Medical Center Groningen (UMCG), one of the largest hospitals in the Netherlands with more than 10 000 employees and almost 1 400 beds. Between 2 April 2018 and 8 April 2018, data were collected in a 32-bed general hospital ward, for stomach, gut and liver patients (Figure 4-1: A). The dates were chosen such that they cover a full calendar week from Monday to Sunday and all shifts to increase the representativeness of the parameter estimates. The ward's floor plan was divided into 33 rooms of which 14 were patient rooms, with between one and four beds, eight storage areas and ten other rooms, including a doctor's office and a medicine room. All facilities are located in the centre of the ward to minimise distances to crucial parts of the ward, including the rinsing kitchen and medication rooms. Single patient rooms are near the entrance of the ward to enable easy isolation of potentially infectious patients.

Data were collected using RFID sensors worn by the HCWs working in the ward during the study period. Seven HCW occupation groups were identified, namely *doctor*, *nurse*, *cleaner*, *department assistant*, *department co-assistant*, *consultant* and *feeding assistant*. The RFID tags were assigned to specific occupation groups. HCWs randomly selected an RFID tag at the start of their shift according to their occupation. The RFID tags (Figure 4-1: B) emit radio signals with unique identifying information and RFID readers (Figure 4-1: C), on the ceiling of the rooms, register those signals. The RFID reader's range was set to the size of the rooms and they continuously monitored the uniquely identifiable RFID tags in their range. HCWs moving in and out of the rooms were registered and the data were generated and stored. The data consist of a room ID, an RFID sensor ID and a DateTime stamp corresponding to the RFID tag movement into and out of a room. The spatial resolution is at the room level and is defined by the set of rooms inside the ward. The temporal resolution equals the second at which the observation signal was received.

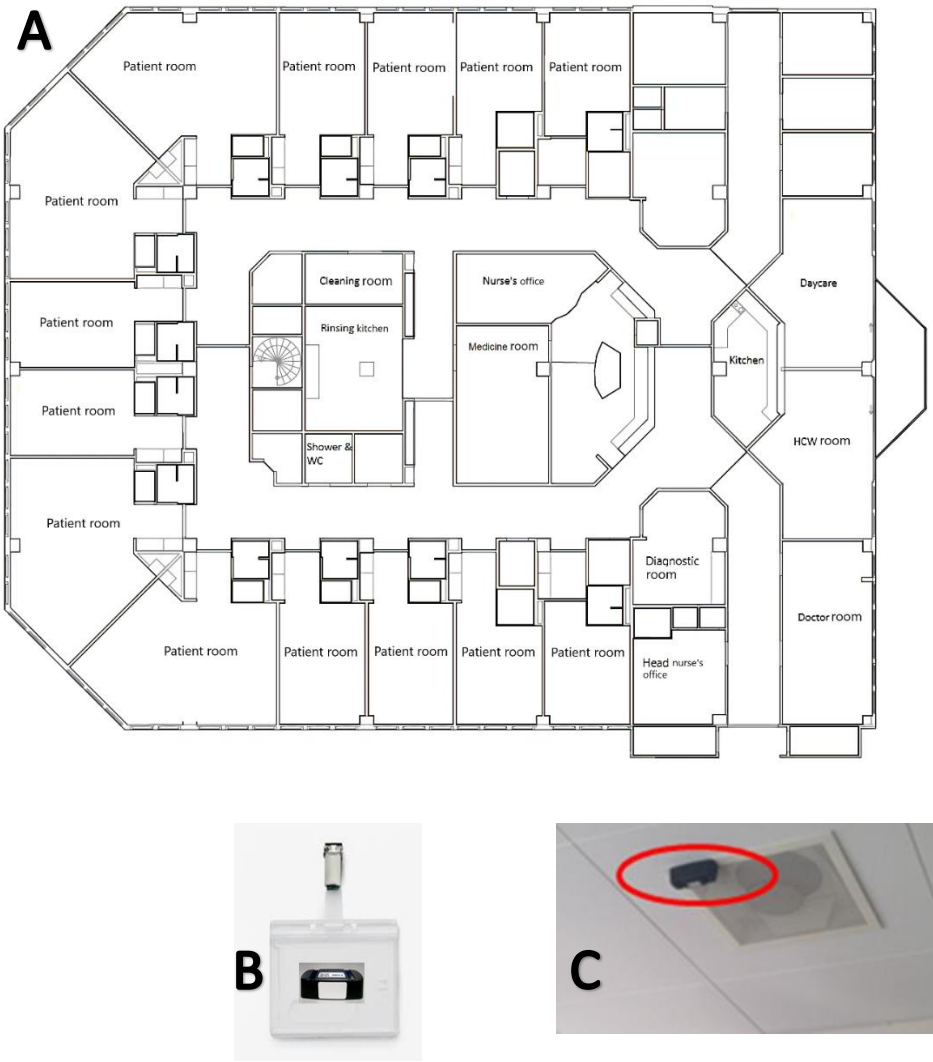

Figure 4-1: Floorplan of the 32-bed general hospital ward for stomach, gut and liver patients. **A** = Floorplan of the 32-bed general hospital ward, for stomach, gut and liver patients where sample data were collected using **B** = RFID tags worn by HCWs during data collection using **C** = RFID readers placed on the ceiling of the rooms inside the ward.

The sampled data were divided into two subsets. Data in subset 1 contained the sampled data collected during weekdays (7 am – 5 pm) and data in subset 2 those collected during the evenings (5 pm – 7 am) and over the weekend.

Table 4-1: Example of data collected using the RFID sensors and readers.

Room	From	To	Sensor
54.3.35A	03/04/2018 16:49	03/04/2018 16:50	58007
54.3.45	03/04/2018 16:51	03/04/2018 17:00	58007
54.3.14	03/04/2018 17:00	03/04/2018 17:37	58007
54.3.17	03/04/2018 17:37	03/04/2018 17:41	58007

Contact data were extracted from the empirical spatiotemporal data. They are generated by the underlying contact network between HCWs and determines the possible pathways over which the spread of HMO occurs over space and time [175,176]. Since the spatial resolution of the collected data was at the room level and not at the face-to-face level, an assumption was needed for the contact definition. Depending upon the data collection context, co-occurrence data can serve as a proxy for face-to-face contact data [177]. In this study, co-occurrence can occur inside the limited space defined by the ward rooms, increasing the probability of HCWs to enter the face-to-face close-range proximity (1.5 m) of other HCW or patients. For this reason, we define a contact as the physical co-occurrence of two HCWs or a HCW and a patient in a ward room. For example, if an HCW enters a patient room, then the HCW and the patient are assumed to be in contact with each other for the time over which they co-occur in that room.

A guideline to identify super-spreaders is to identify the 20% of the people contributing to at least 80% of the transmission potential [178]. We define the transmission potential as the number of 30-second intervals (*contact moment*) of contact with other HCWs or patients. The transmission potential is estimated for all HCW occupation groups and compared to identify disproportionality and thus potential super-spreaders.

We estimated the effect of the transmission and spread of an HMO by a colonised HCW from the potential super-spreading occupation group for varying levels of HHC defining five risk outcomes. The risk outcomes are defined in five variables: the amount of time (minutes) spent colonised (RO1), the amount of time (minutes) spent with HCWs or patients (RO2), the number of HCWs or patients encountered (RO3), the number of transitions made from one room to another (RO4) and the expected number of HMO transmissions to other HCWs or patients before successfully performing hand hygiene (RO5).

To estimate RO1 – RO5, we constructed a dynamic agent-based transition simulation model [179]. To simulate the underlying distribution from the sampled data, we first

estimated this distribution, followed by resampling to generate more samples. This simulation process aids us to explore the consequences of initial simulation assumptions. The simulation follows a four-part (A-D) workflow (Figure 4-2). We assumed that RO1 – RO5 depend upon the order in which an HCW moves between the different rooms in the hospital ward (A), the likelihood of the HCW performing hand hygiene and the efficacy of doing so (B), the amount of time an HCW spends in each room with other HCWs or patients (C) and the transmission dynamics of the HMO (D). Parts A and C are based on statistics from the sampled data, while parts B and D are based on assumptions from the literature.

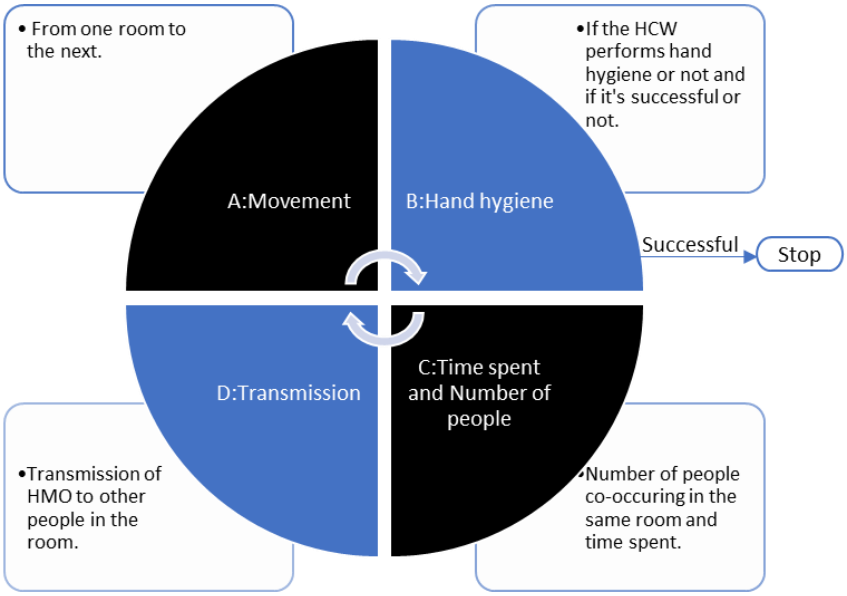

Figure 4-2: The four-part of the simulation workflow. A and C depend upon the sampled data and B and D on initial assumptions from literature. The simulation ends when the HCW successfully performs hand hygiene in part B.

For part A in the simulation workflow, we used continuous Markov chains. They allowed us to model the movement of HCWs from one of the n rooms to the next [180]. If R is the set of n rooms, i.e. $R = \{R_1, R_2, \dots, R_n\}$, then the transition probability p_{ij} (Formula 4.1) in row i and column j of the $n \times n$ transition probability matrix $\boldsymbol{P}$ (Formula 4.2) is the probability that an HCW will transit from room R_i to room R_j during the next transition. Since an HCW will either stay in the same room or move to another room after the next transition, the rows of the matrix $\boldsymbol{P}$ add up to 1 i.e., $\sum_{i=1}^{n} p_{ij} = 1$ for $i = 1, \dots, n$. Each element p_{ij} is between 0 and 1 inclusively, i.e., $0 \leq p_{ij} \leq 1$ for $i, j \in (1, \dots, n)$. An estimate for p_{ij} is obtained by dividing the number of transitions from R_i to R_j by the total number of transitions from R_i. Using

only the transition data of the potential super-spreading occupation group, we obtain a transition probability matrix $\boldsymbol{P}$.

$$p_{ij} = P\left(\text{Next room} = R_j \middle| \text{Current room} = R_i\right) \text{ for i, j} \in (1, \dots, \text{n}) \qquad \text{Formula 4.1}$$

$$\boldsymbol{P} = \begin{array}{c} \\ R_1 \\ \vdots \\ R_n \end{array} \begin{array}{c} \begin{array}{ccc} R_1 & \dots & R_n \end{array} \\ \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \end{array} \qquad \text{Formula 4.2}$$

We assume that the length of time spent in each room (ψ_{R_i}) is exponentially distributed with parameter η with mean $1/\eta$ and variance $1/\eta^2$ [30]. The estimated values of η and the average number of HCWs or patients co-occurring inside each room, together with the corresponding estimated variance, are obtained at room level from the sampled data. We assumed that the number of HCWs or patients co-occurring within each room follows either a Gaussian distribution or a Poisson distribution with mean and standard deviation equal to the estimates obtained from the sampled data. Since no patient location data were available, we assumed that all patient rooms are occupied.

The performance and efficacy of hand hygiene (B) compliance and the transmission of an HMO (C) are simulated using agent-based modelling and the corresponding model assumptions in Table 4-2, based upon a study by Thomas Hornbeck et al. [158].

Table 4-2: Agent-based model parameters (Thomas Hornbeck et al.).

Symbol	Definition	Range
P	Probability of transmission per 30 s of contact	$0.0005, 0.005$ and 0.05
λ	Hand hygiene efficacy using alcohol rub	0.83
γ	Hand hygiene compliance level	$\mu = 0.05, 0.25, 0.5, 0.75, 0.95$ and $\sigma = 0.1$

The simulation starts with one colonised HCW from the potential super-spreading occupation group in a random room inside the hospital ward. It ends when the HCW successfully performed hand hygiene. One thousand simulations were performed for the three different rates of transmission (P) for each of the five HHC distributions. The result of the simulation consists of 15 (3×5) scenarios with outputs RO1 – RO5.

The simulations were repeated for the subsets 1 and 2, both separately and combined.

We summarise the simulation assumptions made by the workflow section as follows:

Workflow A: Movement

1. Contact definition is based upon HCWs or patients co-occurring in the same room.
2. Patient rooms are always occupied by at least 1 patient , being the reason for the HCW to visit the room.

Workflow B: Hand hygiene

1. A colonised HCW can perform hand hygiene once during every transition between rooms.
2. For a colonised HCW to be decolonized, hand hygiene needs to be performed and it needs to be successful. The former depends upon the action of the HCW with probability γ and the latter on the efficacy of the solution used to perform hand hygiene with probability λ.

Workflow C: Time spent and Number of people

1. The number of minutes an infected HCW spends in a room R_i is given by $\psi_{R_i} \sim Exp(\eta)$, where η is the sample average of ψ_{R_i}.
2. The number of HCWs or patients co-occurring in room R_i with the infected HCW is given by $\omega_{R_i} \sim Poisson(\nu)$, where ν is the average ω_{R_i} from the sampled data.

Workflow D: Transmission

1. Only colonised nurses can transmit an HMO.
2. HCW and patients only have two states: susceptible and colonised.
3. Number of colonised HCW or patients after co-occurring with a colonised HCW for m contact moments with a probability of transmission P for each 30 s of co-occurrence is distributed as $I \sim Bin(m, P)$.

There are two key moments in the model. The first key moment is when the colonised HCW enters a room – that is when an opportunity is given to perform hand hygiene with probability γ, corresponding to part B of the simulation workflow. Five probability distributions are used to simulate HHC for simulation. A Gaussian distribution with mean 0.05 represents very low HHC, while means equal to 0.25 and 0.5 show the effect of low to average HHC and 0.75 to 0.95 for high to near-perfect HHC levels, respectively. Should colonised HCW perform hand hygiene, then the probability that the hand hygiene was successfully performed, meaning that all traces of the HMO were eradicated, equals λ. The probability of successful use of

hand hygiene is based upon Girou et al [157] and the different compliance levels (low, medium and high) are based upon Temime et al [159].

The second key moment is when more than one HCW or patient co-occurs with g in the same room. The colonised HCW has a probability of transferring microbes to all HCWs or patients in the room every 30 s with probability P. The probability of transmitting an HMO from one person to another, results in a probability equal to 1.5% - 13.5% of transmission for every 15 minutes spent together. [159] We assume that transmissions between all HCWs and patients are equally likely for each contact moment in the same room. For example, HCWs or patients in contact with a potential super-spreader will be subject to the probability stated in Formula 4.3 during the first contact moment. The last two terms in Formula 4.1 decrease the probability of transmission because of the chance that the potential super-spreader will effectively perform hand hygiene and not carry the HMO anymore. Only the parameter P remains for subsequent contact moment because the potential super-spreader only performs hand hygiene when entering the room.

$$P[Susceptible \rightarrow Infected | n = 1] = P \times \gamma \times (1 - \lambda) \quad \text{Formula 4.3}$$

For successive contact moments, we assume that the probability that a colonised HCW transfers microbes to an uncolonized HCW or patient follows a binomial distribution with parameter m indicating the number of contact moments and parameter P indicating the probability of transmission. To model this as a binomial distribution, we assume that there are only two outcomes, i.e., colonised and uncolonized, that each contact moment is independent of the other and that the probability of transmission stays constant.

The effect of P is positively correlated with the number of transmissions, meaning that more transmissions should take place if the rate of infection increases. However, model parameters λ and γ have an inverse relationship with the expected number of transmissions. Some simulation parameters are positively correlated, and some are negatively correlated with the expected number of transmissions. Opposite parameter correlations make it possible to create scenarios where the expected number of transmissions is mitigated and even entirely off-set. These scenarios provide further insight into the effects of the initial simulation assumptions propagated through the sampled spatiotemporal data.

4.3 Results

During the seven days of data collection, a total of 2 631 observations were recorded of which 58 had to be removed because of spurious measurements detected using outlier detection and identifying aberrant movement patterns in the collected data.

During the seven days, 2 432 co-occurrences were derived from the 2 573 sampled observations which equate to 504 hours (30 272 minutes) of contact data. Nurses and doctors were together responsible for 81.13% and 80.19% of all contacts and time spent in contact, respectively (Table 4-3). Nurses made up 70.68% and 68.06% of these percentages, five times more than the second higher HCW occupation group, *doctor* (10.44% and 12.13%). Therefore, a colonised nurse has a disproportionately high potential of transmitting and HMO based on the amount of contact and time spent with HCWs or patients. For these reasons, we investigate the *nurse* HCW occupation group as potential super-spreaders in this study.

Table 4-3: The number of contacts and duration of those contacts by occupation group.

(Occupation) Group	Number of Contacts (% of total)	Number of Contact Minutes (% of total)	Average of Contact Minutes (SD)
Cleaner	71 (2.9%)	443 (1.5%)	6.24 (10.27)
Co-assistant	29 (1.2%)	290 (1.0%)	10.00 (12.81)
Consultant	144 (5.9%)	2 230 (7.4%)	15.49 (23.27)
Department assistant	200 (8.2%)	2 953 (9.8%)	14.77 (16.17)
Doctor	254 (10.4%)	3 671 (12.1%)	14.45 (20.59)
Feeding assistant	15 (0.6%)	82 (0.3%)	5.47 (8.26)
Nurse	1 719 (70.7%)	20 603 (68.1%)	11.99 (20.13)
Total	2 432 (100.0%)	30 272 (100.0%)	12.45 (19.80)

Individual percentages may not add up to 100% because they are rounded to the first decimal place.

The estimated transition probability matrix ($\boldsymbol{P}$) for *nurse* summarises the transitions of *nurse* between rooms observed in the sampled data (Figure 4-3). According to $\boldsymbol{P}$, *nurse* is most likely to transit to either a patient room, the medicine room or the nurse's office.

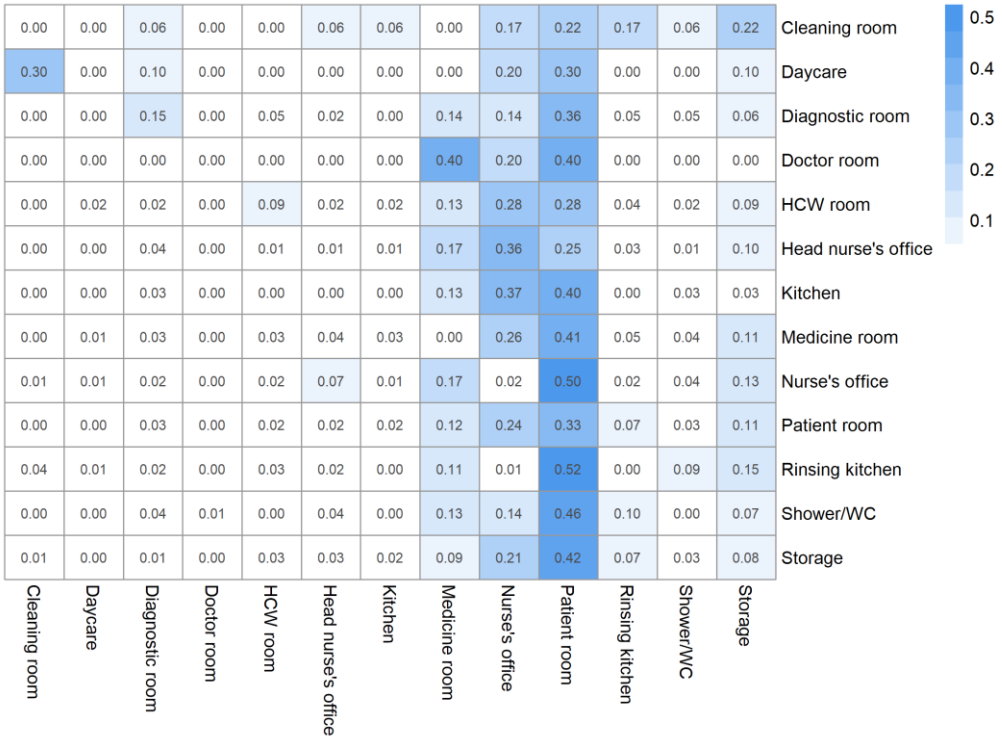

Figure 4-3: Transition probability matrix $\boldsymbol{P}$ for movement of *nurse* between ward rooms. The transmission probabilities are given as p_{ij} in the i^{th} row and j^{th} column for the movement of *nurse* between rooms. Each element is the estimated probability that a nurse will transition from the room i to room j after the next transition.

Nurses spend the most time (19.39 m) and co-occur with the most HCWs (2.16) per visit in the nurse's office (Table 4-4). For this reason, the relatively high estimated probability that a *nurse* will transit to the nurse's office implies that an HCW of the occupation group *nurse* spends a large portion of their time here while co-occurring with a relatively large number of people. Nurses spend less time in patient rooms than the nurse's office (11.5 m vs. 19.39 m), but the average number of people co-occurring is almost the same as in nurse's office (2.09 vs. 2.16). Since we assumed that there is at least one patient in the patient room, the expected number of HCW and patients in contact in patient rooms is more than two by definition.

Table 4-4: Number of HCWs or patients and the time they co-occurred in each room.

Room	Average number of co-occurrences (SD)	Average minutes spent co-occurring (SD)
Cleaning room	1.45 (0.4)	8.05 (12.5)
Daycare	1.10 (0.3)	8.55 (11.1)
Diagnostic room	1.11 (0.6)	27.57 (45.0)
Doctor room	1.00 (0.0)	4.82 (5.2)
HCW room	1.04 (0.3)	18.17 (22.7)
Head nurse's office	1.82 (1.6)	18.64 (33.1)
Kitchen	1.15 (0.5)	9.51 (10.0)
Medicine room	1.16 (0.5)	10.02 (15.6)
Nurse's office	2.16 (1.6)	19.39 (26.7)
Patient room	2.09 (0.4)	11.55 (19.0)
Rinsing kitchen	1.00 (0.1)	0.08 (0.2)
Shower/WC	1.10 (0.4)	10.74 (20.9)
Storage	1.08 (0.4)	7.73 (13.6)

Average number of co-occurrences = time weighted average number of people co-occurring in the room, Average minutes spent co-occurring = average number of minutes spent co-occurring in the room.

4.3.1 Simulation results

The ($P = 0.05$; $\lambda = 0.05$) scenario in Table 4-5 corresponds to the highest probability of transmission (P) and the lowest HHC level (λ). For this scenario, a colonised nurse can transit through 23 wardrooms ($RO4 = 22.03$) for more than one and a half hours ($RO1 = 98.40$) while making contact with 17.41 HCWs or patients ($RO3 = 17.41$), resulting in 83 contacts opportunities to transmit HMO ($RO2 = 83.13$). This scenario also resulted in the highest amount of expected transmissions ($RO5 = 3.36$). Reducing the transmission rate results in an exponential decrease in the number of expected transmissions as expected.

In the ($P = 0.005$; $\lambda = 0.75$) scenario, where the level of HHC is highest and the transmission probability is lowest, the expected time that a colonised NUR would spend carrying an HMO is just more than 9 minutes even though the alcohol rub's effectiveness is 83%. Note that the (P = 0.005; $\lambda = 0.05$) scenario results in a similar amount of expected number of infections as the scenario where $P = 0.05$ and $\lambda = 0.5$ (0.41 vs. 0.51) even though he transmission probability differs by a factor of ten.

Table 4-5: Simulated HMO transmissions potential of a colonised nurse in a hospital ward under various assumed transmission rates and hand hygiene compliance levels.

Transmission probability (P)	HHC (λ)	RO1: Minutes spent colonised (SD)	RO2: Contacts (SD)	RO3: People contacted (SD)	RO4: Room transitions (SD)	RO5: Expected transmissions (SD)
0.05	0.05	98.40 (73.43)	83.13 (70.69)	17.41 (13.65)	23.09 (22.03)	3.36 (2.79)
0.005	0.05	101.60 (76.20)	83.46 (69.22)	17.76 (13.50)	24.03 (22.84)	0.41 (0.34)
0.0005	0.05	98.27 (75.87)	80.56 (69.54)	17.01 (13.59)	22.66 (21.80)	0.04 (0.03)
0.05	0.25	25.24 (26.29)	21.51 (26.53)	4.21 (4.30)	4.66 (4.11)	0.86 (1.01)
0.005	0.25	25.91 (25.68)	22.11 (26.83)	4.36 (4.27)	4.73 (4.08)	0.11 (0.13)
0.0005	0.25	25.87 (26.09)	22.76 (28.74)	4.53 (4.97)	4.78 (4.40)	0.01 (0.01)
0.05	0.5	13.68 (13.29)	12.85 (17.51)	2.42 (2.67)	2.50 (2.00)	0.51 (0.65)
0.005	0.5	13.53 (12.79)	12.68 (17.13)	2.37 (2.38)	2.41 (1.76)	0.06 (0.08)
0.0005	0.5	13.01 (13.73)	13.08 (20.20)	2.37 (2.43)	2.36 (1.81)	0.01 (0.01)
0.05	0.75	9.16 (9.06)	9.43 (14.00)	1.75 (1.79)	1.66 (1.04)	0.36 (0.49)
0.005	0.75	9.12 (8.97)	9.38 (14.08)	1.68 (1.73)	1.64 (1.04)	0.04 (0.07)
0.0005	0.75	9.12 (8.88)	9.39 (14.54)	1.71 (1.73)	1.66 (1.02)	0.00 (0.01)
0.05	0.95	7.34 (6.86)	8.23 (12.66)	1.43 (1.47)	1.32 (0.65)	0.31 (0.46)
0.005	0.95	7.12 (6.71)	8.21 (12.64)	1.47 (1.50)	1.27 (0.62)	0.04 (0.06)
0.0005	0.95	7.26 (6.79)	8.57 (13.51)	1.50 (1.50)	1.29 (0.64)	0.00 (0.01)

Sampled data for three different transmission assumptions and five levels of HHC for one colonised *nurse* starting in a random room in the hospital ward, a hand hygiene efficacy (γ) of 0.83. P = probability of transmission, λ = HHC level, $RO1$ = amount of time spent colonised, $RO2$ = number of contact

moments, $RO3 =$ number of HCWs or patients made contact with, $RO4 =$ number of transitions between hospital ward rooms, $RO5 =$ expected number of HMO transmissions.

The simulation results based upon subset 1 (Table 4-6) show that, for the ($P = 0.05;\ \lambda = 0.05$) scenario, a colonised nurse is expected to spend less time colonised while transiting through the wardrooms during weekdays than during weeknights or weekends (81.46 vs. 114.30 minutes) even though more HCWs or patients are expected to be encountered (19.87 vs. 16.39) by the colonised nurse. Table 4-6 and Table 4-7 show that the difference between the expected number of transitions by a colonised nurse for subset 1 and 2 is less than 10% for all scenarios. The difference in the expected number of transmissions between subset 1 and 2 equals 22.7% for the ($P = 0.05, \lambda = 0.05$) scenario and equals 66.7% for the ($P = 0.005, \lambda = 0.95$) scenario. These differences result from the change of spatiotemporal and social mixing patterns of the HCWs observed during the weekdays and weeknights or weekends.

Table 4-6: Simulated HMO transmissions potential of a colonised nurse in a hospital ward under various assumed transmission rates and hand hygiene compliance levels (between 7am-5pm on weekdays).

Transmission probability	HHC	RO1: Minutes spent colonised (SD)	RO2: Contacts (SD)	RO3: People contacted (SD)	RO4: Room transitions (SD)	RO5: Expected infections (SD)
0.05	0.05	81.46 (58.71)	73.86 (60.06)	19.87 (15.11)	23.18 (21.34)	3.16 (2.52)
0.005	0.05	79.51 (60.02)	72.85 (62.42)	19.41 (15.64)	22.82 (21.77)	0.36 (0.31)
0.0005	0.05	81.64 (59.81)	75.58 (61.32)	20.00 (15.14)	23.34 (21.97)	0.04 (0.03)
0.05	0.25	21.47 (20.41)	20.11 (23.74)	4.87 (5.12)	4.76 (4.24)	0.83 (0.96)
0.005	0.25	21.13 (19.93)	19.87 (23.43)	4.96 (5.11)	4.74 (4.14)	0.10 (0.11)
0.0005	0.25	20.64 (19.69)	18.73 (22.14)	4.65 (5.03)	4.65 (4.30)	0.01 (0.01)
0.05	0.5	11.54 (11.53)	11.00 (14.90)	2.45 (2.68)	2.45 (1.93)	0.44 (0.58)
0.005	0.5	11.39 (10.78)	10.65 (13.22)	2.42 (2.50)	2.42 (1.84)	0.05 (0.06)
0.0005	0.5	11.48 (11.28)	11.60 (15.69)	2.49 (2.60)	2.40 (1.85)	0.01 (0.01)

0.05	0.75	7.67 (7.37)	8.15 (11.75)	1.82 (1.92)	1.61 (1.03)	0.32 (0.45)
0.005	0.75	8.05 (7.62)	8.62 (12.58)	1.83 (1.90)	1.70 (1.10)	0.04 (0.06)
0.0005	0.75	7.44 (7.41)	7.08 (10.21)	1.64 (1.70)	1.61 (0.99)	0.00 (0.01)
0.05	0.9	6.25 (5.60)	7.07 (10.29)	1.45 (1.47)	1.31 (0.66)	0.28 (0.39)
0.005	0.9	6.15 (5.92)	6.84 (10.31)	1.44 (1.46)	1.30 (0.65)	0.03 (0.05)
0.0005	0.9	6.01 (5.76)	6.39 (9.69)	1.44 (1.51)	1.31 (0.68)	0.00 (0.00)

Sampled data for three different transmission assumptions and five levels of HHC for one colonised *nurse* starting in a random room in the hospital ward, a hand hygiene efficacy (γ) of 0.83. $P =$ probability of transmission, $\lambda =$ HHC level, $RO1 =$ amount of time spent colonised, $RO2 =$ number of contact moments, $RO3 =$ number of HCWs or patients made contact with, $RO4 =$ number of transitions between hospital ward rooms, $RO5 =$ expected number of HMO transmissions.

Table 4-7: Simulated HMO transmissions potential of a colonised nurse in a hospital ward under various assumed transmission rates and hand hygiene compliance levels (between 6pm and 6am or on weekends).

Transmission probability	HHC	RO1: Minutes spent colonised (SD)	RO2: Contacts (SD)	RO3: People contacted (SD)	RO4: Room transitions (SD)	RO5: Expected infections (SD)
0.05	0.05	114.30 (89.05)	104.29 (90.80)	16.39 (12.52)	23.14 (21.74)	3.88 (3.26)
0.005	0.05	115.67 (90.79)	104.93 (93.49)	16.44 (12.95)	23.85 (22.56)	0.51 (0.45)
0.0005	0.05	110.68 (87.39)	98.95 (88.37)	15.73 (12.42)	23.11 (22.72)	0.05 (0.04)
0.05	0.25	26.29 (27.74)	27.08 (34.68)	4.00 (4.18)	4.56 (3.96)	0.98 (1.21)
0.005	0.25	28.14 (31.78)	28.35 (37.61)	4.30 (4.36)	4.87 (4.27)	0.13 (0.18)
0.0005	0.25	26.66 (27.13)	26.71 (32.87)	4.13 (4.21)	4.71 (4.25)	0.01 (0.02)
0.05	0.5	13.69 (14.25)	14.73 (22.87)	2.24 (2.32)	2.41 (1.83)	0.53 (0.72)

0.005	0.5	13.58 (14.10)	14.87 (23.00)	2.18 (2.34)	2.42 (1.83)	0.07 (0.11)
0.0005	0.5	14.47 (15.32)	16.10 (23.31)	2.42 (2.56)	2.61 (2.12)	0.01 (0.01)
0.05	0.75	9.66 (9.23)	11.81 (18.08)	1.68 (1.74)	1.67 (1.06)	0.42 (0.59)
0.005	0.75	9.78 (9.31)	11.76 (18.39)	1.61 (1.63)	1.58 (0.94)	0.05 (0.09)
0.0005	0.75	9.55 (9.24)	12.56 (21.62)	1.64 (1.67)	1.58 (0.93)	0.01 (0.01)
0.05	0.95	8.27 (7.90)	9.91 (16.48)	1.39 (1.44)	1.29 (0.63)	0.35 (0.51)
0.005	0.95	8.56 (7.99)	9.94 (16.90)	1.32 (1.38)	1.29 (0.65)	0.05 (0.08)
0.0005	0.95	8.68 (8.42)	10.29 (18.04)	1.32 (1.42)	1.31 (0.69)	0.01 (0.01)

Sampled data for three different transmission assumptions and five levels of HHC for one colonised *nurse* starting in a random room in the hospital ward, a hand hygiene efficacy (γ) of 0.83. P = probability of transmission, λ = HHC level, $RO1$ = amount of time spent colonised, $RO2$ = number of contact moments, $RO3$ = number of HCWs or patients made contact with, $RO4$ = number of transitions between hospital ward rooms, $RO5$ = expected number of HMO transmissions.

RO5 is expressed as a percentage RO5 worst-case scenario ($P = 0.05;\ \lambda = 0.05$) RO5 for subset 1 ($RO5 = 3.16$) and 2 ($RO5 = 3.88$) and the combination of the two ($RO5 = 3.36$) in Figure 4-4. This pivoted view of RO5 shows similar changes in the expected numbers of transmissions for both subsets even though the nominal values of RO5 are different. A likely explanation is that during weekdays (Figure 4-4:B) increasing hand hygiene from 0.05 to 0.75 has a similar effect as decreasing the transmission probability by a factor 10 (0.05 vs 0.005)

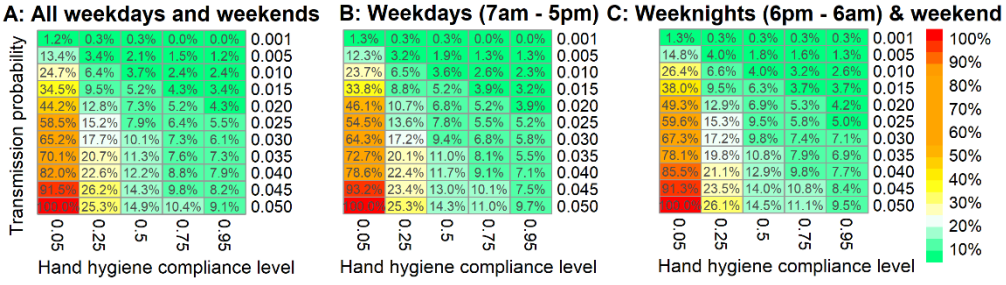

Figure 4-4: The expected number of transmissions expressed as a percentage of the worst-case scenario. For A, B and C: the highest number of expected transmissions (worst-case scenario) occur for the scenario where the transmission probability is 0.05 and hand hygiene compliance is 0.05. The expected number of transmissions is expressed as a percentage of the worst-case scenario.

4.4 Discussion

This study identified nurses as a potential super-spreader healthcare worker (HCW) occupation group in a healthcare setting. Nurses have a disproportionately high potential to transmit hand-transmittable harmful microorganisms (HMO) to other HCWs or patients as compared to the other HCW occupation groups. The expected number of transmissions caused by a colonised nurse increases exponentially as the level of hand hygiene compliance (HHC) deteriorates or the transmission probability increases. These results are due to the spatiotemporal behaviour and social mixing patterns of HCWs.

Five risk outcomes were defined to quantify the spatiotemporal effects of varying levels of HHC on the transmission and spread of HMO. These were: 1) the time that a colonised super-spreader is expected to be colonised; 2) the number of contact moments with other HCWs or patients; 3) the number of HCWs or patients encountered; 4) the number of ward rooms frequented while colonised and 5) the expected number of HCWs or patients a super-spreader will transfer microbes to before performing proper hand hygiene. The risk outcomes were quantified for various levels of hand hygiene compliance and probabilities of transmission. The expected change in the number of transmissions for different levels of HHC may encourage approval for healthcare interventions such as increased education and awareness about HHC and strategic accessibility to alcohol dispensers in healthcare settings.

The simulation results are based upon empirical social mixing patterns of HCWs and highlight one colonised nurse's impact as the super-spreader. These results are applicable when an HMO is transmittable by hand and can be eliminated by hand hygiene using an alcoholic rub. Depending upon the HMO, the probability of transmission may differ, resulting in a change in the expected number of transmissions for various levels of hand hygiene compliance. Such simulations can be used in educational materials to emphasize personal control and responsibility to perform HHC. Normal HHC levels of 50% may deteriorate to 25% during busy periods in a healthcare setting because of reduced healthcare worker capacity or time pressure. The simulation results allow for "what if?" questions to be answered under different assumed levels of HHC and transmission probabilities in terms of the five risk outcomes. HCWs are then able to simulate the impact of the initial assumptions on the expected number of transmissions caused by a super-spreader based on empirical spatiotemporal behaviour and social mixing patterns of HCWs in a real healthcare setting.

The results are consistent with other work done on super-spreaders in healthcare facilities [168]. Our contribution is that we quantified the potential consequences of the spatiotemporal behaviour of HCWs for varying levels of hand hygiene compliance and different transmission probabilities. The simulation results showed that, for the same transmission rates and HHC levels, the number of transmissions is higher during weeknights and weekends. An explanation is that HCWs spent more time with fewer HCWs or patients during weeknights and weekends but had more contact moments for every minute spent colonised. An increase in the time that a super-spreader navigates through the hospital ward results in an increase in the number of encountered HCWs or patients, allowing for more opportunity to transmit the HMO. HHC may vary over time because of varying ward occupancy levels or different days of the week: the simulation results show that for an HMO with a transmission rate of 0.05 and with the average level of HHC of 50% during the week and 25% over the weekend, that the expected number of HCWs and patients to whom a colonised nurse transfers microbes will almost double. Simulation scenarios were identified with equal risk outcomes for different initial conditions. They illustrate that infection prevention and control interventions can use combinations of strategies and bundles of interventions to fight the transmission and spread of HMO to achieve the same results.

The expected number of HCWs or patients to whom a super-spreader will transfer microbes before performing proper hand hygiene (risk outcome 5) is controlled by managing spatiotemporal behaviour (risk outcomes 1 – 4) and the level of HHC. A possible intervention based upon these results is to limit the number of room transitions, contact and contact duration during periods of low expected levels of HHC. If, for example, on busy Friday evenings the levels of HHC change, a possible preventative intervention might be to optimise the number of HCWs, as well as their routes and logistics according to an algorithm based upon sampled spatiotemporal movement data. Such an algorithm should then specifically be designed to minimise the potential transmission and spread of harmful microorganisms. Risk outcomes can thus be monitored over time, for instance, allowing one to determine the seasonality of trends or the effects of spatiotemporal interventions or policy changes. The five risk outcomes may then be addressed as spatiotemporal safety behaviours in hospital wards and in the formulation of healthcare policy to minimise the transmission of hand-transmittable HMOs.

This study contributes to infection prevention and control by highlighting five risk outcomes essential to describing the possible spread of an HMO on an individual temporal level and the spatial level in a healthcare setting. These insights apply to hand-transmittable HMOs and can be used to develop better informed preventative

strategies, for heterogeneous hand hygiene education, feedback, work-place reminders and other interventions.

4.4.1 Limitations and future work

Our sample is taken over seven days, giving a unique sample with good coverage for a single week. Differences may exist, however, with other weeks throughout the year and even between years. The data used in this study may further be biased towards HCWs who were diligent in wearing the RFID badges. The data were carefully checked for any inconsistencies; some loss in data quality caused by incorrect room classification because of overlapping RFID reader areas could still be present in the data. This study's hospital ward is similar to hospital wards found in most healthcare facilities in most aspects. Our results are based upon the sampled RFID tracking data for one specific ward. It is a future challenge to generalise these results to other wards in other hospitals.

The spatial resolution of our data is the room level and an assumption was made regarding the proximity and interaction between people, thus adding uncertainty in the simulation results. The transmission probability may be different during the day than during the evening shifts due to the difference of care provided during different times of the day. Future opportunities include collecting data of a higher spatial resolution that will allow us to identify the proximity between people, within room locations in the hospital and interaction with objects like hand hygiene dispensers and mobile (diagnostic) equipment like computers on wheels. An increase in spatial resolution will enable a more accurate event classification and result in more accurate simulation results. For instance, interaction with a hand hygiene dispenser does not require any assumption about the level of HHC, but only on its efficacy. Interaction of an HCW with objects and equipment inside different architectural designs and room layouts allows one to refine the transmission models, thus improving the transmission scenarios. The spatiotemporal risk outcomes defined should be further investigated to identify the relationship between them. For example, how the expected number of infections change should if the average contact duration contact decreases. These relationships can be used to determine how the risk outcomes should be addressed to reduce the hand transmission of HMOs efficiently.

4.5 Conclusion

This study defines five risk outcomes in terms of the number of contact moments, the duration of contacts and the number of ward rooms frequented while colonised and uses them to quantify the transmission and spread of harmful microorganisms. It shows that nurses are potential super-spreaders of harmful microorganisms due to their spatiotemporal movement and social mixing patterns in a healthcare setting.

The expected number of healthcare workers and patients, to whom a super-spreader transfers microbes, increases exponentially as the level of hand hygiene compliance deteriorates. The performed simulations increase our insight into the consequences of varying levels of adherence to spatiotemporally specific healthcare policies such as hand hygiene compliance. The simulations further show that a change in spatiotemporal movement and social mixing patterns of healthcare workers will affect the expected number of transmissions in a closed healthcare setting. The risk outcomes may be further addressed in terms of spatiotemporal safety behaviour in healthcare settings to reduce the spread of HMO. The adherence level is to be further investigated to improve the information to policymakers and further educate healthcare workers about the risks of their spatiotemporal behaviour.

5. Spatiotemporal prediction of the occurrence of vancomycin-resistant Enterococcus

Abstract

Vancomycin-resistant enterococci (VRE) is the cause of severe public health and monetary burdens. Antibiotic use is usually included as a possible confounding effect to predict VRE in patients, but the antibiotic use of patients who may have frequented the same ward as the patient in question is often neglected. This study investigated how the occurrence and spread of VRE can be explained by intrahospital patient movements (IPM) between hospital wards and their antibiotic use. Retrospective IPM, antibiotic use and PCR screening data were used from a hospital in the Netherlands. A dynamic directed spatiotemporal graph was developed, and together with the PageRank algorithm used to calculate two daily centrality measures to summarise the flow of patients and antibiotics at the ward level. The daily occurrence of VRE for every ward was predicted using a decision tree and random forest model. The models' performance was compared using a 30% test sample. The decision tree model produced a simple set of rules that can determine the daily probability of VRE occurrence for each hospital ward. The decision tree model achieved an acceptable area under the ROC curve (AUC) of 0.755 and the random forest model an excellent AUC of 0.883 on the test set. These results confirm that the random forest model performs better than a single decision tree for all model sensitivity and specificity levels at the cost of model simplicity. An early warning system for VRE can be developed and inform infection prevention plans and outbreak strategies further using these results.

This chapter was submitted to the BMC Infectious Disease journal and is undergoing minor revision. van Niekerk JM, Lokate M, Braakman-Jansen LM, van Gemert-Pijnen JE, Stein A. Spatiotemporal Prediction of the Occurrence of Vancomycin-resistant Enterococcus. Available at Research Square [https://doi.org/10.21203/rs.3.rs-860519/v1]

5.1 Background

Vancomycin-resistant enterococci (VRE) was first reported in Europe in 1986 [47] and since then has been the cause of severe public health and monetary burdens [46]. The prevalence of VRE and VRE outbreaks have increased over the past 20 years in Europe [181]. *Enterococcus faecalis* and *Enterococcus faecium* are the Enterococci species typically found in humans' gastrointestinal tracts, which could lead to bacteraemia, endocarditis, intra-abdominal and pelvic infections and urinary tract infections [47]. Patients are more than twice as likely to die from bloodstream infections caused by VRE as compared to a susceptible strain of Enterococcus [182]. Enterococci have properties that make them naturally resistant to the most used antimicrobial, and in particular, they can quickly become resistant to any new last-resort antimicrobials introduced.

Enterococci can survive on hospital surfaces and they can spread between patients using hands and surfaces as vectors [183]. In addition to direct patient-patient and HCW-HCW transmission pathways, there are five main transmission pathways for VRE inside a hospital: 1) patient to healthcare worker (HCW); 2) patient to the environment; 3) HCW to patient; 4) environment to patient; 5) environment to HCW [184]. Since the VRE can survive on dry environmental surfaces for months, it could be a constant source for new outbreaks [185]. These reservoirs may persist despite routine cleaning procedures [186].

The immediate surroundings of a patient with VRE are likely to contain VRE reservoirs [187] and the odds of a patient acquiring VRE increase when prior room occupants had VRE [188,189]. The risk of colonization increases as the number and proportion of patients with VRE in the same unit increases [190]. Patients also face increased odds of VRE colonization the more days they spend hospitalized [191]. Antibiotic use and immunosuppressing comorbidities such as leukaemia have been identified as risk factors for VRE colonization [182,191].

When a VRE outbreak occurs in a hospital, positive patients are isolated, the extent of the outbreak is estimated and additional control measures are implemented if necessary [181]. Estimating the extent of an outbreak involves determining the contact group, usually at the ward level. The contact group consists of the patients who could potentially have been colonized during the outbreak. Contact tracing is typically used to determine the patients at risk. A screening process can be carried out to verify which patients were indeed colonized, which can be expensive [192]. The benefits of improving the estimation accuracy of these contact groups are: 1) control measures are more effective, which translates into fewer transmissions and ultimately less

infections; 2) fewer patients are burdened by the screening process; 3) less testing reduces the financial burden.

Even though estimation of the extent of an outbreak plays a critical role in outbreak management, few studies have investigated the relationship between the patient movements between hospital departments and the spread of microorganisms. Reasons for patients to move from one department to another include deterioration of health; surgery after which they are moved to intensive care and afterwards to general care or more specialized care department; hospital logistics due to limited capacity. One study used centrality measures of intrahospital patient movements to predict the onset of *clostridium difficile* at the ward level [30]. The centrality of hospital antibiotic use, however, was not considered. *Clostridium difficile* can survive on hospital surfaces and patients are at risk from environmental vectors. Recent studies have shown that each intrahospital transfer increases a patient's odds of contracting *clostridium difficile* by 7% (95% CI 1.02 – 1.13). To our knowledge, no similar studies exist for the VRE.

The effects of intrahospital patient movements and antibiotic usage in hospitals are usually studied separately in antimicrobial resistance (AMR) research. The use of antibiotics is usually included as a possible confounding effect to predict VRE in patients, but the use of antibiotics of other patients who may have frequented the same ward as the patient in question is often neglected. Hospitals are dynamic systems with many moving objects and each of those objects has a surface that can act as a vector for VRE. Furthermore, antibiotic use can increase the number of VRE in patients due to selection pressure which can then spread between patients [186,193]. For these reasons, VRE should be studied using covariates which include spatiotemporal patterns of patients and antibiotics use in the hospital.

This study investigated how the occurrence and spread of VRE can be explained by patient movements and their antibiotic use between hospital wards. We estimated the probability of VRE at the ward level using intrahospital movement data and antibiotic usage data. We estimated this probability using a decision tree model and a random forest model and compared the model performance as a sub-objective. This study is important because it allows infection prevention and control specialists and outbreak management staff to determine which wards are at risk of a VRE outbreak using commonly available data.

5.2 Methods

5.2.1 Patient movement and antibiotic data

We used retrospective patient movement data from the University Medical Center Groningen (UMCG), one of the largest hospitals in the Netherlands with more than 12 000 employees and almost 1 400 beds. Antibiotic usage and patient movement data are stored in an electronic health record (EHR) database. The period under study is January 2018 until December 2019. The anonymised data consist of admission and discharge dates for each department within the hospital and antibiotic administration times during admission. These data were used to calculate two covariates for each day during the period of study: 1) the number of patients in each ward (pat_num); 2) the number of patients using antibiotics in each ward (pat_num_ant).

5.2.2 Spatiotemporal graph

The intrahospital patient movements data can be used to construct a dynamic directed spatiotemporal graph (DG) [194]. The graph nodes are the wards and the edges between the nodes are the patients moving between the wards. The DG is spatiotemporal and dynamic since it presents the location of patients using a node structure over time. We created two DGs using the patient movement data and the antibiotics data. The first graph includes all patient movement between all wards. The second graph only includes the movements of patients using antibiotics.

5.2.3 PageRank algorithm

The PageRank (PR) algorithm aims to determine the centrality or “importance” of nodes given the number of other “important” nodes with vectors directed towards it [50]. In the context of this study, the PR algorithm estimates the probability distribution of an arbitrary patient ending up in a particular ward. We calculated the daily PageRank probabilities for both DGs using a 30-day rolling time window: 1) PageRank of patient movements between wards (PR_pat_num) and 2) PageRank of patient movements currently using antibiotics (PR_pat_num_ant). The PR_pat_num and PR_pat_num_ant represent the centrality of wards in terms of patients and antibiotics, respectively.

5.2.4 VRE screening data

The number of VRE tests per week fluctuated between 100 to 300 per week during the study period. There was a VRE outbreak in the second half of 2018 (Figure 5-1). Outbreak procedures were implemented and hospital ward screening continued. Between July - December 2018, 141 positive VRE tests were reported, with a peak of 25 positive tests in one week. In total, 48 patients tested positive for VRE over the

study period. These data were used to calculate the binary outcome variable for this study (Formula 5-1).

$$Y = \begin{cases} 1, & \text{number of VRE positive patients in ward} > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{(Formula 5-1)}$$

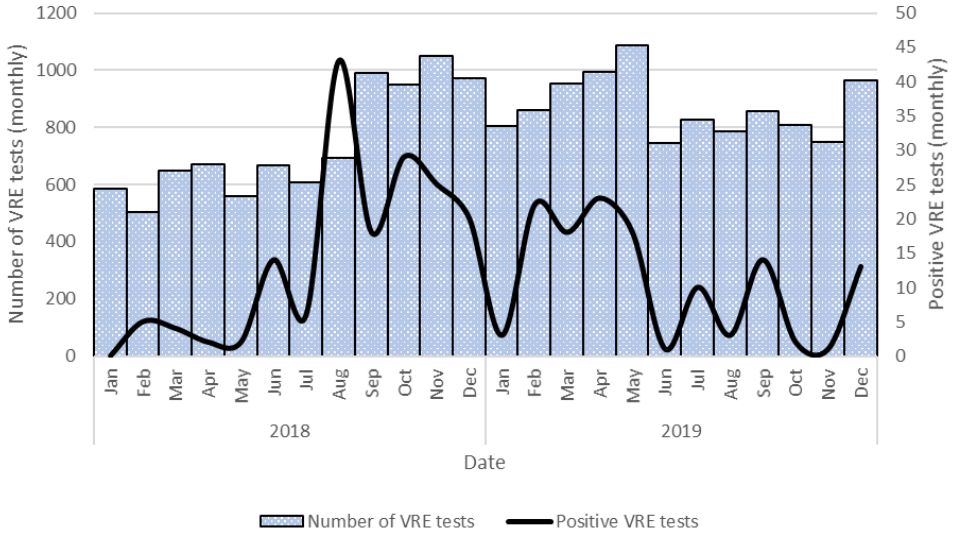

Figure 5-1: VRE tests and the number of positive VRE test results during 2018 - 2019.

5.2.5 Modelling

We estimated the probability that there is at least one patient with VRE in a specific ward (Y) given the covariates pat_num, pat_num_ant, PR_pat_num and PR_pat_num_ant (Formula 5-2).

$$P(Y = 1 | \, pat\_num, pat\_num\_ant, PR\_pat\_num, PM\_pat\_num\_ant) \quad \text{(Formula 5-2)}$$

5.2.6 Decision trees

A decision tree was used to determine a simple set of rules based on the covariates to estimate the probability of Y [195]. The decision tree was grown using a 70% random training sample of the complete set of data. The data were split incrementally by adding question nodes. The question nodes consider the ability of each covariate to discriminate between the observed binary outcomes and formulates the question using the one that can discriminate best [51]. We used the Gini index to quantify the discriminatory ability of each covariate at the question

nodes [195]. Continuing in this way, a tree branch structure is created, leading to the final decision or leaves of the tree.

5.2.7 Random forest

The model performance of decision trees was improved by creating an ensemble of decision trees and using them in unison to predict the outcome variable [51]. We used the same 70% randomly sampled training samples used to train the decision tree model. To build the random forest (RF) model, 500 random samples with replacement (bootstrap sample) were drawn from the training data and two random outcome variables were used to build a decision tree for each of the bootstrap sample. The probability of Y was determined by calculating the proportion of the 500 trees that predicted $Y = 1$.

We compared the model performance of the decision tree and random forest models using the remaining 30% data as a test sample. The area under the receiver operating characteristic curve (AUC) was used to measure model performance as it provides a holistic view of how well the model predicts the outcome variable for different levels of sensitivity and specificity [196]. An AUC between 0.7 and 0.8 is considered as acceptable and between 0.8 and 0.9 excellent [125].

5.2.8 Software

The R statistical programming language was used to perform the analyses in this study [78]. Graphs were created and evaluated using igraph [197]. The decision trees and random forest models were fitted using the R packages rpart and randomForest packages [198,199]. In addition, the tidyverse R package was used to clean and structure the data [81].

5.3 Results

In total, 48 distinct wards were occupied over the 730 days in the study period (2018 – 2019). Of the possible 35 040 observations, if all the wards were occupied every day, only 31 649 observations were collected, of which 1 377 (5.45%) had at least one patient with VRE.

5.3.1 Covariates

The pat_num and pat_num_ant covariates are shown with the number of positive VRE patients during the VRE breakout period in 2018 in Figure 5-2. We highlight the covariate associated with a general care ward with many VRE patients during this outbreak in Figure 5-3. These results show a higher level of variation at the ward level, which conforms better to the number of patients with VRE. The highest number of positive VRE patients were observed in the last week of August 2018. At the hospital level, the relationship between the pat_num_ant, pat_num and the number

of positive VRE patients is not evident. When the same data are shown at the ward level for the general care ward, these covariates are correlated with the number of VRE patients.

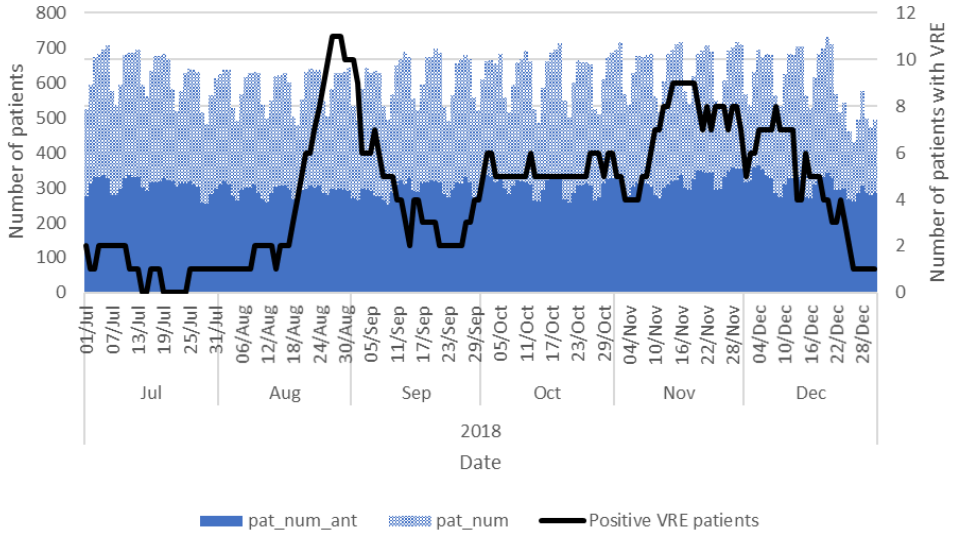

Figure 5-2: Number of patient and patients using antibiotics. pat_num_ant = the number of patients using antibiotics in each ward; pat_num = the number of patients in each ward.

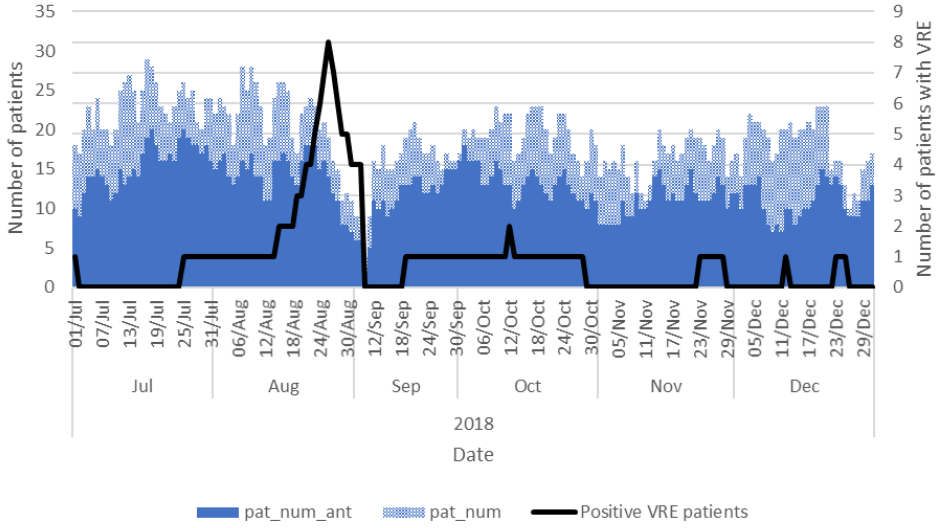

Figure 5-3: Number of patient and patients using antibiotics in example general care ward. pat_num_ant = the number of patients using antibiotics in each ward; pat_num = the number of patients in each ward.

Comparing the two PR_pat_num and PR_pat_num_ant reveal that during this period, PR_pat_num_ant was higher than PR_pat_num (Figure 5-4). This means that, on average, the probability of a patient using antibiotics to visit an occupied ward was higher than for the total patient population. The same covariates are shown for the example general care ward in Figure 5-5. The general care ward experienced a significant increase in PR_pat_num_ant during July and October, which lasted for four weeks and yet PR_pat_num did not show a similar pattern. These results show that the two centrality covariates provide different information of the patient and antibiotics flow in a hospital at the ward level.

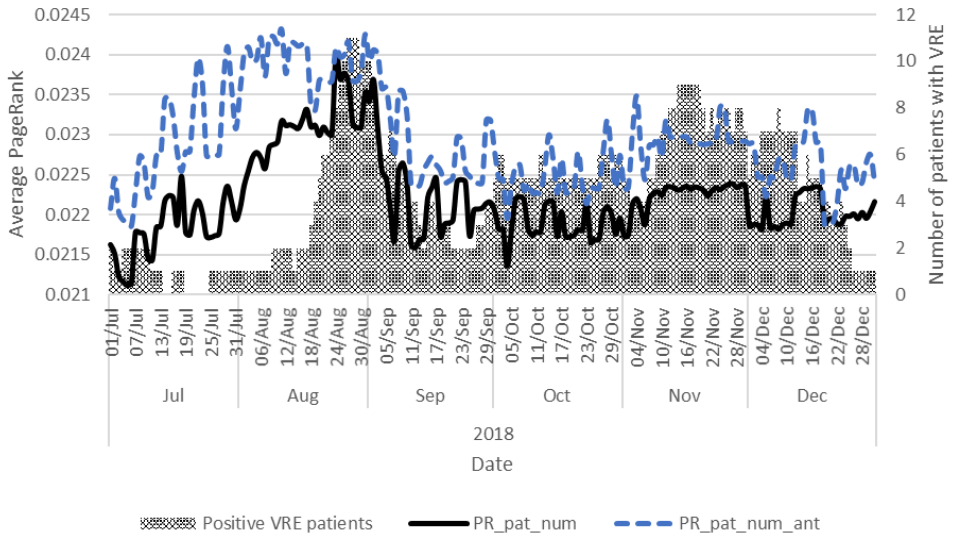

Figure 5-4: Average daily PageRank covariate and the number of VRE positive patients. PR_pat_num = PageRank of patient movements between wards; PR_pat_num_ant = PageRank of patient movements using antibiotics.

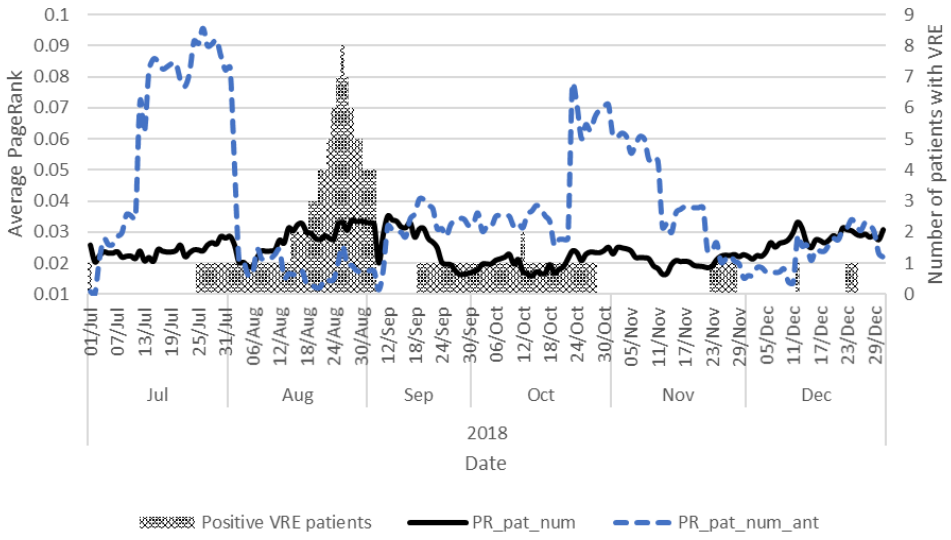

Figure 5-5: Average daily PageRank covariate and the number of VRE positive patients in example ward general care ward. PR_pat_num = PageRank of patient movements between wards; PR_pat_num_ant = PageRank of patient movements using antibiotics.

5.3.2 Decision tree

The 70% training sample had a 4.3% positive VRE percentage of the root node (Figure 5-6). The pat_num_ant covariate splits the first nodes. If the number of patients is less than six, which is the case for 40% of the training sample, then there is a 0.098% probability that the ward has a VRE patient. If the number of patients in a ward is six or more, but less than 13, we continue to the next node to consider the PT_pat_num_ant covariate. After dividing the training sample by the five nodes, we arrive at the seven leaves of the tree. The probabilities of the leave population range between 0.98% and 15.68%. According to the order in which the covariates were used in the model, the pat_num_ant is the most important covariate to estimate the probability of a hospital ward having at least one VRE patient or not. The PR covariates are next in the order of importance to determine the final leaves of the tree. The decision tree results can be written and executed as a simple set of rules provided in (Formula 5-3).

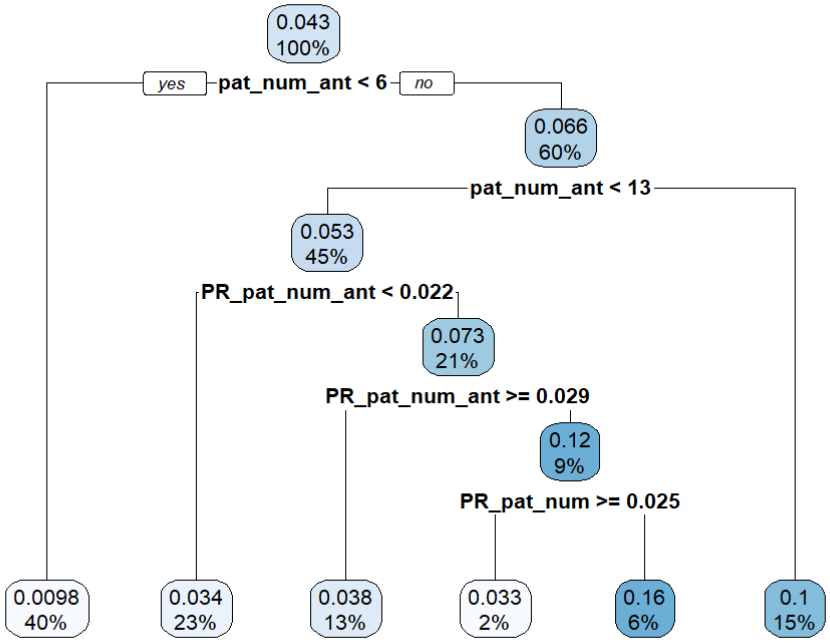

Figure 5-6: Decision tree for the daily VRE occurrence in a hospital ward using PageRank and traditional covariates. pat_num_ant = the number of patients using antibiotics in each ward; PR_pat_num_ant = PageRank of patient movements currently using antibiotics; PR_pat_num = PageRank of patient movements between wards. In each node, the percentage of ward with at least one VRE positive patient is shown above the sample distribution of the node.

$$P(Y = 1|\,\text{pat\_num}, \text{pat\_num\_ant}, \text{PR\_pat\_num}, \text{PM\_pat\_num\_ant}) = \quad \text{(Formula 5-3)}$$

0.0098 if $\text{pat\_num\_ant} < 6$,

0.0326 if $\text{pat\_num\_ant} \in [6,13]$ AND $\text{PR\_pat\_num\_ant} \in [0.022, 0.029)$ AND $\text{PR\_pat\_num} \geq 0.025$,

0.0340 if $\text{pat\_num\_ant} \in [6,13]$ AND $\text{PR\_pat\_num\_ant} < 0.22$,

0.0384 if $\text{pat\_num\_ant} \in [6,13]$ AND $\text{PR\_pat\_num\_ant} \geq 0.29$,

0.1030 if $\text{pat\_num\_ant} \geq 13$,

0.1568 if pat_num_ant $\in [6,13]$ AND PR_pat_num_ant $\in [0.022, 0.029)$ AND PR_pat_num < 0.025

5.3.3 Random forest

The minimal depth provides insight into where a covariate occurs for the first time in the decision trees for the random forest and quantified variable importance. Covariates with lower minimal average depth are used to split larger proportions of the population due to higher discriminatory power. The results show that pat_num_ant has the lowest average depth (0.61) and is most likely to be used in the root node. This result is consistent with our single decision tree model (Figure 5-7). PR_pat_num was not used as a root node for any of the 500 decision trees. It has the largest average depth (1.93) in the trees, which means that it was generally used in nodes appearing lower in the decision trees.

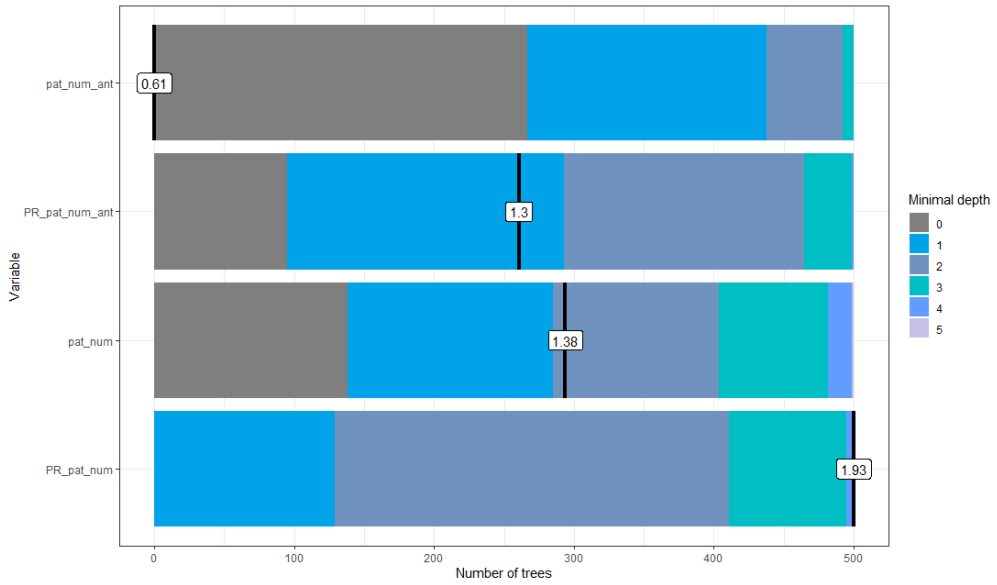

Figure 5-7: Minimal depth for each covariate in the 500 random forest decision trees. pat_num_ant = the number of patients using antibiotics in each ward; PR_pat_num_ant = PageRank of patient movements currently using antibiotics; pat_num = the number of patients in each ward; PR_pat_num = PageRank of patient movements between wards.

We determined the covariate importance in the RF model by calculating the percentage increase in the mean square error (MSE) and the change in the residual sum of squares (RSS) of the model should random information replace the values of

the model covariates. The results show that the PR covariates are the most important ones in terms of the MSE (Figure 5-8) and RSS (Figure 5-9) reductions.

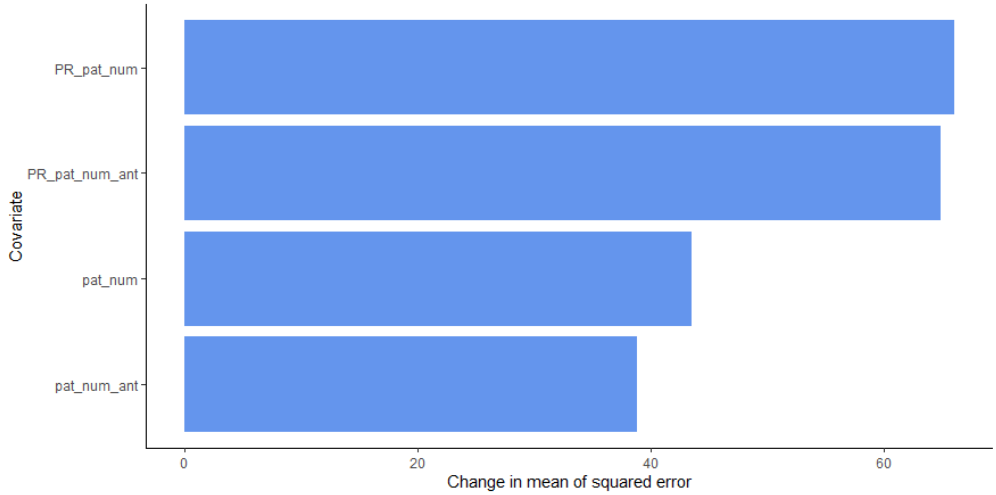

Figure 5-8: The change in mean squared error when covariate values are replaced with random values. PR_pat_num = PageRank of patient movements between wards; PR_pat_num_ant = PageRank of patient movements currently using antibiotics; pat_num = the number of patients in each ward; pat_num_ant = the number of patients using antibiotics in each ward.

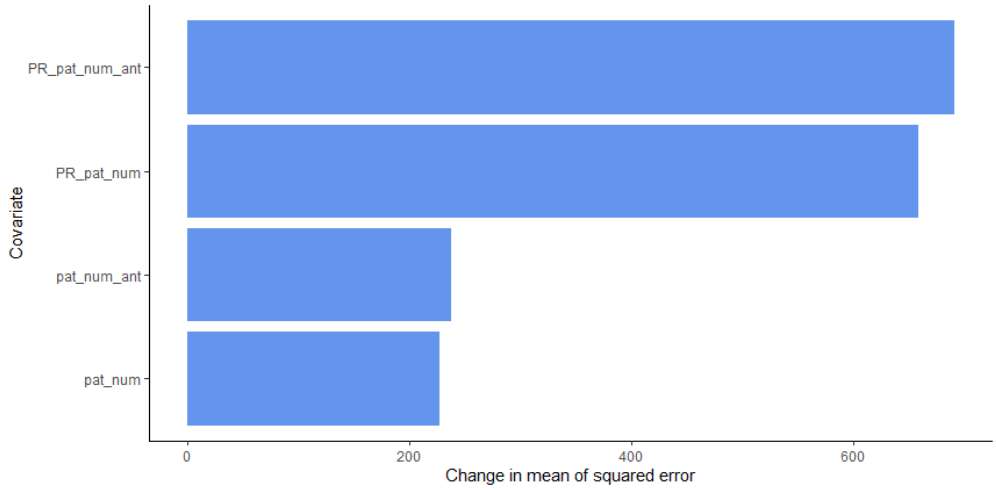

Figure 5-9: The change in residual sum of squares when covariate values are replaced with random values. PR_pat_num_ant = PageRank of patient movements currently using antibiotics; PR_pat_num = PageRank of patient movements between wards; pat_num_ant = the number of patients using antibiotics in each ward; pat_num = the number of patients in each ward.

5.3.4 Model performance

The performance of the models is compared to the Lorenz curves shown in Figure 5-10. The Lorenz curve of the RF model is consistently higher than for the decision tree model. The RF model achieved an excellent AUC of 0.883 and the decision tree model an acceptable AUC of 0.755 on the 30% test set. This result confirms that the random forest model performs better than a single decision tree for all levels of model sensitivity and specificity on data not used to estimate the models. This is important to estimate the loss in model performance when choosing to use the simple set of rules produced by the decision tree model to calculate the probability of Y rather than using the RF model.

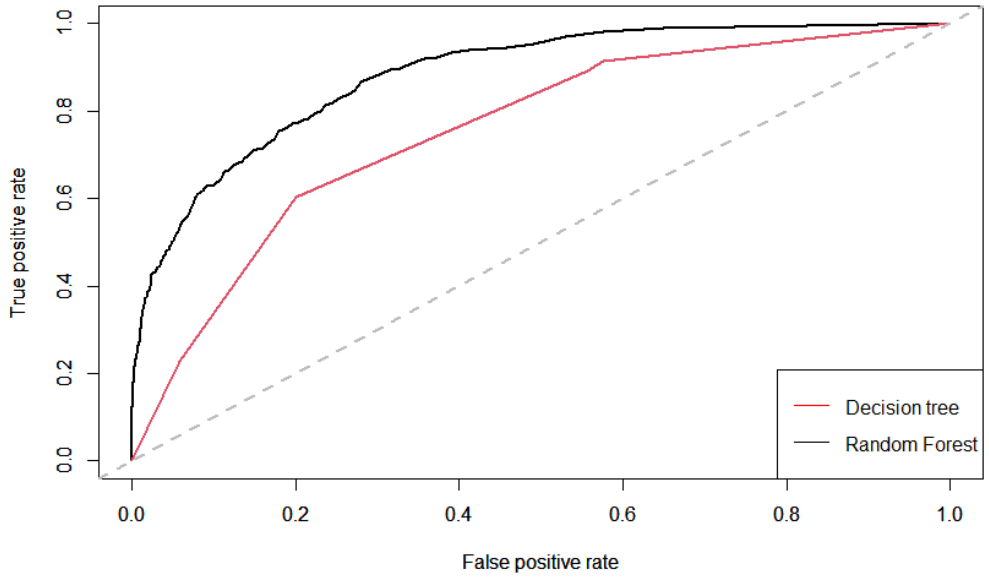

Figure 5-10: Lorenz curves of the decision tree and random forest models.

5.4 Discussion

This study showed how the movements of patients inside hospitals and their use of antibiotics could predict the occurrence of VRE at the ward level. Two daily centrality measures were proposed to summarise the flow of patients and antibiotics at the ward level. A simple set of rules were produced which can be used to monitor the risk of VRE in hospital wards. Using an ensemble method, a more accurate but more complicated model was developed, which can be applied to the same effect should resources allow for it.

The two PageRank covariates proposed offered new insight into the centrality of wards regarding patient and antibiotic movements and their interaction. This study

used the covariates to predict VRE, but they can be used in many other studies concerning antimicrobial resistance in hospitals. Institutional surveillance monitors the usage of antibiotics but not the flow and concentration thereof. The proposed PR covariates can be used in conjunction with existing institutional surveillance metrics to monitor the risks for VRE and AMR in general.

The decision tree model resulted in six simple questions and provided the probability that a ward has at least one patient with VRE as an answer. This model enables hospitals to use passive data collected in their electronic health records to calculate this probability. To improve the accuracy of this model, a random forest model was built, which outperforms the decision tree model. The random forest model results were not as easily interpretable as that of the decision tree as it uses 500 smaller decision trees every time a probability is calculated. In practice, the model used will depend on the skills and resources of the hospital and its infection prevention and control specialists.

5.4.1 Future work

The results of this study can be used to develop an early warning system for VRE and other microorganisms with similar transmission mechanisms. The probabilities produced by the models presented can be used to classify VRE according to the desired level of sensitivity and specificity for such a system. The results can then be updated daily or as frequently as the covariates can be calculated and evaluated by the infection prevention specialists to decide on the best course of action.

Our results showed that the value of the patient movement and antibiotic PR covariates sometimes move in the opposite direction over time. This divergence suggests that the proportion of patients using antibiotics is changing over time. These covariates can be used together to determine if emerging divergences increase the risk of VRE occurrence.

5.4.2 Limitation

The study period was limited by the amount of data available for intrahospital patient movement, antibiotic use and VRE screening. UMCG migrated to a new electronic healthcare system in 2017, resulting in the antibiotic data not being available at the time of publication. There was a VRE outbreak in 2017, which would have allowed us to build these models on the 2017 outbreak and validate them on the 2018 outbreak. Once these data become available, this could be a future research opportunity.

Even though this study can determine if a patient were using antibiotics at a particular time, we could not distinguish between the types of antibiotics used. Some antibiotics target specific bacteria and can have a more significant effect on the risk

of acquiring VRE. A future research opportunity is to create antibiotics centrality measure for antibiotics targeting different bacteria.

The ideas behind this study can be further expanded to the patient level. This expansion will require additional patient data regarding demographics and comorbidities affecting the risk of contracting VRE. A prediction model for VRE at the patient level using the proposed spatiotemporal centrality measures and patient-level data will improve the efficiency with which infection prevention specialists can control AMR in hospital.

5.5 Conclusion

This study showed how the movements of patients inside hospitals and their use of antibiotics could predict the occurrence of VRE at the ward level. Two daily centrality measures were proposed to summarise the flow of patients and antibiotics at the ward level. A simple set of rules was produced which can be used to monitor the risk of VRE in hospital wards. A random forest model was compared with a decision tree model to improve the prediction performance at the cost of simplicity. An early warning system for VRE can be developed to test and further develop infection prevention plans and outbreak strategies using these results.

6. Synthesis

6.1 Findings

This thesis investigates how statistical models can improve our understanding of the occurrence and spread of harmful microorganisms in hospitals with the additional complication of antimicrobial resistance (AMR).

In this chapter, the main findings are summarised and discussed for each of the four research questions in terms of significance, limitation and potential for future research.

RQ 1: How can knowledge gaps in AMR research be identified objectively and automatically?

Potential knowledge gaps in AMR research were identified using a data-driven statistical methodology and the key knowledge gaps were highlighted. AMR scientific research over the past 20 years was grouped into 88 topics. These topics were clustered into seven larger research areas. In total, 421 potential knowledge gaps were identified between the AMR research topics and larger research areas and 2 663 between individual topics. From these potential knowledge gaps, specific knowledge gaps could be highlighted. A knowledge gap between the clinical AMR research area and topics related to molecular and laboratory research was identified. Topics related to the water and the environment and surveillance were found to be unrelated in AMR research and identified as a knowledge gap. Furthermore, the results showed that a knowledge gap exists between *Data modelling and estimation* topic and the *Resistance patterns on hospital level* AMR research topic. These results show that a semi-automated data-driven statistical methodology can be used to identify potential knowledge gaps in AMR research that AMR researchers may consider for further research and suggest that it can function as an alternative or in conjunction with the existing expert methodology.

RQ 2: What are the risk factors for the occurrence of SSI when using data-driven cut-off values for continuous variables?

Risk factors related to surgical site infection were identified using standard medical cut-off values and data-driven cut-off values for continuous variables. Although the standard medical cut-offs were confirmed by the data-driven cut-offs for most continuous variables, the data-driven cut-offs were different for pre-operative patient temperature, CRP and patient age and better explain the outcome value by up to 19.5%. A preoperative body temperature of $\geq$38 ^{o}C and antibiotic use are risk factors for surgical site infection (SSI) after digestive, orthopaedic and thoracic system surgeries. The duration of surgery and patient age are risk factors for SSI after

orthopaedic and thoracic system surgeries, respectively. SSI is more likely to occur in children (age < 18) than in adults after thoracic system surgeries. The results show that data-driven cut-offs for continuous variables may differ from standard medical cut-offs and that they can be effective to identify risk factors for the occurrence of SSI.

RQ 3: How can the spatiotemporal movements of healthcare workers identify potential a super-spreader occupation group of harmful microorganisms in a closed healthcare setting?

Spatiotemporal data were collected from healthcare workers using radio frequency identification (RFID) technology to simulate the spread of harmful microorganisms in a hospital ward for different hand hygiene compliance levels. The results showed that nurses are potential super-spreaders of harmful microorganisms due to their spatiotemporal movement and social mixing patterns in a healthcare setting. The expected number of healthcare workers and patients to whom a super-spreader transfers microbes increased exponentially as the level of hand hygiene compliance deteriorates. Five risk outcomes were defined: 1) the time that a colonised super-spreader is expected to be colonised; 2) the number of contact moments with other healthcare workers (HCW) or patients; 3) the number of HCWs or patients encountered; 4) the number of ward rooms frequented while colonised and 5) the expected number of HCWs or patients a super-spreader will transfer microbes to before performing proper hand hygiene. They were used to quantify the transmission and spread of harmful microorganisms. The results further show that a change in spatiotemporal movement and social mixing patterns of healthcare workers will affect the expected number of transmissions in a closed healthcare setting.

RQ 4: How can the occurrence of vancomycin-resistant Enterococcus (VRE) in a hospital be predicted using intrahospital patient movements and antibiotic usage?

This study shows how commonly available intrahospital patient movement data from EHR could be used to predict the occurrence of VRE at the hospital ward level. Two daily centrality measures summarised the flow of patients based on and antibiotics at the hospital ward level. The result is a simple set of rules that can monitor the risk of VRE in hospital wards. A random forest model improved the model prediction performance (area under the curve (AUC) = 0.883) compared to a decision tree model (AUC = 0.755) at the cost of model simplicity. These results showed how centrality covariates summarising the flow of patients and their antibiotic use between hospital wards could be used to predict the daily occurrence of VRE at the hospital ward level.

6.2 Significance and prospects

This thesis shows how statistical models and novel spatiotemporal data could enrich existing risk factors and identify new risk factors to predict the occurrence and spread of harmful microorganisms and the complication of AMR. Furthermore, this thesis contributes to clinical practice by providing new statistical tools to prevent and control the occurrence and spread of AMR in closed healthcare settings. Although much has been done in the thesis, many future research opportunities were identified.

For the first time, the scientific AMR literature was quantitively classified into topics and assessed for knowledge gaps. It is now clear what the current standing is of AMR research, what the topics are and how they are related or not related to each other. This thesis provides a complete list of potential knowledge gaps in AMR research, of which the most important and urgent ones were highlighted. Technical advisory groups across sectors and industries can use these results to inform national and international research agendas about the current shortcomings in understanding AMR. In addition to answering RQ1, this new insight may change how knowledge gaps are identified in the future. The search for knowledge gaps was limited to AMR research, but the same methodology can be applied to any research field with similar results. A future research opportunity is to fully automate the methodology used in Chapter 2 and generalise it to other research areas. These additional steps may lead to a time when researchers may not have to look for knowledge gaps to fill but choose from a list of gaps automatically generated and available to everyone. The final step may be to determine the importance of those knowledge gaps automatically.

The process followed to identify risk factors in healthcare has evolved over the last century into a standardised process. So much so that the definition of the potential risk factors themselves is rarely questioned. It is intuitive to assume that a cut-off for age seems reasonable at, maybe, 18, since society regards this age as highly significant in other aspects of life. Does that mean that it should be used as a cut-off when considering age as a risk factor for getting an infection after having surgery? Before the research performed in Chapter 3 of this thesis, this question was neither asked nor answered. RQ2 questioned the established standard medical cut-offs used in risk factor identification research and Chapter 3 illustrates the importance of data-driven methodologies when identifying risk factors for the occurrence of SSI. The results showed how a statistical model could be used to determine the cut-off values of potential risk factors and how the subsequent risk factors identified differ from those identified when using standard medical cut-off values. Even though it may be convenient to use existing standard medical cut-off values because they are widely

accepted and easily comparable, key insights are likely to be missed due to using them. These results can inform the methodology of future studies that aim to identify risk factors for adverse patient-related outcomes. Future research may suggest best-practices to evaluate both standard medical and data-driven cut-off values when identifying risk factors in healthcare.

This thesis identified nurses as potential super-spreaders of harmful microorganisms to answer RQ3. This result was confirmed using empirical spatiotemporal RFID data to study the social mixing patterns of healthcare workers. Using the five-risk outcome defined in Chapter 4, hospitals can estimate the change in the expected number of transmissions for different levels of hand hygiene compliance based on the social mixing pattern of healthcare workers. These results are instrumental in informing better infection prevention and control measures and providing relevant training based on empirical data. In addition, this research confirmed the importance of spatiotemporal data and the understanding of social mixing patterns to explain the spread of harmful microorganisms. Future research may use these risk outcomes together with patient-level data to determine the risk of transmission for each patient. Another opportunity is to collect higher spatial resolution data to identify the proximity between people, within room locations in the hospital and interaction with objects like hand hygiene dispensers and mobile (diagnostic) equipment like computers on wheels. Using technologies like active RFID tags to increase spatial resolution may enable a more accurate event classification and result in more accurate simulation results [200].

The final study combined the learnings of the previous chapters to predict the occurrence of AMR at the hospital ward level using spatiotemporal data of patients and their antibiotic use to answer RQ4. Using the PageRank algorithm to encapsulate the flow of patients and antibiotics is a novel and elegant way to summarise the relevant information in a dynamic directed spatiotemporal graph. The same methodology may also be used to model the flow of other attributes and objects over the same network. The simple set of rules is easily interpretable by healthcare workers and infection prevention specialists and can be used to strategic screening policies for VRE. The random forest model can be incorporated into a sophisticated early warning system for VRE to enable state of the art infection prevention and control measures at the hospital ward level. Although the model is not at the patient level, the results may help infection prevention and control specialists to identify wards at high risk of AMR and take preventative measures earlier.

Furthermore, the intrahospital movement data used in this model should be available in most hospitals, making it more straightforward and cost-effective to develop and implement than other real-time location technologies. Future studies

may use the methodology developed in this research in combination with patient-level data to develop patient-level models to predict AMR at the patient level. This expansion will require additional patient data regarding demographics and comorbidities affecting the risk of contracting VRE. These patient-level data may be more challenging to obtain as they may not be available in a structured format but rather in free-text documents. Hospitals must aim to store all potential patient-related risk factor data in a structured format to optimise the use of statistical models to identify risk factors and predict future outcomes. A prediction model for VRE at the patient level using the proposed spatiotemporal centrality measures and patient-level data will dramatically improve the efficiency with which infection prevention specialists can control AMR in hospitals. Such a model can be used to develop an early warning system for VRE to inform strategic screening and cleaning strategies to combat the occurrence and spread of VRE. To further improve the generalisability of Chapter 5, data from different hospitals in The Netherlands and other countries can be used to obtain comparable results. Specifically, this study can be repeated for a university medical hospital in Germany on the Dutch-German border to compare the effect of different healthcare policies on the occurrence and spread of AMR in hospitals.

Future research may consider more advanced statistical models to increase model performance, stability and generalisability. Bayesian statistical methods can determine the probability that the data observed were generated by the estimated model. This ability can determine if our model is suitable to use in different healthcare settings such as other hospitals in the Netherlands or even healthcare facilities in developing countries with different AMR prevalence and infection prevention policies. Should the estimated model not be suitable, then the probability estimate from the initial model can be used as a prior probability and updated by observing the new data. Bayesian statistics can be used to increase the stability of the decision tree model built in Chapter 5. Bayesian decision trees have many advantages of random forests with the additional benefit of an easily interpretable result [201].

Statistical research using complex networks, such as the dynamic directed spatiotemporal graph used in Chapter 5, is still in its infancy [202]. Dynamic graph neural networks (DGNN) use deep neural network architecture to encode the network structure of complex networks and aggregate local and global features of neighbouring nodes over continuous time. Streaming Graph Neural Networks (SGNN) is the state-of-the-art DGNN approach [202,203]. A future opportunity is to use SGNN to take full advantage of the embedded structure of the IPM data while incorporating both ward-specific and patient-related data to predict the occurrence and spread of VRE.

6.3 Limitations

Several obstacles were faced while conducting the research comprising this thesis and some lead to limitations in this research. Most of these obstacles were concerning the availability and integrity of the data needed. Data privacy is of utmost importance in the Netherlands and is enforced by the Dutch Personal Data Protection Act (WBP). This act places restrictions on when and what personal data may be used, where and by whom personal data are processed and when processing is allowed. Although these restrictions are imperative to protect the privacy of Dutch residents, important data may be neglected in healthcare research because of it. For the studies included in this thesis, research protocols were drafted to inform the respective hospitals' ethical committees that the proposed studies would use anonymised retrospective data not subject to the Medical Research Involving Human Subjects Act (WMO). Unfortunately, this process took several months and was necessary as the WMO does not cater specifically for healthcare research. These are examples where data privacy policies may hamper healthcare research. To overcome this challenge, the studies performed in this thesis were performed in parallel as much as possible to research downtime but only became effective during the middle of the total research period. Research centres perpetually doing research in this research field should get the full benefit of this strategy.

Chapter 3 is a retrospective, single-centre study, and therefore the data were not collected for this study. The models built in this study were not externally validated. To overcome this limitation, cross-validation was performed to estimate model performance on new data. Surgeries were aggregated into three broad groups of surgical procedures, which serve as a proxy for the reason for surgery but leads to the loss of information regarding the exact reasons for the surgery. The administration of prophylaxis and the optimal timing thereof is an important risk factor for the occurrence of SSI. However, these data were not available. The data were stratified according to surgery type to minimise the impact of missing risk factor data.

Patient management systems are constantly changing and evolving in healthcare. In Chapters 3 and 5, delays were experienced due to data system migrations and data format changes. Some of the patient comorbidity data used in Chapter 3 were stored in free text in different formats depending on the system used at the time. These data had to be extracted using regular expressions, which may be prone to error. In Chapter 5, some data were not available due to system migrations, resulting in the lack of validation in the study. The challenge of missing data from the free text was overcome using multiple imputation strategies in Chapter 3. This strategy meant that the variables extracted from the free text could still be used in the statistical models,

although their probability of being found significantly decreased due to an understated variability. The migration to new patient management systems showed how rapidly hospitals in the Netherlands transitioned to digital solutions. The new systems store essential data in structured databases, which will reduce research complications in the future.

In Chapter 4, the RFID data used for the study were data collected during a pilot study performed at UMCG. There were some irregularities in the data, and it was decided to collect these data again using different proximity settings. Unfortunately, the company responsible for the data collection closed before this could be achieved. This situation forced us to take a serious look at the data collected during the pilot and devise ways to check it and clean it to be viable for the suggested study. In this way, we worked more efficiently with the data that would probably have been discarded. The data were collected over seven days, giving a unique sample with good coverage for a single week. Differences may exist, however, with other weeks throughout the year and even between years. The data used may further be biased towards HCWs who were diligent in wearing the RFID badges. The data were carefully checked for any inconsistencies; some loss in data quality was caused by incorrect room classification because overlapping RFID reader areas could still be present in the data. The results are based upon the sampled RFID tracking data for one specific ward. Even though the study's hospital ward is similar to hospital wards found in most hospitals in terms of layout and specialism, it is a future challenge to generalise these results to other wards in other hospitals.

The model built in Chapter 5 was not validated on another VRE outbreak. Even though the model performance was compared on a 30% test sample of the data, it was not possible to determine how the model would perform in a different time, a different hospital or a different country. This limitation may affect the generalisability of the model and should be tested in the future. The type of antibiotics used was not considered in this study. Some antibiotics target specific bacteria and can have a more significant effect on the risk of acquiring VRE. A future research opportunity is to create antibiotics centrality measure for antibiotics targeting different bacteria to further increase the discriminatory power of the predictive models.

During 2020 and 2021, SARS-CoV-2 and the pandemic of COVID-19 increased the challenges faced by the research in this thesis. Chapter 2 was already published by that time, but Chapters 1, 4 and 5 were still in progress and relied heavily on the inputs of infection prevention specialists and microbiologists. Their research capacity decreased during this time, which resulted in several delays in the research performed in this thesis. During this time, good communication was extremely important to use time sparingly and efficiently. The pandemic also resulted in a

massive influx of new studies being performed on the transmission of harmful microorganisms. Even though the transmission mechanisms may differ between SARS-CoV-2 and AMR, the pandemic may provide more incentive to understand better the spatiotemporal risk of the transmission of harmful microorganisms.

6.4 Reflection

My research was performed from a statistician's perspective with limited experience in the medical sciences as part of the interdisciplinary and cross-sectoral EurHealth-1Health project. This inexperience has been a blessing and a curse. The blessing was that I could study the AMR research field from an objective and fresh perspective while questioning established concepts and methodologies. A curse was to some degree that I had to rely on data collected by others to determine where the gaps are in the current knowledge of AMR research and how to fill those gaps. Consulted healthcare professionals were needed to interpret the results and to confirm the practical relevance of my questions and ideas.

My experience as a statistician with identifying risk factors and building predictive models helped guide my thoughts on approaching this research. Logistic regression modelling is still widely used in medical modelling, while much more advanced methods are available. The idea to test the medical cut-off values for the continuous variables in Chapter 3, for instance, was inspired by a similar procedure that is widely used in credit scorecard development. I also identified other similarities between studies conducted in the healthcare and financial research fields and saw how both domains could benefit from each other. Personally, I gained valuable statistical knowledge while performing this research because of the different research environments. I would recommend the experience for anyone specialising in applied statistics.

Bibliography

1. O'Neill J. Tackling drug-resistant infections globally: final report and recommendations. 2016.

2. Bengtsson-Palme J, Kristiansson E, Larsson DGJ. Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiology Reviews*. 2018;42(1):68-80. doi:10.1093/femsre/fux053

3. Wall S. Prevention of antibiotic resistance - an epidemiological scoping review to identify research categories and knowledge gaps. *Global health action*. 2019;12(1):1756191. doi:10.1080/16549716.2020.1756191

4. Durand GA, Raoult D, Dubourg G. Antibiotic discovery: history, methods and perspectives. *International journal of antimicrobial agents*. 2019;53(4):371-382.

5. WHO. Global Action Plan on Antimicrobial Resistance. *World Health Organisation*. 2015:28. doi:ISBN 978 92 4 150976 3

6. Antimicrobial JPI on. Innovation Agenda on Antimicrobial Resistance. Strategic Research and Innovation Agenda on Antimicrobial Resistance. https://www.imi.europa.eu/sites/default/files/uploads/documents/About-IMI/research-agenda/IMI2_SRA_March2014.pdf. Published 2014. Accessed May 24, 2021.

7. Yohann Lacotte, Marie-Cécile Ploy CÅ. Gathering of national research priorities from at least five European countries and gap identification. European Union Joint Action on Antimicrobial Resistance and HealthcareAssociated Infections (EU-JAMRAI). https://eu-jamrai.eu/wp-content/uploads/2019/03/EUjamrai_MS9.1_Report_WP9_2019.02.28.pdf. Published 2019. Accessed May 24, 2021.

8. Lacotte Y, Årdal C, Ploy MC. Infection prevention and control research priorities: What do we need to combat healthcare-associated infections and antimicrobial resistance? Results of a narrative literature review and survey analysis. *Antimicrobial Resistance and Infection Control*. 2020;9(1):1-10. doi:10.1186/s13756-020-00801-x

9. The right prevention and treatment for the right patient at the right time. Strategic Research Agenda for Innovative Medicines Initiative. Innovative Medicines Initiative. https://www.imi.europa.eu/sites/default/files/uploads/documents/About-

IMI/research-agenda/IMI2_SRA_March2014.pdf. Published 2014. Accessed May 24, 2021.

10. Rivm. Strategic Research Agenda The challenge To enhance the prevention, detection and control of zoonoses and antimicrobial resistance. The One Health European Joint Programme. https://onehealthejp.eu/wp-content/uploads/2018/12/One-Health-EJP-Strategic-Research-Agenda.pdf. Published 2019. Accessed May 24, 2021.

11. Burgers JS, Wittenberg J, Keuken DG, et al. Development of a research agenda for general practice based on knowledge gaps identified in Dutch guidelines and input from 48 stakeholders. *European Journal of General Practice*. 2019;25(1):19-24. doi:10.1080/13814788.2018.1532993

12. Uzzi B, Mukherjee S, Stringer M, et al. Atypical Combinations and Scientific Impact Published by : American Association for the Advancement of Science Linked references are available on JSTOR for this article : Atypical Combinations and Scientific Impact. 2020;342(6157):468-472.

13. Schilling MA, Green E. Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. *Research Policy*. 2011;40(10):1321-1331. doi:10.1016/j.respol.2011.06.009

14. Larsson DGJ, Andremont A, Bengtsson-Palme J, et al. Critical knowledge gaps and research needs related to the environmental dimensions of antibiotic resistance. *Environment International*. 2018;117(April):132-138. doi:10.1016/j.envint.2018.04.041

15. Luz CF, Niekerk JM van, Keizer J, et al. Mapping Twenty Years of Antimicrobial Resistance Research Trends. *bioRxiv*. 2021. doi:10.1101/2021.03.01.433375

16. Birkegård AC, Halasa T, Toft N, et al. Send more data: a systematic review of mathematical models of antimicrobial resistance. *Antimicrobial Resistance & Infection Control*. 2018;7(1):1-12. doi:10.1186/s13756-018-0406-1

17. van Kleef E, Robotham J V., Jit M, et al. Modelling the transmission of healthcare associated infections: A systematic review. *BMC Infectious Diseases*. 2013;13(1). doi:10.1186/1471-2334-13-294

18. Gibbons C, Bruce J, Carpenter J, et al. Identification of risk factors by systematic review and development of risk-adjusted models for surgical site

infection. *Health Technology Assessment*. 2011;15(30):3-156. doi:10.3310/hta15300

19. Allegranzi B, Pittet D. Role of hand hygiene in healthcare-associated infection prevention. *Journal of Hospital Infection*. 2009;73(4):305-315. doi:10.1016/j.jhin.2009.04.019

20. Toney-Butler TJ, Carver N. *Hand Washing (Hand Hygiene)*. StatPearls Publishing, Treasure Island (FL); 2019. http://europepmc.org/books/NBK470254.

21. Marra AR, Edmond MB. New technologies to monitor healthcare worker hand hygiene. *Clinical Microbiology and Infection*. 2014;20(1):29-33. doi:10.1111/1469-0691.12458

22. Misra P, Rajaraman V, Aishwarya SN, et al. CleanHands. *Proceedings of the 14th International Conference on Information Processing in Sensor Networks - IPSN '15*. 2015:348-349. doi:10.1145/2737095.2742928

23. Pineles LL, Morgan DJ, Limper HM, et al. Accuracy of a radiofrequency identification (RFID) badge system to monitor hand hygiene behavior during routine clinical activities. *American Journal of Infection Control*. 2014;42(2):144-147. doi:10.1016/j.ajic.2013.07.014

24. Abugabah A, Nizamuddin N, Abuqabbeh A. A review of challenges and barriers implementing RFID technology in the Healthcare sector. In: *Procedia Computer Science*. Vol 170. Elsevier B.V.; 2020:1003-1010. doi:10.1016/j.procs.2020.03.094

25. Najafi M, Laskowski M, de Boer PT, et al. The Effect of Individual Movements and Interventions on the Spread of Influenza in Long-Term Care Facilities. *Medical Decision Making*. 2017;37(8):871-881. doi:10.1177/0272989X17708564

26. Kong F, Paterson DL, Whitby M, et al. A hierarchical spatial modelling approach to investigate MRSA transmission in a tertiary hospital. *BMC Infectious Diseases*. 2013;13(1). doi:10.1186/1471-2334-13-449

27. Haddara M, Staaby A. RFID applications and adoptions in healthcare: A review on patient safety. *Procedia Computer Science*. 2018;138:80-88. doi:10.1016/j.procs.2018.10.012

28. Id HT, Lo LE, Xia H, et al. Relevance of intra-hospital patient movements for the spread of healthcare- associated infections within hospitals - a mathematical modeling study. 2021:1-23. doi:10.1371/journal.pcbi.1008600

29. McHaney-Lindstrom M, Hebert C, Miller H, et al. Network analysis of intra-hospital transfers and hospital onset clostridium difficile infection. *Health Information and Libraries Journal*. 2020;37(1):26-34. doi:10.1111/hir.12274

30. Bush K, Barbosa H, Farooq S, et al. Inpatient mobility to predict hospital-onset Clostridium difficile: A network approach. *bioRxiv*. 2018;265. doi:10.1101/404160

31. Naylor NR, Atun R, Zhu N, et al. Estimating the burden of antimicrobial resistance: a systematic literature review. *Antimicrobial resistance and infection control*. 2018;7:58. doi:10.1186/s13756-018-0336-y

32. FAO, OIE, WHO, et al. Contributing to One World, One Health. 2008;(October):3-67.

33. Kim DW, Cha CJ. Antibiotic resistome from the One-Health perspective: understanding and controlling antimicrobial resistance transmission. *Experimental and Molecular Medicine*. 2021. doi:10.1038/s12276-021-00569-z

34. Burgman MA. *Trusting Judgements: How to Get the Best out of Experts*. Cambridge University Press; 2016.

35. Banks GC, Woznyj HM, Wesslen RS, et al. A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App). *Journal of Business and Psychology*. 2018;33(4):445-459. doi:10.1007/s10869-017-9528-3

36. Zar JH. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*. 1972;67(339):578-580. doi:10.1080/01621459.1972.10481251

37. Haque M, Sartelli M, McKimm J, et al. Health care-associated infections – An overview. *Infection and Drug Resistance*. 2018;11:2321-2333. doi:10.2147/IDR.S177247

38. Badia JM, Casey AL, Petrosillo N, et al. Impact of surgical site infection on healthcare costs and patient outcomes: a systematic review in six European countries. *Journal of Hospital Infection*. 2017;96(1):1-15. doi:10.1016/j.jhin.2017.03.004

39. Strobl C, Malley J, Gerhard Tutz. Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol Methods*. 2009;14(4):323-348. doi:10.1037/a0016973.An

40. Streefkerk RHRA, Moorman PW, Parlevliet GA, et al. An Automated Algorithm to Preselect Patients to Be Assessed Individually in Point Prevalence Surveys for Hospital-Acquired Infections in Surgery. *Infection Control and Hospital Epidemiology*. 2014;35(7):886-887. doi:10.1086/676868

41. In Lee K, Koval JJ. Determination of the best significance level in forward stepwise logistic regression. *Communications in Statistics - Simulation and Computation*. 2007;26(2):559-575. doi:10.1080/03610919708813397

42. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees. 2019.

43. Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*. 2015;48(9):2839-2846.

44. Sickbert-Bennett EE, Dibiase LM, Schade Willis TM, et al. Reduction of healthcare-associated infections by exceeding high compliance with hand hygiene practices. *Emerging Infectious Diseases*. 2016;22(9):1628-1630. doi:10.3201/eid2209.151440

45. Hornbeck T, Naylor D, Segre AM, et al. Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *Journal of Infectious Diseases*. 2012;206(10):1549-1557. doi:10.1093/infdis/jis542

46. Datta R, Juthani-Mehta M. Burden and management of multidrug-resistant organisms in palliative care. *Palliative Care*. 2017;10. doi:10.1177/1178224217749233

47. O'Driscoll T, Crank CW. Vancomycin-resistant enterococcal infections: Epidemiology, clinical manifestations, and optimal management. *Infection and Drug Resistance*. 2015;8:217-230. doi:10.2147/IDR.S54125

48. Kramer A, Schwebke I, Kampf G. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. *BMC Infectious Diseases*. 2006;6:1-8. doi:10.1186/1471-2334-6-130

49. McHaney-Lindstrom M, Hebert C, Miller H, et al. Network analysis of intra-hospital transfers and hospital onset clostridium difficile infection. *Health Information and Libraries Journal*. 2020;37(1):26-34. doi:10.1111/hir.12274

50. Gleich DF. PageRank Beyond the Web. *SIAM Review*. 2015;57(3):321-363. doi:10.1137/140976649

51. Kingsford C, Salzberg SL. What are decision trees? *Nature Biotechnology*. 2008;26(9):1011-1012. doi:10.1038/nbt0908-1011

52. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. *The Mathematical Intelligencer*. 2009;27(2):764. doi:10.1007/b94608

53. Jasovský D, Littmann J, Zorzet A, et al. Antimicrobial resistance—a threat to the world's sustainable development. *Upsala Journal of Medical Sciences*. 2016;121(3):159-164. doi:10.1080/03009734.2016.1195900

54. Bryan-Wilson J. No time to wait. *Artforum International*. 2016;54(10):113-114.

55. Dunachie SJ, Day NP, Dolecek C. The challenges of estimating the human global burden of disease of antimicrobial resistant bacteria. *Current Opinion in Microbiology*. 2020;57:95-101. doi:10.1016/j.mib.2020.09.013

56. Walsh TR. A one-health approach to antimicrobial resistance. *Nature Microbiology*. 2018;3(8):854-855. doi:10.1038/s41564-018-0208-5

57. Alghamdi R, Alfalqi K. A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*. 2015;6(1):147-153. doi:10.14569/ijacsa.2015.060121

58. Subbaswamy KR. Brillouin Scattering From Thermal Fluctuations in Superionic Conductors. *Solid State Communications*. 1977;21(4):371-372. doi:10.1016/0038-1098(77)91248-0

59. Westgate MJ, Barton PS, Pierson JC, et al. Text analysis tools for identification of emerging topics and research gaps in conservation science. *Conservation Biology*. 2015;29(6):1606-1614. doi:10.1111/cobi.12605

60. Richter RR, Austin TM. Using MeSH (Medical Subject Headings) to enhance PubMed search strategies for evidence-based practice in physical therapy. *Physical Therapy*. 2012;92(1):124-132. doi:10.2522/ptj.20100178

61. Porter MF. Snowball: A language for stemming algorithms. 2001.

62. Google Developers. Google Maps Platform. https://developers.google.com/maps/documentation. Accessed November 13, 2020.

63. Kumar S. Analyzing the Facebook workload. *Proceedings - 2012 IEEE International Symposium on Workload Characterization, IISWC 2012*. 2012:111-112. doi:10.1109/IISWC.2012.6402911

64. Roberts ME, Stewart BM, Tingley D, et al. The structural topic model and applied social science. *NIPS 2013 Workshop on Topic Models*. 2013:2-5.

65. Roberts ME, Stewart BM, Airoldi EM. A model of text for experimentation in the social sciences ACCEPTED MANUSCRIPT A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*. 2016;1459(March):988-1003. doi:10.1080/01621459.2016.1141684

66. Roberts ME, Stewart BM, Tingley D, et al. Structural topic models for open-ended survey responses. *American Journal of Political Science*. 2014;58(4):1064-1082. doi:10.1111/ajps.12103

67. Roberts M. Structural Topic Models Topic models Methods of unsupervised text analysis. 2017.

68. Roberts ME, Stewart BM, Airoldi EM. A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*. 2016;111(515):988-1003. doi:10.1080/01621459.2016.1141684

69. Roberts ME, Stewart BM, Tingley D. Navigating the Local Modes of Big Data: The Case of Topic Models. *Computational Social Science*. 2016:51-97. doi:10.1017/cbo9781316257340.004

70. Roberts ME, Stewart BM, Tingley D. Stm: An R package for structural topic models. *Journal of Statistical Software*. 2019;91(2). doi:10.18637/jss.v091.i02

71. Lee M, Mimno D. Low-dimensional embeddings for interpretable anchor-based topic inference. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2014:1319-1328. doi:10.3115/v1/d14-1138

72. Chen X, Zou D, Cheng G, et al. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. *Computers and*

Education. 2020;151(September 2019):103855. doi:10.1016/j.compedu.2020.103855

73. Arora S, Ge R, Halpern Y, et al. A practical algorithm for topic modeling with provable guarantees. *30th International Conference on Machine Learning, ICML 2013*. 2013;28(PART 2):939-947.

74. Lau JH, Newman D, Karimi S, et al. Best topic word selection for topic labelling. *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*. 2010;2(August):605-613.

75. Ghoshdastidar D, Perrot M, Von Luxburg U. Foundations of comparison-based hierarchical clustering. *arXiv*. 2018;(NeurIPS).

76. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2001;63(2):411-423. doi:10.1111/1467-9868.00293

77. Robinson KA, Saldanha IJ, Mckoy NA. Frameworks for determining research gaps during systematic reviews. 2011.

78. Team RC. R: A language and environment for statistical computing. 2019;3.

79. Package T, Kovalchik AS. Package ' RISmed .' 2014.

80. Fantini D. easyPubMed: Search and Retrieve Scientific Publication Records from PubMed. 2019.

81. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *Journal of Open Source Software*. 2019;4(43):1686.

82. Moore P, Kyne L, Martin A, et al. Germination efficiency of clinical Clostridium difficile spores and correlation with ribotype, disease severity and therapy failure. *Journal of Medical Microbiology*. 2013;62:1405-1413. doi:10.1099/jmm.0.056614-0

83. Eggers S, Safdar N, Malecki KM. Heavy metal exposure and nasal Staphylococcus aureus colonization: analysis of the National Health and Nutrition Examination Survey (NHANES). *Environmental health : a global access science source*. 2018;17(1):2. doi:10.1186/s12940-017-0349-7

84. Dyar OJ, Huttner B, Schouten J, et al. What is antimicrobial stewardship? *Clinical Microbiology and Infection*. 2017;23(11):793-798. doi:10.1016/j.cmi.2017.08.026

85. Morjaria S, Chapin KC. Who to Test, When, and for What: Why Diagnostic Stewardship in Infectious Diseases Matters. *Journal of Molecular Diagnostics*. 2020;22(9):1109-1113. doi:10.1016/j.jmoldx.2020.06.012

86. Karaiskos I, Giamarellou H. Carbapenem-sparing strategies for ESBL producers: When and how. *Antibiotics*. 2020;9(2). doi:10.3390/antibiotics9020061

87. Walker KJ, Lee YR, Klar AR. Clinical Outcomes of Extended-Spectrum Beta-Lactamase-Producing Enterobacteriaceae Infections with Susceptibilities among Levofloxacin, Cefepime, and Carbapenems. *Canadian Journal of Infectious Diseases and Medical Microbiology*. 2018;2018. doi:10.1155/2018/3747521

88. John R, Colley P, Nguyen HL, et al. Outcomes analysis in patients with extended-spectrum beta-lactamase bacteremia empirically treated with piperacillin/tazobactam versus carbapenems. *Baylor University Medical Center Proceedings*. 2019;32(2):187-191. doi:10.1080/08998280.2019.1582466

89. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*. 2019;20(6):356-370. doi:10.1038/s41576-019-0108-4

90. Angers-Loustau A, Petrillo M, Bengtsson-Palme J, et al. The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. *F1000Research*. 2018;7:459. doi:10.12688/f1000research.14509.1

91. Hendriksen RS, Munk P, Njage P, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications*. 2019;10(1). doi:10.1038/s41467-019-08853-3

92. Lien LTQ, Hoa NQ, Chuc NTK, et al. Antibiotics in wastewater of a rural and an urban hospital before and after wastewater treatment, and the relationship with antibiotic use-a one year study from Vietnam. *International Journal of Environmental Research and Public Health*. 2016;13(6):1-13. doi:10.3390/ijerph13060588

93. Volkoff SJ, McCumber AW, Anderson DJ, et al. Antibiotic-resistant bacteria on personal devices in hospital intensive care units: Molecular approaches to quantifying and describing changes in the bacterial community of

personal mobile devices. *Infection Control and Hospital Epidemiology*. 2019;40(6):717-720. doi:10.1017/ice.2019.56

94. Spicknall IH, Foxman B, Marrs CF, et al. A modeling framework for the evolution and spread of antibiotic resistance: Literature review and model categorization. *American Journal of Epidemiology*. 2013;178(4):508-520. doi:10.1093/aje/kwt017

95. Hebert C, Root ED. Repurposing Geographic Information Systems for Routine Hospital Infection Control. *Advances in health care management*. 2019;18(1):1-13. doi:10.1108/S1474-823120190000018003

96. Opatowski L, Guillemot D, Boëlle PY, et al. Contribution of mathematical modeling to the fight against bacterial antibiotic resistance. *Current Opinion in Infectious Diseases*. 2011;24(3):279-287. doi:10.1097/QCO.0b013e3283462362

97. Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. *Science*. 2013;342(6164):1337-1342. doi:10.1126/science.1245200

98. European Centre for Disease Prevention and Control. *Surveillance of Surgical Site Infections in European Hospitals – HAISSI Protocol*.; 2012. doi:10.2900/12819

99. Zarb P, Coignard B, Griskeviciene J, et al. *Point Prevalence Survey of Healthcare-Associated Infections and Antimicrobial Use in European Acute Care Hospitals*. Vol 17.; 2012. doi:10.2900/86011

100. Roy S, Patkar A, Daskiran M, et al. Clinical and Economic Burden of Surgical Site Infection in Hysterectomy. *Surgical Infections*. 2014;15(3):266-273. doi:10.1089/sur.2012.163

101. Leung Wai Sang S, Chaturvedi R, Alam A, et al. Preoperative hospital length of stay as a modifiable risk factor for mediastinitis after cardiac surgery. *Journal of Cardiothoracic Surgery*. 2013;8(1):45. doi:10.1186/1749-8090-8-45

102. Triantafyllopoulos G, Stundner O, Memtsoudis S, et al. Patient, surgery, and hospital related risk factors for surgical site infections following total hip arthroplasty. *Scientific World Journal*. 2015;2015:1-9. doi:10.1155/2015/979560

103. Abuzaid A, Zaki M, Al Tarief H. Potential risk factors for surgical site infection after isolated coronary artery bypass grafting in a Bahrain Cardiac Centre: A retrospective, case-controlled study. *Heart Views*. 2015;16(3):79. doi:10.4103/1995-705x.164457

104. Schimmel JJP, Horsting PP, De Kleuver M, et al. Risk factors for deep surgical site infections after spinal fusion. *European Spine Journal*. 2010;19(10):1711-1719. doi:10.1007/s00586-010-1421-y

105. Guohua X, cheng keping, Li J, et al. Risk factors for surgical site infection in a teaching hospital: a prospective study of 1,138 patients. *Patient Preference and Adherence*. 2015;9:1171. doi:10.2147/ppa.s86153

106. Gomila A, Carratalà J, Biondo S, et al. Predictive factors for early- and late-onset surgical site infections in patients undergoing elective colorectal surgery. A multicentre, prospective, cohort study. *Journal of Hospital Infection*. 2018;99(1):24-30. doi:10.1016/j.jhin.2017.12.017

107. Martin ET, Kaye KS, Knott C, et al. Diabetes and risk of surgical site infection: A systematic review and meta-analysis. *Infection Control and Hospital Epidemiology*. 2016;37(1):88-99. doi:10.1017/ice.2015.249

108. Takahashi Y, Takesue Y, Fujiwara M, et al. Risk factors for surgical site infection after major hepatobiliary and pancreatic surgery. *Journal of Infection and Chemotherapy*. 2018;24(9):739-743. doi:10.1016/j.jiac.2018.05.007

109. Silvestri M, Dobrinja C, Scomersi S, et al. Modifiable and non-modifiable risk factors for surgical site infection after colorectal surgery: a single-center experience. *Surgery Today*. 2018;48(3):338-345. doi:10.1007/s00595-017-1590-y

110. Fukuda H. Patient-related risk factors for surgical site infection following eight types of gastrointestinal surgery. *Journal of Hospital Infection*. 2016;93(4):347-354. doi:10.1016/j.jhin.2016.04.005

111. Kokudo T, Uldry E, Demartines N, et al. Risk factors for incisional and organ space surgical site infections after liver resection are different. *World Journal of Surgery*. 2015;39(5):1185-1192. doi:10.1007/s00268-014-2922-3

112. Isik O, Kaya E, Dundar HZ, et al. Surgical Site Infection: Re-assessment of the Risk Factors. *Chirurgia (Bucharest, Romania : 1990)*. 2015;110(5):457-461.

113. Moreno Elola-Olaso A, Davenport DL, Hundley JC, et al. Predictors of surgical site infection after liver resection: A multicentre analysis using National Surgical Quality Improvement Program data. *Hpb*. 2012;14(2):136-141. doi:10.1111/j.1477-2574.2011.00417.x

114. Lola I, Levidiotou S, Petrou A, et al. Are there independent predisposing factors for postoperative infections following open heart surgery? *Journal of Cardiothoracic Surgery*. 2011;6(1):151. doi:10.1186/1749-8090-6-151

115. Chen LF, Arduino JM, Sheng S, et al. Epidemiology and outcome of major postoperative infections following cardiac surgery: Risk factors and impact of pathogen type. *American Journal of Infection Control*. 2012;40(10):963-968. doi:10.1016/j.ajic.2012.01.012

116. Jolivet S, Lescure FX, Armand-Lefevre L, et al. Surgical site infection with extended-spectrum β-lactamase-producing Enterobacteriaceae after cardiac surgery: incidence and risk factors. *Clinical Microbiology and Infection*. 2018;24(3):283-288. doi:10.1016/j.cmi.2017.07.004

117. Di Gennaro F, Marotta C, Pisani L, et al. Maternal caesarean section infection (MACSI) in Sierra Leone: a case-control study. *Epidemiology and Infection*. 2020:1-6. doi:10.1017/S0950268820000370

118. European Centre for Disease Prevention and Control. Healthcare-associated infections: surgical site infections. *ECDC Annual epidemiological report for 2016 Stockholm: ECDC*. 2018;(May).

119. KUNUTSOR SK, WHITEHOUSE MR, BLOM AW, et al. Systematic review of risk prediction scores for surgical site infection or periprosthetic joint infection following joint arthroplasty. *Epidemiology and Infection*. 2017;145(9):1738-1749. doi:10.1017/s0950268817000486

120. Fiorini N, Canese K, Starchenko G, et al. Best Match: New relevance search for PubMed. *PLoS Biology*. 2018;16(8):1-12. doi:10.1371/journal.pbio.2005343

121. Aldiabat KM, Le Navenec CL. Data saturation: The mysterious step in grounded theory methodology. *Qualitative Report*. 2018;23(1):245-261.

122. Streefkerk RHRA, Borsboom GJJM, van der Hoeven CP, et al. Evaluation of an Algorithm for Electronic Surveillance of Hospital-Acquired Infections Yielding Serial Weekly Point Prevalence Scores. *Infection Control & Hospital Epidemiology*. 2014;35(07):888-890. doi:10.1086/676869

123. Asendorpf JB, Van De Schoot R, Denissen JJA, et al. Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation. *International Journal of Behavioral Development*. 2014;38(5):453-460. doi:10.1177/0165025414542713

124. van Buuren S. Multiple imputation of discrete and continuous. *Statistical Methods in Medical Research*. 2007;16(3):219-242.

125. Hosmer DW, Lemeshow S. *Applied Logistic Regression*.; 2000. doi:10.1080/00401706.1992.10485291

126. R Core Team. R: A Language and Environment for Statistical Computing. 2020.

127. van Buuren S, Groothuis-Oudshoorn K. {mice}: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45(3):1-67.

128. Jopia H. smbinning: Scoring Modeling and Optimal Binning. 2019.

129. Wickham H, François R, Henry L, et al. dplyr: A Grammar of Data Manipulation. 2019.

130. Harrison E, Drake T, Ots R. finalfit: Quickly Create Elegant Regression Results Tables and Plots when Modelling. 2019.

131. Xie S. scorecard: Credit Risk Scorecard. 2020.

132. Araki T, Okita Y, Uchino M, et al. Risk factors for surgical site infection in Japanese patients with ulcerative colitis: A multicenter prospective study. *Surgery Today*. 2014;44(6):1072-1078. doi:10.1007/s00595-013-0809-9

133. Ben-Ami E, Levy I, Katz J, et al. Risk factors for sternal wound infection in children undergoing cardiac surgery: a case-control study. *Journal of Hospital Infection*. 2008;70(4):335-340. doi:10.1016/j.jhin.2008.08.010

134. Ejaz A, Schmidt C, Johnston FM, et al. Risk factors and prediction model for inpatient surgical site infection after major abdominal surgery. *Journal of Surgical Research*. 2017;217:153-159. doi:10.1016/j.jss.2017.05.018

135. Giri S, Kandel BP, Pant S, et al. Risk factors for surgical site infections in abdominal surgery: A study in Nepal. *Surgical Infections*. 2013;14(3):313-318. doi:10.1089/sur.2012.108

136. Hijas-Gómez AI, Egea-Gámez RM, Martínez-Martín J, et al. Surgical Wound Infection Rates and Risk Factors in Spinal Fusion in a University Teaching

Hospital in Madrid, Spain. *Spine*. 2017;42(10):748-754. doi:10.1097/BRS.0000000000001916

137. Liu S, Miao J, Wang G, et al. Risk factors for postoperative surgical site infections in patients with Crohn's disease receiving definitive bowel resection. *Scientific Reports*. 2017;7(1):1-6. doi:10.1038/s41598-017-10603-8

138. McKenzie Stancu S, Iordache F. Thrombocytosis Is a Risk Factor for Surgical Site Infections after Colon Resection: A Prospective Observational Study. *Surgical infections*. 2019;20(1):39-44. doi:10.1089/sur.2018.146

139. Mehta PA, Cunningham CK, Colella CB, et al. Risk factors for sternal wound and other infections in pediatric cardiac surgery patients. *Pediatric Infectious Disease Journal*. 2000;19(10):1000-1004. doi:10.1097/00006454-200010000-00012

140. Morikane K, Honda H, Suzuki S. Factors associated with surgical site infection following gastric surgery in Japan. *Infection Control and Hospital Epidemiology*. 2016;37(10):1167-1172. doi:10.1017/ice.2016.155

141. Perkins JD. Techniques to ensure adequate portal flow in the presence of splenorenal shunts. *Liver Transplantation*. 2007;13(5):767-768. doi:10.1002/lt

142. Shao J, Zhang H, Yin B, et al. Risk factors for surgical site infection following operative treatment of ankle fractures: A systematic review and meta-analysis. *International Journal of Surgery*. 2018;56(May):124-132. doi:10.1016/j.ijsu.2018.06.018

143. Su J, Cao X. Risk factors of wound infection after open reduction and internal fixation of calcaneal fractures. *Medicine (United States)*. 2017;96(44):e8411. doi:10.1097/MD.0000000000008411

144. Xing D, Ma JX, Ma XL, et al. A methodological, systematic review of evidence-based independent risk factors for surgical site infections after spinal surgery. *European Spine Journal*. 2013;22(3):605-615. doi:10.1007/s00586-012-2514-6

145. Zhang JF, Zhu HY, Sun YW, et al. Pseudomonas aeruginosa infection after pancreatoduodenectomy: Risk factors and clinic impacts. *Surgical Infections*. 2015;16(6):769-774. doi:10.1089/sur.2015.041

146. Isik O, Kaya E, Sarkut P, et al. Factors affecting surgical site infection rates in hepatobiliary surgery. *Surgical Infections*. 2015;16(3):281-286. doi:10.1089/sur.2013.195

147. Berríos-Torres SI, Umscheid CA, Bratzler DW, et al. Centers for Disease Control and Prevention Guideline for the Prevention of Surgical Site Infection, 2017. *JAMA Surgery*. 2017;152(8):784-791. doi:10.1001/jamasurg.2017.0904

148. Leaper DJ, Edmiston CE. World Health Organization: global guidelines for the prevention of surgical site infection. *Journal of Hospital Infection*. 2017;95(2):135-136. doi:10.1016/j.jhin.2016.12.016

149. Werkgroep infectiepreventie. Preventie van postoperatieve wondinfecties. 2011:1-26.

150. Mulani MS, Kamble EE, Kumkar SN, et al. Emerging strategies to combat ESKAPE pathogens in the era of antimicrobial resistance: A review. *Frontiers in Microbiology*. 2019;10(APR). doi:10.3389/fmicb.2019.00539

151. Santajit S, Indrawattana N. Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens. *BioMed Research International*. 2016;2016. doi:10.1155/2016/2475067

152. Weber KL, Lesassier DS, Kappell AD, et al. Simulating transmission of ESKAPE pathogens plus C. difficile in relevant clinical scenarios. *BMC Infectious Diseases*. 2020;20(1):1-15. doi:10.1186/s12879-020-05121-4

153. Pittet D, Allegranzi B, Sax H, et al. Evidence-based model for hand transmission during patient care and the role of improved practices. *Lancet Infectious Diseases*. 2006;6(10):641-652. doi:10.1016/S1473-3099(06)70600-4

154. Cheng CH, Kuo YH, Zhou Z. Tracking Nosocomial Diseases at Individual Level with a Real-Time Indoor Positioning System. *Journal of Medical Systems*. 2018;42(11). doi:10.1007/s10916-018-1085-4

155. Chassin MR, Mayer C, Nether K. Improving hand hygiene at eight hospitals in the United States by targeting specific causes of noncompliance. *Joint Commission Journal on Quality and Patient Safety*. 2015;41(1):4-12. doi:10.1016/S1553-7250(15)41002-5

156. Hofmann DA, Staats BR, Dai H, et al. The Impact of Time at Work and Time off from Work on Rule Compliance: The Case of Hand Hygiene in Healthcare. *Journal of Applied Psychology*. 2015;100(3):846-862.

157. Girou E, Loyeau S, Legrand P, et al. Efficacy of handrubbing with alcohol based solution versus standard handwashing with antiseptic soap: randomised clinical trial. *BMJ (Clinical research ed)*. 2002;325(7360):362. doi:10.1136/bmj.325.7360.362

158. Hornbeck T, Naylor D, Segre AM, et al. Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *Journal of Infectious Diseases*. 2012;206(10):1549-1557. doi:10.1093/infdis/jis542

159. Temime L, Opatowski L, Pannet Y, et al. Peripatetic health-care workers as potential superspreaders. *Proceedings of the National Academy of Sciences*. 2009;106(43):18420-18425. doi:10.1073/pnas.0900974106

160. Lopez-Garcia M, Aruru M, Pyne S. Health analytics and disease modeling for better understanding of healthcare-associated infections. *BLDE University Journal of Health Sciences*. 2018;3(2):69-74. doi:10.4103/bjhs.bjhs_36_18

161. Ciavarella C, Fumanelli L, Merler S, et al. School closure policies at municipality level for mitigating influenza spread: A model-based evaluation. *BMC Infectious Diseases*. 2016;16(1):1-11. doi:10.1186/s12879-016-1918-z

162. Salathe M, Kazandjieva M, Lee JW, et al. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*. 2010;107(51):22020-22025. doi:10.1073/pnas.1009094108

163. Smieszek T, Castell S, Barrat A, et al. Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: Method comparison and participants' attitudes. *BMC Infectious Diseases*. 2016;16(1):1-14. doi:10.1186/s12879-016-1676-y

164. Ozella L, Gesualdo F, Tizzoni M, et al. Close encounters between infants and household members measured through wearable proximity sensors. *PLoS ONE*. 2018;13(6):1-16. doi:10.1371/journal.pone.0198733

165. Voirin N, Payet C, Barrat A, et al. Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care

hospital. *Infection Control and Hospital Epidemiology*. 2015;36(3):254-260. doi:10.1017/ice.2014.53

166. English KM, Langley JM, McGeer A, et al. Contact among healthcare workers in the hospital setting: Developing the evidence base for innovative approaches to infection control. *BMC Infectious Diseases*. 2018;18(1):1-12. doi:10.1186/s12879-018-3093-x

167. Mastrandrea R, Soto-Aladro A, Brouqui P, et al. Enhancing the evaluation of pathogen transmission risk in a hospital by merging hand-hygiene compliance and contact data: A proof-of-concept study Public Health. *BMC Research Notes*. 2015;8(1):426. doi:10.1186/s13104-015-1409-0

168. Vanhems P, Barrat A, Cattuto C, et al. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS one*. 2013;8(9):e73970. doi:10.1371/journal.pone.0073970

169. Ozella L, Gauvin L, Carenzo L, et al. Wearable Proximity Sensors for Monitoring a Mass Casualty Incident Exercise: Feasibility Study. *Journal of medical Internet research*. 2019;21(4):e12251. doi:10.2196/12251

170. Machens A, Gesualdo F, Rizzo C, et al. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. *BMC Infectious Diseases*. 2013;13(1):185. doi:10.1186/1471-2334-13-185

171. Isella L, Romano M, Barrat A, et al. Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS ONE*. 2011;6(2). doi:10.1371/journal.pone.0017144

172. Najafi M, Laskowski M, De Boer PT, et al. The Effect of Individual Movements and Interventions on the Spread of Influenza in Long-Term Care Facilities. *Medical Decision Making*. 2017;37(8):871-881. doi:10.1177/0272989X17708564

173. Assab R, Nekkab N, Crépey P, et al. Mathematical models of infection transmission in healthcare settings: Recent advances from the use of network structured data. *Current Opinion in Infectious Diseases*. 2017;30(4):410-418. doi:10.1097/QCO.0000000000000390

174. Kim B, Barrat A, Khanafer N, et al. Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE*. 2013;8(9):e73970. doi:10.1371/journal.pone.0073970

175. Chen H, Yang B, Pei H, et al. Next Generation Technology for Epidemic Prevention and Control: Data-Driven Contact Tracking. *IEEE Access*. 2019;7:2633-2642. doi:10.1109/ACCESS.2018.2882915

176. Iozzi F, Trusiano F, Chinazzi M, et al. Little italy: An agent-based approach to the estimation of contact patterns- fitting predicted matrices to serological data. *PLoS Computational Biology*. 2010;6(12). doi:10.1371/journal.pcbi.1001021

177. Génois M, Barrat A. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*. 2018;7(1):11. doi:10.1140/epjds/s13688-018-0140-1

178. Woolhouse MEJ, Dye C, Etard J-F, et al. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences*. 2002;94(1):338-342. doi:10.1073/pnas.94.1.338

179. Jit M, Brisson M. Modelling the epidemiology of infectious diseases for decision analysis. *Pharmacoeconomics*. 2011;29(5):371-386.

180. Banks HT, Broido A, Canter B, et al. Simulation algorithms for continuous time Markov chain models. *Studies in Applied Electromagnetics and Mechanics*. 2012;37:3-18. doi:10.3233/978-1-61499-092-5-3

181. Satilmis L, Vanhems P, Bénet T. Outbreaks of Vancomycin-resistant enterococci in hospital settings: A systematic review and calculation of the basic reproductive number. *Infection Control and Hospital Epidemiology*. 2015;37(3):289-294. doi:10.1017/ice.2015.301

182. Mutters NT, Mersch-Sundermann V, Mutters R, et al. Control of the spread of vancomycin-resistant enterococci in hospitals: Epidemiology and clinical relevance. *Deutsches Arzteblatt International*. 2013;110(43):725-732. doi:10.3238/arztebl.2013.0725

183. García Martínez de Artola D, Castro B, Ramos MJ, et al. Outbreak of vancomycin-resistant enterococcus on a haematology ward: management and control. *Journal of Infection Prevention*. 2017;18(3):149-153. doi:10.1177/1757177416687832

184. Blanco N, O'hara LM, Harris AD. Transmission pathways of multidrug-resistant organisms in the hospital setting: A scoping review. *Infection*

Control and Hospital Epidemiology. 2019;40(4):447-456. doi:10.1017/ice.2018.359

185. Frakking FNJ, Bril WS, Sinnige JC, et al. Recommendations for the successful control of a large outbreak of vancomycin-resistant Enterococcus faecium in a non-endemic hospital setting. *Journal of Hospital Infection*. 2018;100(4):e216-e225. doi:10.1016/j.jhin.2018.02.016

186. Drees M, Snydman DR, Schmid CH, et al. Prior Environmental Contamination Increases the Risk of Acquisition of Vancomycin-Resistant Enterococci. *Clinical Infectious Diseases*. 2008;46(5):678-685. doi:10.1086/527394

187. Mcdermott H, Skally M, O'rourke J, et al. Vancomycin-Resistant Enterococci (VRE) in the intensive care unit in a nonoutbreak setting: Identification of potential reservoirs and epidemiological associations between patient and environmental VRE. *Infection Control and Hospital Epidemiology*. 2018;39(1):40-45. doi:10.1017/ice.2017.248

188. Huang SS, Datta R, Platt R. Risk of Acquiring Antibiotic-Resistant Bacteria From Prior Room Occupants. *JAMA Internal Medicine*. 2006;166(18):1945-1951. doi:10.1001/archinte.166.18.1945

189. Mitchell BG, Dancer SJ, Anderson M, et al. Risk of organism acquisition from prior room occupants: A systematic review and meta-analysis. *Journal of Hospital Infection*. 2015;91(3):211-217. doi:10.1016/j.jhin.2015.08.005

190. Ford CD, Lopansri BK, Gazdik MA, et al. Room contamination, patient colonization pressure, and the risk of vancomycin-resistant Enterococcus colonization on a unit dedicated to the treatment of hematologic malignancies and hematopoietic stem cell transplantation. *American Journal of Infection Control*. 2016;44(10):1110-1115. doi:10.1016/j.ajic.2016.03.044

191. Pan SC, Wang JT, Chen YC, et al. Incidence of and Risk Factors for Infection or Colonization of Vancomycin-Resistant Enterococci in Patients in the Intensive Care Unit. *PLoS ONE*. 2012;7(10):1-8. doi:10.1371/journal.pone.0047297

192. Dik JWH, Hendrix R, Poelman R, et al. Measuring the impact of antimicrobial stewardship programs. *Expert Review of Anti-Infective Therapy*. 2016;14(6):569-575. doi:10.1080/14787210.2016.1178064

193. Kaya A, Kaya SY, Balkan II, et al. Risk factors for development of vancomycin-resistant enterococcal bacteremia among VRE colonizers: A retrospective

case control study. *Wiener Klinische Wochenschrift*. 2021;133(9-10):478-483. doi:10.1007/s00508-020-01733-7

194. Rozenshtein P, Gionis A, Prakash BA, et al. Reconstructing an epidemic over time. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016;13-17-Augu(Ic):1835-1844. doi:10.1145/2939672.2939865

195. Breiman L, Friedman J, Stone CJ, et al. *Classification and Regression Trees*. CRC press; 1984.

196. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747

197. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal*. 2006;Complex Systems:1695.

198. Paluszynska A, Biecek P, Jiang Y, et al. Package 'randomForestExplainer.' *Explaining and visualizing random forests in terms of variable importance*. 2017.

199. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3):18-22.

200. Xu S, Zhou H, Wu C, et al. Spatial signal attenuation model of active RFID tags. *IEEE Access*. 2018;6:6947-6960. doi:10.1109/ACCESS.2018.2794556

201. Linero AR. A review of tree-based Bayesian methods. 2017;24(6):543-559.

202. Skarding J, Gabrys B, Member S. Foundations and Modeling of Dynamic Networks Using Dynamic Graph Neural Networks : A Survey. 2021:79143-79168. doi:10.1109/ACCESS.2021.3082932

203. Ma Y, Guo Z, Ren Z, et al. Streaming Graph Neural Networks. 2020:719-728.

Summary

Statistical models are essential to understand the occurrence and spread of microbes to support decision-making in microbiology and epidemiology. Antimicrobial resistance (AMR) is a multifaceted global problem and a significant threat to sustainable modern healthcare. This thesis aims to identify knowledge gaps in the AMR research field and explain the added value of using statistical models and novel spatiotemporal data to predict and identify risk factors for the occurrence and spread of AMR.

Strategic action plans to tackle the increasing international threat of AMR are based upon research agendas that are informed using knowledge gaps in the AMR research field. Currently, these knowledge gaps are identified manually and are often subjective. Chapter 2 describes how bibliometric data-driven methodology can be used to identify knowledge gaps in AMR research. To this end, twenty years of AMR related articles were extracted using the PubMed search engine. With structural topic modelling I identified the topics comprising the AMR research field, while topic clusters were created using hierarchical clustering on the topic proportions. Potential AMR knowledge gaps were obtained using Spearman's correlation between topic clusters and topics and between individual topics. A total of 88 topics and seven topic clusters were identified from 158 616 scientific AMR research articles. In total, 421 potential knowledge gaps were identified between the topic clusters and topics and 2 663 between individual topics. Key knowledge gaps between molecular and laboratory AMR research were highlighted. The knowledge gaps between AMR research regarding water and the environment and both institutional and international surveillance topics were highlighted at the topic level. These results provide an innovative, data-driven way to identify knowledge gaps in AMR research.

Surgical site infections (SSI) make up 19.6% of healthcare-associated infections (HAIs) in Europe [98]. Risk factor identification studies for the occurrence of SSI do not usually specify how continuous variables cut-offs are determined. In most cases, they use standard medical cut-offs without considering the data being studied. Chapter 3 identifies the risk factors for the occurrence of SSI for digestive, thoracic and orthopaedic system surgeries using standard medical and data-driven cut-off values. Retrospective surgical procedure data, individual electronic health records, pharmaceutical data and laboratory data were used from the Erasmus MC University Medical Centre in The Netherlands. Risk factors for the occurrence of SSI were identified using a multivariate forward-step logistic regression model. Standard medical cut-off values were compared with cut-offs determined from the data. For digestive, orthopaedic and thoracic system surgical procedures, the risk factors identified for the occurrence of SSI were preoperative temperature of 38 ^{o}C and

antibiotics used at the time of surgery. C-reactive protein (CRP) and the duration of the surgery were identified as risk factors for digestive surgical procedures. Being an adult (age ≥ 18) was identified as a protective effect for thoracic surgical procedures. Data-driven cut-off values identified for temperature, age, and CRP, explained the occurrence of SSI outcome up to 19.5% better than standard medical cut-off values. Future studies should investigate if data-driven cut-offs can add value to explain the clinical outcome being modelled and not solely rely on standard medical cut-off values for continuous variables to identify risk factors.

Transmission of harmful microorganisms (HMO) poses a major threat to patients and healthcare workers in healthcare settings. The most effective countermeasure against these transmissions is the adherence to hand hygiene policies, but adherence rates are relatively low and vary over space and time. The spatiotemporal effects of varying levels hand hygiene compliance on the transmission and spread of hand-transmitted HMO in a closed healthcare setting must still be quantified. Chapter 4 describes how identifies healthcare worker occupation group of potential super-spreaders and the spatiotemporal effects on the hand transmission of HMO quantified for varying levels of hand hygiene compliance (HHC) caused by this group using their spatiotemporal movements. Spatiotemporal data were collected in the University Medical Center Groningen (UMCG) using radio frequency identification technology. The effects of five probability distributions of HHC and three harmful microorganism transmission rates were simulated using a dynamic agent-based simulation model. The effects of initial simulation assumptions on the simulation results were quantified using five risk outcomes. Nurses were identified as the potential super-spreader healthcare worker occupation group. During lack of HHC (5%) and high transmission rates (5% per contact moment), a colonised nurse can transfer microbes to three of the 17 healthcare worker or patients encountered during the 98.4 minutes of visiting 23 rooms while colonised. The HMO transmission potential for nurses is higher during weeknights (5 pm – 7 am) and weekends as compared to weekdays (7 am – 5 pm). Spatiotemporal behaviour and social mixing patterns of healthcare can change the expected number of hand transmissions and spread HMO by super-spreaders in a closed healthcare setting. These insights can be used to evaluate spatiotemporal safety behaviours and develop infection prevention and control strategies.

Vancomycin-resistant enterococci (VRE) is can cause severe patient health and monetary burdens. The odds of a hospital patient acquiring VRE increases when using antibiotics and when prior room occupants had VRE, but the antibiotic use of prior room occupants are often neglected. Chapter 5 describes how the occurrence and spread of VRE can be explained using intrahospital patient movements (IPM) and their antibiotic use between hospital wards. Retrospective IPM, antibiotic use and

PCR screening data were used from a hospital in the Netherlands. A dynamic directed spatiotemporal graph was developed, and together with the PageRank algorithm used to calculate two daily centrality measures to summarise the flow of patients and antibiotics at the ward level. With a decision tree and random forest model I predicted the daily occurrence of VRE for every ward and compared the models' performance using a 30% test sample. The decision tree model produced a simple set of rules that can be used to determine the daily probability of VRE occurrence for each hospital ward. The decision tree model achieved an area under the curve of 0.685 and the random forest model 0.886 on the test set. These results confirm that the random forest model performs better than a single decision tree for all levels of model sensitivity and specificity at the cost of model simplicity. An early warning system for VRE can be developed and inform infection prevention plans and outbreak strategies further using these results.

In summary, this thesis showed that data-driven statistical models can improve our understanding of antimicrobial resistance. It considers how different sources of spatiotemporal data may be used to predict its occurrence and spread of AMR in hospitals.

Samenvatting

Antimicrobiële resistentie (AMR) vormt een belangrijke bedreiging voor de volksgezondheid en is een wereldwijd domein-overstijgend probleem. Voor de aanpak van AMR binnen de humane gezondheidszorg kunnen statistische modellen een belangrijke bijdrage leveren voor besluitvorming in de microbiologie en epidemiologie.

Dit proefschrift heeft tot doel om kennishiaten in het AMR-onderzoeksveld op te sporen door gebruik te maken van data gedreven methodes. Daarnaast wordt de toegevoegde waarde onderzocht van statistische modellen met tijdruimtelijke gegevens voor het voorspellen en identificeren van risicofactoren voor het voorkomen en verspreiding van microben en AMR.

De internationale strategische onderzoek en Innovatieagenda (SRIA) geeft een overzicht van recente ontwikkelingen en toekomstige aandachtspunten in AMR-onderzoek en is gebaseerd op kennishiaten in het AMR-onderzoeksveld. Momenteel worden deze kennishiaten handmatig geïdentificeerd door expertconsultatie en zijn daarom subjectief. In hoofdstuk 2 is beschreven in hoeverre kennislacunes in AMR-onderzoek objectief en automatisch kunnen worden gesignaleerd. Hiervoor is een bibliometrische datagedreven methodologie gebruikt. Met behulp van de PubMed-zoekmachine zijn twintig jaar aan AMR-gerelateerde artikelen geëxtraheerd. Met structurele onderwerpmodellering zijn de onderwerpen die het AMR-onderzoeksveld omvatten in kaart gebracht. Vervolgens werden onderwerpclusters gecreëerd met behulp van hiërarchische clustering op de onderwerpverhoudingen. Potentiële AMR-kennishiaten werden verkregen met behulp van Spearman's correlatie tussen onderwerpclusters en onderwerpen en tussen individuele onderwerpen. In totaal werden 88 onderwerpen en zeven onderwerpclusters geïdentificeerd uit 158616 wetenschappelijke AMR-onderzoeksartikelen. In totaal zijn 421 potentiële kennislacunes geïdentificeerd tussen de themaclusters en thema's en 2663 tussen de afzonderlijke thema's. Belangrijke hiaten in de kennis tussen moleculair en laboratorium AMR-onderzoek werden benadrukt. De kennishiaten tussen AMR-onderzoek met betrekking tot water en milieu en zowel institutionele als internationale surveillance-onderwerpen werden op onderwerpniveau benadrukt. Deze resultaten bieden een innovatieve, datagestuurde manier om kennishiaten in AMR-onderzoek te identificeren.

Postoperatieve wondinfecties (POWI’s) vormen 19,6% van de zorg-gerelateerde infecties (HAI's) in Europa. In publicaties waarin risicofactoren voor POWI’s in kaart worden gebracht, wordt meestal niet gespecifieerd hoe de grenswaarden voor continue variabelen zijn bepaald. Meestal wordt gebruik gemaakt van standaard

medische grenswaarden. In hoofdstuk 3 wordt beschreven welke risicofactoren kunnen worden geïdentificeerd voor het optreden van POWI's bij operaties aan het spijsverteringsstelsel, de borstkas en orthopedische verrichtingen door gebruik te maken van standaard medische versus datagedreven grenswaarden. Voor dit onderzoek is gebruik gemaakt van retrospectieve operatiegegevens uit individuele elektronische medische dossiers, farmaceutische gegevens en laboratoriumgegevens van het Erasmus MC Universitair Medisch Centrum in Nederland. Risicofactoren voor het optreden van POWI's werden geïdentificeerd met behulp van een Multivariabele Logistische Regressie Analyse. Standaard medische grenswaarden werden vergeleken met grenswaarden bepaald uit de data. Voor chirurgische ingrepen aan het spijsverteringsstelsel, orthopedische en thoracale ingrepenzijn de geïdentificeerde risicofactoren een preoperatieve temperatuur van 38 ^{o}C en het gebruik van antibiotica op het moment van de operatie. Voor chirurgische ingrepen aan het spijsverteringstelsel werden C-reactief proteïne (CRP) en de duur van de operatie geïdentificeerd als risicofactoren. Een volwassen leeftijd (leeftijd $\geq$ 18 jaar) werd geïdentificeerd als een beschermende factor voor thoracale chirurgische ingrepen. Datagedreven grenswaarden voor de variabelen lichaamstemperatuur, leeftijd en CRP, verklaarden het optreden van POWI's tot 19,5% beter dan standaard medische grenswaarden. Toekomstige studies moeten onderzoeken of datagedreven grenswaarden van toegevoegde waarde zijn om de klinische uitkomst te voorspellen en niet alleen te vertrouwen op standaard medische grenswaarden voor het identificeren van risicofactoren.

Overdracht van schadelijke micro-organismen (SMO) vormt een grote bedreiging voor patiënten en gezondheidswerkers in de gezondheidszorg. De meest effectieve maatregel om deze overdrachten te voorkomen is de naleving van het handhygiënebeleid, maar de nalevingspercentages zijn relatief laag en variëren in tijd en ruimte. De tijdruimtelijke effecten van verschillende niveaus van naleving van handhygiëne op de overdracht en verspreiding van hand overdraagbare SMO in een gesloten zorgomgeving moeten nog worden gekwantificeerd. In hoofdstuk 4 is beschreven in hoeverre tijdruimtelijke bewegingen van zorgprofessionals potentiële superverspreiders kunnen identificeren binnen een gesloten zorgomgeving. Hiervoor is de beroepsgroep van potentiële superverspreiders in de gezondheidszorg geïdentificeerd en zijn de tijdruimtelijke effecten op de handtransmissie van SMO gekwantificeerd voor verschillende niveaus van naleving van handhygiëne (NH) veroorzaakt door deze groep. In het Universitair Medisch Centrum Groningen (UMCG) zijn tijdruimtelijke gegevens verzameld met behulp van radiofrequentie-identificatietechnologie. De effecten van vijf kansverdelingen van NH en drie transmissiesnelheden van schadelijke micro-organismen werden gesimuleerd met behulp van een dynamisch agent-gebaseerd model. De effecten van initiële

aannames op de simulatieresultaten werden gekwantificeerd met behulp van vijf risico-uitkomsten. Verpleegkundigen werden geïdentificeerd als de potentiële super verspreidende beroepsgroep van gezondheidswerkers. Bij beperkte naleving van handhygiëne (NH5%) en hoge transmissiesnelheden (5% per contactmoment), kan een gekoloniseerde verpleegster microben overbrengen naar drie van de 17 gezondheidswerkers of van de patiënten waar ze contact mee hebben tijdens de 98,4 minuten van 23 kamers die gekoloniseerd zijn. De tijdsperiode van mogelijke SMO-transmissie door verpleegkundigen is hoger tijdens doordeweekse avonden (17 uur – 7 uur) en in het weekend in vergelijking met weekdagen (7 uur – 17 uur). Tijdsruimtelijk gedrag en sociale mengpatronen van de gezondheidszorg kunnen het verwachte aantal handtransmissies veranderen en SMO verspreiden door superverspreiders in een gesloten zorgomgeving. Deze inzichten kunnen worden gebruikt om tijdruimtelijke veiligheidsgedragingen te evalueren en strategieën voor infectiepreventie en -bestrijding te ontwikkelen.

Een uitbraak met Vancomycine-resistente enterokokken (VRE) is geassocieerd met ernstige gezondheidslasten- voor kwetsbare patiënten en hoge kosten voor de zorg. De kans dat een ziekenhuispatiënt VRE krijgt, neemt toe bij gebruik van antibiotica en wanneer eerdere kamergenoten VRE hadden. Het antibioticagebruik van eerdere kamergenoten wordt echter vaak verwaarloosd. In hoofdstuk 5 is beschreven hoe het optreden en de verspreiding van VRE verklaard kan worden met behulp van patiëntbewegingen binnen het ziekenhuis (IPB) en het antibioticagebruik tussen verschillende ziekenhuisafdelingen. Er is gebruik gemaakt van retrospectieve IPB-, antibioticagebruik- en PCR-screeningsgegevens van een ziekenhuis in Nederland. Er werd een dynamisch gestuurde tijdruimtelijke grafiek ontwikkeld, die samen met het PageRank-algoritme werd gebruikt om twee dagelijkse centraliteitsmaatregelen te berekenen om de stroom van patiënten en antibiotica op afdelingsniveau samen te vatten. Met een Decision Tree en Random Forest model werd een voorspelling berekend van het dagelijkse optreden van VRE voor elke afdeling. In een 30% teststeekproef werden de prestaties van de modellen vergeleken. Het Decision Tree model leverde een eenvoudige set regels op die kunnen worden gebruikt om de dagelijkse kans op het optreden van VRE voor elke ziekenhuisafdeling te bepalen. Het Decision Tree model behaalde een oppervlakte onder de curve van 0,685 en het Random Forest model 0,886 op de testset. Deze resultaten bevestigen dat het Random Forest model beter presteert dan een enkele Decision Tree voor alle niveaus van modelgevoeligheid en specificiteit, ten koste van de eenvoud van het model. Met behulp van deze resultaten kan een systeem voor vroegtijdige waarschuwing voor VRE worden ontwikkeld en kunnen plannen voor infectiepreventie en uitbraakstrategieën worden gebruikt.

Samenvattend wordt in dit proefschrift beschreven dat datagestuurde statistische modellen ons begrip en voorspellen van antimicrobiële resistentie kunnen verbeteren. Er wordt uitleg gegeven hoe verschillende bronnen van tijdruimtelijke gegevens kunnen worden gebruikt om het optreden en de verspreiding van AMR in ziekenhuizen te voorspellen.

Acknowledgements

This research would not have been possible without a great deal of support and assistance from numerous groups and individuals.

My sincere thanks to my co-supervisors, Prof. Alfred Stein and Prof. Lisette van Gemert-Pijnen.

First, thank you to Alfred, who first presented me with the opportunity to move to the Netherlands and work here as a PhD candidate. Alfred taught me how to defend my work while still being open to improvements. He constantly pushed me to keep writing, provided fast feedback and encouraged me to work efficiently. Most of all, he showed me the difference between being a consultant and being a scientist.

Lisette was able to provide the connection between the hospitals and the university. Her profound insights into how resources and research can complement each other was invaluable to this research. On top of that, she was always willing to discuss my ideas and concerns with an open mind and help me see the bigger picture.

Thank you to my daily supervisor, Annemarie, who was always willing to empathise and discuss the standing of my research and provide perspective and guidance to navigating this long treacherous PhD journey.

I would like to extend a special thanks to the healthcare workers at Erasmus MC and UMCG, who supported me even when facing a pandemic. Anne, Mariëtte and Edwina, thank you for patiently teaching me the aspects of infection prevention and control, making yourselves available to discuss this research and helping me to obtain and understand the necessary data. Also, thank you to Margreet and Corinna, who provided insight and guidance while performing this research at the respective hospitals.

To my fellow PhD candidates who took part in this journey with me, we made it! Julia, Roberto and Christian: we shared many highs and lows, but it was always a pleasure to work with you. I have high hopes for what you will accomplish in the future.

During my research, I became part of a community of people specialising in fighting the threat of antimicrobial resistance. I would like to thank my comrades from the EurHealth-1Health project for their support and willingness to discuss problems faced and potential solutions during our regular meetings.

During my time at ITC, I met several people who made my day-to-day life enjoyable. A special thanks to Roelof for all the welcoming morning greetings and to Theresa, who provided a friendly ear that was always willing to discuss the meaning of life and

efficiently deal with any administrative issues that would arise. To my office mate, Fashuai, thank you for listening to my venting and keeping things interesting. Frank, thank you for being my soundboard for everything statistical and more. Thank you and all the best to all the other colleagues that I shared coffee or lunch with.

To the love of my life, Valentina, you are my rock, and I love you to the stars and back. Thank you for the endless love and support you gave me during my PhD journey. Your continuous passion, encouraging words and constant supply of delicious pasta were instrumental to achieving this goal. This achievement is so much more significant because I can share it with you.

Finally, I would like to thank my family. It is difficult to express the tremendous appreciation I have for my parents. They nurtured my curiosity for science from an early age and lovingly gave me the opportunity and support to realise my dreams. Thank you to my sister, Thea, who was always willing to accept a call to share my joy and frustration. To my late grandmother, Ouma Mara, thank you for your love and persistent prayers. I am here, we made it and I miss you.

Overview of publications

van Niekerk JM, Lokate M, Braakman-Jansen LM, van Gemert-Pijnen JE, Stein A. Spatiotemporal Prediction of the Occurrence of Vancomycin-resistant Enterococcus. Available at Research Square [https://doi.org/10.21203/rs.3.rs-860519/v1]

van Niekerk JM, Stein A, Doting MH, Lokate M, Braakman-Jansen LM, van Gemert-Pijnen JE. A spatiotemporal simulation study on the transmission of harmful microorganisms through connected healthcare workers in a hospital ward setting. BMC infectious diseases. 2021 Dec;21(1):1-4.

van Niekerk JM, Vos MC, Stein A, Braakman-Jansen LM, Voor in 't holt AF, van Gemert-Pijnen JE. Risk factors for surgical site infections using a data-driven approach. PloS one. 2020 Oct 28;15(10):e0240995.

Luz C, van Niekerk JM, Keizer J, Beerlage-de Jong N, Braakman-Jansen A, Stein A, Sinha B, van Gemert-Pijnen L, Glasner C. Mapping twenty years of antimicrobial resistance research trends. Available at SSRN 3792901. 2021 Jan 1.

Cruz-Martínez RR, Wentzel J, Asbjørnsen RA, Noort PD, van Niekerk JM, Sanderman R, van Gemert-Pijnen JE. Supporting self-management of cardiovascular diseases through remote monitoring technologies: metaethnography review of frameworks, models, and theories used in research and development. Journal of medical Internet research. 2020 May 21;22(5):e16157.

Cruz-Martínez RR, Noort PD, Asbjørnsen RA, van Niekerk JM, Wentzel J, Sanderman R, van Gemert-Pijnen L. Frameworks, models, and theories used in electronic health research and development to support self-management of cardiovascular diseases through remote monitoring technologies: protocol for a metaethnography review. JMIR research protocols. 2019;8(7):e13334.